NCC-WCH

Version 1.4

Menopause

Appendices I - K

Clinical guideline

Methods, evidence and recommendations

22 October 2015

Final

Commissioned by the National Institute for Health and Clinical Excellence

Disclaimer
Healthcare professionals are expected to take NICE clinical guidelines fully into account when exercising their clinical judgement. However, the guidance does not override the responsibility of healthcare professionals to make decisions appropriate to the circumstances of each patient, in consultation with the patient and/or their guardian or carer.

Funding
Registered charity no. 213280

Contents

Appendices

Appendix A: Scope

The scope is presented in a separate document

Appendix B: Stakeholders

The list of stakeholders are presented in a separate document

Appendix C: Declarations of interest

The declarations of interest are presented in a separate document

Appendix D: Review protocols

Review protocols are presented in a separate document

Appendix E: Search strategies

Search strategies are presented in a separate document

Appendix F:PRISMA flow charts

The PRISMA flow charts are presented in a separate document

Appendix G: Excluded studies

Excluded studies are presented in a separate document

Appendix H: Evidence tables

The evidence tables are presented in a separate file.

Appendix I: GRADE profiles

I.1 Diagnosis of perimenopause and menopause

Table 1: GRADE profile: diagnosis of menopause in perimenopausal women

Number of studies	Number of women	Measure of diagnostic accuracy				Quality	Quality assessment					
		Sensitivity	Specificity	Positive likelihood ratio	Negative likelihood ratio		Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Age												
≥ 45												
1 (Williams 2008)	3135	95 (94 to 96)	9 (7 to 12)	1.04 (1.01 to 1.08)	0.55 (0.39 to 0.77)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
≥ 50												
1 (Williams 2008)	3135	84 (83 to 85)	47 (43 to 52)	1.60 (1.46 to 1.75)	0.34 (0.30 to 0.38)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
≥ 55												
1 (Williams 2008)	3135	62 (60 to 64)	89 (85 to 91)	5.44 (4.17 to 7.09)	0.43 (0.41 to 0.46)	Low	Prospective case series	Serious[1]	No serious	No serious	Serious[2]	None
≥ 60												
1 (Williams 2008)	3135	33 (31 to 35)	98 (96 to 99)	15.84 (8.28 to 30.30)	0.68 (0.66 to 0.71)	Low	Prospective case series	Serious[1]	No serious	No serious	Serious[3]	None
Vasomotor symptoms												
Hot flushes currently												
1 (El Shafie 2011)	282	55 (48 to 61)	51 (39 to 63)	1.11 (0.85 to 1.44)	0.90 (0.68 to 1.18)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Hot flushes in the previous 2 weeks												
3 (Dennerstein 1993, Ho 1999, Punyahotra 1997)	1657	24 (21 to 27)	69 (65 to 73)	0.77 (0.41 to 1.41)	1.06 (0.84 to 1.34)	Very low	Prospective case series	Serious[1,4]	Very serious[5]	Serious[6]	No serious	None
Hot flushes in the past 12 months												
1 (Brown 2002)	2669	55 (51 to 59)	56 (54 to 58)	1.25 (1.15 to 1.36)	0.80 (0.73 to 0.89)	Moderate	Prospective case series	Serious[4]	No serious	No serious	No serious	None
Hot flushes (time not specified)												
1 (Chompootweep 1993)	2669	6 (5 to 7)	78 (73 to 82)	0.26 (0.19 to 0.35)	1.21 (1.14 to 1.29)	Low	Prospective case series	Serious[1]	No serious	Serious[7]	No serious	None

Number of studies	Number of women	Measure of diagnostic accuracy				Quality	Quality assessment					
		Sensitivity	Specificity	Positive likelihood ratio	Negative likelihood ratio		Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
1 (Maartens 2001)	1924	66 (62 to 70)	51 (49 to 54)	1.36 (1.26 to 1.47)	0.66 (0.59 to 0.74)	Moderate	Prospective case series	Serious[4]	No serious	No serious	No serious	None
Night sweats in the past 2 weeks												
1 (Punyahotra 1997)	121	32 (23 to 42)	73 (50 to 89)	1.19 (0.57 to 2.48)	0.93 (0.70 to 1.24)	Moderate	Prospective case series	Serious[1,4]	No serious	No serious	No serious	None
Night sweats in the past 4 weeks												
1 (Williams 2008)	3135	44 (42 to 46)	44 (39 to 49)	0.79 (0.72 to 0.86)	1.27 (1.14 to 1.42)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Night sweats in the past 12 months												
1 (Brown 2002)	2669	39 (35 to 43)	67 (65 to 69)	1.18 (1.05 to 1.33)	0.91 (0.85 to 0.98)	Moderate	Prospective case series	Serious[4]	No serious	No serious	No serious	None
Night sweats (time not specified)												
1 (Chompootweep 1993)	1619	5 (4 to 7)	83 (78 to 87)	0.30 (0.21 to 0.42)	1.15 (1.09 to 1.21)	Low	Prospective case series	Serious[1]	No serious	Serious[7]	No serious	None
1 (Maartens et al 2001)	1924	58 (54 to 61)	50 (47 to 52)	1.14 (1.05 to 1.24)	0.86 (0.77 to 0.95)	Moderate	Prospective case series	Serious[4]	No serious	No serious	No serious	None
Cold sweats in the past 2 weeks												
2 (Dennerstein 1993, Ho 1999)	1536	4 (3 to 6)	91 (89 to 93)	0.44 (0.05 to 4.10)	1.04 (0.94 to 1.15)	Very low	Prospective case series	Serious[1,4]	Very serious[5]	Serious[6]	No serious	None
Hot flushes or night sweats currently												
2 (Blümel et al 2012, Chuni and Sreemareddy 2011)	6180	66 (65 to 68)	37 (35 to 39)	1.06 (0.99 to 1.14)	0.62 (0.24 to 1.59)	Low	Prospective case series	Serious[1,4]	Serious[8]	No serious	No serious	None
Severe hot flushes or night sweats currently												
1 (Blümel 2012)	5718	12 (11 to 13)	89 (88 to 91)	1.10 (0.93 to 1.29)	0.99 (0.97 to 1.01)	Moderate	Prospective case series	Serious[1,4]	No serious	No serious	No serious	None
Hot flushes or night sweats during the past 2 weeks												
1 (Gold 2000)	5911	49 (46 to 51)	60 (59 to 62)	1.22 (1.15 to 1.30)	0.85 (0.81 to 0.90)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None

Number of studies	Number of women	Measure of diagnostic accuracy				Quality	Quality assessment					
		Sensitivity	Specificity	Positive likelihood ratio	Negative likelihood ratio		Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Hot flushes or night sweats during the past 4 weeks												
1 (Williams 2008)	3135	60 (58 to 62)	25 (21 to 29)	0.80 (0.75 to 0.85)	1.60 (1.35 to 1.90)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Palpitations in the past 2 weeks												
4 (Dennerstein 1993, Gold 2000, Ho 1999, Punyahotra 1997)	7568	18 (16 to 19)	81 (80 to 82)	0.95 (0.85 to 1.05)	1.02 (0.99 to 1.04)	Moderate	Prospective case series	Serious[1,4]	No serious	No serious	No serious	None
Palpitations (time not specified)												
1 (Chompootweep 1993)	1619	15 (13 to 17)	66 (60 to 71)	0.44 (0.36 to 0.54)	1.29 (1.19 to 1.41)	Low	Prospective case series	Serious[1]	No serious	Serious[7]	No serious	None
1 (Maartens 2001)	1924	38 (35 to 42)	66 (64 to 69)	1.14 (1.01 to 1.29)	0.93 (0.87 to 1.00)	Moderate	Prospective case series	Serious[4]	No serious	No serious	No serious	None
Endocrine tests												
FSH: cut point ≥ 38 IU/L												
1 (Stellato 1998)	246	63 (50 to 74)	64 (57 to 71)	1.75 (1.34 to 2.30)	0.58 (0.42 to 0.81)	Moderate	Prospective case series	Serious[4]	No serious	No serious	No serious	None
FSH: cut point 45 IU/L												
1 (Henrich 2006)	272	74 (60 to 84)	71 (52 to 84)	2.54 (1.83 to 3.53)	0.37 (0.28 to 0.49)	Moderate	Prospective case series	Serious[4]	No serious	No serious	No serious	None
Inhibin A: undetectable level												
1 (Burger 1998)	82	96 (78 to 100)	39 (27 to 53)	1.57 (1.26 to 1.96)	0.11 (0.02 to 0.78)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Inhibin B: undetectable level												
1 (Burger 1998)	82	43 (23 to 66)	54 (41 to 68)	0.95 (0.55 to 1.64)	1.04 (0.68 to 1.60)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None

1. HRT use status of participants not clearly reported at enrolment/or a significant proportion of participants were on HRT use at enrolment;
2. Evidence was downgraded by 1 due to 95% confidence interval for positive likelihood ratio ranges from not useful (<5) to moderately useful (5 to 10)
3. Evidence was downgraded by 1 due to 95% confidence interval for positive likelihood ratio ranges from moderately useful (5 to 10) to very useful (>10)
4. Selection bias as no clear methods are described in the recruitment of sample;
5. Evidence was downgraded by 2 due to very serious heterogeneity (chi-squared p<0.1, I-squared inconsistency statistic of >74.99%)
6. Perimenopause defined only as 3-11 months amenorrhoea, not including irregular menstruation (Ho et al 1999)
7. All irregular menses defined as perimenopause (Chompootweep et al 1993)
8. Evidence was downgraded by 1 due to serious heterogeneity (chi-squared p<0.1, I-squared inconsistency statistic of 50%-74.99%)

Table 2: GRADE profile: Diagnosis of menopause in premenopausal women

Number of studies	Number of women	Measure of diagnostic accuracy				Quality	Quality assessment					
		Sensitivity	Specificity	Positive likelihood ratio	Negative likelihood ratio		Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Age												
≥ 45												
1 (Williams 2008)	3970	95 (94 to 96)	53 (50 to 56)	2.03 (1.92 to 2.16)	0.09 (0.08 to 0.11)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
≥ 50S												
1 (Williams 2008)	3970	84 (83 to 85)	88 (86 to 90)	6.92 (5.96 to 8.03)	0.18 (0.17 to 0.20)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
≥ 55												
1 (Williams 2008)	3970	62 (60 to 64)	99 (98 to 99)	45.99 (28.66 to 73.81)	0.39 (0.37 to 0.41)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
≥ 60												
1 (Williams 2008)	3970	33 (31 to 35)	100 (99 to 100)	69.69 (31.31 to 155.10)	0.67 (0.65 to 0.69)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Vasomotor symptoms												
Hot flushes currently												
1 (El Shafie 2011)	399	55 (48 to 61)	74 (67 to 80)	2.07 (1.59 to 2.71)	0.62 (0.52 to 0.73)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Hot flushes in the previous 2 weeks												
3 (Dennerstein 1993, Ho 1999, Punyahotra 1997)	2695	24 (21 to 27)	90 (89 to 92)	2.17 (1.07 to 4.41)	0.81 (0.61 to 1.08)	Very low	Prospective case series	Serious[1, 2]	Very serious[3]	No serious	No serious	None
Hot flushes in the past 12 months												
1 (Brown 2002)	5148	55 (51 to 59)	84 (83 to 85)	3.44 (3.11 to 3.79)	0.54 (0.49 to 0.59)	Moderate	Prospective case series	Serious[2]	No serious	No serious	No serious	None
Hot flushes (time not specified)												
1 (Chompootweep 1993)	2062	6 (5 to 7)	90 (87 to 92)	0.55 (0.41 to 0.75)	1.05 (1.02 to 1.08)	Low	Prospective case series	Serious[1]	No serious	Serious[4]	No serious	None
1 (Maartens 2001)	1200	66 (62 to 70)	88 (85 to 91)	5.51 (4.35 to 6.99)	0.39 (0.35 to 0.43)	Low	Prospective case series	Serious[2]	No serious	No serious	Serious[5]	None

Number of studies	Number of women	Measure of diagnostic accuracy				Quality	Quality assessment					
		Sensitivity	Specificity	Positive likelihood ratio	Negative likelihood ratio		Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Night sweats in the past 2 weeks												
1 (Punyahotra 1997)	226	32 (23 to 42)	83 (75 to 89)	1.87 (1.16 to 3.00)	0.82 (0.70 to 0.96)	Moderate	Prospective case series	Serious[1, 2]	No serious	No serious	No serious	None
Night sweats in the past 4 weeks												
1 (Williams 2008)	3970	44 (42 to 46)	70 (67 to 76)	1.47 (1.33 to 1.61)	0.80 (0.76 to 0.84)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Night sweats in the past 12 months												
1 (Brown 2002)	5148	39 (35 to 43)	88 (87 to 89)	3.25 (2.86 to 3.69)	0.69 (0.65 to 0.74)	Moderate	Prospective case series	Serious[2]	No serious	No serious	No serious	None
Night sweats (time not specified)												
1 (Chompootweep 1993)	2062	5 (4 to 7)	93 (91 to 95)	0.80 (0.56 to 1.14)	1.01 (0.99 to 1.04)	Low	Prospective case series	Serious[1]	No serious	Serious[4]	No serious	None
1 (Maartens 2001)	1200	58 (54 to 61)	74 (70 to 78)	2.23 (1.90 to 2.61)	0.57 (0.52 to 0.63)	Moderate	Prospective case series	Serious[2]	No serious	No serious	No serious	None
Cold sweats in the past 2 weeks												
2 (Dennerstein 1993, Ho 1999)	2469	4 (3 to 6)	96 (95 to 97)	1.12 (0.61 to 2.07)	1.00 (0.97 to 1.02)	Low	Prospective case series	Serious[1,2]	No serious	Serious[6]	No serious	None
Hot flushes or night sweats currently												
2 (Blümel 2012, Chuni and Sreemreddy 2011)	7239	66 (65 to 68)	64 (62 to 66)	2.71 (1.10 to 6.65)	0.11 (0.00 to 4.06)	Very low	Prospective case series	Serious[1, 2]	Very serious[3]	No serious	Serious[5]	None
Severe hot flushes or night sweats currently												
1 (Blümel 2012)	6725	12 (11 to 13)	95 (94 to 95)	2.16 (1.81 to 2.58)	0.93 (0.92 to 0.95)	Moderate	Prospective case series	Serious[1,2]	No serious	No serious	No serious	None
Hot flushes or night sweats during the past 2 weeks												
1 (Gold 2000)	6250	49 (46 to 51)	81 (79 to 82)	2.52 (2.33 to 2.72)	0.64 (0.61 to 0.67)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Hot flushes or night sweats during the past 4 weeks												
1 (Williams 2008)	3970	60 (58 to 62)	60 (57 to 63)	1.50 (1.39 to 1.61)	0.67 (0.63 to 0.71)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None

Number of studies	Number of women	Measure of diagnostic accuracy				Quality	Quality assessment					
		Sensitivity	Specificity	Positive likelihood ratio	Negative likelihood ratio		Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Palpitations in the past 2 weeks												
4 (Dennerstein 1993, Gold 2000, Ho 1999, Punyahotra 1997)	8945	18 (16 to 19)	86 (85 to 86)	1.22 (0.93 to 1.61)	0.97 (0.91 to 1.02)	Low	Prospective case series	Serious[1,2]	Serious[7]	No serious	No serious	None
Palpitations (time not specified)												
1 (Chompootweep 1993)	2062	15 (13 to 17)	77 (74 to 80)	0.65 (0.54 to 0.78)	1.11 (1.06 to 1.16)	Low	Prospective case series	Serious[1]	No serious	Serious[4]	No serious	None
1 (Maartens 2001)	1200	38 (35 to 42)	75 (71 to 79)	1.53 (1.28 to 1.83)	0.82 (0.76 to 0.89)	Moderate	Prospective case series	Serious[2]	No serious	No serious	No serious	None
Endocrine tests												
FSH: cut point > 22.3IU/L												
1 (Shin 2008)	144	99 (89 to 100)	97 (92 to 99)	33.04 (11.47 to 95.21)	0.01 (0.00 to 0.33)	Low	Prospective case control	Serious[2]	No serious	Serious[8]	No serious	None
AMH: cut point < 3.57pmol/litre												
1 (Shin 2008)	144	92 (80 to 98)	97 (92 to 99)	30.88 (10.62 to 89.83)	0.08 (0.03 to 0.26)	Low	Prospective case control	Serious[2]	No serious	Serious[8]	No serious	None
Oestradiol: cut point <126.6pmol/litre												
1 (Shin 2008)	144	84 (68 to 93)	97 (92 to 99)	28.23 (9.65 to 82.58)	0.17 (0.08 to 0.36)	Very low	Prospective case control	Serious[2]	No serious	Serious[8]	Serious[9]	None
Inhibin A: cut point undetectable												
1 (Burger 1998)	51	96 (78 to 100)	54 (34 to 72)	2.06 (1.37 to 3.10)	0.08 (0.01 to 0.57)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Inhibin B undetectable												
1 (Burger 1998)	51	43 (23 to 66)	78 (58 to 91)	1.96 (0.84 to 4.56)	0.73 (0.48 to 1.10)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Inhibin B: cut point < 0.4 ng/litre												
1 (Shin 2008)	144	91 (80 to 98)	100 (97 to 100)	∞ (NC)	0.09 (0.03 to 0.27)	Low	Prospective case control	Serious[2]	No serious	Serious[8]	No serious	None

1. HRT use status of participants not clearly reported at enrolment/or a significant proportion of participants were on HRT use at enrolment;
2. Selection bias as no clear methods are described in the recruitment of sample;
3. Evidence was downgraded by 2 due to very serious heterogeneity (chi-squared p<0.1, I-squared inconsistency statistic of >74.99%)
4. All irregular menses defined as perimenopause (Chompootweep et al 1993)

5. Evidence was downgraded by 1 due to 95% confidence interval for positive likelihood ratio ranges from not useful (<5) to moderately useful (5 to 10)
6. Perimenopause defined only as 3-11 months amenorrhoea, not including irregular menstruation (Ho et al 1999)
7. Evidence was downgraded by 1 due to serious heterogeneity (chi-squared p<0.1, I-squared inconsistency statistic of 50%-74.99%)
8. More than 50% of premenopausal women were aged less than 40 (Shin et al 2008)
9. Evidence was downgraded by 1 due to 95% confidence interval for positive likelihood ratio ranges from moderately useful (5 to 10) to very useful (>10)

Table 3: GRADE profile: diagnosis of menopause in all other women

Number of studies	Number of women	Measure of diagnostic accuracy				Quality	Quality assessment					
		Sensitivity	Specificity	Positive likelihood ratio	Negative likelihood ratio		Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Age												
None												
1 (Williams 2008)	4402	95 (94 to 96)	42 (40 to 44)	1.64 (1.57 to 1.71)	0.12 (0.10 to 0.14)	Moderate	Prospective case series	Serious [1]	No serious	No serious	No serious	None
≥ 48												
1 (Giacobbe 2004)	192	79 (68 to 88)	76 (67 to 83)	3.29 (2.34 to 4.62)	0.28 (0.18 to 0.44)	Moderate	Prospective case series	Serious[2]	No serious	No serious	No serious	None
≥ 50												
2 (Giacobbe 2004, Williams 2008)	4594	84 (82 to 85)	79 (77 to 81)	6.23 (2.06 to 18.87)	0.26 (0.16 to 0.43)	Very low	Prospective case series	Serious [1,2]	Serious[3]	No serious	Very serious[4]	None
≥ 55												
1 (Williams 2008)	4402	62 (60 to 64)	96 (95 to 97)	15.89 (12.52 to 20.16)	0.40 (0.38 to 0.42)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
≥ 60												
1 (Williams 2008)	4402	33 (31 to 35)	99 (99 to 100)	37.38 (22.52 to 62.04)	0.68 (0.66 to 0.69)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Vasomotor symptoms												
Hot flushes currently												
1 (El Shafie 2011)	472	55 (48 to 61)	67 (61 to 73)	1.67 (1.35 to 2.06)	0.68 (0.57 to 0.80)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Hot flushes in the previous 2 weeks												
3 (Dennerstein 1993, Ho 1999, Punyahotra 1997)	3358	24 (21 to 27)	84 (83 to 86)	1.47 (1.19 to 1.82)	0.88 (0.73 to 1.05)	Low	Prospective case series	Serious[1,2]	Serious[3]	No serious	No serious	None
Hot flushes in the past 12 months												
1 (Brown 2002)	8236	55 (51 to 59)	75 (74 to 76)	2.22	0.60	Moderate	Prospective case series	Serious[2]	No serious	No serious	No serious	None

Number of studies	Number of women	Measure of diagnostic accuracy				Quality	Quality assessment					
		Sensitivity	Specificity	Positive likelihood ratio	Negative likelihood ratio		Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
				(2.04 to 2.41)	(0.55 to 0.66)							
Hot flushes (time not specified)												
1 (Chompootweep 1993)	2354	6 (4 to 7)	86 (84 to 88)	0.42 (0.32 to 0.54)	1.09 (1.06 to 1.12)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
1 (Maartens 2001)	2450	66 (62 to 70)	62 (60 to 65)	1.75 (1.61 to 1.90)	0.55 (0.49 to 0.61)	Moderate	Prospective case series	Serious[2]	No serious	No serious	No serious	None
Night sweats in the past 2 weeks												
1 (Punyahotra 1997)	268	32 (23 to 42)	81 (74 to 87)	1.72 (1.11 to 2.67)	0.83 (0.71 to 0.97)	Moderate	Prospective case series	Serious[1, 2]	No serious	No serious	No serious	None
Night sweats in the past 4 weeks												
1 (Williams 2008)	4402	44 (42 to 46)	63 (61 to 66)	1.20 (1.11 to 1.30)	0.88 (0.84 to 0.93)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Night sweats in the past 12 months												
1 (Brown 2002)	8236	39 (35 to 43)	81 (80 to 82)	2.09 (1.87 to 2.34)	0.75 (0.70 to 0.80)	Moderate	Prospective case series	Serious[2]	No serious	No serious	No serious	None
Night sweats (time not specified)												
1 (Chompootweep 1993)	2354	5 (4 to 7)	90 (88 to 92)	0.54 (0.40 to 0.73)	1.05 (1.02 to 1.07)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
1 (Maartens 2001)	2450	58 (54 to 61)	57 (54 to 59)	1.33 (1.23 to 1.45)	0.75 (0.68 to 0.82)	Moderate	Prospective case series	Serious[2]	No serious	No serious	No serious	None
Cold sweats in the past 2 weeks												
2 (Dennerstein 1993, Ho 1999)	3110	4 (3 to 6)	95 (94 to 96)	0.54 (0.08 to 3.75)	1.02 (0.95 to 1.10)	Very low	Prospective case series	Serious[1, 2]	Very serious[5]	No serious	No serious	None
Hot flushes or night sweats currently												
2 (Blümel 2012, Chuni and Sreemareddy,	9102	66 (65 to 68)	54 (52 to 55)	1.59 (1.25 to 2.01)	0.16 (0.01 to 3.27)	Very low	Prospective case series	Serious[1,2]	Very serious[5]	No serious	No serious	None
Severe hot flushes or night sweats currently												
1 (Blümel 2012)	8373	12 (11 to 13)	92 (92 to 93)	1.58 (1.38 to 1.80)	0.95 0.94 to 0.97)	Moderate	Prospective case series	Serious[1,2]	No serious	No serious	No serious	None
Hot flushes or night sweats during the past 2 weeks												

Number of studies	Number of women	Measure of diagnostic accuracy				Quality	Quality assessment					
		Sensitivity	Specificity	Positive likelihood ratio	Negative likelihood ratio		Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
1 (Gold 2000)	10408	49 (46 to 51)	71 (70 to 72)	1.67 (1.58 to 1.77)	0.72 (0.69 to 0.76)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Hot flushes or night sweats during the past 4 weeks												
1 (Williams 2008)	4402	60 (58 to 62)	51 (47 to 53)	1.23 (1.16 to 1.30)	0.78 (0.73 to 0.84)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Palpitations in the past 2 weeks												
4 (Dennerstein 1993, Gold 2000, Ho 1999, Punyahotra 1997)	13766	18 (16 to 19)	83 (83 to 84)	1.07 (0.87 to 1.32)	0.99 (0.95 to 1.04)	Low	Prospective case series	Seriou[1,2]	Serious[3]	No serious	No serious	None
Palpitations (time not specified)												
1 (Chompootweep 1993)	2354	15 (13 to 17)	74 (71 to 76)	0.57 (0.48 to 0.67)	1.15 (1.10 to 1.20)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
1 (Maartens 2001)	2450	38 (35 to 42)	69 (67 to 71)	1.23 (1.09 to 1.39)	0.89 (0.84 to 0.96)	Moderate	Prospective case series	Serious[2]	No serious	No serious	No serious	None
Endocrine tests												
Inhibin A: cut point undetectable												
1 (Burger 1998)	110	96 (78 to 100)	44 (33 to 55)	1.70 (1.38 to 2.08)	0.10 (0.01 to 0.69)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Inhibin B: cut point undetectable												
1 (Burger 1998)	110	43 (23 to 66)	62 (51 to 72)	1.14 (0.67 to 1.96)	0.91 (0.61 to 1.36)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Ovarian ultrasound features												
Antral follicle count cut point ≤ 2 follicles												
1 (Giacobbe 2004)	204	89 (79 to 95)	42 (33 to 51)	1.53 (1.29 to 1.82)	0.27 (0.13 to 0.53)	Moderate	Prospective case series	Serious[2]	No serious	No serious	No serious	None
Ovarian volume <4cm3												
1 (Giacobbe 2004)	204	73 (61 to 83)	81 (73 to 88)	3.85 (2.60 to 5.71)	0.33 (0.22 to 0.49)	Low	Prospective case series	Serious[2]	No serious	No serious	Serious[6]	None
Combination tests												
Menstrual algorithm												
1 (Johnson 2004)	507	90 (70 to 99)	98 (93 to 99)	36.19	0.09	Low	Prospective case series	Serious[2]	No serious	Serious[7]	No serious	None

Number of studies	Number of women	Measure of diagnostic accuracy				Quality	Quality assessment					
		Sensitivity	Specificity	Positive likelihood ratio	Negative likelihood ratio		Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
				(11.74 to 111.58)	(0.03 to 0.37)							
Hormonal algorithm												
1 (Johnson 2004)	507	90 (70 to 99)	100 (97 to 100)	∞ (NC)	0.10 (0.03 to 0.36)	Very low	Prospective case series	Serious[2]	No serious	Serious[7]	Serious[8]	None
Historical algorithm												
1 (Johnson 2004)	507	90 (70 to 99)	98 (93 to 99)	36.19 (11.74 to 111.58)	0.09 (0.03 to 0.37)	Low	Prospective case series	Serious[2]	No serious	Serious[7]	No serious	None

1. HRT use status of participants not clearly reported at enrolment/or a significant proportion of participants were on HRT use at enrolment;
2. Selection bias as no clear methods are described in the recruitment of sample;
3. Evidence was downgraded by 1 due to serious heterogeneity (chi-squared p<0.1, I-squared inconsistency statistic of 50%-74.99%)
4. Evidence was downgraded by 2 due to 95% confidence interval for positive likelihood ratio ranges from not useful (<5) to very useful (>10)
5. Evidence was downgraded by 2 due to very serious heterogeneity (chi-squared p<0.1, I-squared inconsistency statistic of >74.99%)
6. Evidence was downgraded by 1 due to 95% confidence interval for positive likelihood ratio ranges from not useful (<5) to moderately useful (5 to 10)
7. Only women with suspected myocardial ischaemia and without hysterectomy included
8. 95% confidence interval not able to be calculated

Table 4: GRADE profile: diagnosis of perimenopause in postmenopausal women

Number of studies	Number of women	Measure of diagnostic accuracy				Quality	Quality assessment					
		Sensitivity	Specificity	Positive likelihood ratio	Negative likelihood ratio		Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Age												
< 45												
1 (Williams 2008)	3135	9 (7 to 12)	95 (94 to 96)	1.82 (1.29 to 2.56)	0.96 (0.93 to 0.99)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
< 50												
1 (Williams 2008)	3135	47 (43 to 52)	84 (83 to 85)	2.98 (2.61 to 3.40)	0.62 (0.57 to 0.68)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
< 55												
1 (Williams 2008)	3135	89 (85 to 91)	62 (60 to 64)	2.32 (2.18 to 2.46)	0.18 (0.14 to 0.24)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
< 60												
1 (Williams 2008)	3135	98 (96 to 99)	33 (31 to 35)	1.46 (1.42 to 1.51)	0.06 (0.03 to 0.12)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None

Number of studies	Number of women	Measure of diagnostic accuracy				Quality	Quality assessment					
		Sensitivity	Specificity	Positive likelihood ratio	Negative likelihood ratio		Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Vasomotor symptoms												
Hot flushes currently												
1 (El Shafie 2011)	282	49 (37 to 61)	45 (39 to 52)	0.90 (0.69 to 1.18)	1.12 (0.85 to 1.46)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Hot flushes in the previous 2 weeks												
3 (Dennerstein 1993, Ho 1999, Punyahotra 1997)	1657	31 (27 to 35)	76 (74 to 79)	1.31 (0.71 to 2.41)	0.94 (0.75 to 1.19)	Very low	Prospective case series	Serious[1, 2]	Very serious[3]	Serious[4]	No serious	None
Hot flushes in the past 12 months												
1 (Brown 2002)	2669	44 (42 to 46)	45 (41 to 49) a	0.80 (0.73 to 0.87)	1.24 (1.13 to 1.37)	Moderate	Prospective case series	Serious [2]	No serious	No serious	No serious	None
Hot flushes (time not specified)												
1 (Chompootweep 1993)	1619	22 (18 to 27)	94 (93 to 95)	3.89 (2.86 to 5.28)	0.82 (0.77 to 0.88)	Very low	Prospective case series	Serious[1]	No serious	Serious[5]	Serious[6]	None
1 (Maartens 2001)	1924	49 (46 to 51)	34 (30 to 38)	0.74 (0.68 to 0.80)	1.51 (1.35 to 1.70)	Moderate	Prospective case series	Serious [2]	No serious	No serious	No serious	None
Night sweats in the past 2 weeks												
1 (Punyahotra 1997)	121	27 (11 to 50)	68 (58 to 77)	0.84 (0.40 to 1.77)	1.07 (0.80 to 1.44)	Moderate	Prospective case series	Serious[1, 2]	No serious	No serious	No serious	None
Night sweats in the past 4 weeks												
1 (Williams 2008)	3135	56 (51 to 61)	56 (54 to 58)	1.27 (1.16 to 1.40)	0.79 (0.70 to 0.88)	Moderate	Prospective case series	Serious [1]	No serious	No serious	No serious	None
Night sweats in the past 12 months												
1 (Brown 2002)	2669	33 (31 to 35)	61 (57 to 65)	0.85 (0.75 to 0.95)	1.10 (1.02 to 1.18)	Moderate	Prospective case series	Serious [2]	No serious	No serious	No serious	None
Night sweats (time not specified)												
1 (Chompootweep 1993)	1619	17 (13 to 22)	95 (93 to 96)	3.36 (2.39 to 4.71)	0.87 (0.82 to 0.92)	Low	Prospective case series	Serious [1]	No serious	Serious[5]	No serious	None
1 (Maartens 2001)	1924	50 (48 to 53)	42 (39 to 46)	0.88 (0.81 to 0.95)	1.17 (1.05 to 1.30)	Moderate	Prospective case series	Serious [2]	No serious	No serious	No serious	None
Cold sweats in the past 2 weeks												

Number of studies	Number of women	Measure of diagnostic accuracy				Quality	Quality assessment					
		Sensitivity	Specificity	Positive likelihood ratio	Negative likelihood ratio		Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
2 (Dennerstein 1993, Ho 1999)	1536	9 (7 to 11)	96 (94 to 97)	2.27 (0.24 to 21.10)	0.96 (0.87 to 1.07)	Very low	Prospective case series	Serious [1, 2]	Very serious[3]	Serious[4]	Very serious[7]	None
Hot flushes or night sweats currently												
2 (Blümel 2012, Chuni and Sreemareddy 2011)	6180	63 (61 to 65)	34 (33 to 35)	0.94 (0.88 to 1.01)	1.62 (0.63 to 4.16)	Low	Prospective case series	Serious [1,2]	Serious[8]	No serious	No serious	None
Severe hot flushes or night sweats currently												
1 (Blümel 2012)	5718	11 (9 to 12)	88 (87 to 89)	0.91 (0.77 to 1.07)	1.01 (0.99 to 1.03)	Moderate	Prospective case series	Serious [1,2]	No serious	No serious	No serious	None
Hot flushes or night sweats during the past 2 weeks												
1 (Gold 2000)	5911	40 (38 to 41)	51 (49 to 54)	0.82 (0.77 to 0.87)	1.17 (1.12 to 1.24)	Moderate	Prospective case series	Serious [1]	No serious	No serious	No serious	None
Hot flushes or night sweats during the past 4 weeks												
1 (Williams 2008)	3135	75 (71 to 79)	40 (38 to 42)	1.25 (1.17 to 1.33)	0.63 (0.53 to 0.74)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Palpitations in the past 2 weeks												
4 (Dennerstein 1993, Gold 2000, Ho 1999, Punyahotra 1997)	7568	19 (18 to 20)	83 (81 to 84)	1.06 (0.95 to 1.17)	0.99 (0.96 to 1.01)	Moderate	Prospective case series	Serious [1,2]	No serious	No serious	No serious	None
Palpitations (time not specified)												
1 (Chompootweep et al 1993)	1619	34 (29 to 40)	85 (83 to 87)	2.28 (1.86 to 2.80)	0.77 (0.71 to 0.84)	Low	Prospective case series	Serious [1]	No serious	Serious[5]	No serious	None
1 (Maartens et al 2001)	1924	34 (31 to 36)	62 (58 to 65)	0.88 (0.78 to 0.99)	1.08 (1.00 to 1.16)	Moderate	Prospective case series	Serious [2]	No serious	No serious	No serious	None
Endocrine tests												
Inhibin A: cut point undetectable												
1 (Burger 1998)	82	61 (47 to 73)	4 (0 to 22)	0.64 (0.51 to 0.80) a	8.97 (1.28 to 62.60)	Moderate	Prospective case series	Serious [1]	No serious	No serious	No serious	None
Inhibin B: cut point undetectable												
1 (Burger 1998)	82	46 (32 to 59)	57 (34 to 77)	1.05 (0.61 to 1.81)	0.96 (0.63 to 1.48)	Moderate	Prospective case series	Serious [1]	No serious	No serious	No serious	None

1. HRT use status of participants not clearly reported at enrolment/or a significant proportion of participants were on HRT use at enrolment;
2. Selection bias as no clear methods are described in the recruitment of sample;
3. Evidence was downgraded by 2 due to very serious heterogeneity (chi-squared p<0.1, I-squared inconsistency statistic of >74.99%)
4. Perimenopause defined only as 3-11 months amenorrhoea, not including irregular menstruation (Ho et al 1999)
5. All irregular menses defined as perimenopause (Chompootweep et al 1993)
6. Evidence was downgraded by 1 due to 95% confidence interval for positive likelihood ratio ranges from not useful (<5) to moderately useful (5 to 10)
7. Evidence was downgraded by 2 due to 95% confidence interval for positive likelihood ratio ranges from not useful (<5) to very useful (>10)
8. Evidence was downgraded by 1 due to serious heterogeneity (chi-squared p<0.1, I-squared inconsistency statistic of 50%-74.99%)

Table 5: GRADE profile: diagnosis of perimenopause in premenopausal women

Number of studies	Number of women	Measure of diagnostic accuracy				Quality	Quality assessment					
		Sensitivity	Specificity	Positive likelihood ratio	Negative likelihood ratio		Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Age												
≥ 42												
1 (Cooper and Baird 1995)	280	90 (76 to 97)	29 (23 to 35)	1.26 (1.10 to 1.45)	0.36 (0.14 to 0.93)	Moderate	Prospective case series	Serious [1]	No serious	No serious	No serious	None
≥ 45												
1 (Williams 2008)	1699	91 (88 to 94)	53 (50 to 56)	1.95 (1.82 to 2.08)	0.17 (0.13 to 0.23)	Moderate	Prospective case series	Serious [2]	No serious	No serious	No serious	None
≥ 46												
1 (Cooper and Baird 1995)	280	54 (37 to 70)	73 (67 to 79)	2.00 (1.40 to 2.85)	0.63 (0.45 to 0.89)	Moderate	Prospective case series	Serious [1]	No serious	No serious	No serious	None
≥ 50												
1 (Williams 2008)	1699	53 (48 to 57)	88 (86 to 90)	4.32 (3.64 to 5.14)	0.54 (0.49 to 0.60)	Low	Prospective case series	Serious[2]	No serious	No serious	Serious[3]	None
≥ 55												
1 (Williams 2008)	1699	11 (9 to 15)	99 (98 to 99)	8.45 (4.92 to 14.52)	0.90 (0.87 to 0.93)	Very low	Prospective case series	Serious[2]	No serious	No serious	Very serious[4]	None
≥ 60												
1 (Williams 2008)	1699	2 (1 to 4)	100 (99 to 100)	4.40 (1.58 to 12.29)	0.98 (0.97 to 1.00)	Very low	Prospective case series	Serious[2]	No serious	No serious	Very serious[4]	None
Hot flushes currently												
1 (El Shafie 2011)	263	49 (37 to 61)	74 (67 to 80)	1.87 (1.34 to 2.61)	0.69 (0.54 to 0.88)	Moderate	Prospective case series	Serious[2]	No serious	No serious	No serious	None
Hot flushes in the previous 2 weeks												

Number of studies	Number of women	Measure of diagnostic accuracy				Quality	Quality assessment					
		Sensitivity	Specificity	Positive likelihood ratio	Negative likelihood ratio		Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
3 (Dennerstein 1993, Ho 1999, Punyahotra 1997)	2364	31 (27 to 35)	90 (89 to 92)	2.94 (2.31 to 3.76)	0.78 (0.69 to 0.89)	Very low	Prospective case series	Serious[1,2]	Very serious[5]	Serious [6]	No serious	None
Hot flushes in the past 12 months												
1 (Brown 2002)	6663	44 (42 to 46)	84 (83 to 85)	2.75 (2.53 to 2.98)	0.67 (0.64 to 0.69)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Hot flushes (time not specified)												
1 (Chompootweep 1993)	1027	22 (18 to 27)	90 (87 to 92)	2.15 (1.59 to 3.87)	0.87 (0.81 to 0.93)	Low	Prospective case series	Serious[2]	No serious	Serious[7]	No serious	None
1 (Maartens 2001)	1776	49 (46 to 51)	88 (85 to 91)	4.05 (3.19 to 5.15)	0.58 (0.55 to 0.62)	Low	Prospective case series	Serious[1]	No serious	No serious	Serious[3]	None
Night sweats in the past 2 weeks												
1 (Punyahotra 1997)	149	27 (11 to 50)	83 (75 to 89)	1.57 (0.72 to 3.44)	0.88 (0.67 to 1.15)	Moderate	Prospective case series	Serious[1,2]	No serious	No serious	No serious	None
Night sweats in the past 4 weeks												
1 (Williams 2008)	1699	56 (52 to 61)	70 (67 to 73)	1.87 (1.66 to 2.10)	0.63 (0.56 to 0.70)	Moderate	Prospective case series	Serious [2]	No serious	No serious	No serious	None
Night sweats in the past 12 months												
1 (Brown 2002)	6663	33 (31 to 35)	88 (87 to 89)	2.75 (2.49 to 3.03)	0.76 (0.74 to 0.79)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Night sweats (time not specified)												
1 (Chompootweep 1993)	1027	17 (13 to 22)	93 (91 to 95)	2.67 (1.85 to 3.87)	0.88 (0.83 to 0.93)	Low	Prospective case series	Serious[2]	No serious	Serious[7]	No serious	None
1 (Maartens 2001)	1776	50 (48 to 53)	74 (70 to 78)	1.96 (1.67 to 2.28)	0.67 (0.62 to 0.72)	Moderate	Prospective case series	Serious[1]	No serious	Serious	No serious	None
Cold sweats in the past 2 weeks												
2 (Dennerstein 1993, Ho 1999)	2215	9 (7 to 11)	96 (95 to 97)	2.13 (0.48 to 9.41)	0.96 (0.89 to 1.04)	Very low	Prospective case series	Serious[1,2]	No serious	Serious[6]	Serious [3]	None
Hot flushes or night sweats currently												
2 (Blümel 2012, Chuni and	4785	63 (61 to 65)	64 (62 to 66)	2.55 (0.99 to 6.59)	0.21 (0.02 to 2.30)	Very low	Prospective case series	Serious[1,2]	Very serious[5]	No serious	Serious[3]	None

Number of studies	Number of women	Measure of diagnostic accuracy				Quality	Quality assessment					
		Sensitivity	Specificity	Positive likelihood ratio	Negative likelihood ratio		Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Sreemareddy 2011)												
Severe hot flushes or night sweats currently												
1 (Blümel 2012)	4303	11 (9 to 12)	95 (94 to 95)	1.96 (1.59 to 2.42)	0.94 (0.93 to 0.96)	Moderate	Prospective case series	Serious[1,2]	No serious	No serious	No serious	None
≥ 1 hot flush/night sweat per day for the last 6 months												
1 (Cooper and Baird 1995)	280	29 (15 to 43)	97 (95 to 99)	9.43 (3.90 to 22.80)	0.73 (0.60 to 0.90)	Very low	Prospective case series	Serious[1]	No serious	No serious	Very serious[4]	None
Hot flushes or night sweats during the past 2 weeks												
1 (Gold 2000)	8655	40 (38 to 41)	81 (79 to 82)	2.05 (1.91 to 2.20)	0.75 (0.73 to 0.77)	Moderate	Prospective case series	Serious[2]	No serious	No serious	No serious	None
Hot flushes or night sweats during the past 4 weeks												
1 (Williams 2008)	1699	75 (71 to 79)	60 (57 to 63)	1.87 (1.72 to 2.04)	0.42 (0.35 to 0.49)	Moderate	Prospective case series	Serious[2]	No serious	No serious	No serious	None
Palpitations in the past 2 weeks												
4 (Dennerstein 1993, Gold 2000, Ho 1999, Punyahotra 1997)	11019	19 (18 to 20)	86 (85 to 86)	1.38 (1.26 to 1.50)	0.94 (0.92 to 0.96)	Moderate	Prospective case series	Serious[1,2]	No serious	No serious	No serious	None
Palpitations (time not specified)												
1 (Chompootweep 1993)	1027	34 (29 to 40)	77 (74 to 80)	1.48 (1.20 to 1.82)	0.86 (0.78 to 0.94)	Low	Prospective case series	Serious[2]	No serious	Serious[7]	No serious	None
1 (Maartens 2001)	1776	33 (31 to 36)	75 (71 to 79)	1.35 (1.14 to 1.59)	0.88 (0.83 to 0.94)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Endocrine tests												
FSH: cut point 13 IU/L												
1 (Henrich 2006)	397	67 (50 to 81)	88 (81 to 92)	5.72 (4.08 to 8.01)	0.37 (0.28 to 0.49)	Low	Prospective case series	Serious[1]	No serious	No serious	Serious[3]	None
FSH: cut point ≥ 24 IU/L												
1 (Stellato 1998)	278	65 (57 to 72)	69 (59 to 78)	2.07 (1.52 to 2.82)	0.51 (0.41 to 0.65)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Inhibin A: cut point undetectable level												

Number of studies	Number of women	Measure of diagnostic accuracy				Quality	Quality assessment					
		Sensitivity	Specificity	Positive likelihood ratio	Negative likelihood ratio		Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
1 (Burger 1998)	87	61 (47 to 73)	54 (34 to 72)	1.31 (0.84 to 2.06)	0.73 (0.45 to 1.16)	Moderate	Prospective case series	Serious[2]	No serious	No serious	No serious	None
Inhibin B: cut point undetectable level												
1 (Burger 1998)	87	46 (32 to 59)	78 (58 to 91)	2.05 (0.96 to 4.39)	0.70 (0.51 to 0.96)	Moderate	Prospective case series	Serious[2]	No serious	No serious	No serious	None
Combination tests												
At least one of the following: started HRT when periods became irregular, ≥ 1 hot flush/night sweat per day for the past 6 months or last menstrual cycle longer than 60 days												
1 (Cooper and Baird 1995)	280	56 (41 to 72)	95 (93 to 98)	12.36 (6.52 to 23.44)	0.46 (0.32 to 0.65)	Low	Prospective case series	Serious [1]	No serious	No serious	Serious [8]	None
At least one of the following: started HRT when periods became irregular, ≥ 1 hot flush/night sweat per day for the past 6 months, last menstrual cycle longer than 60 days or menstrual cycles longer or more variable during the past 5 years												
1 (Cooper and Baird 1995)	280	69 (55 to 84)	75 (70 to 81)	2.78 (2.05 to 3.77)	0.41 (0.25 to 0.66)	Moderate	Prospective case series	Serious [1]	No serious	No serious	No serious	None

1. Selection bias as no clear methods are described in the recruitment of sample;
2. HRT use status of participants not clearly reported at enrolment/or a significant proportion of participants were on HRT use at enrolment;
3. Evidence was downgraded by 1 due to 95% confidence interval for positive likelihood ratio ranges from not useful (<5) to moderately useful (5 to 10)
4. Evidence was downgraded by 2 due to 95% confidence interval for positive likelihood ratio ranges from not useful (<5) to very useful (>10)
5. Evidence was downgraded by 2 due to very serious heterogeneity (chi-squared p<0.1, I-squared inconsistency statistic of >74.99%)
6. Perimenopause defined only as 3-11 months amenorrhoea, not including irregular menstruation (Ho et al 1999)
7. All irregular menses defined as perimenopause (Chompootweep et al 1993)
8. Evidence was downgraded by 1 due to 95% confidence interval for positive likelihood ratio ranges from moderately useful (5 to 10) to very useful (>10)

Table 6: GRADE profile: diagnosis of perimenopause in all other women

Number of studies	Number of women	Measure of diagnostic accuracy				Quality	Quality assessment					
		Sensitivity	Specificity	Positive likelihood ratio	Negative likelihood ratio		Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Age												
No evidence identified												
Vasomotor symptoms												
Hot flushes currently												
1 (El Shafie 2011)	479	49 (37 to 61)	59 (54 to 64)	1.20 (0.92 to 1.56)	0.86 (0.68 to 1.09)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Hot flushes in the previous 2 weeks												

Number of studies	Number of women	Measure of diagnostic accuracy				Quality	Quality assessment					
		Sensitivity	Specificity	Positive likelihood ratio	Negative likelihood ratio		Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
3 (Dennerstein 1993, Ho 1999, Punyahotra 1997)	3358	31 (27 to 35)	85 (84 to 87)	1.80 (1.12 to 2.89)	0.88 (0.79 to 0.98)	Very low	Prospective case series	Serious[1,2]	Very serious[3]	Serious[4]	No serious	None
Hot flushes in the past 12 months												
1 (Brown 2002)	8236	44 (42 to 46)	80 (79 to 81)	2.16 (2.01 to 2.32)	0.70 (0.68 to 0.73)	Moderate	Prospective case series	Serious[2]	No serious	No serious	No serious	None
Hot flushes (time not specified) (similar findings reported by Legorreta et al. 2013)												
1 (Chompootweep 1993)	2354	22 (18 to 27)	93 (91 to 94)	3.04 (2.34 to 3.96)	0.84 (0.79 to 0.89)	Low	Prospective case series	Serious[1]	No serious	Serious[4]	No serious	None
1 (Maartens 2001)	2450	49 (46 to 51)	58 (55 to 60)	1.15 (1.05 to 1.25)	0.89 (0.83 to 0.96)	Moderate	Prospective case series	Serious[2]	No serious	No serious	No serious	None
Night sweats in the past 2 weeks												
1 (Punyahotra 1997)	248	27 (11 to 50)	77 (70 to 82)	1.16 (0.57 to 2.39)	0.95 (0.73 to 1.24)	Moderate	Prospective case series	Serious[1, 2]	No serious	No serious	No serious	None
Night sweats in the past 4 weeks												
1 (Williams 2008)	4402	56 (52 to 61)	60 (59 to 62)	1.42 (1.29 to 1.55)	0.72 (0.65 to 0.81)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Night sweats in the past 12 months												
1 (Brown 2002)	8236	33 (31 to 35)	85 (84 to 86)	2.20 (2.01 to 2.40)	0.79 (0.76 to 0.81)	Moderate	Prospective case series	Serious[2]	No serious	No serious	No serious	None
Night sweats (time not specified)												
1 (Chompootweep et al 1993)	2354	17 (13 to 22)	94 (93 to 95)	3.08 (2.27 to 4.18)	0.88 (0.83 to 0.92)	Low	Prospective case series	Serious[1]	No serious	Serious[4]	No serious	None
1 (Maartens 2001)	2450	50 (48 to 53)	56 (53 to 59)	1.16 (1.06 to 1.26)	0.88 (0.82 to 0.95)	Moderate	Prospective case series	Serious[2]	No serious	No serious	No serious	None
Cold sweats in the past 2 weeks												
2 (Dennerstein 1993, Ho 1999)	3110	9 (7 to 11)	96 (95 to 97)	2.28 (0.39 to 13.40)	0.96 (0.88 to 1.05)	Very low	Prospective case series	Serious[1, 2]	Serious[5]	Serious[6]	Very serious[7]	None
Hot flushes or night sweats currently												

Number of studies	Number of women	Measure of diagnostic accuracy				Quality	Quality assessment					
		Sensitivity	Specificity	Positive likelihood ratio	Negative likelihood ratio		Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
2 (Blümel 2012, Chuni and Sreemareddy 2011)	9102	63 (61 to 65)	46 (45 to 47)	1.33 (0.91 to 1.95)	0.34 (0.05 to 2.48)	Very low	Prospective case series	Serious[1, 2]	Very serious[3]	No serious	No serious	None
Severe hot flushes or night sweats currently												
1 (Blümel 2012)	8373	11 (9 to 12)	91 (90 to 91)	1.15 (0.99 to 1.35)	0.98 (0.97 to 1.00)	Moderate	Prospective case series	Serious[1, 2]	No serious	No serious	No serious	None
Hot flushes or night sweats during the past 2 weeks												
1 (Gold 2000)	10408	40 (38 to 41)	72 (71 to 73)	1.44 (1.36 to 1.52)	0.83 (0.81 to 0.86)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Hot flushes or night sweats during the past 4 weeks												
1 (Williams 2008)	4402	75 (71 to 79)	46 (45 to 48)	1.40 (1.31 to 1.49)	0.54 (0.46 to 0.64)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Palpitations in the past 2 weeks												
4 (Dennerstein 1993, Gold 2000, Ho 1999, Punyahotra 1997)	13766	19 (18 to 20)	85 (84 to 85)	1.26 (1.17 to 1.37)	0.95 (0.94 to 0.97)	Moderate	Prospective case series	Serious[1,2]	No serious	No serious	No serious	None
Palpitations (time not specified)												
1 (Chompootweep 1993)	2354	34 (29 to 40)	82 (80 to 84)	1.91 (1.59 to 2.30)	0.80 (0.74 to 0.87)	Low	Prospective case series	Serious[1]	No serious	Serious[4]	No serious	None
1 (Maartens 2001)	2450	34 (31 to 36)	67 (65 to 70)	1.04 (0.93 to 1.16)	0.98 (0.93 to 1.04)	Moderate	Prospective case series	Serious[2]	No serious	No serious	No serious	None
Endocrine tests												
Inhibin A: cut point undetectable level												
1 (Burger 1998)	110	61 (47 to 73)	31 (19 to 46)	0.89 (0.67 to 1.17)	1.24 (0.74 to 2.08)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Inhibin B: cut point undetectable												
1 (Burger 1998)	110	46 (32 to 59)	68 (54 to 80)	1.43 (0.87 to 2.34)	0.80 (0.59 to 1.08)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
Combination tests												
Menstrual algorithm												

Number of studies	Number of women	Measure of diagnostic accuracy				Quality	Quality assessment					
		Sensitivity	Specificity	Positive likelihood ratio	Negative likelihood ratio		Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
1 (Johnson 2004)	507	96 (78 to 100)	98 (94 to 100)	56.43 (14.24 to 223.63)	0.04 (0.01 to 0.30)	Low	Prospective case series	Serious[2]	No serious	Serious[8]	No serious	None
Hormonal algorithm												
1 (Johnson 2004)	507	91 (72 to 99)	98 (94 to 100)	53.87 (13.55 to 214.11)	0.09 (0.02 to 0.33)	Low	Prospective case series	Serious[2]	No serious	Serious[8]	No serious	None

1. HRT use status of participants not clearly reported at enrolment/or a significant proportion of participants were on HRT use at enrolment;
2. Selection bias as no clear methods are described in the recruitment of sample;
3. Evidence was downgraded by 2 due to very serious heterogeneity (chi-squared p<0.1, I-squared inconsistency statistic of >74.99%)
4. All irregular menses defined as perimenopause (Chompootweep et al 1993)
5. Evidence was downgraded by 1 due to serious heterogeneity (chi-squared p<0.1, I-squared inconsistency statistic of 50%-74.99%)
6. Perimenopause defined only as 3-11 months amenorrhoea, not including irregular menstruation (Ho et al 1999)
7. Evidence was downgraded by 2 due to 95% confidence interval for positive likelihood ratio ranges from not useful (<5) to very useful (>10)
8. Only women with suspected myocardial ischaemia and without hysterectomy included (Johnson et al 2004)

I.2 Information and advice

Table 7: Areas of information needs for women in menopause (summary of findings and quality assessment of qualitative evidence, *italics represent direct quotations of women. Non-italics represent field-workers' reporting of women's words*)

Studies	Summary of information needs	Quality of evicence
Mahon 2000 N = 161	N (%) who found knowing what tests to expect at menopause valuable N (%): 29 (19); and who wanted to know the definition of menopause: 11 (7) What does 'menopause' mean?	Very low quality[1]
Thewes 2003 N = 24 (Women with breast cancer history)	Questions which women thought were important on reflection after treatment Will my periods stop? How will that affect my life? How do I know if I'm menopausal or not? What tests diagnose menopause? How do I manage symptoms? What does 'menopause' mean? How will treatment affect my bone density? What does a hot flush feel like? Can I have children during menopause? What effect does menopause have on my body? Who do I talk to about sexuality issues?	Low quality[2]
Perceived menopause symptoms[1]		
Connelly 1999	Percentage of 114 women who wanted the following advice topics recommended to doctors: Topics which women felt should be included in guidelines for menopause counselling (ranked by popularity) %:	Moderate quality[3]

Studies	Summary of information needs	Quality of evicence
	Risk of breast cancer: 77 Medication: 73 Osteoporosis: 69 Prevention of heart disease: 58 Insomnia: 54 Living with medical uncertainty: 54 Genitourinary symptoms: 50	
Mahon 2000	N (%) of 161 women who found knowing the following valuable: Physical and emotional changes at menopause: 19 (12) Risk factors for heart disease: 10 (6)	Very low quality[1]
Mingo 2000 N = 165	Women felt they needed information on more than the 'core' symptoms of menopause (change in menstrual pattern, hot flushes, vaginal dryness, urinary incontinence). They would like HPs to give them information on memory loss, changes in skin, 'feeling blue', tender breasts, metallic taste, hot feet, burning head, mental lapses, formication ('bugs crawling'), chills, shape-changing, weight-gain, moodiness ('hating your husband'), change in libido and muscle pain (including waist).	Moderate quality
Thewes 2003. N=24 (Women with breast cancer history)	How do I manage symptoms? What does a hot flush feel like? What effect does menopause have on my body?	Low quality[2]
Hallowell 2000 N= 23 Women post-oophorectomy	Women needed to have known that their oestrogen would fluctuate and they might have menopausal symptoms following (surgical menopause) as none were told this.	Moderate quality
Alfred et al., 2006 N = 31	Women wanted information from their doctors about incontinence as it was embarrassing to bring it up.	Low quality[4]
Clinkingbeard 1999	Questions women wanted their HCP to answer: When will periods end with HRT? Why do I feel so lousy when I'm taking hormones? What does one believe with all the conflicting reports one hears? Will all my questions be answered?	Low quality[5]
HRT: benefits, risks and length of treatment[2]		
Alfred 2006 N = 31	Which treatments can be combined (e.g. complementary and conventional): 1=2 (0.5); 2=1 (0.2); 3=11 (2.7); 4=49 (12.0); 5=344 (84.5)	Low quality[4]
Fox-Young 1995 N = 148	Women needed information that was clear and not contradictory: *"You hear such divergent opinions"*	Very low quality[6]
Mahon 2000 N=161	N (%) of 161 women who found knowing the following valuable: Risks of HRT: 45 (71) Benefits of HRT: 54 (35) Expected tests at menopause: 29 (19) Risk factors for breast cancer: 24 (15)	Very low quality[1]
Mingo 2000	Women found it helpful to have a gynaecologist who gave information about coming off HRT. Some did not give information on discontinuing and some did.	Moderate quality
Thewes 2003 (Women with breast cancer history)	Women who had had total hysterectomies felt their doctors had not prepared them for menopause beforehand: *"I was very angry about the lack of preparation for the (menopausal) changes I experienced after my operation"*	Low quality[2]
Hallowell 2000 N=23 (post-oophorectomy only)	Women needed to have known how long to take HRT for (some HCPs did not know this). They would also like to have been informed of the likely cost of prescriptions for HRT as money was an issue and they had assumed it would be free. Although most women (with surgical menopause) were informed that they would have to take HRT following surgery, many said this was the only information they received:	Moderate quality

Studies	Summary of information needs	Quality of evicence
	"My information from the hospital was about the operation ...it just tells you what it does. That was it. It didn't say - it said a bit about, you will be given HRT, and that was it." Only 1 woman recalled being given a choice about the different forms of HRT.	
Roberts 1991	37% of women wanting information would like to have known the long term effects of HRT, and 26% would have liked information about the optimal duration of therapy. When asked what worries about HRT they had (in an information-receiving context), 2% said weight gain. No other specific worries were mentioned.	Low quality[7]
Self-management strategies		
Armitage 2007 N = 413	Women wanted comprehensive information on self-management practices; alternative options; acknowledgement of therapy risks and referral to reliable sources.	Very low quality[8]
Doubova 2012	*"I learnt that we do not have to leave everything up to the doctor";* *"It is very important to start working with ourselves: taking care, exercising. (If) we are not aware of this we will always continue living for others."*	Moderate quality
Mingo 2000	23/155 (15%) of surveyed women thought self-management strategies were important to have known.	Moderate quality
Theroux 2007 N = 7	Information women thought important: Lifestyle changes they could make to manage symptoms, and facts that empowered them to make choices.	Moderate quality
Wathen 2006	A proportion of women cited “themselves” as their main source of information.	Moderate quality
Walter 2004	Women wanted the information to make the decision for themselves. A woman with local oestrogen implanted during oophorectomy had to delay decision-making by 6 months.	Low to moderate quality
Hallowell 2000 N = 23	Women wanted the information to make the decision for themselves.	Moderate quality
Non-hormonal treatments		
Alfred 2006 N = 31	Women wanted information from their doctors on 'natural' treatments.	Low quality[4]
Armitage 2007 N = 413 (does not add up to 100)	Relevance of the following information, n(%): 1 – 5 on Likert scale: Not important (1) - very important (5): Which treatments relate to which symptoms: 1=0 (0); 2=0 (0); 3=7 (1.7); 4=40 (9.9); 5=358 (88.4) How a therapy works: 1=3 (0.7); 2=5 (1.2); 3=32 (7.8); 4=99 (24.2); 5=270 (66.0) How long it takes to work: 1=2 (0.5); 2=6 (1.5); 3=41 (10.1); 4=122 (30.0); 5=235 (68.0) How long should I take the treatment after seeing results: 1=2 (0.5); 2=4 (1.0); 3=34 (8.3); 4=91 (22.2); 5=279 (68.0) Side-effects: 1=0 (0); 2=0 (0); 3=4 (1.0); 4=16 (3.9); 5=388 (95.1)	Very low quality[8]
Fertility		
Thewes 2003 (Women with breast cancer history) N = 24	Women wanted clarity about their fertility and menopause status following treatment: "There was no clear answer on anything." *"There was no clear answer on anything."* They wanted to know if tests could be performed to establish these parameters: *"Even if there are no answers to my questions, well then I want to read information which says at this stage we don't know x, y, z."* Fertility became a bigger issue for women over time (a year was mentioned). This was because the cancer took priority until it was abated. Women wanted doctors to take seriously their need for fertility and menopause information. They had experienced 'discord' with doctors over this issue. "Aggressive" and "blasé" were adjectives used. *"They (doctors) have their priorities in curing you but they just thought it (menopause/fertility) wasn't that important."* Women wanted menopause information prior to treatment.	Low quality[2]

1. A convenience sample was used, high attrition and outcomes were subjective.
2. 60% participation rate, and under-reporting of method, though data is rich.
3. Very well reported, with saturation value given.
4. No quotations in results, just summaries in bullet points.

5. Not many direct quotations from women, and no record of unreturned studies.
6. Very poor reporting of method. It was not clear how many researchers were involved in the data collection or analysis. No standardised analytical method was reported. In spite of the above limitation, thorough descriptions of women's views are reported.
7. Data were not rich, and analysis was unreliable.
8. Serious under-reporting of method.

Table 8: GRADE profile: Effectiveness of information provision methods: (quantitative outcomes)

Number of studies	Design	Risk of bias	Indirectness	Imprecision	Number of women		Effect	Quality
					Intervention	Control	Absolute	
Decision conflict score (higher scores reflect greater decision conflict)								
Becker 2009 (Women with disabilities)	Randomised trials	Serious [1]	No serious	No serious	Booklet N=86 Mean=2.14	Menopause guidebook N=90 Mean=1.99	- MD: 0.15 [-0.03, 0.33]	Moderate
Deschamps 2004	Randomised trials	Serious [1]	No serious	Serious due to non calculable MID [2]	Booklet N=56 Mean=1.9	Pharmacist N=49 Mean=2.0	- P > 0.05	Low
Legare 2008	Randomised trials	No serious	No serious	No serious	Booklet N=44 Mean=1.92	Control N=41 Mean=2.08	- MD: -0.16 [-0.41, 0.09]	Moderate
Murray 2001	Randomised trials	Serious [3]	No serious	No serious	Interactive multimedia programme & booklet N=102 Mean=2.5	Control N=102 Mean=2.8	- MD: -0.30 [-0.15, -0.45]	Moderate
Rothert 1997	Randomised trials	Serious [4]	No serious	No serious	Booklet N=89 Mean=3.0	Lecture with Q&A N=80 Mean=2.7	- MD: 0.30 [0.01, 0.59]	Moderate
Knowledge score (higher scores reflect greater knowledge)								
Becker 2009	Randomised trials	No serious	No serious	Very serious [5]	Booklet N=86 Mean=14.77	Control N=90 Mean=15.03	- MD: -0.26 [-1.27, 0.75]	Low
Kiatpongsan 2014	Randomised trials	Serious [6]	No serious	No serious	DVD & booklet N=188 Mean=63.3%	Control N=213 Mean=57.5%	- MD: 5.80 [2.37, 9.23]	Moderate
Legare 2008	Randomised trials	No serious	No serious	Serious [7]	Booklet N=44 Mean (improvement) =0.51	Control N=41 Mean (improvement) =0.86	- MD: -0.35 [-1.04, 0.34]	Moderate
Rostom 2002	Randomised trial	No serious	No serious	No serious	Computer programme N=25 Mean (improvement) =8.4	Audio-booklet N=26 Mean (improvement) =17.5	- MD: 9.10 [1.77, 16.43]	Moderate
Hunter 1999	Randomised trials	Serious[8,]	Serious [9]	Serious [7]	Educational programme (2 x 90 minute sessions) N=34 Mean knowledge score (10 multiple choice Qs): score=5.16	Control N=34 Mean knowledge score (10 multiple choice Qs): score=3.74	MD: 1.42 (0.39-2.45) P<0.01	Low

Number of studies	Design	Risk of bias	Indirectness	Imprecision	Number of women		Effect	Quality
					Intervention	Control	Absolute	
Liao 1998	Randomised trials	No serious	No serious	No serious	Education programme N=45 3 points: Baseline; 3 months; 15 months. Knowledge score: 2.58; 5.56; 5.19	Control N=41 Knowledge score: 3 points: Baseline; 3 months; 15 months: 2.71; 3.05; 3.03	Baseline MD: -0.13 (-0.95-0.69); 3 month MD: 2.51 (1.52-3.50); 15 month MD: 2.16 (1.32-3.00) P<0.001	Moderate
Quality of life score								
Forouhari 2010	Randomised trials	Serious [10]	serious [11]	Serious due to non calculable MID [2]	Intervention n = 31 Pre- course / 3 months post course: Study group 81.7 / 75.3 SD (within group change) = 6.4 T = 7.6	Control group n=31 Pre- course / 3 months post course: 74.8 / 75.8 SD (within group change) = 1.4 t=-3.7	P=0.001	Very low
Percent of intervention group who found an educational course helpful in experience of menopause :								
Hunter 1999	RCT with post follow up	Serious[8]	Serious [9]	Serious due to non calculable MID [2]	Health eduction intervention (2x90 minutes sessions) (N=34) aspects of menopause: 87	Control (n=34)	P < 0.01	Very low
Supllementary information; Studies with no intervention using results from a questionnaire								
Women who found doctors a very useful source about CAM alternatives to HRT (The remaining responses were: somewhat or not useful)								
Wathen 2006 N=20	Cross- sectional	Very serious[13]	No serious	Not calculable	38%	Percentage rounded up, so cannot produce fraction.	- -	Very low
Women who found other health professionals a very useful source about CAM alternatives to HRT								
Wathen 2006 N=	Cross- sectional	Very serious[13]	No serious	Not calculable	46%	Percentage rounded up, so cannot produce fraction.	- -	Very low
Women who found the Internet a very useful source about CAM alternatives to HRT								
Wathen 2006 N=20	Cross- sectional	Very serious[13]	No serious	Not calculable	47.5%	Percentage rounded up, so cannot produce fraction.	- -	Very low
Women who found magazines and other media very useful								
Wathen 2006 N=20	Cross- sectional	Very serious[13]	No serious	Not calculable	27%	Percentage rounded up, so cannot produce fraction.	- -	Very low

1. Under-reporting of intervention and survey methods
2. Unable to calculate 95% CI as MD and SD not reported therefore the confidence in the precision of results is compromised.
3. Possible bias from part-private funding. Subjective data collection. Non-blinded study.
4. Possible selection and performance bias as reporting unclear.
5. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (-/+0.5 times SD)
6. 42 participants lost to follow-up in the control arm, and 72 participants lost to follow-up in the intervention arm.
7. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (-/+0.5 times SD)
8. Although the attrition rate and sample-heterogeneity were low, there was potential bias in that the educational experience of the control group was unmeasured.
9. There was a risk of indirectness in how 'influence of programme' outcomes were reported. What did 'influence' mean in the contexts of both groups

10. Under-reporting of intervention and data-collection method (questionnaire was translated from English with no record of how this may have compromised the standardised version). Also, exclusion-criteria under-reported. Intervention not described sufficiently.
11. Study carried out in Iran
12. Not RCTs – and under-reporting of intervention
13. Cross sectional study; under-reporting of method or/and inclusion/exclusion criteria.

Table 9: Methods of presenting risk information to menopausal women (summary of supplementary descriptive information and quality assessment)

Studies	Preferences for how risk information is presented		Quality of evidence
Rating of graphic displays of risk information (SD)			
Fortin 2001 N=40	Bar graph: 4±1; Linegraph: 3.1±0.9; "100 faces" (visual Lickert): 2.4±1.5; Survival curves: 2.5±1.1 Thermometer chart: 2.6±1.1	Real time worksheet results prior to focus group discussion.	Low quality[1]
Time horizons			
Fortin 2001 N=40	First choice: 10-year: 23% 20-year: 58% Lifetime: 27% Second choice: 10-year: 12% 20-year :58% Lifetime: 27% No response 3%	Real time worksheet results prior to focus group discussion.	Low quality[1]
Multiple diseases, multiple time:			
Fortin 2001 N=40	1 disease over 3 time horizons: 53% 3 diseases over I time horizon 43% No response 5%	Real time worksheet results prior to focus group discussion.	Low quality[1]
Relative vs. absolute risk:			
Fortin 2001 N=25	Preference for graph: Absolute risk: 72% Relative risk: 28% Preference for text: Absolute risk: 65% Relative risk: 30% No response: 0% / 5%	Real time worksheet results prior to focus group discussion.	Low quality[1]

1. Results were inconsistently reported. No description of what 'worksheet' entailed re data collection.

Table 10: Information provision methods: (summary of supplementary qualitative information and quality assessment) (Italics represent direct quotations of women. Non-italics represent field-workers' reporting of women's words)

Studies	Healthcare professionals	Quality
Andrist 1998	One woman (who happened to be a professor of nursing) said that even academic HCPs feel confused because *"I notice that some people have very strong opinions on it when I've asked professional people."*	Very low quality[1]
Legare 2007 N = 40	Women were ambivalent regarding doctors as sources of information. Sometimes women were given all the information they needed from their physician, but they did not understand it. Women wanted information from doctors to be free from the doctor's own strong opinions. They wanted information to be "objective, reliable and credible".	Low[2] to moderate quality
Bravata 2002 N = 23	*"I would like the doctor to be strong one way or the other. Not to waver too much. So I think scientific data is important, but also the doctor should take a position."*	Low quality[2]
Wathen 2006	Medical sources were the most influential in terms of decision making, though women did consult a number of other sources including books, libraries, or local information sessions (n=9), media stores or the Internet (n=8). Some women found the medical perspective from a doctor troubling because of the many related diseases to consider e.g. heart, breast cancer and osteoporosis: *"Well, maybe we shouldn't be doing this… the breast cancer problems are minor compared to the other things that might develop if you didn't take it".*	Low[2] to moderate quality
Clickingbeard 1999 N = 668	68% of 668 women preferred their HCP to provide information. 36% of 668 women felt their questions were not answered by HCP. Reassurance was needed that: Male doctors were not seen as well-informed as female ones. Women did not appreciate denigrating comments such as *"It's not such a big deal*", and *"You're like an old chicken that's not laying eggs anymore."*	Low quality[2]
Walter 2002	The vast majority of women talked about…wanting an input into the decision-making: *"statistics on other people and just go from my own experience."*	Moderate quality
Thewes 2003 (women with breast cancer history)	Most women had been given information orally by their HCP which left them feeling 'bombarded' and 'overwhelmed' when it happened immediately after diagnosis.	Moderate quality
Studies	**Internet, TV, magazines**	
Legare 2007 N = 40	Internet not considered a useful source of information because women needed help to distinguish what information is science from information that is marketing (especially re internet). The sheer volume of information was confusing. Informal sources, and often the media, were not particularly helpful compared with medical sources and books etc.: *"I read things and I get frustrated when I hear things on the TV and then see it in the paper and it's twisted around or you don't get all, you never get all the facts"* 2/5 focus groups agreed they wanted a trustworthy website as a way of providing information.	Low[2] to moderate quality
Wathen 2006	The internet was seen as untrustworthy, inaccurate and contradictory: *"I did a few times go into the Internet but not knowing how reliable the sites were that I was looking at...and there's so much contradiction."*	Low[2] to moderate quality
Roberts 1991 N = 64	The largest proportion of women (61%) sourced information from the Media (TV, magazines, newspapers etc.), but women often find this inaccurate, and that doctors should be aware of what women are reading. In 3/6 focus groups a Women were affected by the WH1 from the TV News: *"If I stop taking oestrogen, because of the possibility after what I saw in the news report on the television last night"*	Very low quality[3]
Studies	**Other women (peers) as educators**	
Armitage 2007 N = 413	Good information includes *"personal accounts of women"*	Very low quality[1]
Doubova 2012 N = 121	Peer discussion was as a way of learning how to approach the menopause as it was information which women found empowering: *"I learnt that we do not have to leave everything up to the doctor".* Peer sessions motivated women to transmit acquired knowledge of menopause to others.	Moderate quality

Studies	Healthcare professionals	Quality
	"By myself, I would not know what to do. Hearing others, I have another perspective to do other things." On group-work: *"We get to know ourselves through others."*	
Mingo 2000	*"What's worked for us is that we tell our story to the rest. Then everyone opens up and builds trust and confidence. Then they realise that (friends) have the same problem, but they never talked about it. The thing is (non white) women are more submissive...we have many taboos. We haven't woken up."*	Low[2] to moderate quality

1. Under-reporting, and results do not quite answer the outcome-question.
2. Under-reporting
3. Data were not rich, and analysis was unreliable.

I.3 Managing short-term symptoms

I.3.1 Results for the outcomes of low mood, anxiety, musculoskeletal symptoms and frequency of sexual intercourse

Table 11: GRADE profile: Oestrogen versus no treatment/placebo for the outcomes of low mood, anxiety and musculoskeletal symptoms

Quality assessment							No of patients		Effect			
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Oestrogen	Placebo	Relative (95% CI)	Absolute	Quality	Importance
Anxiety: final score (follow-up mean 2 months; measured with: Hamilton Anxiety Score; Better indicated by lower values)												
1 (Thomson 1977)	Randomised trials	Very serious[1,2,3,4]	No serious	No serious	Very serious[5]	None	17	17	-	MD 0.2 higher (2.88 lower to 3.28 higher)	Very low	CRITICAL
Anxiety: change in scores from baseline, measured by Greene Scale (Estradiol 50 mcg/day), 13-wk (Better indicated by lower values)												
1 (Speroff 2003)	Randomised trials	Serious[2]	No serious	No serious	N/A	None	113 MD (CI): -2.56 (not reported)	108 MD (CI): -1.94 (not reported)	-	Significant difference p<0.002	Moderate	CRITICAL
Anxiety: change in scores from baseline, measured by Greene Scale (Estradiol 100 mcg/day), 13-wk (Better indicated by lower values)												
1 (Speroff 2003)	Randomised trials	Serious[2]	No serious	No serious	N/A	None	112 MD (CI): -2.86 (not reported)	108 MD (CI): -1.94 (not reported)	-	Significant difference p<0.002	Moderate	CRITICAL
Anxiety: prevalence of self-reported anxiety after intervention												

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Oestrogen	Placebo	Relative (95% CI)	Absolute		
1 (Hachul 2008)	Randomised trials	Very serious[1,2,3,6]	No serious	No serious	Very serious[10]	None	3/14 (21.4%)	7/19 (36.8%)	RR 0.58 (0.18 to 1.86)	155 fewer per 1000 (from 302 fewer to 317 more)	Very low	CRITICAL
Low mood: final score measured by various scales (Better indicated by lower values)												
2 (Schmidt 2000; Thomson 1977)	Randomised trials	Very serious[1,2,3]	Very serious[7]	No serious	Very serious[5]	None	33	35	-	SMD 0.54 lower (2.09 lower to 1.01 higher)	Very low	CRITICAL
Low mood: final score measured by Montagomery-Asberg scale, 4-wk (Better indicated by lower values)												
1 (De, NovaeSoares 2001)	Randomised trials	Serious[2]	No serious	No serious	Serious[8]	None	25	25	-	MD 2.08 lower (4.95 lower to 0.79 higher)	Low	CRITICAL
Low mood: final score measured by Montgomery-Asberg scale, 8-wk (Better indicated by lower values)												
1 (De, NovaeSoares 2001)	Randomised trials	Serious[2]	No serious	No serious	Serious[8]	None	25	25	-	MD 5.12 lower (7.97 to 2.27 lower)	Low	CRITICAL
Low mood: final score measured by Montgomery-Asberg scale, 12-wk (Better indicated by lower values)												
1 (De, NovaeSoares 2001)	Randomised trials	Serious[2]	No serious	No serious	No serious	None	25	25	-	MD 7.74 lower (10.89 to 4.59 lower)	Moderate	CRITICAL
Anxiety/low mood: mood changes measured by Women's Health Questionnaire (WHQ), 2-yr, Estradiol 150 mcg/d (Better indicated by lower values)												
1 (Nielsen 2006)	Randomised trials	Serious[1,2]	No serious	No serious	No serious	None	114	118	-	MD 1.1 higher (1.92 lower to 4.12 higher)	Moderate	CRITICAL

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	**Design**	**Risk of bias**	**Inconsistency**	**Indirectness**	**Imprecision**	**Other considerations**	**Oestrogen**	**Placebo**	**Relative (95% CI)**	**Absolute**		
Anxiety/low mood: mood changes measured by WHQ, 2-yr, Estradiol 300 mcg/d (Better indicated by lower values)												
1 (Nielsen 2006)	Randomised trials	Serious[1,2]	No serious	No serious	Serious[8]	None	103	118	-	MD 3.5 higher (0.5 to 6.5 higher)	Low	CRITICAL
Low mood: mean changes measured by Hamilton Low mood scale, 8-wk (Better indicated by lower values)												
1 (Morrison 2004)	Randomised trials	Serious[1,2]	No serious	No serious	Serious[8]	None	31	26	-	MD 2.4 higher (0.17 to 4.63 higher)	Low	CRITICAL
Low mood: mean changes measured by Centre Epi studies Low mood scale, 8-wk (Better indicated by lower values)												
1 (Morrison 2004)	Randomised trials	Serious[1,2]	No serious	No serious	Serious[8]	None	31	26	-	MD 2.4 higher (0.97 lower to 5.77 higher)	Low	CRITICAL
Low mood: mean changes from baseline measured by Greene Scale, 13-wk, Estradiol 50 mcg/d (Better indicated by lower values)												
1 (Speroff 2003)	Randomised trials	Serious[2]	No serious	No serious	N/A	None	113 (MD: -2.10)	108 (MD: -0.97)	-	Significant difference p<0.002	Moderate	CRITICAL
Low mood: mean changes from baseline measured by Greene scale, 13-wk, Estradiol 100 mcg/d (Better indicated by lower values)												
1 (Speroff 2003)	Randomised trials	Serious[2]	No serious	No serious	N/A	None	113 (MD: -1.88)	108 (MD: -0.97)	-	Significant difference p<0.002	Moderate	CRITICAL
Low mood: prevalence of low mood after intervention												
1 (Hachul 2008)	Randomised trials	Very serious[1,2,3,4]	No serious	No serious	Very serious[10]	None	8/14 (57.1%)	13/19 (68.4%)	RR 0.84 (0.48 to 1.44)	109 fewer per 1000 (from 356 fewer to 301 more)	Very low	CRITICAL
Risk of musculoskeletal symptoms: among those without joint pain present at baseline, 1-year follow-up												

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Oestrogen	Placebo	Relative (95% CI)	Absolute		
1 (Brunner 2010)	Randomised trials	Serious[9]	No serious	No serious	No serious	None	522/3261 (16%)	596/3,333 (17.9%)	RR 0.91 (0.81 to 1.01)	3 fewer per 100 (from 7 to 0.1 fewer)	Moderate	CRITICAL
Risk of musculoskeletal symptoms: among those with joint pain present at baseline, 1-year follow-up												
1 (Brunner 2010)	Randomised trials	Serious[9]	No serious	No serious	No serious	None	968/1,467 (66%)	1028/1,520 (37.6)	RR 0.98 (0.93 to 1.03)	2 fewer per 100 (from 6 to 1.9 fewer)	Moderate	CRITICAL

1. Unclear how randomisation was performed
2. Unclear how concealment of allocation was conducted
3. Unclear how double-blinding was conducted
4. Unclear whether the two groups were comparable at baseline
5. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed2 default MIDs (-/+0.5 times SD)
6. Detection bias: self-reported outcome (complaints about anxiety);
7. Evidence was downgraded by 2 due to very serious heterogeneity (chi-squared p<0.1, I-squared inconsistency statistic of >75%)
8. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (-/+0.5 times SD)
9. High attrition bias: about 40% of women in the intervention and 38% of women in the placebo group stopped taking the study drugs during follow-up
10. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)

Table 12: GRADE profile: Oestrogen plus progestogen versus no treatment/placebo for the outcomes of low mood and anxiety

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Oestrogen and Progestogen	Placebo/no treatment	Relative (95% CI)	Absolute		
Low mood: final scores (measured with: 4 different scales across studies [1]; Better indicated by lower values)												
5 (Derman 1995, Purdie 1995, Rudolph 2004, Veerus 2008,	Randomised trials	Very serious[2]	Very serious[3]	No serious	Serious[4]	None	852	839	-	SMD 0.35 lower (0.66 to 0.44 lower)	Very low	CRITICAL

Quality assessment							No of patients		Effect			
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Oestrogen and Progestogen	Placebo/no treatment	Relative (95% CI)	Absolute	Quality	Importance
Polisseni 2013)												
Low mood: change scores (measured with: HAMD; 24-wk, Better indicated by lower values)												
1 (Rudolph 2004)	Randomised trials	No serious	No serious	No serious	Serious[4]	None	64	64	-	MD -3.30 lower (5.72 lower to 0.88lower)	Moderate	CRITICAL
Anxiety: final scores (measured with: WHQ (2 studies) and CCEI; Better indicated by lower values)												
3 (Veerus 2008, Polisseni 2013, Purdie 1995)	Randomised trials	Very serious[2]	No serious	No serious	No serious	None	747	733	-	SMD 0.01 lower (0.11 lower to 0.09 higher)	Low	CRITICAL
Anxiety: change scores (measured with: Greene scale; 1 year, Better indicated by lower values)												
1 (Geller 2009)	Randomised trials	Serious[5]	No serious	No serious	No serious	None	23	21	-	Difference in mean reduction in both groups, p= 0.29	Moderate	CRITICAL

1. Scales used: Beck Depression Inventory, Crown - Crisp experiential index (CCEI), The Hamilton Rating Scale for Depression (HAMD), The Women's Health Questionnaire (WHQ)
2. The highest weighted study, Veerus 2008, did not report randomisation process and blinding was broken.
3. Evidence was downgraded by 2 due to very serious heterogeneity (chi-squared p<0.1, I-squared inconsistency statistic of >75%)
4. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (-/+0.5 times SD)
5. Allocation concealment unclear

Table 13: GRADE profile: HRT (tibolone) versus no treatment/placebo for the outcomes of low mood and anxiety

Quality assessment							No of patients		Effect			
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Tibolone	Placebo/no treatment	Relative (95% CI)	Absolute	Quality	Importance
Low mood: final scores (measured with: WHQ scale; 1 year, Better indicated by lower values)												

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Tibolone	Placebo/no treatment	Relative (95% CI)	Absolute		
1 (Polissen i 2013)	randomised trials	no serious	no serious	no serious	very serious[1]	none	42	44	-	MD 0.42 lower (2.22 lower to 1.38 higher)	Low	CRITICAL
Anxiety: final scores (measured with: WHQ scale; 1 year, Better indicated by lower values)												
1 (Polissen i 2013)	randomised trials	no serious	no serious	no serious	no serious	none	42	44	-	MD 0.06 higher (-1.01 lower to 1.13 higher)	High	CRITICAL

1. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence intervals crossed 2 default MIDs (-/+0.5 times SD)

Table 14: GRADE profile: Testosterone versus no treatment/placebo for the outcomes frequency of sexual activity and low mood

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Testosterone	Placebo/no treatment	Relative (95% CI)	Absolute		
Frequency of satisfying sexual activity: final frequency at endpoint (24 week, Better indicated by higher values)												
1 (Simon 2005)	randomised trials	very serious[1,2]	no serious	no serious	no serious	none	283	279	-	MD 1.00 higher (0.17 to 1.83 higher)	Low	CRITICAL
Frequency of satisfying sexual activity (4 week, Better indicated by higher values)												
1 (Davis 2008)	randomised trials	serious[3]	no serious	no serious	no serious	none	254	265	-	Increase of 2.1 episodes vs 0.7, p<0.001	Moderate	CRITICAL
Low mood: final score (measured with: PGWB; Better indicated by lower values)												
1 (Nathorst-Boos 2006)	randomised trials	very serious[4,5]	no serious	no serious	no serious	none	27	26	-	p = 0.382	Low	CRITICAL

1. Allocation concealment not reported
2. Detection bias
3. Randomisation method unclear
4. Allocation concealment unclear
5. Comparability of groups at baseline unclear

Table 15: GRADE profile: Tibolone versus CEE plus MPA for the outcomes of low mood and anxiety

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Tibolone (2.5mg)	CEE plus MPA (Oestrogen plus Progestogen)	Relative (95% CI)	Absolute		
Anxiety (measured with: Greene scale change score at 3 months; Better indicated by lower values)												
1 (Wu 2001)	randomised trials	very serious[1,2]	no serious	serious[3]	serious[4]	none	18	18	-	MD 0.39 lower (1.27 lower to 0.49 higher)	Very low	CRITICAL
Low mood (measured with: Greene scale change score at 3 months; Better indicated by lower values)												
1 (Wu 2001)	randomised trials	very serious[1,2]	no serious	serious[3]	serious[4]	none	18	18	-	MD 0.78 lower (1.76 lower to 0.2 higher)	Very low	CRITICAL

1. single-blind
2. allocation concealment unclear
3. study used Taiwanese women only
4. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (-/+0.5 times SD)

Table 16: GRADE profile: CEE plus MPA versus oestrogen plus progestogen (E2/NETA) for the outcome of low mood

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	CEE/MPA	E2/NETA	Relative (95% CI)	Absolute		
Low mood (measured with: Cyclicity Diagnoser Scale, final score at 1 month; Better indicated by lower values)												
1 (Odmark 2004)	randomised trials	no serious	no serious	no serious	no serious	none	123	123	-	MD 0.2 lower (0.25 to 0.15 lower)	High	CRITICAL

Table 17: GRADE profile: SSRI (non-hormonal pharmaceutical treatment) versus oestrogen/progestogen (hormonal treatment) for the outcome of low mood

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	SSRI-Escitalopram	Oestrogen/Progestogen	Relative (95% CI)	Absolute		
Low mood (measured with: Montgomery-Asberg Low mood Rating Scale, change at 8 week; Better indicated by lower values)												
1 (Soares 2006)	randomised trials	very serious [1]	no serious	no serious	no serious	none	16	16	-	Median decline of 19.2 in SSRI group compared with 9.4 in oestrogen + progestogen (p = 0.03)	Low	CRITICAL

1. open label study-no concealment or blinding

Table 18: GRADE profile: SNRI versus SSRI for the outcome of low mood and anxiety

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	SNRI-desvenlafaxine	SSRI-escitalopram	Relative (95% CI)	Absolute		
Anxiety (measured with: HAM-A, change at 8 months; Better indicated by lower values)												
1 (Soares 2010)	randomised trials	serious[1]	no serious	no serious	no serious	none	110	124	-	MD 0.08 lower (1.94 lower to 1.78 higher)3	Moderate	CRITICAL
Low mood (measured with: HAMD, change at 8 months; Better indicated by lower values)												
1 (Soares 2010)	randomised trials	serious[1]	no serious	no serious	serious[2]	none	110	124	-	MD 0.94 lower (2.29 lower to 0.41 higher)[3]	Low	CRITICAL

1. Groups contained both blinded and open-labelled participants
2. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (-/+0.5 times SD)
3. from mixed effects model

Table 19: GRADE profile: Tibolone versus oestrogen plus progestogen (E2/NETA) for the outcome of sexual activity

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Tibolone	Oestrogen/Progestogen (E2/NETA)	Relative (95% CI)	Absolute		
Total sexual activity (measured with: daily diary (in 4 weeks); Better indicated by higher values)												
1 (Nijland 2008)	randomised trials	very serious[1,2]	no serious	no serious	no serious	none	199	201	-	Mean change from baseline: Tibolone: 0.66, E2/NETA: 5.6, p-value = not significant	Low	CRITICAL

1. Allocation concealment unclear
2. Attrition bias unclear

Table 20: GRADE profile: Tibolone versus oestradiol for the outcome of anxiety

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Tibolone	Estradiol	Relative (95% CI)	Absolute		
Anxiety: final scores (measured with: Greene; Better indicated by lower values)												
1 (Somunkiran 2007)	randomised trials	very serious[1,2]	no serious	no serious	Serious[3]	none	20	20	-	MD 0.57 lower (1.20 lower to 0.06 higher)	Very low	CRITICAL

1. Allocation concealment unclear
2. Single blinded study
3. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (-/+0.5 times SD)

Table 21: GRADE profile: Herbal versus oestradiol plus progesterone treatment for the outcomes of low mood and anxiety

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Black cohosh	Oestradiol plus progesterone	Relative (95% CI)	Absolute		
Anxiety (final) (follow-up mean 3 months[1]; measured with: HADS scale (lower is better); range of scores: 0-42; Better indicated by lower values)												
1 (Zheng 2013)	randomised trials	serious[2]	no serious	no serious	serious[3]	none	31	30	-	MD 0.58 lower (2.16 lower to 1 higher)	Low	CRITICAL
Low mood (final) (follow-up mean 3 months[1]; measured with: HAD score; range of scores: 0-42; Better indicated by lower values)												
1 (Zheng 2013)	randomised trials	serious[2]	no serious	no serious	serious[3]	none	31	30	-	MD 0.13 higher (1.47 lower to 1.73 higher)	Low	CRITICAL

1. Data was reported after treatment of three months
2. Risk of bias was high across all domains
3. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (-/+0.5 times SD)

Table 22: GRADE profile: Herbal versus oestradiol plus MPA treatment for the outcomes of low mood and anxiety

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Black cohosh	Oestradiol plus MPA	Relative (95% CI)	Absolute		
Anxiety (final) (follow-up mean 3 months[2]; measured with: HAD score; range of scores: 0-42; Better indicated by lower values)												
1(Zheng 2013)	randomised trials	serious[3]	no serious	no serious	serious[1]	none	31	28	-	MD 0.37 lower (1.97 lower to 1.23 higher)	Low	CRITICAL
Low mood (final) (follow-up mean 3 months[2]; measured with: HAD score; range of scores: 0-42; Better indicated by lower values)												
1(Zheng 2013)	randomised trials	serious[3]	no serious	no serious	serious[1]	none	31	28	-	MD 0.62 lower (2.43 lower to 1.19 higher)	Low	CRITICAL

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed1 default MID (-/+0.5 times SD)
2. Data reported after three months of treatment
3. Risk of bias was high across all domains

Table 23: GRADE profile: Oestradiol plus progesterone versus oestradiol plus MPA treatment for low mood and anxiety symptoms

Quality assessment							No of patients		Effect			
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Oestradiol plus Progesterone	Oestradiol plus MPA	Relative (95% CI)	Absolute	Quality	Importance
Anxiety (final) (follow-up mean 3 months[1]; measured with: HAD score; range of scores: 0-42; Better indicated by lower values)												
1 (Zheng 2013)	Randomised trials	Serious[2]	No serious	No serious	Very serious[3]	none	30	28	-	MD 0.21 higher (1.4 lower to 1.82 higher)	Very low	CRITICAL
Low mood (final) (follow-up mean 3 months[1]; measured with: HAD score; range of scores: 0-42; Better indicated by lower values)												
1 (Zheng 2013)	Randomised trials	Serious[2]	No serious	No serious	Serious[4]	none	30	28	-	MD 0.75 lower (2.56 lower to 1.06 higher)	Low	CRITICAL

1. Data reported after three months of treatment
2. Risk of bias was high across all domains
3. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (-/+0.5 times SD)
4. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (-/+0.5 times SD)

Table 24: GRADE profile: Herbal treatment versus placebo for the outcomes of low mood and anxiety

Quality assessment							No of patients		Effect			
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Herbal treatment	Placebo	Relative (95% CI)	Absolute	Quality	Importance
Anxiety (final scores at endpoint 16 weeks) Ginseng/black cohosh/pycogneal (measured with: PGWB, HAMA, Greene Climacteric scale for anxiety; Better indicated by lower values)												
3 (Wiklund 1999, van Die 2009, Yang 2007)	randomised trials	serious[1]	very serious[2]	no serious	very serious[3]	none	330	316	-	SMD 0.93 higher (0.01 higher to 1.86 higher)	Very low	CRITICAL
Anxiety (change scores at endpoint 12 to 16 weeks) Ginseng/black cohosh/St. John's Wort plus Chaste (measured with: PWGB, HAMD, Greene Climacteric scale; Better indicated by lower values)												
3 (Amsterdam 2009, van Die 2009, Wiklund 1999)	randomised trials	no serious	very serious[2]	no serious	very serious[3]	none	258	254	-	SMD 0.48 lower (1.57 lower to 0.62 higher)	Very low	CRITICAL
Anxiety (mean reduction difference at endpoint 12 months) Black cohosh (measured with: Greene Climacteric scale for anxiety; Better indicated by lower values)												
1 (Geller 2009)	randomised trials	serious[1]	no serious	no serious	no serious	none	21	21	-	MD 0.47 (0.81)	Moderate	CRITICAL

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	**Design**	**Risk of bias**	**Inconsistency**	**Indirectness**	**Imprecision**	**Other considerations**	**Herbal treatment**	**Placebo**	**Relative (95% CI)**	**Absolute**		
Low mood (final scores at endpoint 16 weeks) Ginseng/black cohosh/black cohosh plus St. John's Wort/pycogneal (measured with: WHQ, Greene Climacteric scale, HAM-D; Better indicated by lower values)												
4 (Wiklund 1999, van Die 2009, Uebelhack 2006, Yang 2007)	randomised trials	serious[1]	very serious[2]	no serious	very serious[3]	none	474	459	-	SMD 0.16 higher (0.88 lower to 1.2 higher)	Very low	CRITICAL
Low mood (change scores at endpoint 12 to 16 weeks) Ginseng/black cohosh/black cohoshplus St. John's Wort/St. John's Wort plus Chaste (measured with: WHQ, Greene Climacteric scale, HAM-D; Better indicated by lower values)												
4 (Amsterdam 2009, Uebelhack 2006, van Die 2009, Wiklund 1999)	randomised trials	serious[1]	very serious[2]	no serious	serious[4]	none	409	397	-	SMD 0.39 lower (1.13 lower to 0.36 higher)	Very low	CRITICAL

1. Risk of bias due to selection and performance
2. Evidence was downgraded by 2 due to very serious heterogeneity (chi-squared p<0.1, I-squared inconsistency statistic of >75%)
3. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (-/+0.5 times SD)
4. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (-/+0.5 times SD)

Table 25: GRADE profile: Phytoestrogen versus placebo for the outcome of low mood and anxiety

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	**Design**	**Risk of bias**	**Inconsistency**	**Indirectness**	**Imprecision**	**Other considerations**	**Phytoestrogens**	**Placebo**	**Relative (95% CI)**	**Absolute**		
Low mood (final scores at endpoint 12-16 weeks) Genistein/isoflavones (measured with: Greene Climacteric scale, CES-D scale; Better indicated by lower values)												
2 studies (Evans 2011, de Sousa-Munoz 2009)	randomised trials	serious[1]	no serious	no serious	serious[2]	none	82	84	-	SMD -0.23 lower (-0.54 lower to 0.07 higher)	Low	CRITICAL

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Phytoestrogens	Placebo	Relative (95% CI)	Absolute		
Low mood (change scores at endpoint 12 weeks) Promensil (measured with: Greene Climacteric scale; Better indicated by lower values)												
1 study (Tice 2003)	randomised trials	serious[1]	no serious	no serious	no serious	none	84	85	-	MD 0.4 lower (1.1 lower to 0.2 higher)	Moderate	CRITICAL
Low mood (change scores at endpoint 12 weeks) Rimostil (measured with: Greene Climacteric scale; Better indicated by lower values)												
1 study (Tice 2003)	randomised trials	serious[1]	no serious	no serious	no serious	none	83	85	-	MD 0.1 lower (0.9 lower to 0.7 higher)	Moderate	CRITICAL
Anxiety (change scores at endpoint 12 weeks) Promensil (measured with: Greene Climacteric scale for anxiety; Better indicated by lower values)												
1 study (Tice 2003)	randomised trials	serious[1]	no serious	no serious	very serious[3]	none	83	85	-	MD 1.1 lower (1.6 lower to 0.6 higher)	Very low	CRITICAL
Anxiety (change scores at endpoint 12 weeks) Rimostil (measured with: Greene Climacteric scale for anxiety; Better indicated by lower values)												
1 study (Tice 2003)	randomised trials	serious[1]	no serious	no serious	serious[2]	none	82	85	-	MD 0.8 lower (1.3 lower to 0.3 higher)	Low	CRITICAL
Anxiety (final scores at endpoint 12 weeks) Genistein (measured with: Greene Climacteric scale for anxiety; Better indicated by lower values)												
1 study (Evans 2011)	randomised trials	no serious	no serious	no serious	serious[2]	none	42	42	-	MD 1.32 lower (2.54 to 0.1 lower)	Moderate	CRITICAL
Anxiety (mean reduction difference at endpoint 12 months) Red clover (measured with: Greene Climacteric scale ; Better indicated by lower values)												
1 study (Geller 2009)	randomised trials	serious[4]	no serious	no serious	no serious	none	22	21	-	MD 1.64 (0.8)	Moderate	CRITICAL

1. Risk of bias due to unclear selection and performance
2. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (-/+0.5 times SD)
3. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (-/+0.5 times SD)
4. Risk of bias due to unclear allocation concealment

Table 26: GRADE profile: Acupuncture versus sham acupuncture for the outcome of low mood

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Acupuncture	Sham acupuncture	Relative (95% CI)	Absolute		
Low mood (measured with: CESD Scale, median change scores at 8 weeks; Better indicated by lower values)												
1 (Bao 2014)	randomised trials	serious[1]	no serious	no serious	no serious	none	24	23	-	P= 0.442	Moderate	CRITICAL

1. No adequate concealment

Table 27: GRADE profile: Citalopram versus placebo for the outcome of anxiety and low mood

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Citalopram	Placebo	Relative (95% CI)	Absolute		
Anxiety (measured with: mean change scores with Profile of Mood Scale-anxiety; 6 week, Better indicated by higher values)-10mg												
1 (Barton 2010)	randomised trials	no serious	no serious	no serious	serious[1]	none	54 Mean change score=5.8	28 Mean change score=3.3	-	-	Moderate	CRITICAL
Anxiety (measured with: mean change scores with Profile of Mood Scale-anxiety; 6 week, Better indicated by higher values)-20 mg												
1 (Barton 2010)	randomised trials	no serious	no serious	no serious	serious[1]	none	56 Mean change score=12.9	27 Mean change score=3.3	P <0.01	-	Moderate	CRITICAL
Anxiety (measured with: mean change scores with Profile of Mood Scale-anxiety; 6 week, Better indicated by higher values)-30 mg												
1 (Barton 2010)	randomised trials	no serious	no serious	no serious	serious[1]	none	55 Mean change score=4.1	28 Mean change score=3.3	-	-	Moderate	CRITICAL
Low mood (measured with: mean change scores with Profile of Mood Scale-low mood; 6 week, Better indicated by higher values)-10mg												
1 (Barton 2010)	randomised trials	no serious risk of bias	no serious	no serious	serious[1]	none	54 Mean change score=6.0	28 Mean change score=-0.1	-	-	Moderate	CRITICAL
Low mood (measured with: mean change scores with Profile of Mood Scale-low mood; 6 week, Better indicated by higher values)-20 mg												
1 (Barton 2010)	randomised trials	no serious	no serious	no serious	serious[1]	none	56 Mean change score=5.2	27 Mean change score=-0.1	-	-	Moderate	CRITICAL

Quality assessment							No of patients		Effect			
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Citalopram	Placebo	Relative (95% CI)	Absolute	Quality	Importance
Low mood (measured with: mean change scores with Profile of Mood Scale-low mood; 6 week, Better indicated by higher values)-30mg												
1 (Barton 2010)	randomised trials	no serious	no serious	no serious	serious[1]	none	55 Mean change score=6.5	28 Mean change score=-0.1	-	-	Moderate	CRITICAL

1. N/A-SD not reported so magnitude of the effect was unclear

Table 28: GRADE profile: Sertraline versus placebo for the outcome of low mood

Quality assessment							No of patients		Effect			
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Sertraline	Placebo	Relative (95% CI)	Absolute	Quality	Importance
Low mood (measured with: Final CESD score at 6 week; Better indicated by lower values)												
1 (Kimmick 2006)	randomised trials	very serious[1]	no serious	no serious	very serious[2]	none	25	22	-	MD 0.5 higher (4.02 lower to 5.02 higher)	Very low	CRITICAL

1. Unclear selection, attrition and detection bias
2. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (-/+0.5 times SD)

Table 29: GRADE profile: Gabapentin versus placebo for the outcome of anxiety

Quality assessment							No of patients		Effect			
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Gabapentin	Placebo	Relative (95% CI)	Absolute	Quality	Importance
Anxiety (measured with: Profile of Mood Scale-anxiety change scores at 12 week; Better indicated by higher values)												
1 (Guttuso 2003)	randomised trials	no serious	no serious	no serious	serious[1]	none	30	29	-	MD 1.7 lower (4.32 lower to 0.92 higher)	Moderate	CRITICAL

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (-/+0.5 times SD)

Table 30: GRADE profile: Psychological treatments versus usual care for the outcomes of low mood and anxiety

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	CBT	Usual Care	Relative (95% CI)	Absolute		
Anxiety (Final scores) (follow-up mean 26 weeks; measured with: WHQ (anxiety); range of scores: 0-1; Better indicated by lower values)												
1 (Mann 2012)	randomised trials	no serious	no serious i	no serious	serious[1]	none	43	45	-	MD 0.15 lower (0.24 to 0.06 lower)	Moderate	CRITICAL
Low mood (Final scores) (follow-up mean 26 weeks; measured with: WHQ (low mood); range of scores: 0-1; Better indicated by lower values)												
1 (Mann 2012)	randomised trials	no serious	no serious	no serious	serious[1]	none	43	45	-	MD 0.15 lower (0.28 to 0.02 lower)	Moderate	CRITICAL

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed one default MID (-/+0.5 times SD)

I.3.2 Results on pair-wise comparisons for studies excluded from the NMA for purely statistical reasons

I.3.2.1 Women without a uterus

Table 31: GRADE profile: Gabapentin versus placebo for the outcome of vaginal bleeding

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Gabapentin	Placebo	Relative (95% CI)	Absolute		
Vaginal bleeding (follow-up mean 17 weeks)												
Guttuso 2003	randomised trials	very serious[1]	no serious	no serious	very serious[2]	none	2/30 (6.7%)	3/29 (10.3%)	RR 0.64 (0.12 to 3.58)	37 fewer per 1000 (from 91 fewer to 267 more)	Very low	CRITICAL

1. Very high risk due to selection, performance, attrition and detection bias
2. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)

Table 32: GRADE profile: Acupuncture versus sham acupuncture for the outcome of vaginal bleeding

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Acupuncture	Sham acupuncture	Relative (95% CI)	Absolute		
Vaginal bleeding (follow-up mean 7 weeks)												
Nir 2006	randomised trials	very serious[1]	no serious	no serious	no serious	none	8/12 (66.7%)	1/17 (5.9%)	RR 11.33 (1.62 to 79.11)	608 more per 1000 (from 36 more to 1000 more)	Low	CRITICAL

1. Very high risk due to selection, performance, attrition and detection bias

I.3.2.2 Women with a uterus

Table 33: GRADE profile: Gabapentin versus placebo for the outcome of vaginal bleeding

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Gabapentin	Placebo	Relative (95% CI)	Absolute		
Vaginal bleeding (follow-up mean 17 weeks)												
Guttuso 2003	randomised trials	very serious	no serious i	no serious	very serious[1]	none	2/30 (6.7%)	3/29 (10.3%)	RR 0.64 (0.12 to 3.58)	37 fewer per 1000 (from 91 fewer to 267 more)	Very low	CRITICAL

1. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)

Table 34: GRADE profile: Acupuncture versus sham acupuncture for the outcome of vaginal bleeding

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Acupuncture	Sham acupuncture	Relative (95% CI)	Absolute		
Vaginal bleeding												
Nir 2006	randomised trials	very serious[1]	no serious	no serious	no serious	none	8/12 (66.7%)	1/17 (5.9%)	RR 11.33 (1.62 to 79.11)	608 more per 1000 (from 36 more to 1000 more)	Low	CRITICAL

1. Very high risk due to selection, performance, attrition and detection bias

Table 35: GRADE profile: 17β-oestradiol 0.5mg plus dydrogesterone 2.5mg versus placebo for the outcome of vaginal bleeding

Quality assessment							No of patients		Effect			
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Oestradiol 0.5mg plus dydrogesterone 2.5mg	Placebo	Relative (95% CI)	Absolute	Quality	Importance
Vaginal bleeding (follow-up mean 13 weeks[1])												
Stevenson 2010	randomised trials	serious[2]	no serious	no serious	very serious[3]	none	0/122 (0%)	4/124 (3.2%)	RR 0.11 (0.01 to 2.08)	29 fewer per 1000 (from 32 fewer to 35 more)	Very low	CRITICAL

1. Bleeding or spotting was reported at any time during the study
2. Risk due to attrition and detection bias
3. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)

Table 36: GRADE profile: 17β-oestradiol 1mg plus dydrogesterone 5mg versus placebo for the outcome of vaginal bleeding

Quality assessment							No of patients		Effect			
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Oestradiol 1mg plus dydrogesterone 5mg	Placebo	Relative (95% CI)	Absolute	Quality	Importance
Vaginal bleeding (follow-up mean 13 weeks[1])												
Stevenson 2010	randomised trials	serious[2]	no serious	no serious	very serious[3]	none	5/59 (8.5%)	4/124 (3.2%)	RR 2.63 (0.73 to 9.43)	53 more per 1000 (from 9 fewer to 272 more)	Very low	CRITICAL

1. Bleeding and spotting was reported at any time during the study
2. Risk due to attrition and detection bias
3. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)

I.3.3 Urogenital atrophy

Table 37: GRADE profile: local oestrogens versus placebo for the outcomes of decrease in vaginal dryness, maturation index, symptom improvement, assessment of endometrial stimulation, breast pain, adverse events, treatment withdrawal, treatment adherence, treatment acceptability, and health related quality of life at 12 weeks for short term symptoms

Quality assessment							Number of patients		Effect		
Number of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Intervention	Comparator	Relative (95% CI)	Absolute (95% CI)	Quality
Decrease in vaginal pH (better indicated by lower value) (12 weeks)											
4 (Cano 2012; Karp 2012; Griesser 2012; Dessole 2004)	randomised trials	Serious[1]	no serious	no serious	no serious	none	293	169	-	MD 0.95 lower (1.19 lower to 0.71 lower)	Moderate
Maturation index (better indicated by higher value) (12 weeks)											
5 (Bachmann 2008; Cano 2012; Karp 2012; Griesser 2012; Dessole 2004)	randomised trials	Serious[1]	very serious[2]	no serious	no serious	none	436	205	-	MD 17.73 higher (7.66 higher to 27.00 higher)	Very low
Patient assessment of symptom improvement at 12 weeks											
4 (Eriksen 1992; Griesser 2012; Casper 1999)	randomised trial	Serious[1]	Serious[3]	no serious	no serious	none	123/270 (45.6%)	47/210 (22.4%)	RR 2.23 (1.4 to 3.57)	275 more per 1000 (from 90 more to 575 more)	Low
Assessment of endometrial stimulation at 12 weeks											
2 (Bachmann 2008; Simon 2008)	randomised trials	Serious[4]	no serious	no serious	no serious	none	2/257 (0.78%)	0/122 (0%)	RR 1.28 (0.14 to 12.08)	NC	Moderate
Breast pain at 12 weeks											
1 (Cano 2012)	randomised trials	Serious[1]	no serious	no serious	no serious	none	0/114 (0%)	1/53 (1.9%)	RR 0.16 (0.01 to 3.78)	16 fewer per 1000 (from 19 fewer to 52 more)	Moderate
Adverse events at 12 weeks											
2 (Cano 2012; Eriksen 1992)	randomised trial	Serious[1]	no serious	no serious	no serious	none	64/189 (33.9%)	35/132 (26.5%)	RR 1.09 (0.77 to 1.53)	24 more per 1000 (from 61 fewer to 141 more)	Moderate

Quality assessment							Number of patients		Effect		
Number of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Intervention	Comparator	Relative (95% CI)	Absolute (95% CI)	Quality
Withdrawal due to adverse events at 12 weeks											
8 (Bachmann 2008; Bachmann 2009; Cano 2012; Casper 1999; Dessole 2004; Griesser 2012; Simon 2008; Eriksen 1992)	Randomised trials	Serious[1]	no serious	no serious	no serious	none	42/995 (4.2%)	21/658 (3.2%)	RR 1.23 (0.72 to 2.11)	7 more per 1000 (from 9 fewer to 35 more)	Moderate
Treatment adherence at 12 weeks											
1 (Karp 2012)	Randomised trials	Serious[4]	no serious	no serious indirectness	no serious	none	19/22 (86.4%)	18/21 (85.7%)	RR 1.01 (0.79 to 1.28)	9 more per 1000 (from 180 fewer to 240 more)	Moderate
Treatment acceptability at 12 weeks											
2 (Cano 2012; Griesser 2012)	Randomised trials	Serious[1]	very serious[2]	no serious	no serious	none	207/256 (80.9%)	131/200 (65.5%)	RR 1.38 (0.93 to 2.04)	249 more per 1000 (from 46 fewer to 681 more)	Very low
Health related quality of life at 12 weeks											
No evidence available											

1. Detection and selection bias
2. Evidence was downgraded by 2 due to very serious heterogeneity (chi-squared p<0.1, I-squared inconsistency statistic of >75%)
3. Evidence was downgraded by 1 due to serious heterogeneity (chi-squared p<0.1, I-squared inconsistency statistic of 50%-74.99%)
4. Detection bias

Table 38: GRADE profile: local oestrogens versus placebo for the outcomes of improvement in vaginal dryness, dyspareunia, itching/discomfort, endometrial hyperplasia, treatment withdrawal, treatment acceptability (duration 12 months) for long term symptoms

Quality assessment							Number of patients		Effect		
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Oestrogen	Control	Relative (95% CI)	Absolute	Quality
Improvement in vaginal dryness at 12 months' treatment duration											
1 (Simunic 2003)	randomised trials	Serious[1]	no serious	no serious	no serious	None	472/560 (84.3%)	143/504 (28.4%)	RR 2.97 (2.57 to 3.43)	559 more per 1000 (from 445 more to 689 more)	Moderate
Improvement in dyspareunia at 12 months' treatment duration											
1 (Simunic 2003)	randomised trials	Serious[1]	no serious	no serious	no serious	None	265/361 (73.4%)	80/298 (26.8%)	RR 2.73 (2.24 to 3.33)	464 more per 1000 (from 333 more to 626 more)	Moderate
Improvement in itching and/or discomfort at 12 months' treatment duration											
1 (Simunic 2003)	randomised trials	Serious[1]	no serious	no serious	no serious	None	329/410 (80.2%)	132/361 (36.6%)	RR 2.19 (1.9 to 2.53)	435 more per 1000 (from 329 more to 559 more)	Moderate
Endometrial hyperplasia or cancer, confirmed by biopsy, at 12 months' treatment duration											
1 (Simon 2008)	randomised trials	Serious[1]	no serious	no serious	serious imprecision[2]	None	1/205 (0.49%)	0/104 (0%)	RR 1.53 (0.06 to 37.21)	-	Low
Withdrawal due to adverse effects at 12 months' treatment duration											
1 (Simon 2008)	randomised trials	Serious[1]	no serious	no serious	serious imprecision[2]	None	11/205 (5.4%)	5/104 (4.8%)	RR 1.12 (0.4 to 3.13)	6 more per 1000 (from 29 fewer to 102 more)	Low
Acceptability of treatment to women at 12 months' treatment duration											
1 (Simunic 2003)	randomised trials	Serious[1]	no serious	no serious	serious imprecision[2]	None	700/828 (84.5%)	675/784 (86.1%)	RR 0.98 (0.94 to 1.02)	17 fewer per 1000 (from 52 fewer to 17 more)	Low

1. Detection bias
2. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)

Table 39: GRADE profile: ospemifene versus placebo (short term treatment)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Ospemifene	Placebo	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Maturation index: Percentage change in Parabasal cells (better indicated by lower value; treatment of less than 1 year)											
60 mg ospemifene											
5 (Bachmann, 2010; Portman, 2014; Portman, 2013; Rutanen, 2003; and Goldstein, 2014)	1142	827	-	MD35.54 lower (41.25 to 29.82 lower)	Low	Randomised trials	Serious[1]	Serious[2]	No serious	No serious	None
25 mg ospemifene											
1 (Voipio, 2002)	8	8	-	MD 47.20 lower (75.04 to 19.36 lower)	Very low	Randomised trials	Very serious[3]	No serious	No serious	Serious[4]	None
50 mg ospemifene											
1 (Voipio, 2002)	7	8	-	MD 97.40 lower (130.09 to 64.71 lower)	Low	Randomised trials	Very serious[3]	No serious	No serious	No serious	None
100 mg ospemifene											
1 (Voipio, 2002)	8	8	-	MD 64.70 lower (99.52 to 29.88 lower)	Very low	Randomised trials	Very serious[3]	No serious	No serious	No serious	None
200 mg ospemifene											
1 (Voipio, 2002)	8	8	-	MD 85.30 lower (117.69 to 52.91 lower)	Low	Randomised trials	Very serious[3]	No serious	No serious	No serious	None
Maturation index: Percentage change in Superficial cells (better indicated by higher value; treatment of less than 1 year)											
60 mg ospemifene											
5 (Bachmann, 2010; Portman, 2014; Portman, 2013; Rutanen, 2003; and Goldstein, 2014)	1142	827	-	MD 8.33 higher (7.43 to 9.22 higher)	Very low	Randomised trials	Serious[1]	Very serious[5]	No serious	No serious	None
25 mg ospemifene											
1 (Voipio, 2002)	8	8	-	MD 11.40 higher (3.29 to 19.51 higher)	Very low	Randomised trials	Very serious[3]	No serious	No serious	Serious[4]	None
50 mg ospemifene											
1 (Voipio, 2002)	7	8	-	MD 15.40 higher (3.87 to 26.93 higher)	Very low	Randomised trials	Very serious[3]	No serious	No serious	Serious[4]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Ospemifene	Placebo	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
100 mg ospemifene											
1 (Voipio, 2002)	8	8	-	MD 18.30 higher (5.02 to 31.58 higher)	Very low	Randomised trials	Very serious[3]	No serious	No serious	Serious[4]	None
200 mg ospemifene											
1 (Voipio, 2002)	8	8	-	MD 10.10 higher (2.96 to 17.24 higher)	Very low	Randomised trials	Very serious[3]	No serious	No serious	Serious[4]	None
Maturation index: Percentage change in Intermediate cells (treatment of less than 1 year)											
25 mg ospemifene											
1 (Voipio, 2002)	8	8	-	MD 28.10 lower (55.15 to 1.05 lower)	Very low	Randomised trials	Very serious[3]	No serious	No serious	Serious[4]	None
50 mg ospemifene											
1 (Voipio, 2002)	7	8	-	MD 24.30 lower (49.20 lower to 0.60 higher)	Very low	Randomised trials	Very serious[3]	No serious	No serious	Serious[4]	None
100 mg ospemifene											
1 (Voipio, 2002)	8	8	-	MD 26.10 lower (52.18 to 0.02 lower)	Very low	Randomised trials	Very serious[3]	No serious	No serious	Serious[4]	None
200 mg ospemifene											
1 (Voipio, 2002)	8	8	-	MD 32.20 lower (58.99 to 5.41 lower)	Very low	Randomised trials	Very serious[3]	No serious	No serious	Serious[4]	None
Patient assessment of symptoms improvement: Change in dyspareunia, severity score (60 mg; better indicated by lower value; treatment of less than 1 year)											
2 (Bachmann, 2010; and Portman, 2013)	579	570	-	SMD 0.30 lower (0.39 to 0.21 lower)	Moderate	Randomised trials	Serious[1]	No serious	No serious	No serious	None
Measurement of vaginal pH: Change in vaginal pH (60 mg; better indicated by lower value; treatment of less than 1 year)											
4 (Bachmann, 2010; Portman, 2014; Portman, 2013; and Goldstein, 2014)	1102	787	-	MD 0.87 lower (0.95 to 0.79 lower)	Moderate	Randomised trials	Serious[1]	No serious	No serious	No serious	None
Patient assessment of symptoms improvement: Change in vaginal dryness, severity score (60 mg; treatment of less than 1 year)											
2 (Bachman 2010; Portman, 2014)	436	422	-	SMD 0. 20 lower (0.33 lower to 0.06 lower)	Moderate	Randomised trials	Serious[1]	No serious	No serious	No serious	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Ospemifene	Placebo	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Assessment of endometrial stimulation: Change in endometrial thickness from baseline (mm) (treatment of less than one year)											
25 mg ospemifene											
1 (Voipio, 2002)	8	8	-	MD 0.28 lower (0.78 lower to 0.22 higher)	Very loow	Randomised trials	Very serious[3]	No serious	No serious	Serious[4]	None
30 mg ospemifene											
2 (Rutanen, 2003 and Bachmann, 2010)	322	308	-	MD 0.48 higher (0.30 to 0.66 higher)	Low	Randomised trials	Serious[1]	No serious	No serious	Serious[4]	None
50 mg ospemifene											
1 (Voipio, 2002)	7	8	-	MD 1.53 higher (1.18 lower to 4.24 higher)	Very low	Randomised trials	Very serious[3]	No serious	No serious	Serious[4]	None
60 mg ospemifene											
5 (Portman, 2013; Portman, 2014; Bachmann, 2010; Goldstein 2014; Rutanen, 2003;)	1142	827	-	SMD 0.41 higher (0.20 to 0.63 higher)	Low	Randomised trials	Serious[1]	No serious[7]	No serious	Serious[4]	None
90 mg ospemifene											
1 (Rutanen, 2003)	40	40	-	MD 0.43higher (0.10 to 0.76 higher)	Low	Randomised trials	Serious[1]	No serious	No serious	Serious[4]	None
100 mg ospemifene											
1 (Voipio, 2002)	10	10	-	MD 0.45 higher (0.20 lower to 1.10 higher)	Very low	Randomised trials	Very serious[3]	No serious	No serious	Serious[4]	None
200 mg ospemifene											
1 (Voipio, 2002)	10	8	-	MD 1.25 higher (0.45 to 2.05 higher)	Very low	Randomised trials	Very serious[3]	No serious	No serious	Serious[4]	None
Endometrial hyperplasia (treatment of less than one year) (25mg)											
1 (Voipio 2012)	8	8		No cases	Low	Randomised trials	Very Serious[3]	No serious	No serious	No serious	None
Endometrial hyperplasia (treatment of less than one year) (30mg)											
1 (Rutanen, 2003)	40	40		No cases	Moderate	Randomised trials	Serious[1]	No serious	No serious	No serious	None
Endometrial hyperplasia (treatment of less than one year) (50mg)											
1 (Voipio 2012)	10	10		No cases	Moderate	Randomised trials	Serious[3]	No serious	No serious	No serious	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Ospemifene	Placebo	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Endometrial hyperplasia (treatment of less than one year) (60mg)											
4 (Bachmann, 2010; Portman, 2014; Portman, 2013; Rutanen, 2003;)	406	403		No cases	Moderate	Randomised trials	Serious[1]	No serious	No serious	No serious	None
Endometrial hyperplasia (treatment of less than one year) (90mg)											
1 (Rutanen, 2003)	40	40		No cases	Moderate	Randomised trials	Serious[1]	No serious	No serious	No serious	None
Endometrial hyperplasia (treatment of less than one year) (100mg)											
1 (Voipio 2012)	10	10		No cases	Low	Randomised trials	Very serious[3]	No serious	No serious	No serious	None
Endometrial hyperplasia (treatment of less than one year) (200mg)											
1 (Voipio 2012)	10	10		No cases	Low	Randomised trials	Very serious[3]	No serious	No serious	No serious	None
Frequency of adverse events relating to treatment (treatment of less than one year)											
60 mg ospemifene											
3 (Bachmann, 2010; Portman, 2014; and Portman, 2013)	739	724	RR 1.60 (1.04 - 2.46)	167 more per 1000 (from 11 more to 407 more)	Very low	Randomised trials	Serious[1]	Very serious[5]	No serious	Serious[8]	None
30 mg ospemifene											
1 (Bachmann, 2010)	276	268	RR 1.26 (1.09 - 4.46)	136 more per 1000 (from 47 more to 1000 more)	Low	Randomised trials	Serious[1]	No serious	No serious	Serious[8]	None
Withdrawal due to treatment related adverse events (treatment of less than one year)											
60 mg Ospemifene											
4 (Bachmann, 2010; Portman, 2014; Portman, 2013; Rutanen, 2003)	779	764	RR 1.59 (0.94 – 2.68)	17 more per 1000 (from 2 fewer to 48 more)	Low	Randomised trials	Serious[1]	No serious	No serious	Serious[4]	None
30 mg ospemifene											
1 (Bachmann, 2010)	276	268	RR 1.12 (0.54 – 2.31)	6 more per 1000 (from 22 fewer to 64 more)	Very low	Randomised trials	Serious[1]	No serious	No serious	Very serious[9]	None

1. Unclear allocation concealment in all trials. Rutanen 2003 did not provide details on randomization and no subgroup analysis on the women with intact uterus (140/160).
2. Evidence was downgraded by 1 due to serious heterogeneity (ch i-squared p<0.1, I-squared inconsistency statistic of 50%-74.99%). Goldstein 2013 study gave results on 95% confidence interval and mean values were approximated based on other results.
3 .Risk of bias was unclear in all aspects of the domain "Selection bias". Small sample size (N=15)
4. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed1e default MID (-/+0.5 times SD)

5. Evidence was downgraded by 2 due to very serious heterogeneity (chi-squared p<0.1, I-squared inconsistency statistic of >75%)
6. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (-/+0.5 times SD)
7. There is inconsistency (chi-squared p<0.1, I-squared inconsistency statistics of 50%-74.99%) but it does not matter as all studies favour control
8. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
9. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)

Table 40: GRADE profile: ospemifene versus placebo (long-term treatment)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Ospemifene	Placebo	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Assessment of endometrial stimulation: Change in endometrial thickness, mm (treatment duration of more than one year)											
30mg ospemifene											
1 (Simon, 2013)	62	49	-	MD (CI): 0.72 (0.00-1.44)	Low	Randomised trials	Serious[1]	No serious	No serious	Serious[2]	None
60mg ospemifene											
2 (Goldstein 2014, Simon, 2013)	432	112		SMD (CI): 0.59 higher (0.16-1.02 higher)	Low	Randomised trials	Serious[1]	No serious	No serious	Serious[2]	None
Endometrial hyperplasia or carcinoma (treatment duration of more than one year)											
30mg ospemifene											
1 (Simon, 2013)	0/62	0/49		No cases	Moderate	Randomised trials	Serious[1]	No serious	No serious	No serious	None
60mg ospemifene											
2 (Simon, 2013; Goldstein 2014)	1/432	0/112	RR 0.52 (0.02-12.57)	Not estimable	Very low	Randomised trials	Serious[1]	No serious	No serious	Very serious[4]	None
Frequency of adverse events relating to treatment (treatment duration of more than one year)											
60 mg ospemifene											
2 (Goldstein , 2014 and Simon, 2013)	352/432	69/112	RR 1.16 (1.01 – 1.33)	99 more per 1000 (from 6 more to 205 more)	Low	Randomised trials	Serious[1]	No serious	No serious	Serious[3]	None
30 mg ospemifene											
1 (Simon, 2013)	38/62	22/49	RR 1.34 (0.93 – 1.94)	153 more per 1000 (from 31 fewer to 422 more)	Low	Randomised trial	Serious[1]	No serious	No serious	Serious[3]	None
Withdrawal due to treatment related adverse events (treatment duration of more than one year)											
60 mg ospemifene											
1	53/432	7/112	RR 1.52 (0.71 – 3.22)		Moderate	Randomised trials	Serious[1]	No serious	No serious	No serious	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Ospemifene	Placebo	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
(Goldstein 2014; Simon, 2013)											
30 mg ospemifene											
1 (Simon, 2013)	4/62	1/49	RR 2.16 (0.23 – 20.17)	24 more per 1000 (from 16 fewer to 391 more)	Very low	Randomised trials	Serious[1]	No serious	No serious	Very serious[4]	None

1. Unclear allocation concealment
2. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (-/+0.5 times SD)
3. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25).
4. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)

I.4 Starting and stopping HRT

Table 41: GRADE profile: tapered discontinuation versus abrupt discontinuation of HRT during tapering regime

Number of studies	Number of women		Effect		Quality	Design	Quality assessment				
	Tapered discontinuation	Abrupt discontinuation	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Occurrence of menopausal symptoms*											
Blatt Kupperman score (at 2 months, during 2 month tapering process)											
1 (Cunha 2010)	18	17	-	MD 4.10 lower (from 8.44 lower to 0.24 higher)	Low	Randomised trials	Serious[1]	No serious	No serious	Serious[2]	None
Blatt Kupperman score (at 4 months, during 4 month tapering process)											
1 (Cunha 2010)	19	17	-	MD 4.30 lower (from 8.91 lower to 0.31 higher)	Low	Randomised trials	Serious[1]	No serious	No serious	serious[2]	None
Hot flush component of Blatt Kupperman score (at 2 months, during 2 month tapering process)											
1 (Cunha 2010)	18	17	-	MD 5.00 lower (from 7.18 lower to 2.82 lower)	Moderate	Randomised trials	Serious[1]	No serious	No serious	serious[2]	None
Hot flush component of Blatt Kupperman score (at 4 months, during 4 month tapering process)											

Number of studies	Number of women		Effect		Quality	Design	Quality assessment				
	Tapered discontinuation	Abrupt discontinuation	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
1 (Cunha 2010)	19	17	-	5.00 lower (from 7.80 lower to 2.20 lower)	Moderate	Randomised trials	Serious[1]	No serious	No serious	serious[2]	None
Hot flush score (at 2 weeks, during 2 week tapering process)											
1 (Aslan 2007)	35	35	-	1.09 lower (from 3.64 lower to 1.46 higher)	Low	Randomised trials	Serious[3]	No serious	No serious	Serious[2]	None
Number of women with no vasomotor symptoms (at 2 weeks, during 2 week tapering process)											
1 (Aslan 2007)	19/35	17/35	RR 1.12 (0.71 to 1.79)	58 more per 1000 (from 141 fewer to 384 more)	Very low	Randomised trials	Serious[3]	No serious	No serious	very serious[4]	None
Number of women with mild vasomotor symptoms (at 2 weeks, during 2 week tapering process)											
1 (Aslan 2007)	13/35	15/35	RR 0.87 (0.49 to 1.54)	56 fewer per 1000 (from 219 fewer to 231 more)	Very low	Randomised trials	Serious[3]	No serious	No serious	very serious[4]	None
Number of women with moderate vasomotor symptoms (at 2 weeks, during 2 week tapering process)											
1 (Aslan 2007)	2/35	1/35	RR 2.00 (0.19 to 21.06)	29 more per 1000 (from 23 fewer to 573 more)	Very low	Randomised trials	Serious[3]	No serious	No serious	very serious[4]	None
Number of women with severe vasomotor symptoms (at 2 weeks, during 2 week tapering process)											
1 (Aslan 2007)	1/35	2/35	RR 0.50 (0.05 to 5.27)	29 fewer per 1000 (from 54 fewer to 244 more)	Very low	Randomised trials	Serious[3]	No serious	No serious	very serious[4]	None
Total Greene Climacteric Score (at 1 month, during 6 month tapering process)											
1 (Haimov-Kochman 2006)	41	50	Reduced score in taper group (p = 0.001)	-	Low	Randomised trials	Serious[3]	No serious	No serious	Serious[5]	None
Total Greene Climacteric Score (at 3 months, during 6 month tapering process)											
1 (Haimov-Kochman 2006)	41	50	Reduced score in taper group (p = 0.047)	-	Low	Randomised trials	Serious[3]	No serious	No serious	Serious[5]	None

Number of studies	Number of women		Effect		Quality	Design	Quality assessment				
	Tapered discontinuation	Abrupt discontinuation	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Total Greene Climacteric Score (at 6 months, during 6 month tapering process)											
1 (Haimov-Kochman 2006)	41	50	No significant difference between groups.	-	Low	Randomised trials	Serious[3]	No serious	No serious	Serious[5]	None
Vasomotor Greene Climacteric Score (at 1 month, during 6 month tapering process)											
1 (Haimov-Kochman 2006)	41	50	Reduced score in taper group (p = 0.0001)	-	Low	Randomised trials	Serious[3]	No serious	No serious	Serious[5]	None
Vasomotor Greene Climacteric Score (at 3 months, during 6 month tapering process)											
1 (Haimov-Kochman 2006)	41	50	Reduced score in taper group (p = 0.001)	-	Low	Randomised trials	Serious[3]	No serious	No serious	Serious[5]	None
Vasomotor Greene Climacteric Score (at 6 months, during 6 month tapering process)											
1 (Haimov-Kochman 2006)	41	50	Increased score in taper group (p = 0.001)		Low	Randomised trials	Serious[3]	No serious	No serious	Serious[5]	None

1. The study was double-blinded by design, but it was unclear whether the investigators and participants were properly blinded;
2. Unable to calculate confidence interval for the SMD as mean in each group not reported;
3. The study was open-label trial in design;
4. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25);
5. Unable to calculate 95% CI as mean and SD not reported

Table 42: GRADE profile: tapered discontinuation versus abrupt discontinuation of HRT after tapering regime complete

Number of studies	Number of women		Effect		Quality	Design	Quality assessment				
	Tapered discontinuation	Abrupt discontinuation	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Occurrence of menopausal symptoms*											
Blatt Kupperman index (at 6 months, following tapering over 2 or 4 months)											
1 (Cunha 2010)	37	17	-	2.57 points higher (from 2.05 points lower to 7.19 points higher)	Low	Randomised trials	Serious[1]	No serious	No serious	Serious[2]	None
Hot flush component of Blatt Kupperman index (at 6 months, following tapering over 2 or 4 months)											

Number of studies	Number of women		Effect		Quality	Design	Quality assessment				
	Tapered discontinuation	Abrupt discontinuation	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
1 (Cunha 2010)	37	17	-	0.25 points lower (from 2.97 points lower to 2.47 points higher)	Low	Randomised trials	Serious[1]	No serious	No serious	Serious[2]	None
Hot flush score (at 4 weeks, following tapering over 2 weeks)											
1 (Aslan 2007)	35	35	-	0.40 points lower (from 3.37 points lower to 2.57 points higher)	Very low	Randomised trials	Serious[3]	No serious	No serious	Serious[2]	None
Number of women with no vasomotor symptoms (at 4 weeks, following tapering for 2 weeks)											
1 (Aslan 2007)	18/35	18/35	RR 1.00 (0.63 to 1.58)	0 fewer per 1000 (from 190 fewer to 298 more	Very low	Randomised trials	Serious[3]	No serious	No serious	Very serious[4]	None
Number of women with mild vasomotor symptoms (at 4 weeks, following tapering for 2 weeks)											
1 (Aslan 2007)	15/35	13/35	RR 1.15 (0.65 to 2.05)	56 more per 1000 (from 130 fewer to 390 more)	Very low	Randomised trials	Serious[3]	No serious	No serious	Very serious[4]	None
Number of women with moderate vasomotor symptoms (at 4 weeks, following tapering for 2 weeks)											
1 (Aslan 2007)	0/35	2/35	RR 5.00 (0.25 to 100.53)	229 more per 1000 (from 43 fewer to 1000 more)	Very low	Randomised trials	Serious[3]	No serious	No serious	Very serious[4]	None
Number of women with severe vasomotor symptoms (at 4 weeks, following tapering for 2 weeks)											
1 (Aslan 2007)	2/35	2/35	RR 1.00 (0.15 to 6.71)	0 fewer per 1000 (from 49 fewer to 326 more)	Very low	Randomised trials	Serious[3]	No serious	No serious	Very serious[4]	None
Frequency of hot flushes in 24 hours (at 6 weeks, following tapering for 4 weeks)											
1 (Lindh-Åstrand et al. 2010)	45	36	P=0.50	-	Low	Randomised trials	Serious[3]	No serious	No serious	Serious[5]	None
Severity of hot flushes in 24 hours (at 6 weeks, following tapering for 4 weeks)											

Number of studies	Number of women		Effect		Quality	Design	Quality assessment				
	Tapered discontinuation	Abrupt discontinuation	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
1 (Lindh-Åstrand et al. 2010)	45	36	P = 0.75	-	Low	Randomised trials	Serious[3]	No serious	No serious	Serious[5]	None
Total Greene Climacteric Score (at 9 months, following 6 month tapering process)											
1 (Haimov-Kochman 2006)	41	50	No significant difference between groups.	-	Low	Randomised trials	Serious[3]	No serious	No serious	Serious[5]	None
Total Greene Climacteric Score (at 12 months, following 6 month tapering process)											
1 (Haimov-Kochman 2006)	41	50	No significant difference between groups.	-	Low	Randomised trials	Serious[3]	No serious	No serious	Serious[5]	None
Vasomotor Greene Climacteric Score (at 9 months, following 6 month tapering process)											
1 (Haimov-Kochman 2006)	41	50	No significant difference between groups.	-	Low	Randomised trials	Serious[3]	No serious	No serious	Serious[5]	None
Vasomotor Greene Climacteric Score (at 12 months, following 6 month tapering process)											
1 (Haimov-Kochman 2006)	41	50	No significant difference between groups.	-	Low	Randomised trials	Serious[3]	No serious	No serious	Serious[5]	None
Health related quality of life (at 6 weeks, following tapering over 4 weeks)											
1 (Lindh-Åstrand 2010)	45	36	P = 0.50		Low	Randomised trials	Serious[3]	No serious	No serious	Serious[5]	None
Recommencing HRT treatment by 12 months (following tapering over 4 weeks or 6 months)											
2 (Haimov-Kochman 2006, Lindh-Åstrand 2010)	85	86	RR 1.11 (0.78 to 1.58)	45 more per 1000 (from 90 fewer to 236 more)	Low	Randomised trial	Serious[3]	No serious	No serious	Serious[6]	None

1. The study was double-blinded by design, but it was unclear whether the investigators and participants were properly blinded;
2. Unable to calculate confidence interval for the SMD as mean in each group not reported;
3. The study was open-label trial in design;
4. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25);
5. Unable to calculate 95% CI as mean and SD not reported;
6. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25);

I.5 Long-term benefits and risks of HRT

I.5.1 Venous thromboembolism (VTE)

Table 43: GRADE profile: HRT use versus placebo for the outcome of VTE

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	Placebo	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
VTE (current HRT use, oral route)											
7 (Cherry 2002, Holmberg 2008, Høibraaten 2000, Manson 2013 Nachtigall 1979, Vickers 2007, Whiteman 1999)	387/ 17604	211/ 16775	RR 1.78 (1.51 to 2.10)	10 more per 1000 (from 6 more to 14 more)	Low	Randomised trials	Serious[1]	No serious	No serious	No serious	None
VTE (oestrogen alone)											
2 (Cherry 2002, Manson 2013)	142/ 5823	102/ 5933	RR 1.42 (1.1 to 1.83)	7 more per 1000 (from 2 more to 14 more)	Low	Randomised trials	Serious[2]	No serious	No serious	Serious[3]	None
VTE (oestrogen plus progesterone)											
4 (Høibraaten 2000, Nachtigall 1979, Manson 2013, Vickers 2007)	239/ 10857	107/ 10444	RR 2.13 (1.70 to 2.67)	12 more per 1000 (from 7 more to 17 more)	Low	Randomised trials	Serious[1]	Serious[4]	No serious	No serious	None
VTE (current use of any HRT for 1 year or less)											
1 (Vickers 2007)	22/ 2196	3/ 2189	RR 7.31 (2.19 to 24.39)	9 more per 1000 (from 2 more to 32 more)	Moderate	Randomised trials	Serious[5]	No serious	No serious	No serious	None
VTE (current use of any HRT for between 1 and 5 years)											

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	Placebo	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
4 (Cherry 2002, Holmberg 2008, Høibraaten 2000, Whiteman 1999)	19/ 1508	7/ 971	RR 2.12 (0.90 to 4.99)	8 more per 1000 (from 1 fewer to 29 more)	Low	Randomised trials	Serious[1]	No serious	No serious	Serious[3]	None
VTE (current use of any HRT for over 5 years)											
2 (Manson 2013, Nachtigall 1979)	346/ 13900	201/ 13615	RR 1.68 (1.42 to 2.00)	10 more per 1000 (from 5 more to 13 more)	Moderate	Randomised trials	Serious[2]	No serious	No serious	No serious	None
VTE (oestrogen plus progesterone, women aged 50-59 years at baseline)											
1 (Manson 2013)	32/ 2837	13/ 2683	HR 2.27 (1.19 to 4.33)d	6 more per 1000 (from 1 more to 16 more)	Low	Randomised trials	Serious[2]	No serious	No serious	Serious[3]	None
VTE (oestrogen alone, women aged 50-59 years at baseline)											
1 (Manson 2013)	20/ 1639	15/ 1674	HR 1.37 (0.70 to 2.68)d	3 more per 1000 (from 3 fewer to 15 more)	Very low	Randomised trials	Serious[2]	No serious	No serious	Very serious[6]	. None
VTE (time since menopause (oestrogen plus progesterone, < 10 years)											
1 (Canonico 2014)	33/2758	10/2694	HR 3.4 (1.6-7.2)	9 more per 1000 (from 2 more to 23 more)	Moderate	Randomised trials	Serious[7,8]	No serious	No serious	No serious	None
VTE (timesince menopause (oestrogen alone, < 10 years)											
1 (Canonico 2014)	9/817	8/802	HR 1.1 (0.4-2.9)	1 more per 1000 (from 6 fewer to 19 m)	Very low	Randomised trials	Serious[7,8]	No serious	No serious	Very serious[6]	None

1. Risk of biases across studies included open-label trial, breaking of blinding, high and/or unbalanced drop-out rates, highly selected participants, and small sample size;
2. High rates of blinding breaking, high drop-out rates in studies included in the analysis;
3. Evidence was downgraded by 1 due to serious imprecision as 95%CI crossed 1 default MID (0.75 to 1.25);
4. Evidence was downgraded by 1 due to serious heterogeneity (chi-squared p<0.1, I-squared inconsistency statistic of 50%-74.99%);
5. Un-proportional drop-out rates between the two arms in the study;
6. Evidence was downgraded by 2 due to very serious imprecision as 95%CI crossed 2 default MIDs (0.75 to 1.25);
7. Self-reported information on HRT initiation year which could lead to misclassification;
8. Stratified analyses included subgroup with a relatively low number of cases, especially for PE, resulting in low statistical power;

Table 44: GRADE profile: HRT use versus no HRT use for the outcome of VTE (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
VTE (current HRT use)											
1 (Grodstein 1996)	Not reported	Not reported	RR 2.1 (1.2 to 3.8)	Not calculable	Very low	Prospective cohort	Serious[1]	No serious	Serious[2]	Serious[3]	None
1 (Ohira 2010)	30/ 1439	120/ 5025	RR 1.60 (1.06 to 2.36)	14 more per 1000 (from 1 more to 32 more)	Low	Prospective cohort	Serious[4]	No serious	No serious	Serious[3]	None
1 (Benson 2012)	909/ 380033	965/ 476711	RR 1.59 (1.45 to 1.75)	1 more per 1000 (from 1 more to 2 more)	Moderate	Prospective cohort	Serious[5]	No serious	No serious	No serious	None
VTE (current HRT use, oral route)											
1 (Canonico 2010)	Not reported	Not reported	HR 1.7 to 2.8)	Not calculable	Low	Prospective cohort	Serious[6]	No serious	No serious	Serious[3]	None
1 (Olie 2011)	Not reported	68/ 893	HR 6.4 (1.5 to 27.3)	321 more per 1000 (from 36 more to 809 more)	Low	Retrospective cohort	Serious[6]	No serious	Serious[7]	No serious	None
VTE (current use of HRT, transdermal route)											
1 (Canonico 2010)	Not reported	Not reported	HR 1.1 (0.8 to 1.8)	Not calculable	Low	Prospective cohort	Serious[6]	No serious	No serious	Serious[3]	None
1 (Olie 2011)	Not reported	68/ 893	HR 1.0 (0.4 to 2.4)	0 fewer per 1000 (from 45 fewer to 97 more)	Very low	Retrospective cohort	Serious[6]	No serious	Serious[7]	Very serious[8]	None
VTE (current use of oestrogen, transdermal route)											
1 (Benson 2012)	66/ 51853	965/ 476711	RR 0.82 (0.64 to 1.06)	0 more per 1000 (from 1 fewer to 0 more)	Low	Prospective cohort	Serious[5]	No serious	No serious	Serious[3]	None
VTE (current use of HRT, oral route)											
1 (Canonico 2010)	Not reported	Not reported	HR 1.5 (1.1 to 2.0)	Not calculable	Low	Prospective cohort	Serious[6]	No serious	No serious	Serious[3]	None
VTE (current use of HRT, oral versus transdermal route)											

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
1 (Laliberté 2011)	Oral HRT 164/ 27018	Transdermal HRT 115/ 27018	RR 1.49 (1.07 to 2.04)	2 more per 1000 (from 0 more to 4 more)	Low	Retrospective cohort	Serious[9]	No serious	No serious	Serious[3]	None
VTE (current oestrogen alone, oral route)											
1 (Benson 2012)	194/ 86250	965/ 476711	RR 1.42 (1.22 to 1.66)	1 more per 1000 (from 0 more to 1 more)	Low	Prospective cohort	Serious[5]	No serious	No serious	Serious[3]	None
VTE (current oestrogen alone, transdermal route)											
1 (Benson 2012)	66/ 51853	965/ 476711	RR 0.82 (0.64 to 1.06)	0 fewer per 1000 (from 1 fewer to 0 more)	Very low	Prospective cohort	Serious[5]	No serious	No serious	Very serious[8]	None
1 (Olie 2011)	Not reported	68/ 893	HR 1.1 (0.2 to 8.1)	7 more per 1000 (from 60 fewer to 397 more)	Very low	Retrospective cohort	Serious[6]	No serious	Serious[7]	Very serious[8]	None
Pulmonary embolism (current use of oestrogen alone)											
1 (Su 2012)	Not reported	Not reported	HR 2.75 (0.45 to 16.8)	Not calculable	Very low	Retrospective cohort	No serious	No serious	Serious[10]	Very serious[8]	None
Deep vein thrombosis (current use of oestrogen alone)											
1 (Su 2012)	Not reported	Not reported	HR 3.63 (1.48 to 8.89)	Not calculable	Moderate	Retrospective cohort	No serious	No serious	Serious[10]	No serious	None
VTE (current use of oestrogen plus progesterone, oral route)											
1 (Benson 2012)	542/ 196358	965/ 476711	RR 2.07 (1.86 to 2.32)	2 more per 1000 (from 2 more to 3 more)	Low	Prospective cohort	Serious[5]	No serious	No serious	No serious	None
Pulmonary embolism (oestrogen and progesterone)											
1 (Su 2012)	Not reported	Not reported	HR 0.80 (0.35 to 1.85)	Not calculable	Very low	Retrospective cohort	No serious	No serious	Serious[10]	Very serious[8]	None
Deep vein thrombosis (oestrogen and progesterone)											
1 (Su 2012)	Not reported	Not reported	HR 0.90 (0.51 to 1.60)	Not calculable	Very low	Retrospective cohort	No serious	No serious	Serious[10]	Very serious[8]	None
VTE, (current use of any HRT commenced within the past 2 years, oral route)											

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
1 (Benson 2012)	Not reported	965/ 476711	RR 3.83 (1.91 to 7.71)	6 more per 1000 (from 2 more to 14 more)	Moderate	Prospective cohort	Serious[5]	No serious	No serious	No serious	None
VTE (current use of oestrogen and progesterone, oral route, commenced within the past 2 years)											
1 (Benson 2012)	Not reported	965/ 476711	RR 3.17 (2.10 to 4.78)	4 more per 1000 (from 2 more to 8 more)	Moderate	Prospective cohort	Serious[5]	No serious	No serious	No serious	None
VTE, current use of oestrogen alone, transdermal route, commenced within the past 2 years)											
1 (Benson 2012)	Not reported	965/ 476711	RR 1.63 (0.41 to 6.53	1 more per 1000 (from 1 fewer to 11 more)	Very low	Prospective cohort	Serious[5]	No serious	No serious	Very serious[8]	None
VTE (current use of any HRT for 5 years or less)											
1 (Grodstein 1996)	Not reported	Not reported	RR 2.6 (1.2 to 5.2)	Not calculable	Very low	Prospective cohort	Serious[1]	No serious	Serious[2]	Serious[3]	None
VTE (current use of oestrogen alone for 5 years or less, oral route)											
1 (Benson 2012)	Not reported	965/ 476711	RR 1.41 (1.19 to 1.67)	1 more per 1000 (from 0 more to 1 more)	Low	Prospective cohort	Serious[5]	No serious	No serious	Serious[3]	None
VTE (current use of oestrogen plus progesterone for 5 years or less, oral route)											
1 (Benson 2012)	Not reported	965/ 476711	**RR 2.00 (1.77 to 2.26)**	2 more per 1000 (from 2 more to 3 more)	Moderate	Prospective cohort	Serious[5]	No serious	No serious	No serious	None
VTE (current use of oestrogen alone for 5 years or less, transdermal route)											
1 (Benson 2012)	Not reported	965/ 476711	RR 0.84 (0.64 to 1.09)	0 fewer per 1000 (from 1 fewer to 0 more)	Low	Prospective cohort	Serious[5]	No serious	No serious	Serious[3]	None
VTE (current use of any HRT for over 5 years)											
1 (Grodstein 1996)	Not reported	Not reported	RR 1.9 (0.9 to 4.0)	Not calculable	Very low	Prospective cohort	Serious[1]	No serious	Serious[2]	Serious[3]	None
VTE (current use of oestrogen for over 5 years, oral route)											
1 (Benson 2012)	Not reported	965/ 476711	RR 1.49 (1.24 to 1.77)	1 more per 1000 (from 0 more to 2 more)	Low	Prospective cohort	Serious[5]	No serious	No serious	Serious[3]	None
VTE (**current use of oestrogen plus progesterone for over 5 years, oral route)**											

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
1 (Benson 2012)	Not reported	965/ 476711	RR 2.05 (1.80 to 2.33)	2 more per 1000 (from 2 more to 3 more)	Moderate	Prospective cohort	Serious[5]	No serious	No serious	No serious	None
VTE (current use of oestrogen for over 5 years, transdermal route)											
1 (Benson et al., 2012)	Not reported	965/ 476711	RR 0.85 (0.63 to 1.13)	0 fewer per 1000 (from 1 fewer to 0 more)	Low	Prospective cohort	Serious[5]	No serious	No serious	Serious[3]	None
VTE (women aged < 50 years at first use of oestrogen alone, oral route)											
1 (Benson 2012)	Not reported	965/ 476711	RR 1.45 (1.17 to 1.80)	1 more per 1000 (from 0 more to 2 more)	Low	Prospective cohort	Serious[5]	No serious	No serious	Serious[3]	None
VTE (women aged < 50 years at first use of oestrogen plus progesterone, oral route)											
1 (Benson 2012)	Not reported	965/ 476711	RR 1.87 (1.59 to 2.21)	2 more per 1000 (from 1 more to 2 more)	Moderate	Prospective cohort	Serious[5]	No serious	No serious	No serious	None
VTE (women aged < 50 years at first use of oestrogen alone, transdermal route)											
1 (Benson 2012)	Not reported	965/ 476711	RR 0.80 (0.55 to 1.15)	0 more per 1000 (from 1 fewer to 0 more)	Low	Prospective cohort	Serious[5]	No serious	No serious	Serious[3]	None
VTE (women aged ≥ 50 years at first use of oestrogen alone, oral route)											
1 (Benson 2012)	Not reported	965/ 476711	RR 1.33 (1.06 to 1.65)	1 more per 1000 (from 0 more to 1 more)	Low	Prospective cohort	Serious[5]	No serious	No serious	Serious[3]	None
VTE (women aged ≥ 50 years at first use of oestrogen plus progesterone, oral route)											
1 (Benson 2012)	Not reported	965/ 476711	RR 2.16 (1.90 to 2.45)	2 more per 1000 (from 2 more to 3 more)	Moderate	Prospective cohort	Serious[5]	No serious	No serious	No serious	None
VTE (women aged ≥ 50 years at first use of oestrogen, transdermal route)											
1 (Benson 2012)	Not reported	965/ 476711	RR 0.85 (0.61 to 1.20)	0 fewer per 1000 (from 1 fewer to 0 more)	Low	Prospective cohort	Serious[5]	No serious	No serious	Serious[3]	None
VTE (past use of HRT)											
1 (Grodstein 1996)	Not reported	Not reported	RR 1.3 (0.7 to 2.4)	Not calculable	Very low	Prospective cohort	Serious[1]	No serious	Serious[2]	Very serious[8]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
1 (Canonico 2010)	Not reported	Not reported	HR 1.1 (0.8 to 1.5)	Not calculable	Low	Prospective cohort	Serious[6]	No serious	No serious	Serious[3]	None
1 (Ohira et al., 2010)	36/ 1579	120/ 5025	RR 1.07 (0.72 to 1.62)	2 more per 1000 (from 7 fewer to 15 more)	Very low	Prospective cohort	Serious[4]	No serious	No serious	Very serious[8]	None
VTE (past use of HRT, oral route)											
1 (Benson 2012)	326/ 201515	965/ 476711	RR 0.95 (0.84 to 1.08)	0 fewer per 1000 (from 0 fewer to 0 more)	Moderate	Prospective cohort	Serious[5]	No serious	No serious	No serious	None
VTE (past use of oestrogen plus progesterone)											
1 (Manson 2013)	44/ 8052	45/ 7678	HR 0.95 (0.63 to 1.44)	0 fewer per 1000 (from 0 fewer to 3 more)	Very low	Cohort follow up from RCT	Serious[11]	No serious	No serious	Very serious[8]	None
VTE (past use of oestrogen alone)											
1 (Manson 2013)	52/ 3778	74/ 3867	HR 0.72 (0.51 to 1.03)	5 fewer per 1000 (from 9 fewer to 1 more)	Low	Cohort follow up from RCT	Serious[11]	No serious	No serious	Serious[3]	None
Recurrence of VTE (women who have had a first VTE, oestrogen alone use)											
1 (Eischer 2014)	22/333	49/297	HR 0.7 (0.3-1.5)		Very low	Prospective cohort	Serious[12]	Not applicable	Serious[7]	Serious[3]	None

1. Selection bias, HRT users were healthier and younger than non-users;
2. Participants were registered nurses only;
3. Evidence was downgraded by 1 due to serious imprecision as 95%CI crossed 1 default MID (0.75 to 1.25);
4. Risk of bias for ascertainment of VTE outcomes in the study;
5. Known risk factors such as family history of VTE not available and not controlled for in the analysis; drop-out rates not clearly reported;
6. Self-reported HRT use; HRT users were healthier and younger than non-users;
7. Participants were women who have had a previous VTE;
8. Evidence was downgraded by 2 due to very serious imprecision as 95%CI crossed 2 default MIDs (0.75 to 1.25);
9. Data on important confounders not available therefore not controlled for in analysis;
10. The study was carried out among Chinese women only;
11. Extended post-stopping follow-up of an RCT (the WHI);
12. Different follow-up time for the HRT and nonusers group, reasons not reported;

Table 45: GRADE profile: HRT use (by preparations) versus no HRT use for the outcome of VTE (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
VTE (current use of conjugated equine oestrogen)											
1 (Benson 2012)	Not reported	965/ 476711	RR 1.46 (1.23 to 1.75)	1 more per 1000 (from 0 more to 2 more)	Low	Prospective cohort	Serious[1]	No serious	No serious	Serious[2]	None
VTE (current use of oestradiol)											
1 (Benson 2012)	Not reported	965/ 476711	RR 1.45 (1.06 to 1.98)	1 more per 1000 (from 0 more to 2 more)	Low	Prospective cohort	Serious[1]	No serious	No serious	Serious[2]	None
VTE (current use of micronized progesterone, in combined oestrogen + progesterone preparations)											
1 (Canonico 2010)	Not reported	Not reported	HR 0.9 (0.6 to 1.5)	Unable to calculate	Very low	Prospective cohort	Serious[3]	No serious	No serious	Very serious[4]	None.
VTE (current us e of of transdermal oestrogen plus micronized progesterone)											
1 (Olie 2011)	3/ 130	68/ 893	HR 1.0 (0.3 to 3.2)	0 fewer per 1000 (from 53 fewer to 148 more)	Very low	Retrospective cohort	Serious[3]	No serious	Serious[5]	Very serious[4]	None
VTE (current use of pregnane derivatives, in combined oestrogen + progesterone preparations)											
1 (Canonico 2010)	Not reported	Not reported	HR 1.3 (0.9 to 2.0)	Not calculable	Low	Prospective cohort	Serious[3]	No serious	No serious	Serious[2]	None
VTE (current use of combined preparations including medroxyprogesterone acetate)											
1 (Benson 2012)	Not reported	965/ 476711	RR 2.67 (2.25 to 3.17)	3 more per 1000 (from 3 more to 4 more)	Moderate	Prospective cohort	Serious[1]	No serious	No serious	No serious	None
VTE (current use of norpregnane derivatives)											
1 (Canonico 2010)	Not reported	Not reported	HR 1.8 (1.2 to 2.7)	Not calculable	Low	Prospective cohort	Serious[1]	No serious	No serious	Serious[2]	None
VTE (current use of transdermal oestrogen plus norpregnane derivatives)											
1 (Olie 2011)	2 /130	68 /893	HR 4.7 (1.1 to 20.0)	235 more per 1000 (from 7 more to 719 more)	Very low	Retrospective cohort	Serious[3]	No serious	Serious[5]	Serious[2]	None
VTE (current use of nortestosterone derivatives ,in combined oestrogen + progesterone preparations)											
1 (Canonico 2010)	Not reported	Not reported	HR 1.4 (0.6 to 2.4)	Not calculable	Very low	Prospective cohort	Serious[3]	No serious	No serious	Very serious[4]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
VTE (current use of Combined preparations including norethisterone)											
1 (Benson 2012)	Not reported	965/ 476711	RR 1.82 (1.52 to 2.17)	2 more per 1000 (from 1 more to 2 more)	Moderate	Prospective cohort	Serious[1]	No serious	No serious	No serious	None
VTE (current use of combined preparations including norgestrel)											
1 (Benson 2012)	Not reported	965/ 476711	RR 1.98 (1.71 to 2.29)	2 more per 1000 (from 1 more to 3 more)	Moderate	Prospective cohort	Serious[1]	No serious	No serious	No serious	None

1. Known risk factors such as family history of VTE not available and not controlled for in the analysis; drop-out rates not clearly reported;
2. Evidence was downgraded by 1 due to serious imprecision as 95%CI crossed 1 default MID (0.75 to 1.25);
3. Self-reported HRT use; HRT users were healthier and younger than nonusers;
4. Evidence was downgraded by 2 due to very serious imprecision as 95%CI crossed 2 default MIDs (0.75 to1.25);
5. Participants were women who have had a previous VTE;

I.5.2 Cardiovascular disease (CVD)

Table 46: GRADE profile: HRT use versus placebo or no HRT use for the outcomes of CHD, stroke and blood pressure change (RCTs)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	Placebo/no HRT	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
CHD (Women aged 45-58 years, 10-year follow-up)											
1 (Schierbeck 2012)	16/502	33/504	HR 0.48 (0.26-0.87)	8 fewer per 1000 (from 11 fewer to 2 fewer)	Low	Randomised trials	Serious[1].	No serious	No serious	Serious[2]	None
CHD (women aged 50-58 years, 10-year follow-up)											
1 (Schierbeck 2012)	N/R	N/R	HR 0.63 (0.29-1.36)	6 fewer per 1000 (from 11 fewer to 5 more)	Very low	Randomised trials	Serious[1]	No serious	No serious	Very serious[3]	None
CHD (women aged 45-49 years, 10-year follow-up)											
1 (Schierbeck 2012)	N/R	N/R	HR 0.35 (0.13-0.89)	10 fewer per 1000 (from 13 fewer to 2 fewer)	Low	Randomised trials	Serious[1]	No serious	No serious	Serious [2]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	Placebo/no HRT	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
CHD (women aged 45-58 years, total 16-year follow-up, 6 -year post-intervention)											
1 (Schierbeck 2012)	33/502	53/504	HR 0.61 (0.39-0.94)	6 fewer per 1000 (from 9 fewer to 1 fewer)	Low	Randomised trial with post-interventional follow up	Serious[1, 4]	No serious	No serious	Serious[2]	None
CHD (women aged 50-58 years, ,total 16-year follow-up , 6-year post-intervention)											
1 (Schierbeck 2012)	N/R	N/R	HR 0.68 (0.38-1.21)	5 fewer per 1000 (from 9 fewer to 3 more)	Low	Randomised trial with post-interventional follow up	Serious[1, 4]	No serious	No serious	Serious[2]	None
CHD (women aged 45-49 years, total 16-year follow-up , 6-year post-intervention)											
1 (Schierbeck 2012)	N/R	N/R	HR 0.55 (0.29-1.05)	7 fewer per 1000 (from 11 fewer to 1 more)	Low	Randomised trial with post-interventional follow up	Serious [1,4]	No serious	No serious	Serious[2]	None
Stroke(women aged 45-58 years, 10-year follow-up)											
1 (Schierbeck 2012)	N/R	N/R	HR 0.77 (0.35-1.70)	3 fewer per 1000 (from 7 fewer to 8 more)	Very low	Randomised trial with post-interventional follow up	Serious[1,4]	No serious	No serious	Very serious[3]	None
Stroke (women aged 45-58 years, total 16-year follow-up, 6-year post-intervention)											
1 (Schierbeck 2012)	19/502	21/504	HR (95% CI): 0.89 (0.48-1.65)	1 fewer per 1000 (from 6 fewer to 7 more)	Very low	Randomised trial with post-interventional follow up	Serious[1,2]	No serious	No serious	Very serious[3]	None
Reduction of systolic BP (mmHg), (HRT use < 5 years)											
1 (Brownley 2004)	N= (19)	N= (23)		(p < 0.001)	Low	Randomised trials	Serious[5]	No serious	Serious[6]	No serious	None

1. Evidence was downgraded due to lack of blinding (open-label RCT) and relatively high attrition (at 5-year follow-up, about 25% of women dropped-out);
2. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
3. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)
4. Evidence was downgraded because this was the six years post-intervention results after discontinuation of the randomised treatment
5. Evidence was downgraded by 1 due to randomisation, blinding and allocation concealment not clearly reported
6. Evidence was downgraded due to incomplete results reported (no measure of relative effect and indication of variation in the effect size)

Table 47: GRADE profile: oestrogen plus progesterone use versus placebo for the outcomes of CHD, MI, stroke and blood pressure change (RCTs)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Oestrogen plus progesterone	Placebo/ no HRT	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
CHD (women aged 50-59 years)											
1 (Manson 2013)	38/2837 (0.23)	27/2683 (0.17)	HR 1.34 (0.82-2.19)	5 more per 1000 (from 3 fewer to 18 more)	Low	Randomised trials	Serious [1,2,3,4]	No serious	No serious	Serious [5]	None
CHD (women aged 50-59 years, ≤ 2 years duration)											
1 (Toh 2010)	16/2839	10/2683	HR 1.60 (0.73-3.55)	9 more per 1000 (from 4 fewer to 38 more)	Very low	Randomised trials	Serious [1,2,3,4]	No serious	No serious	Very Serious[6]	None
CHD (women aged 50-59 years, ≥ 2 years duration)											
1 (Toh 2010)	21/2839	17/2683	HR 1.14 (0.60-2.16)	2 more per 1000 (from 6 fewer to 17 more)	Very low	Randomised trials	Serious [1,2,3,4]	No serious	No serious	Very Serious[6]	None
CHD (women within 5 years from menopause and without prior HRT use)											
1 (Prentice 2009)	N/R	N/R	HR 0.99 (0.49-1.98)	0 fewer per 1000 (from 8 fewer to 15 more)	Very low	Randomised trials	Serious [1,2,3,4]	No serious	No serious	Very Serious[6]	None
CHD (women within 5 years from menopause and with prior HRT use)											
1 (Prentice 2009)	N/R	N/R	HR 1.57 (0.99-2.50)	9 more per 1000 (from 0 fewer to 23 more)	Low	Randomised trials	Serious [1,2,3,4]	No serious	No serious	Serious[5]	None
CHD (women within 10 years since menopause											

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Oestrogen plus progesterone	Placebo/ no HRT	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
1 (Manson 2002; also reported in Wassertheil-Smoller 2003)	31 (0.19)	34 (0.22)	HR 0.89 (0.40-1.51)	2 fewer per 1000 (from 9 fewer to 8 more)	Very low	Randomised trials	Serious [1,2,3,4]	No serious	No serious	Very Serious[6]	None
CHD (women within 10 years since menopause, ≤ 2 years duration)											
1 (Toh 2010)	14/2782	12/2712	HR 1.17 (0.54-2.52)	3 more per 1000 (from 7 fewer to 23 more)	Very low	Randomised trials	Serious [1,2,3,4]	No serious	No serious	Very Serious[6]	None
CHD (women within 10 years since menopause, ≥ 2 years duration											
1 (Toh 2010)	17/2782	22/2712	HR 0.74 (0.39-1.40)	4 fewer per 1000 (from 9 fewer to 6 more)	Very low	Randomised trials	Serious [1,2,3,4]	No serious	No serious	Very Serious[6]	None
CHD (women 50-59 years at baseline, 8.2 years post-intervention follow-up)											
1 (Manson 2013)	93/8506 (0.26)	69/8102 (0.21)	HR 1.27 (0.93-1.74)	4 more per 1000 (from 1 fewer to 11 more)	Low	Randomised trial with post-interventional follow up	Serious [1,2,3,4]	N/A	No serious	Serious [5]	None
MI (women 50-59 years at baseline, 8.2 years post-intervention follow-up)											
1 (Manson 2013)	75/8506 (0.21)	57/8102 (0.17)	HR 1.25 (0.88-1.76)	4 more per 1000 (from 2 fewer to 11 more)	Low	Randomised trial with post-interventional follow up	Serious [1,2,3,4]	N/A	No serious	Serious[5]	None
Stroke (women aged 50-59 years)											

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Oestrogen plus progesterone	Placebo/ no HRT	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
1 (Manson 2013)	26/2839	16/2683	HR 1.51 (0.81-2.82)	6 more per 1000 (from 2 fewer to 21 more)	Low	Randomised trials	Serious [1,2,3,4]	No serious	No serious	Serious[5]	None
Stroke (women within 5 years from menopause and without prior HRT use)											
1 (Prentice 2009)	N/R	N/R	HR 0.92 (0.38-2.24)	1 fewer per 1000 (from 7 fewer to 14 more)	Very low	Randomised trials	Serious [1,2,3,4]	No serious	No serious	Very Serious[6]	None
Stroke (women within 5 years from menopause and with prior HRT use)											
1 (Prentice 2009)	N/R	N/R	HR 1.20 (0.71-2.03)	2 more per 1000 (from 3 fewer to 12 more)	Very low	Randomised trials	Serious [1,2,3,4]	No serious	No serious	Very Serious[6]	None
Stroke (women within 10 years since menopause)											
1 (Rossouw 2007)	24/2782	15/2712	HR 1.59 (0.81-3.05)	7 more per 1000 (from 2 fewer to 23 more)	Low	Randomised trials	Serious [1,2,3,4]	No serious	No serious	Serious[5]	None
Stroke (women 50-59 years at baseline, 8.2 years post-intervention follow-up)											
1 (Manson 2013)	52/8506 (0.15)	35/8102 (0.10)	HR 1.37 (0.89-2.11)	4 more per 1000 (from 1 fewer to 13 more)	Low	Randomised trial with post-intervention follow up	Serious [1,2,3,4]	No serious	No serious	Serious[5]	None
Reduction of Systolic blood pressure (mmHg), (CEE + MPA cyclic)											
1 (The Writing Group for the PEPI trial, 1995)	174	174	-	MD 0.7 higher (0.6 lower to 2.1 higher)	Moderate	Randomised trials	No serious	No serious	No serious	N/A[7]	None
Reduction of systolic blood pressure (mmHg), (CEE + MPA daily)											
1 (The Writing Group for the PEPI trial, 1995)	174	174	-	MD 1.8 higher (0.6 higher to 3.0 higher)	Moderate	Randomised trials	No serious	No serious	No serious	N/A[7]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Oestrogen plus progesterone	Placebo/ no HRT	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Reduction of systolic blood pressure (mmHg), (CEE + MP cyclic)											
1 (The Writing Group for the PEPI trial, 1995)	178	174	-	MD 0.1 higher(0.1 lower to 1.1 higher)	Moderate	Randomised trials	No serious	No serious	No serious	N/A[7]	None
Reduction of diastolic blood pressure (mmHg), (CEE + MPA cyclic)											
1 (The Writing Group for the PEPI trial, 1995)	174	174	-	MD 1.0 lower(1.8 lower to 0.1 lower)	Moderate	Randomised trials	No serious	No serious	No serious	N/A[7]	None
Reduction of diastolic blood pressure (mmHg), (CEE + MPA daily)											
1 (The Writing Group for the PEPI trial, 1995)	174	174	-	MD 0.2 higher (0.5 lower to 0.9 higher)	Moderate	Randomised trials	No serious	No serious	No serious	N/A[7]	None
Reduction of diastolic blood pressure (mmHg),(CEE + MPA)											
1 (The Writing Group for the PEPI trial, 1995)	178	174	-	MD 0.6 lower (1.3 lower to 0.0)	Moderate	Randomised trials	No serious	No serious	No serious	N/A[7]	None

1. Overall breaking of blinding was relatively high because of the need to unmask 40.5% of the gynaecologists in the treatment group (due to vaginal bleeding after taking HRT) compared with 6.8% in the placebo group, though the distribution of that in the age group this review is interested is unclear;
2. High attrition bias among all participants; about 42% of women in the oestrogen plus progesterone and 38% of women in the placebo stopped taking the study drugs during follow-up. The drop-in rates were 6.2% in the intervention group and 10.7% in the placebo group, respectively. However, the distribution of that in the age group this review is interested is unclear;
3. In the WHI CEE plus progesterone trial, about 26% participants in each group had used HRT prior the enrolment of the trial in their lifetime; and about 6% in each group were current HRT users. The distribution of that in the age group this review is interested is unclear.
4. Among all participants in the WHI, about 34% of women had a BMI higher than 30 kg/m2 in each group. The distribution of that in the age group this review is interested is unclear;
5. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
6. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)
7. The study only reported mean change from baseline for each HRT intervention group without SE or SD, absolute difference between intervention and placebo group could not be derived from that.

Table 48: GRADE profile: oestrogen use alone versus placebo for the outcomes of CHD, MI, stroke and blood pressure change (RCTs)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Oestrogen	Placebo/ no HRT	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
CHD (women aged 50-59 years)											
1 (Manson 2013)	21/1639 (0.17)	35/1674 (0.28)	HR 0.60 (0.35-1.04)	6 fewer per 1000 (from 10 fewer to 1 more)	Low	Randomised trials	Serious[1,2,3,4]	No serious	No serious	Serious[5]	None
CHD (women within 5 years from menopause and without prior HRT use)											
1 (Prentice 2009)	N/R	N/R	N/R (less than 4 events among HRT users)	3 more per 1000 (from 2 fewer to 10 more)	Moderate	Randomised trials	Serious[1,2,3,4]	No serious	No serious	N/A	None
CHD (women within 5 years from menopause and with prior HRT use)											
1 (Prentice 2009)	N/R	N/R	HR1.22 (0.89-1.67)	8 fewer per 1000 (from 12 fewer to 3 more)	Low	Randomised trials	Serious[1,2,3,4]	No serious	No serious	Serious[5]	None
CHD (women within 10 years since menopause)											
1 (Rossouw 2007)	8/826	16/817	HR 0.48 (0.20-1.17)	8 fewer per 1000 (from 12 fewer to 3 more)	Low	Randomised trials	Serious[1,2,3,4]	No serious	No serious	Serious[5]	None
CHD (women 50-59 at baseline, median 5.9 years post-intervention follow-up)											
1 (Lacroix 2009)	33/1223 (0.18)	56/1232 (0.31)	HR 0.59 (0.38-0.90)	6 fewer per 1000 (from 9 fewer to 2 fewer)	Low	Randomised trials with post-interventional follow-up	Serious[1,2,3,4]	No serious	No serious	Serious[5]	None
MI (women 50-59 at baseline, median 5.9 years post-intervention follow-up)											
1 (Lacroix 2009)	27/1223 (0.15)	50/1232 (0.27)	HR 0.54 (0.34-0.86)	7 fewer per 1000 (from 10 fewer to 2 fewer)	Low	Randomised trials with post-interventional follow-up	Serious[1,2,3,4]	No serious	No serious	Serious[5]	None
CHD (women aged 50-59 at baseline, median 6.6 years post-intervention follow-up)											
1 (Manson 2013)	42 (0.21)	64 (0.32)	HR 0.65 (0.44-0.96)	5 fewer per 1000 (from 8 fewer to 1 fewer)	Low	Randomised trials with post-interventional follow-up	Serious[1,2,3,4]	No serious	No serious	Serious[5]	None
MI (women aged 50-59 at baseline, median 6.6 years post-intervention follow-up)											

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Oestrogen	Placebo/ no HRT	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
1 (Manson 2013)	35 (0.17)	58 (0.29)	HR 0.60 (0.39-0.91)	6 fewer per 1000 (from 9 fewer to 1 fewer)	Low	Randomised trials with post-interventional follow-up	Serious[1,2,3,4]	No serious	No serious	Serious[5]	None
Stroke (women aged 50-59 years)											
1 (Manson 2013)	19/1639 (0.16)	21/1674 (0.17)	HR 0.99 (0.53-1.85)	0 more per 1000 (from 5 fewer to 10 more)	Low	Randomised trials	Serious[1,2,3,4]	No serious	No serious	Serious[5]	None
Stroke (women within 5 years since menopause and without prior HRT use)											
1 (Prentice 2009)	N/R	N/R	N/R (less than 4 events in the HRT group)	N/C	Low	Randomised trials	Serious[1,2,3,4]	No serious	No serious	No serious	None
Stroke (women within 5 years since menopause and with prior HRT use)											
1 (Prentice 2009)	N/R	N/R	HR 1.36 (0.98-1.90)	N/C	Low	Randomised trials	Serious[1,2,3,4]	No serious	No serious	Serious[5]	None
Stroke (women within 10 years since menopause)											
1 (Rossouw 2007)	17/826	8/817	HR 2.24 (0.92-5.44)	14 more per 1000 (from 1 fewer to 50 more)	Low	Randomised trials	Serious[1,2,3,4]	No serious	No serious	Serious[5]	None
Stroke (women 50-59 at baseline, median 5.9 years post-intervention follow-up)											
1 (Lacroix 2009)	29/1223 (0.16)	28/1232 (0.15)	HR 1.09 (0.65-1.83)	1 more per 1000 (from 4 fewer to 9 more)	Very low	Randomised trials with post-interventional follow-up	Serious[1,2,3,4]	No serious	No serious	Very serious[6]	None
Stroke (women aged 50-59 at baseline, median 6.6 years follow-up)											
1 (Manson 2013)	33 (0.16)	36 (0.18)	HR 0.96 (0.60-1.55)	0 fewer per 1000 (from 5 fewer to 6 more)	Very low	Randomised trials with post-interventional follow-up	Serious[1,2,3,4]	No serious	No serious	Very serious[6]	None
Ischemic heart disease (IHD) death, (women who have had an MI, aged 50-59 years at baseline, 14 year post-intervention follow-up)											
1 (Cherry 2014)	23/167	14/134	HR 1.23 (0.63-2.41)	24 more per 1000 (from 39 fewer to 148 more)	Very low	Randomised trials with post-interventional follow-up	Serious[7]	No serious	No serious	Very serious[6]	None
Reduction of systolic blood pressure											

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Oestrogen	Placebo/ no HRT	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
1 (The Writing Group for the PEPI trial, 1995)	175	174	-	MD 0.5 higher (0.7 lower to 1.8 higher)	Moderate	Randomised trials	No serious	No serious	No serious	N/A[8]	None
Reduction of diastolic blood pressure											
1 (The Writing Group for the PEPI trial, 1995)	175	174	-	MD 0.7 lower (1.5 lower to 0.1 higher)	Moderate	Randomised trials	No serious	No serious	No serious	N/A[8]	None

1. An average follow-up of 6.8 years, the study was terminated earlier than expected;
2. Relatively high drop-out and drop- in rates in both the CEE and placebo groups in the WHI CEE trial. When the CEE trial was terminated, earlier than expected, overall about 54% of women had already stopped taking study medication. About 5.7% in the CEE group and 9.1% in the placebo group initiated HRT use through their own clinicians. However, the distribution of that in the age group this review is interested is unclear;
3. In the WHI CEE trial, about 36% participants in each group had used HRT prior the enrolment of the trial in their lifetime; and about 13% in each group were current HRT users. However, the distribution of that in the age group this review is interested is unclear;
4. BMI was high in both groups at baseline (mean 30.1 ± 6.1 in CEE group and 30.1 ± 6.2 in the placebo group, respectively), not really the "healthy" women at baseline as the authors stated. However, the distribution of that in the age group this review is interested is unclear.
5. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25);
6. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25);
7. Evidence was downgraded by 1 due to selection (participants were originally recruited from an RCT). HRT use or not during post-study intervention phase which this study examined was not followed up or ascertained;
8. The study reported mean change from baseline for each HRT intervention group without SE or SD, absolute difference between intervention and placebo group could not be derived from that.

Table 49: GRADE profile: HRT use versus no HRT use for the outcomes of CHD, MI, CVD, CHD death, CVD death, IHD, IHD death, stroke, ischemic stroke, haemorrhagic stroke and stroke death (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT (mainly oestrogen)	Non users	Hazard ratio or risk ratio(95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
CHD (current users)											
4 (Hedblad 2002; Lokkegaard	N/A	N/A	HR 0.91 (0.85-0.98)	1 fewer per 1000 (from 2	Very low	Prospective cohort	Serious 1,2	Serious3	Serious4	No serious	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT (mainly oestrogen)	Non users	Hazard ratio or risk ratio(95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
2008; Stram 2011; Grodstein 2000-The NHS-)				fewer to 0 fewer)							
CHD (women aged <55 years)											
2 (Lokkegaard 2008; Weiner 2008)	N/A	N/A	HR 1.18 (0.98-1.41)	3 more per 1000 (from 0 fewer to 6 more)	Very low	Prospective cohort	Serious 1	Serious3	No serious	Serious5	None
CHD (women aged 50-59 years at baseline)											
1 (Rossouw 2007)	59/4479	61/4356	HR 0.93 (0.65-1.33)	1 fewer per 1000 (from 5 fewer to 5 more)	Very low	Prospective cohort	Serious6	No serious	No serious	Very Serious7	None
CHD (duration > 2 years)											
2 (Folsom 1995; Grodstein 2000-the NHS)	N/A	N/A	HR 0.68 (0.59-0.79)	5 fewer per 1000 (from 6 fewer to 3 fewer)	Very low	Prospective cohort	Serious	Serious3	Serious4	Serious5	None
CHD (duration > 5 years)											
2 (Folsom 1995; Grodstein 2000-the NHS)	N/A	N/A	HR 0.68 (0.56-0.83)	5 fewer per 1000 (from 7 fewer to 3 fewer)	Very low	Prospective cohort	Serious1	No Serious	Serious4	Serious5	None
CHD (current users, 4-year follow-up)											
1 (Stampfer 1985)	N/R	N/R	RR 0.30 (0.14-0.64)a	11 fewer per 1000 (from 13 fewer to 5 fewer)	Low	Prospective cohort	Serious1	No serious	Serious8	No serious	None
CHD (past users, 4-year follow-up)											
1 (Stampfer 1985)	N/R	N/R	RR 0.59 (0.33-1.66)a	6 fewer per 1000 (from 10 fewer to 10 more)	Very low	Prospective cohort	Serious1	No serious	Serious8	Very serious7	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT (mainly oestrogen)	Non users	Hazard ratio or risk ratio(95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
MI (current users, 4-year follow-up)											
1 (Stampfer 1985)	N/R	N/R	RR 0.34 (0.14-0.82)	10 fewer per 1000 (from 13 fewer to 3 fewer)	Very low	Prospective cohort	Serious1	No serious	Serious8	Serious5	None
MI (past users, 4-year follow-up)											
1 (Stampfer 1985)	N/R	N/R	RR 0.65 (0.33-1.28)	5 fewer per 1000 (from 10 fewer to 4 more)	Very low	Prospective cohort	Serious1	No serious	Serious8	Very serious7	None
CHD (ever users , 4-year follow-up)											
1 (Stampfer 1985)	30/54,308.7	60/51,477.5	RR (0.5 (0.3-0.8)	8 fewer per 1000 (from 11 fewer to 3 fewer)	Very low	Prospective cohort	Serious1	No serious	Serious8	Serious5	None
CHD (current users, 4-year follow-up)											
1 (Stampfer 1985)	11/29,922.0	60/51,477.5	RR 0.3 (0.2-0.6)	11 fewer per 1000 (from 12 fewer to 6 fewer)	Very low	Prospective cohort	Serious1	No serious	Serious8	No serious	None
CHD (ever users aged 40-44 years, 4-year follow-up)											
1 (Stampfer 1985)	2/5401.9	1/2073.3	RR 0.8 (0.1-4.6)	3 fewer per 1000 (from 14 fewer to 54 more)	Very low	Prospective cohort	Serious1	No serious	Serious8	Very serious7	None
CHD (current users aged 40-44 years, 4-year follow-up)											
1 (Stampfer 1985)	1/3833.0	1/2073.3	RR 0.6 (0.2-2.4)	6 fewer per 1000 (from 12 fewer to 21 more)	Very low	Prospective cohort	Serious1	No serious	Serious8	Very serious7	None
CHD (ever users aged 45-49 years, 4-year follow-up)											
1 (Stampfer 1985)	3/11,064	11/9106.9	RR 0.2 (0.1-0.7)	12 fewer per 1000 (from 14 fewer to 5 fewer)	Low	Prospective cohort	Serious1	No serious	Serious8	No serious	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT (mainly oestrogen)	Non users	Hazard ratio or risk ratio(95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
CHD (current users aged 45-49 years, 4-year follow-up)											
1 (Stampfer 1985)	2/6,890	11/9106.9	RR 0.2 (0.1-0.9)	12 fewer per 1000 (from 14 fewer to 2 fewer)	Very low	Prospective cohort	Serious1	No serious	Serious8	Serious5	None
CHD (ever users aged 50-59 years, 4-year follow-up)											
1 (Stampfer 1985)	323/30,045	40/34,197.6	RR 0.6 (0.4-1.1)	6 fewer per 1000 (from 9 fewer to 2 more)	Very low	Prospective cohort	Serious1	No serious	Serious8	Serious5	None
CHD (current users aged 50-55 years, 4-year follow-up)											
1 (Stampfer 1985)	8/15,239.2	40/34,197.6	RR 0.4 (0.2-0.9)	9 fewer per 1000 (from 12 fewer to 2 fewer)	Very low	Prospective cohort	Serious1	No serious	Serious8	Serious5	None
CHD (ever users aged 56-59 years, 4-year follow-up)											
1 (Stampfer 1985)	2/4837.2	8/5238.7	RR 0.3 (0.1-1.1)	11 fewer per 1000 (from 14 fewer to 2 more)	Very low	Prospective cohort	Serious1	No serious	Serious8	Serious5	None
CHD (current users aged 56-59 years, 4-year follow-up)											
1 (Stampfer 1985)	0/1721.4	8/5238.7	RR 0	N/C	N/A	Prospective cohort	Serious1	No serious	Serious8	N/A	None
CHD (current users, 10-year follow-up)											
1 (Stampfer 1991)	45/75,532	250/179,194	RR 0.56 (0.40-0.80)	7 fewer per 1000 (from 9 fewer to 3 fewer)	Very low	Prospective cohort	Serious1	No serious	Serious8	Serious5	None
CHD (former users, 10-year follow-up)											
1 (Stampfer 1991)	110/85,128	250/179,194	RR 0.83 (0.65-1.05)	3 fewer per 1000 (from 5 fewer to 1 more)	Very low	Prospective cohort	Serious1	No serious	Serious8	Serious5	None
CVD (current users, 10-year follow-up)											

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT (mainly oestrogen)	Non users	Hazard ratio or risk ratio(95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
1 (Stampfer 1991)	21/75,532	129/179,194	RR 0.61 (0.37-1.00)	N/C	Very low	Prospective cohort	Serious1	No serious	Serious8	Serious5	None
CVD (former users, 10-year follow-up)											
1 (Stampfer 1991)	55/85,128	129/179,194	RR 0.79 (0.56-1.10)	N/C	Very low	Prospective cohort	Serious1	No serious	Serious8	Serious5	None
CHD (current users, 16-year follow-up)											
1 (Grodstein 1996)	98/166,371	452/324,748	RR 0.60 (0.47-0.76)	6 fewer per 1000 (from 8 fewer to 4 fewer)	Very low	Prospective cohort	Serious1	No serious	Serious8	Serious5	None
CHD (past users, 16-year follow-up)											
1 (Grodstein 1996)	195/150,238	452/324,748	RR 0.85 (0.71-1.01)	2 fewer per 1000 (from 4 fewer to 0 fewer)	Very low	Prospective cohort	Serious1	No serious	Serious8	Serious5	None
CHD (current users aged < 50 years, 16-year follow-up)											
1 (Grodstein 1996)	33/166,371	79/324,748	RR 0.90 (0.57-1.41)	2 fewer per 1000 (from 6 fewer to 6 more)	Very low	Prospective cohort	Serious1	No serious	Serious8	Very Serious7	None
CHD (current users aged 50-59 years, 16-year follow-up)											
1 (Grodstein 1996)	61/92,922	272/213,636	RR 0.71 (0.52-0.96)	4 fewer per 1000 (from 7 fewer to 1 fewer)	Very low	Prospective cohort	Serious1	No serious	Serious8	Serious5	None
CVD death (current users, 16-year follow-up)											
1 (Grodstein 1996)	No of cases: 43	No. of cases: 289	RR 0.47 (0.32-0.69)	N/C	Very low	Prospective cohort	Serious1	No serious	Serious8	Serious5	None
CVD death (past users, 16-year follow-up)											
1	No of cases: 129	No. of cases: 289	RR 0.99 (0.75-1.30)	N/C	Very low	Prospective cohort	Serious1	No serious	Serious8	Serious5	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT (mainly oestrogen)	Non users	Hazard ratio or risk ratio(95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
(Grodstein 1996)											
CHD (past users, 20-year follow-up)											
1 (Grodstein 2000)	337/185,497	662/358,125	RR 0.82 (0.72-0.94)	3 fewer per 1000 (from 4 fewer to 1 fewer)	Very low	Prospective cohort	Serious1	No serious	Serious8	Serious5	None
CHD (current users, 20-year follow-up)											
1 (Grodstein 2000)	259/265,203	662/358,125	RR 0.61 (0.52-0.71)	6 fewer per 1000 (from 7 fewer to 4 fewer)	Low	Prospective cohort	Serious1	No serious	Serious8	No serious	None
CHD (current users of < 1-year duration, 20-year follow-up)											
1 (Grodstein 2000)	9/20,091	662/358,125	RR 0.40 (0.21-0.77)	9 fewer per 1000 (from 12 fewer to 3 fewer)	Low	Prospective cohort	Serious1	No serious	Serious8	No serious	None
CHD (current users of < 1 to 1.9 year duration, 20-year follow-up)											
1 (Grodstein 2000)	9/19,155	662/358,125	RR 0.41 (0.21-0.80)	9 fewer per 1000 (from 12 fewer to 3 fewer)	Very low	Prospective cohort	Serious1	No serious	Serious8	Serious5	None
CHD (current users of 2 to 4.9 years duration, 20-year follow-up)											
1 (Grodstein 2000)	60/78,928	662/358,125	RR 0.53(0.41-0.70)	7 fewer per 1000 (from 9 fewer to 5 fewer)	Low	Prospective cohort	Serious1	No serious	Serious8	No serious	None
CHD (current users 5 to 9.9 years duration, 20-year follow-up)											
1 (Grodstein 2000)	74/77,435	662/358,125	RR 0.58 (0.45-0.74)	6 fewer per 1000 (from 8 fewer to 4 fewer)	Low	Prospective cohort	Serious1	No serious	Serious8	No serious	None
CHD (current users ≥ 10 year duration, 20-year follow-up)											
1	107/69,594	662/358,125	RR 0.74 (0.59-0.91)	4 fewer per 1000 (from 6	Very low	Prospective cohort	Serious1	No serious	Serious8	Serious5	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT (mainly oestrogen)	Non users	Hazard ratio or risk ratio(95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
(Grodstein 2000)				fewer to 1 fewer)							
CHD death (current users)											
4 (Ettinger 1996; Graff-Iversen 2004; Pentti 2001; Shilpak. 2001)	N/A	N/A	HR 0.66 (0.53-0.8.3)	N/C	Very low	Retrospective and prospective cohort	Serious2	Very serious9	No serious	Serious5	None
CHD death (duration > 5 years)											
2 (Ettinger 1996; Pentti 2006)	N/A	N/A	HR 0.91 (0.48-1.73)	N/C	Very low	Retrospective and prospective cohort	Serious2	Very serious9	No serious	Very serious 7	None
CVD death (current users)											
4 (Ettinger 1996; Graff-Inversen 2004; Laffety 1994; Sourander 1998)	N/A	N/A	HR 0.41 (0.26-0.64)	N/C	Low	Retrospective and prospective cohort	Serious2	Serious3	No serious	No serious	None
CHD (current users, among women with pre-existing heart disease)											
2 (Alexander. 2001; Hernandez 1990)	N/A	N/A	HR 1.40 (1.02-1.91)	9 more per 1000 (from 0 fewer to 20 more)	Low	Retrospective and prospective cohort	Serious 10	No serious	No serious	Serious5	None
CHD (prior and current users of 2 year duration among women with pre-existing heart disease)											
1 (Alexander 2001)	N/R	N/R	HR 0.94 (0.75-1.18)	1 fewer per 1000 (from 5 fewer to 3 more)	Moderate	Prospective cohort	Serious 10	No serious	No serious	No serious	None
CHD death (prior and current users of > 2 year duration among women with pre-existing heart disease)											

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT (mainly oestrogen)	Non users	Hazard ratio or risk ratio(95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
1 (Alexander 2001)	N/R	N/R	HR 0.36 (0.17-0.77)	14 fewer per 1000 (from 18 fewer to 5 fewer)	Low	Prospective cohort	Serious 10	No serious	No serious	Serious5	None
MI (prior and current users > 2 year duration among women who have had an MI)											
1 (Alexander 2001)	N/R	N/R	HR 0.88 (0.58-1.33)	3 fewer per 1000 (from 9 fewer to 7 more)	Very low	Prospective cohort	Serious10	No serious	No serious	Very serious7	None
CHD (ever users without flushing symptoms)											
1 (Gast 2011)	N/R	N/R	HR 1.11 (0.73-1.69)	2 more per 1000 (from 4 fewer to 10 more)	Very low	Prospective cohort	Serious2	No serious	No serious	Very serious7	None
CHD (ever users with flushing symptoms)											
1 (Gast 2011)	N/R	N/R	HR 1.18 (0.78-1.79)	3 more per 1000 (from 3 fewer to 12 more)	Low	Prospective cohort	Serious2	No serious	No serious	Serious5	None
CHD (ever users without night sweat)											
1 (Gast 2011)	N/R	N/R	HR 1.35 (0.91-2.01)	5 more per 1000 (from 1 fewer to 15 more)	Low	Prospective cohort	Serious2	No serious	No serious	Serious5	None
CHD (ever users with night sweat)											
1 (Gast 2011)	N/R	N/R	HR 0.89 (0.57-1.38)	2 fewer per 1000 (from 6 fewer to 6 more)	Very low	Prospective cohort	Serious2	No serious	No serious	Very Serious7	None
CHD (ever users without intense vasomotor symptoms)											
1 (Gast 2011)	N/R	N/R	HR 1.26 (0.92-1.72)	4 more per 1000 (from 1 fewer to 11 more)	Low	Prospective cohort	Serious2	No serious	No serious	Serious5	None
CHD (ever users with intense vasomotor symptoms)											

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT (mainly oestrogen)	Non users	Hazard ratio or risk ratio(95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
1 (Gast 2011)	N/R	N/R	HR 0.51 (0.21-1.23)	7 fewer per 1000 (from 12 fewer to 3 more)	Low	Prospective cohort	Serious2	No serious	No serious	Serious5	None
Ischemic heart disease hospitalisation (IHD) (any route of administration, 7-12 months duration)											
1 (Corrao 2007)	N/R	N/R	HR 1.00 (0.80-1.26)	0 fewer per 1000 (from 3 fewer to 4 more)	Low	Prospective cohort	Serious11, 12	No serious	No serious	Serious5	None
Ischemic heart disease hospitalisation (IHD) (any route of administration, 13-24 months duration)											
1 (Corrao 2007)	N/R	N/R	HR 0.85 (0.65-1.11)	2 fewer per 1000 (from 5 fewer to 2 more)	Low	Prospective cohort	Serious11, 12	No serious	No serious	Serious5	None
Ischemic heart disease hospitalisation (IHD) (any route of administration, 25-36 months duration)											
1 (Corrao 2007)	N/R	N/R	HR 0.83 (0.58-1.20)	3 fewer per 1000 (from 6 fewer to 3 more)	Low	Prospective cohort	Serious11, 12	No serious	No serious	Serious5	None
Ischemic heart disease hospitalisation (IHD) (any route of administration, > 36 months duration)											
1 (Corrao 2007)	N/R	N/R	HR 0.61 (0.37-0.99)	6 fewer per 1000 (from 9 fewer to 0 fewer)	Low	Prospective cohort	Serious11, 12	No serious	No serious	Serious5	None
Ischemic heart disease hospitalisation (IHD) (transdermal administration, 7-12 months duration)											
1 (Corrao 2007)	N/R	N/R	HR 1.03 (0.82-1.30)	0 fewer per 1000 (from 3 fewer to 5 more)	Low	Prospective cohort	Serious11, 12	No serious	No serious	Serious5	None
Ischemic heart disease hospitalisation (IHD) (transdermal administration, 13-24 months duration)											
1 (Corrao 2007)	N/R	N/R	HR 0.79 (0.59-1.05)	3 fewer per 1000 (from 6 fewer to 1 more)	Low	Prospective cohort	Serious11, 12	No serious	No serious	Serious5	None
Ischemic heart disease hospitalisation (IHD) (transdermal administration, 25-36 months duration)											

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT (mainly oestrogen)	Non users	Hazard ratio or risk ratio(95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
1 (Corrao 2007)	N/R	N/R	HR 0.83 (0.56-1.24)	3 fewer per 1000 (from 7 fewer to 4 more)	Low	Prospective cohort	Serious11, 12	No serious	No serious	Serious5	None
Ischemic heart disease hospitalisation (IHD) (transdermal administration, > 36 months duration)											
1 (Corrao 2007)	N/R	N/R	HR 0.59 (0.33-1.05)	6 fewer per 1000 (from 10 fewer to 1 more)	Low	Prospective cohort	Serious11, 12	No serious	No serious	Serious5	None
Ischemic heart disease hospitalisation (IHD) (oral administration, 7-12 months duration)											
1 (Corrao 2007)	N/R	N/R	HR 1.08 (0.75-1.55)	1 more per 1000 (from 4 fewer to 8 more)	Low	Prospective cohort	Serious11, 12	No serious	No serious	Serious5	None
Ischemic heart disease hospitalisation (IHD) (oral administration, 13-24 months duration)											
1 (Corrao 2007)	N/R	N/R	HR 0.60 (0.31-1.14)	6 fewer per 1000 (from 10 fewer to 2 more)	Low	Prospective cohort	Serious11, 12	No serious	No serious	Serious5	None
Ischemic heart disease hospitalisation (IHD) (oral administration, 25-36 months duration)											
1 (Corrao 2007)	N/R	N/R	HR 1.02 (0.38-2.75)	0 fewer per 1000 (from 9 fewer to 26 more)	Very low	Prospective cohort	Serious11, 12	No serious	No serious	Very serious7	None
Ischemic heart disease hospitalisation (IHD) (oral administration, > 36 months duration)											
1 (Corrao 2007)	N/A	N/A	HR 1.80 (0.66-4.88)	12 more per 1000 (from 5 fewer to 58 more)	Very low	Prospective cohort	Serious11, 12	No serious	No serious	Very serious7	None
Ischemic heart disease (IHD) death (former users aged 36-59 years)											
1 (Stram 2011)	4/23,189	23/48,219	HR 0.37 (0.13-1.06)	9 fewer per 1000 (from 13 fewer to 1 more)	Very low	Prospective cohort	Serious1	No serious	Serious13	Serious5	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT (mainly oestrogen)	Non users	Hazard ratio or risk ratio(95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Ischemic heart disease (IHD) death (former users aged 60-64 years)											
1 (Stram 2011)	6/13,042	19/20,983	HR 0.52 (0.21-1.27)	7 fewer per 1000 (from 12 fewer to 4 more)	Very low	Prospective cohort	Serious1	No serious	Serious13	Very serious7	None
Ischemic heart disease (IHD) death (women started HRT at age 45-54 years)											
1 (Stram 2011)	N/R	N/R	HR 0.91 (0.72-1.15)	1 fewer per 1000 (from 4 fewer to 2 more)	Very low	Prospective cohort	Serious1	No serious	Serious13	Serious5	None
Ischemic heart disease (IHD) death (women started HRT at age 55-64 years)											
1 (Stram 2011)	N/R	N/R	HR 1.05 (0.87-1.27)	1 more per 1000 (from 2 fewer to 4 more)	Very low	Prospective cohort	Serious1	No serious	Serious13	Serious5	None
Stroke (current users)											
3 (Grodstein 2000-the NHS; Li 2006; Sourander 1996)	N/A	N/A	HR 1.30 (1.14-1.48)	3 more per 1000 (from 2 more to 5 more)	Very low	Prospective cohort	Serious1	No serious	Serious4	Serious5	None
Stroke (current users, 10-year follow-up)											
1 (Stampfer 1991)	39/75,532	123/179,194	RR 0.97 (0.65-1.45)	0 fewer per 1000 (from 4 fewer to 5 more)	Very low	Prospective cohort	Serious1	No serious	Serious4	Very serious7	None
Stroke (former users, 10-year follow-up)											
1 (Stampfer 1991)	62/85,128	123/179,194	RR 0.99 (0.72-1.36)	0 fewer per 1000 (from 3 fewer to 4 more)	Very low	Prospective cohort	Serious1	No serious	Serious4	Very serious7	None
Ischemic stroke (current users, 10-year follow-up)											

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT (mainly oestrogen)	Non users	Hazard ratio or risk ratio(95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
1 (Stampfer 1991)	23/75,532	56/179,194	RR 1.46 (0.85-2.51)	5 more per 1000 (from 2 fewer to 17 more)	Very low	Prospective cohort	Serious1	No serious	Serious4	Serious5	None
Ischemic stroke (former users, 10-year follow-up)											
1 (Stampfer 1991)	34/85,128	56/179,194	RR 1.19 (0.77-1.86)	2 more per 1000 (from 3 fewer to 10 more)	Very low	Prospective cohort	Serious1	No serious	Serious4	Serious5	None
Subarachnoid haemorrhage (current users, 10-year follow-up)											
1 (Stampfer 1991)	5/75,532	19/179,194	RR 0.53 (0.18-1.57)	5 fewer per 1000 (from 9 fewer to 6 more)	Very low	Prospective cohort	Serious1	No serious	Serious4	Very serious7	None
Subarachnoid haemorrhage (former users, 10-year follow-up)											
1 (Stampfer 1991)	12/85,128	19/179,194	RR 1.03 (0.47-2.25)	0 fewer per 1000 (from 6 fewer to 14 more)	Very low	Prospective cohort	Serious1	No serious	Serious4	Serious5	None
Stroke (current users, 16-year follow-up)											
1 (Grodstein 1996)	121/166,371	279/324,748	RR 1.03 (0.82-1.31)	0 fewer per 1000 (from 2 fewer to 4 more)	Very low	Prospective cohort	Serious1	No serious	Serious4	Serious5	None
Stroke (past users, 16-year follow-up)											
1 (Grodstein 1996)	152/150,238	279/324,748	RR 0.99 (0.80-1.22)	0 fewer per 1000 (from 2 fewer to 2 more)	Very low	Prospective cohort	Serious1	No serious	Serious4	Serious5	None
Ischemic stroke (current users, 16-year follow-up)											
1 (Grodstein 1996)	73/163,371	133/324,748	RR 1.40 (1.02-1.92)	5 more per 1000 (from 0 fewer to 10 more)	Very low	Prospective cohort	Serious1	No serious	Serious4	Serious5	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT (mainly oestrogen)	Non users	Hazard ratio or risk ratio(95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Ischemic stroke (past users, 16-year follow-up)											
1 (Grodstein 1996)	75/150,238	133/324,748	RR 1.19 (0.89-1.57)	2 more per 1000 (from 1 fewer to 6 more)	Very low	Prospective cohort	Serious1	No serious	Serious4	Serious5	None
Subarachnoid stroke (current users, 16-year follow-up)											
1 (Grodstein 1996)	33/166,371	79/324,748	RR 0.90 (0.57-1.41)	0 fewer per 1000 (from 0 fewer to 0 more)	Very low	Prospective cohort	Serious1	No serious	Serious4	Very Serious7	None
Subarachnoid stroke (past users, 16-year follow-up)											
1 (Grodstein 1996)	32/150,238	79/324,748	RR 0.81(0.52-1.25)	0 fewer per 1000 (from 0 fewer to 0 more)	Very low	Prospective cohort	Serious1	No serious	Serious4	Serious5	None
Stroke (past users, 20-year follow-up)											
1 (Grodstein 2000)	217/185,497	312/358,125	RR 1.02 (0.85-1.24)	0 fewer per 1000 (from 2 fewer to 3 more)	Low	Prospective cohort	Serious1	No serious	Serious4	No serious	None
Stroke (current users, 20-year follow-up)											
1 (Grodstein 2000)	238/265,203	312/358,125	RR 1.13 (0.94-1.35)	1 more per 1000 (from 1 fewer to 4 more)	Very low	Prospective cohort	Serious1	No serious	Serious4	Serious5	None
Stroke (Current users of < 1 year duration, 20-year follow-up)											
1 (Grodstein 2000)	13/20,091	312/358,125	RR 1.32 (0.76-2.32)	4 more per 1000 (from 3 fewer to 15 more)	Very low	Prospective cohort	Serious1	No serious	Serious4	Serious5	None
Stroke (Current users of 1 to 1.9 year duration, 20-year follow-up)											
1 (Grodstein 2000)	10/19,155	312/358,125	RR 1.04 (0.55-1.97)	0 fewer per 1000 (from 5 fewer to 11 more)	Very low	Prospective cohort	Serious1	No serious	Serious4	Very serious7	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT (mainly oestrogen)	Non users	Hazard ratio or risk ratio(95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Stroke (Current users of 2 to 4.9 year durationd, 20-year follow-up)											
1 (Grodstein 2000)	61/78,928	312/358,125	RR 1.14 (0.86-1.52)	2 more per 1000 (from 2 fewer to 6 more)	Very low	Prospective cohort	Serious1	No serious	Serious4	Serious5	None
Stroke (current users of 5-9.9 year duration, 20-year follow-up)											
1 (Grodstein 2000)	63/77,435	312/358,125	RR 1.05 (0.79-1.38)	1 more per 1000 (from 2 fewer to 4 more)	Very low	Prospective cohort	Serious1	No serious	Serious4	Serious5	None
Stroke (Current users of ≥ 10 year duration, 20-year follow-up)											
1 (Grodstein 2000)	91/65,594	312/358,125	RR 1.17 (0.91-1.49)	2 more per 1000 (from 1 fewer to 6 more)	Very low	Prospective cohort	Serious1	No serious	Serious4	Serious5	None
Stroke (women aged 50-59 years at baseline) HRT											
1 (Rossouw 2007)	44/4476	37/4356	HR 1.13 (0.73-1.76)	1 more per 1000 (from 3 fewer to 9 more)	Very low	Prospective cohort	Serious6	No serious	No serious	Very Serious7	None
Stroke (women aged < 55 years)											
1 (Weiner 2008)	N/A	N/A	HR 1.46 (1.11-1.92)	5 more per 1000 (from 1 more to 10 more)	Low	Prospective cohort	Serious1	No serious	Np serious	Serious5	None
Stroke (age < 55 years)											
1 (Su 2012)	17 (434)	18/515	HR 0.99 (0.50-1.95)	0 fewer per 1000 (from 6 fewer to 11 more)	Very low	Retrospective cohort	Serious1	No serious	Serious14	Very serious7	None
Stroke (duration > 2 years)											
3 (Folsom 1995; Grodstein	N/A	N/A	HR 1.12 (0.91-1.38)	1 more per 1000 (from 1 fewer to 4 more)	Very low	Prospective cohort	Serious1	No serious	Serious4	Serious5	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT (mainly oestrogen)	Non users	Hazard ratio or risk ratio(95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
2000-The NHS)											
Stroke (duration > 5 years)											
3 (Folsom 1995; Grodstein 2000-The NHS)	N/A	N/A	HR 1.11 (0.89-1.38)	N/C	Very low	Prospective cohort	Serious1	No serious	Serious4	Serious5	None
Hospitalisation of cerebrovascular disease (any route of administration , 7-12 months HRT duration)											
1 (Corrao 2007)	N/R	N/R	HR 0.82 (0.61-1.10)	N/C	Low	Prospective cohort	Serious11, 12	No serious	No serious	Serious5	None
Hospitalisation of cerebrovascular disease (any route of administration, 13-24 months HRT duration)											
1 (Corrao 2007)	N/R	N/R	HR 0.74 (0.53-1.06)	N/C	Low	Prospective cohort	Serious11, 12	No serious	No serious	Serious5	None
Hospitalisation of cerebrovascular disease (any route of administration , 25-36 months HRT duration)											
1 (Corrao 2007)	N/R	N/R	HR (95%CI): 0.57 (0.34-0.94)	N/C	Low	Prospective cohort	Serious11, 12	No serious	No serious	Serious5	None
Hospitalisation of cerebrovascular disease (any route of administration, > 36 months HRT duration)											
1 (Corrao 2007)	N/R	N/R	HR (95%CI): 0.53 (0.30-0.94)	N/C	Low	Prospective cohort	Serious11, 12	No serious	No serious	Serious5	None
Hospitalisation of cerebrovascular disease (transdermal administration , 7-12 months HRT duration)											
1 (Corrao 2007)	N/R	N/R	HR (95%CI): 0.73 (0.53-0.99)	N/C	Low	Prospective cohort	Serious 11,12	No serious	No serious	Serious5	None
Hospitalisation of cerebrovascular disease (transdermal administration, 13-24 months HRT duration)											
1 (Corrao 2007)	N/R	N/R	HR (95%CI): 0.81 (0.58-1.15)	N/C	Low	Prospective cohort	Serious11,12	No serious	No serious	Serious5	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT (mainly oestrogen)	Non users	Hazard ratio or risk ratio(95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Hospitalisation of cerebrovascular disease (transdermal administration , 25-36 months HRT duration)											
1 (Corrao 2007)	N/R	N/R	HR (95%CI): 0.50 (0.29-0.87)	N/C	Low	Prospective cohort	Serious1 1,12	No serious	No serious	Serious5	None
Hospitalisation of cerebrovascular disease (transdermal administration, > 36 months HRT duration)											
1 (Corrao 2007)	N/R	N/R	HR (95%CI): 0.39 (0.18-0.82)	N/C	Low	Prospective cohort	Serious1 1,12	No serious	No serious	Serious5	None
Hospitalisation of cerebrovascular disease (oral administration, 7-12 months HRT duration)											
1 (Corrao 2007)	N/R	N/R	HR (95%CI): 1.21 (0.78-1.90)	N/C	Low	Prospective cohort	Serious1 1,12	No serious	No serious	Serious5	None
Hospitalisation of cerebrovascular disease (oral administration, 13-24 months HRT duration)											
1 (Corrao 2007)	N/R	N/R	HR (95%CI): 1.26 (0.69-2.31)	N/C	Very low	Prospective cohort	Serious 11,12	No serious	No serious	Very serious7	None
Hospitalisation of cerebrovascular disease (oral administration, 25-36 months HRT duration)											
1 (Corrao 2007)	N/R	N/R	HR (95%CI): 0.73 (0.18-2.93)	N/C	Very low	Prospective cohort	Serious 11,12	No serious	No serious	Very serious7	None
Hospitalisation of cerebrovascular disease (oral administration, > 36 months HRT duration)											
1 (Corrao 2007)	N/R	N/R	HR (95%CI): 0.54 (0.08-3.86)	N/C	Very low	Prospective cohort	Serious1 1,12	No serious	No serious	Very serious7	None
Stroke death (current users, 16-year follow-up)											
1 (Grodstein 1996)	No of cases: 28	No. of cases: 91	RR (95% CI): 0.68 (0.39-1.16)	4 fewer per 1000 (from 7 fewer to 2 more)	Very low	Prospective cohort	Serious1	No serious	Serious4	Serious5	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT (mainly oestrogen)	Non users	Hazard ratio or risk ratio(95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Stroke death (past users, 16-year follow-up)											
1 (Grodstein 1996)	No of cases: 48	No. of cases: 91	RR (95% CI): 1.07 (0.68-1.69)	1 more per 1000 (from 4 fewer to 8 more)	Very low	Prospective cohort	Serious1	No serious	Serious4	Very serious7	None

1. Evidence was downgraded by 1 due to selection (HRT users were "healthier" and "younger" than non-users at baseline, with lower BMI, BP, or triglycerides levels);
2. Evidence was downgraded by 1 due to measurement bias (self-reported HRT use information at baseline and was just taken once at baseline in some studies, change of exposure over the follow-up time was not updated)
3. Evidence was downgraded by 1 due to serious heterogeneity (chi-squared p<0.1, I-squared inconsistency statistic of 50% -74.99%);
4. Evidence has only 1 indirect aspect of PICO (population-the NHS was carried out among registered nurses only-);
5. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25);
6. Observational data in nature, the re-analyses (WHI) of data inherited all the risk of biases from the 2 original trials as reported above tables;
7. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)
8. Evidence has only 1 indirect aspect of PICO (population-the study was conducted among registered nurses only-);
9. Evidence was downgraded by 2 due to very serious heterogeneity (chi-squared p<0.1, I-squared statistic of > 75%)
10. Evidence was downgraded by 1 due to selection (participants were subjects enrolled in a prior randomised trial);
11. Evidence was downgraded due to measurement bias (exposure to HRT among some women (especially the older) in their lifetime before the study might not be captured);
12. Evidence was downgraded by 1 due to important confounders not adjusted for in analyses (several lifestyle factors, such as smoking, alcohol drinking, physical exercises were not controlled for in analyses due to lack of data availability);
13. Evidence has only 1 indirect aspect of PICO (population _the study was carried out among teachers only-);
14. Evidence has only 1 indirect aspect of PICO (population _the study was carried out among Chinese women only-);

Table 50: GRADE profile: oestrogen alone versus no HRT use for the outcomes of CHD, stroke, ischemic stroke, haemorrhagic stroke, fatal stroke and non-fatal stroke (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Oestrogen	Non users	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
CHD (current users, 16-year follow-up)											
1 (Grodstein 1996)	47/82,626	43/304,744	RR 0.60 (0.43-0.83)	6 fewer per 1000 (from 9 fewer to 3 fewer)	Very low	Prospective cohort	Serious[1]	No serious	Serious[2]	Serious[3]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Oestrogen	Non users	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
CHD (current users, 24-year follow-up)											
1 (Grodstein 2006)	225/206,383	795/429,032	RR 0.71(0.61-0.83)	4 fewer per 1000 (from 6 fewer to 3 fewer)	Very low	Prospective cohort	Serious[1]	No serious	Serious[2]	Serious[3]	None
CHD (current users, 24-year follow-up, including women with and without existing heart disease)											
1 (Grodstein 2006)	274/220,368	922/449,599	RR 0.72 (0.62-0.82)	4 fewer per 1000 (from 6 fewer to 3 fewer)	Very low	Prospective cohort	Serious[1]	No serious	Serious[2]	Serious[3]	None
Stroke (current users, 16-year follow-up)											
1 (Grodstein 1996)	74/82,626	270/304,744	RR 1.27 (0.95-1.69)	3 more per 1000 (from 1 fewer to 8 more)	Very low	Prospective cohort	Serious[1]	No serious	Serious[2]	Serious[3]	None
Stroke (current users, 28-year follow-up)											
1 (Grodstein 2008)	276/256,437	360/485,987	RR 1.39 (1.18-1.63)	4 more per 1000 (from 2 more to 7 more)	Very low	Prospective cohort	Serious[1]	No serious	Serious[2]	Serious[3]	None
Ischemic stroke (current users, 28-year follow-up)											
1 (Grodstein 2008)	183/256,437	235/485,987	RR 1.43 (1.17-1.74)	5 more per 1000 (from 2 more to 8 more)	Very low	Prospective cohort	Serious[1]	No serious	Serious[2]	Serious[3]	None
Hemorrhagic stroke (Current users, 28-year follow-up)											
1 (Grodstein 2008)	61/256,437	85/485,987	RR 1.37 (0.98-1.91)	0 more per 1000 (from 0 fewer to 0 more)	Very low	Prospective cohort	Serious1	No serious	Serious[2]	Serious[3]	None
Fatal stroke (current users, 28-year follow-up)											
1 (Grodstein 2008)	33/256,437	50/485,987	RR 1.22 (0.78-1.90)	2 more per 1000 (from 2 fewer to 10 more)	Very low	Prospective cohort	Serious[1]	No serious	Serious[2]	Serious[3]	None
Nonfatal stroke (current users, 28-year follow-up)											
1 (Grodstein 2008)	243/256,437	310/485,987	RR 1.41 (1.19-1.68)	5 more per 1000 (from 2 more to 8 more)	Very low	Prospective cohort	Serious[1]	No serious	Serious[2]	Serious[3]	None

1. Evidence was downgraded by 1 due to selection (HRT users tended to have lower BMI and lower blood pressure at baseline);
2. Evidence has only 1 indirect aspect of PICO (population – the study was conducted among registered nurses only-);

3. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (+0.75 to +1.25);

Table 51: GRADE profile: oestrogen plus progesterone use versus no HRT use for the outcomes of non fatal stroke, CHD and IHD (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	oestrogen plus progesterone	Non users	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Non fatal stroke (current users, 28-year follow-up)											
1 (Grodstein 2008)	123/153,192	310/485,987	RR (95%CI): 1.31 (1.05-1.62)	4 more per 1000 (from 1 more to 7 more)	Very low	Prospective cohort	Serious[1]	No serious	Serious[2]	Serious[3]	None

1. Evidence was downgraded by 1 due to selection (HRT users tended to have lower BMI and lower blood pressure at baseline);
2. Evidence has only 1 indirect aspect of PICO (the study was conducted among registered nurses only);
3. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25);

Table 52: GRADE profile: timing of HRT initiation versus no HRT use for the outcomes of CHD, IHD and stroke (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT (mainly oestrogen)	Non users	Hazard ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
CHD (HRT initiated within 10 years (<10 years) since menopause) HRT initiation											
1 (Rossouw 2007)	39/3608	51/3529	HR 0.76 (0.50-1.16)	4 fewer per 1000 (from 8 fewer to 2 more)	Very low	Prospective Cohort	Serious[1]	No serious	No serious	Very Serious[2]	None
IHD (HRT initiated 1-5 years since menopause)											
1 (Stram 2011)	N/R	N/R	HR 1.06 (0.85-1.32)	1 more per 1000 (from 2 fewer to 5 more)	Very low	Prospective Cohort	Serious[3]	No serious	Serious[4]	Serious[5]	None
IHD (HRT initiated 5-10 years since menopause)											
1 (Stram 2011)	N/R	N/R	HR 1.11 (0.85-1.46)	2 more per 1000 (from 2 fewer to 7 more)	Very low	Prospective Cohort	Serious[3]	No serious	Serious[4]	Serious[5]	None
IHD (HRT initiated > 10 years since menopause)											
1 (Stram 2011)	N/R	N/R	HR 0.99 (0.76-1.30)	0 fewer per 1000 (from 4 fewer to 5 more)	Very low	Prospective Cohort	Serious[3]	No serious	Serious[4]	Serious[5]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT (mainly oestrogen)	Non users	Hazard ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
CHD (CEE plus Progesterone initiated within 2 years since menopause, women without prior HRT use)											
1 (Prentice 2009)	N/R	N/R	HR 1.42 (0.76-2.65)	6 more per 1000 (from 4 fewer to 25 more)	Low	Prospective cohort	Serious[1]	No serious	No serious	Serious[5]	None
CHD (CEE plus Progesterone initiated within 2 years since menopause, women with prior HRT use)											
1 (Prentice 2009)	N/R	N/R	HR 1.43 (0.61-3.39)	6 more per 1000 (from 6 fewer to 36 more)	Very low	Prospective cohort	Serious[1]	No serious	No serious	Very serious[2]	None
CHD (CEE plus Progesterone initiated within 2 to 4 years since menopause, women without prior HRT use)											
1 (Prentice 2009)	N/R	N/R	HR 1.37 (0.71-2.67)	6 more per 1000 (from 4 fewer to 25 more)	Very low	Prospective cohort	Serious[1]	No serious	No serious	Very serious[2]	None
CHD (CEE plus Progesterone initiated within 2 to 4 years since menopause, women with prior HRT use)											
1 (Prentice 2009)	N/R	N/R	HR 1.10 (0.46-2.63)	2 more per 1000 (from 8 fewer to 24 more)	Very low	Prospective cohort	Serious[1]	No serious	No serious	Very serious[2]	None
CHD (oestrogen plus progesterone initiated within 4 years since menopause)											
1 (Grodstein 2006)	78/91,985	666/329,604	RR 0.72 (0.56-0.92)	4 fewer per 1000 (from 7 fewer to 1 fewer)	Very low	Prospective cohort	Serious[1]	No serious	Serious[6]	Serious[5]	None
CHD (oestrogen plus progesterone initiated within 10 years since menopause)											
1 (Grodstein 2006)	23/11,945	400/152,205	RR 0.80 (0.53-1.23)	3 fewer per 1000 (from 7 fewer to 3 more)	Very low	Prospective cohort	Serious[1]	No serious	Serious[6]	Serious[5]	None
CHD (oestrogen plus progesterone initiated within 4 years since menopause, women with and without existing heart disease)											
1 (Grodstein 2006)	89/95,847	773/346,219	RR 0.71 (0.56-0.89)	4 fewer per 1000 (from 7 fewer to 2 fewer)	Very low	Prospective cohort	Serious[1]	No serious	Serious[6]	Serious[5]	None
CHD (oestrogen plus progesterone initiated at least 10 years after menopause, women with and without existing heart disease)											
1 (Grodstein 2006)	31/13,133	481/164,537	RR 0.90 (0.62-1.29)	2 fewer per 1000 (from 6 fewer to 4 more)	Very low	Prospective cohort	Serious[1]	No serious	Serious[6]	Very serious[2]	None
CHD (CEE alone initiated within 2 years since menopause, among women with prior HRT use)											
1 (Prentice 2009)	N/R	N/R	HR 1.26 (0.64-2.46)	4 more per 1000 (from 5 fewer to 22 more)	Very low	Prospective cohort	Serious[1]	No serious	No serious	Very serious[2]	None
CHD (CEE alone initiated within 2 years since menopause, among women without prior HRT use)											
1 (Prentice 2009)	N/R	N/R	HR 1.12 (0.55-2.24)	2 more per 1000 (from 7 fewer to 19 more)	Very low	Prospective cohort	Serious[1]	No serious	No serious	Very serious[2]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT (mainly oestrogen)	Non users	Hazard ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
CHD (CEE alone initiated within 2 to 4 years since menopause, among women with prior HRT use)											
1 (Prentice 2009)	N/R	N/R	HR 1.52 (0.81-2.86)	8 more per 1000 (from 3 fewer to 28 more)	Low	Prospective cohort	Serious[1]	No serious	No serious	Serious[5]	None
CHD (CEE alone initiated within 2 to 4 years since menopause, among women without prior HRT use)											
1 (Prentice 2009)	N/R	N/R	HR 0.99 (0.49-2.00)	0 fewer per 1000 (from 8 fewer to 15 more)	Very low	Prospective cohort	Serious[1]	No serious	No serious	Very serious[2]	None
CHD (oestrogen alone initiated within 4 years since menopause)											
1 (Grodstein 2006)	116/133,194	666/329,604	RR 0.66 (0.54-0.80)	5 fewer per 1000 (from 7 fewer to 3 fewer)	Very low	Prospective cohort	Serious[1]	No serious	Serious[6]	Serious[5]	None
CHD (oestrogen alone initiated within 10 years since menopause)											
1 (Grodstein 2006)	59/34,000	400/152,205	RR 0.76 (0.57-1.00)	4 fewer per 1000 (from 6 fewer to 0 fewer)	Very low	Prospective cohort	Serious[1]	No serious	Serious[6]	Serious[5]	None
CHD (oestrogen alone initiated within 4 years since menopause, women with and without existing heart disease)											
1 (Grodstein 2006)	130/140,515	773/346,219	RR 0.62 (0.52-0.76)	6 fewer per 1000 (from 7 fewer to 4 fewer)	Very low	Prospective cohort	Serious[1]	No serious	Serious[6]	Serious[5]	None
CHD (oestrogen alone initiated at least 10 years after menopause, women with and without existing heart disease)											
1 (Grodstein 2006)	84/37,978	481/164,537	RR 0.87 (0.69-1.10)	2 fewer per 1000 (from 5 fewer to 2 more)	Very low	Prospective cohort	Serious[1]	No serious	Serious[6]	Serious[5]	None
Stroke (HRT initiated within 10 years since menopause)											
1 (Rossouw 2007)	41/3608	23/3529	HR 1.77 (1.05-2.98)	9 more per 1000 (from 1 more to 22 more)	Very low	Prospective cohort	Serious[1]	No serious	No serious	Very Serious[2]	None
Stroke (CEE plus progesterone initiated within 2 years since menopause, women without prior HRT use)											
1 (Prentice 2009)	N/R	N/R	HR 1.58 (0.69-3.66)	7 more per 1000 (from 4 fewer to 30 more)	Very low	Prospective cohort	Serious[1]	No serious	No serious	Very serious[2]	None
Stroke (CEE plus progesterone initiated within 2 years since menopause, women with prior HRT use)											
1 (Prentice 2009\0	N/R	N/R	HR 1.73 (0.53-5.59)	8 more per 1000 (from 5 fewer to 52 more)	Very low	Prospective cohort	Serious[1]	No serious	No serious	Very serious[2]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT (mainly oestrogen)	Non users	Hazard ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Stroke (CEE plus progesterone initiated within 2 to 4 years since menopause, women without prior HRT use											
1 (Prentice 2009)	N/R	N/R	HR 2.17 (0.99-4.80)	13 more per 1000 (from 0 fewer to 43 more)	Low	Prospective cohort	Serious[1]	No serious	No serious	Serious[5]	None
Stroke (CEE plus progesterone initiated within 2 to 4 years since menopause, women with prior HRT use)											
1 (Prentice 2009)	N/R	N/R	HR 1.05 (0.45-2.45)	1 more per 1000 (from 6 fewer to 16 more)	Very low	Prospective cohort	Serious[1]	No serious	No serious	Very serious[2]	None
Stroke (oestrogen plus progesterone initiated within 4 years since menopause)											
1 (Grodstein 2008)	93/119,912	312/370,831	RR 1.22 (0.95-1.55)	2 more per 1000 (from 1 fewer to 6 more)	Very low	Prospective cohort	Serious[1]	No serious	Serious[6]	Serious[5]	None
Stroke (oestrogen plus progesterone initiated within 10 years since menopause)											
1 (Grodstein 2008)	93/119,912	240/193,066	RR 1.18 (0.87-1.60)	2 more per 1000 (from 1 fewer to 7 more)	Very low	Prospective cohort	Serious[1]	No serious	Serious[6]	Serious[5]	None
Stroke (oestrogen plus progesterone initiated at age 50-59 years)											
1 (Grodstein 2008)	25/51,904	108/239,967	RR 1.34 (0.84-2.13)	4 more per 1000 (from 2 fewer to 13 more)	Very low	Prospective cohort	Serious[1]	No serious	Serious[6]	Serious[5]	None
Stroke (CEE alone initiated within 2 years since menopause, women without prior HRT use)											
1 (Prentice 2009)	N/R	N/R	HR 1.49 (0.68-3.28)	6 more per 1000 (from 4 fewer to 26 more)	Very low	Prospective cohort	Serious[1]	No serious	No serious	Very serious[2]	None
Stroke (CEE alone initiated within 2 years since menopause, women with prior HRT use)											
1 (Prentice 2009)	N/R	N/R	HR 1.43 (0.61-3.39)	5 more per 1000 (from 4 fewer to 27 more)	Very low	Prospective cohort	Serious[1]	No serious	No serious	Very serious[2]	None
Stroke (CEE alone initiated within 2 to 4 years since menopause, women without prior HRT use)											
1 (Prentice 2009)	N/R	N/R	HR 2.45 (1.06-5.65)	16 more per 1000 (from 1 more to 53 more)	Low	Prospective cohort	Serious[1]	No serious	No serious	Serious[5]	None
Stroke (CEE alone initiated within 2 to 4 years since menopause, women with prior HRT use)											
1 (Prentice 2009)	N/R	N/R	HR 1.56 (0.81-3.03)	6 more per 1000 (from 2 fewer to 23 more)	Low	Prospective cohort	Serious[1]	No serious	No serious	Serious[5]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT (mainly oestrogen)	Non users	Hazard ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Stroke (oestrogen alone initiated within 4 years since menopause)											
1 (Grodstein 2008)	146/163,092	312/370,831	RR 1.29 (1.06-1.58)	3 more per 1000 (from 1 more to 7 more)	Very low	Prospective cohort	Serious[1]	No serious	Serious[6]	Serious[5]	None
Stroke (oestrogen alone initiated within 10 years since menopause)											
1 (Grodstein 2008)	133/87,038	240/193,066	RR 1.31 (1.06-1.63)	4 more per 1000 (from 1 more to 7 more)	Very low	Prospective cohort	Serious[1]	No serious	Serious[6]	Serious[5]	None
Stroke (oestrogen alone initiated at age 50-59 years)											
1 (Grodstein 2008)	31/49,590	108/239,967	RR 1.58 (1.06-2.37)	7 more per 1000 (from 1 more to 15 more)	Very low	Prospective cohort	Serious[1]	No serious	Serious[6]	Serious[5]	None

1. Observational data in nature, the re-analyses of data inherited all the risk of biases from the 2 original trials (WHI) as reported above
2. Evidence was downgraded by 2 due to very serious imprecision as 95% CI crossed 2 default MIDs (0.75 to 1.25)
3. Evidence was downgraded by 1 due to selection (HRT users were "healthier" and "younger" than non-users at baseline, with lower BMI, BP, or triglycerides levels)
4. Evidence has only 1 indirect aspect of PICO (population _the study was carried out among teachers only-)
5. Evidence was downgraded by 1 due to serious imprecision as 95% CI crossed 1 default MID (0.75 to 1.25)
6. Evidence has only 1 indirect aspect of PICO (population-the study was conducted among registered nurses only)

I.5.3 Development of Type 2 diabetes

Table 53: GRADE profile: HRT use versus placebo for the outcome of type 2 diabetes

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	Placebo	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
T2DM (age 50-59) (7.1 years follow-up)											
1 (Bonds 2006)	131/4806	159/4906	HR 0.83 (0.66-1.05)	5 fewer per 1000 (from 11 fewer to 2 more)	Low	Randomised trials with post-intervention follow-up	Serious[1,2,3,4]	No serious	No serious	Serious[5]	None

1. An average follow-up of 6.8 years, the study was terminated earlier than expected;
2. Relatively high drop-out and drop- in rates in both the CEE and placebo groups in the WHI CEE trial. When the CEE trial was terminated, earlier than expected, overall about 54% of women had already stopped taking study medication. About 5.7% in the CEE group and 9.1% in the placebo group initiated HRT use through their own clinicians. However, the distribution of that in the age group this review is interested is unclear;

3. In the WHI CEE trial, about 36% participants in each group had used HRT prior the enrolment of the trial in their lifetime; and about 13% in each group were current HRT users. However, the distribution of that in the age group this review is interested is unclear;
4. BMI was high in both groups at baseline (mean 30.1 ± 6.1 in CEE group and 30.1 ± 6.2 in the placebo group, respectively), not really the "healthy" women at baseline as the authors stated. However, the distribution of that in the age group this review is interested is unclear.
5. Evidence was downgraded by 1 due to serious imprecision as 95% CI crossed 1 default MID (0.75 to 1.25)

Table 54: GRADE profile: current HRT use versus no HRT use for the outcome of type 2 diabetes

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Current HRT	No HRT	Hazard ratio, odds ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
T2DM (12-year follow-up)											
1 (Manson 1992)	160/91,680 person-years	747/225,248 person-years	RR 0.80 (0.67-0.96)	1 fewer per 1000 (from 0 fewer to 1 fewer)	Very low	Cohort study	Very serious[2,8]	No serious	No serious	Serious[5]	None
T2DM (12-year follow-up) (duration of current use of HRT < 1 year) (subgroup analysis)											
1 (Manson 1992)	16/9,206 person-years	747/225,248 person-years	RR 0.84 (0.50-1.40)	1 fewer per 1000 (from 2 fewer to 1 more)	Very low	Cohort study	Very serious[2,8]	No serious	No serious	Very Serious[6]	None
T2DM (12-year follow-up) (duration of current use of HRT 1-3 years) (subgroup analysis)											
1 (Manson 1992)	28/28,193 person-years	747/225,248 person-years	RR 0.47 (0.31-0.69)	7 fewer per 1000 (from 3 fewer to 11 fewer)	Low	Cohort study	Very serious[2,8]	No serious	No serious	No serious	None
T2DM (12-year follow-up) (duration of HRT current use 4-6 years) (subgroup analysis)											
1 (Manson 1992)	39/20,460 person-years	747/225,248 person-years	RR 0.89 (0.64-1.24)	0 fewer per 1000 (from 1 fewer to 1 more)	Very low	Cohort study	Very serious[2,8]	No serious	No serious	Serious[5]	None
T2DM (12-year follow-up) (duration of HRT current use more than 7 years) (subgroup analysis)											
1 (Manson 1992)	72/30,771 person-years	747/225,248 person-years	RR 1.08 (0.84-1.38)	0 more per 1000 (from 1 fewer to 1 more)	Very low	Cohort study	Very serious[2,8]	No serious	No serious	Serious[5]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Current HRT	No HRT	Hazard ratio, odds ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
T2DM (14-year follow-up)											
1 (de Lauzon-Guillain 2009)	702/45,394	518/18,230	HR 0.75 (0.66-0.85)	7 fewer per 1000 (from 4 fewer to 10 fewer)	Very low	Cohort study	Very serious[1,7]	No serious	Serious[4]	Serious[5]	None
T2DM (14-year follow-up) (duration of HRT current use 0-2 years) (subgroup analysis)											
1 (de Lauzon-Guillain 2009)	144/7,300	518/18,230	HR 0.75 (0.61-0.91)	7 fewer per 1000 (from 3 fewer to 11 fewer)	Very low	Cohort study	Very serious[1,7]	No serious	Serious[4]	Serious[5]	None
T2DM (14-year follow-up) (duration of HRT current use 2-5 years) (subgroup analysis)											
1 (de Lauzon-Guillain 2009)	202/11,868	518/18,230	HR 0.84 (0.70-1.00)	4 fewer per 1000 (from 8 fewer to 0 more)	Very low	Cohort study	Very serious[1,7]	No serious	Serious[4]	Serious[5]	None
T2DM (14-year follow-up) (duration of HRT current use more than 5 years) (subgroup analysis)											
1 (de Lauzon-Guillain 2009)	294/23,460	518/18,230	HR 0.70 (0.59-0.82)	8 fewer per 1000 (from 5 fewer to 12 fewer)	Very low	Cohort study	Very serious[1,7]	No serious	Serious[4]	Serious[5]	None
T2DM (for fasting glucose ≥mmol/litre or 2 hour glucose ≥11.1 mmol/L) (average 4 years follow-up)											
1 (Zhang 2002)	N/R	N/R	OR 1.1 (0.62-1.97)	N/C	Very low	Cohort study	Very serious[3,9,10]	No serious	Serious[4]	Very serious[6]	None
T2DM (for 2 hour glucose ≥11.1 mmol/litre) (average 4 years follow-up)											
1 (Zhang 2002)	N/R	N/R	OR 1.58 (0.81-3.1)	N/C	Very low	Cohort study	Very serious[3,9,10]	No serious	Serious[4]	Serious[5]	None
T2DM (any duration) (fasting glucose ≥7.0 mmol/litre) (average 4-years follow-up)											
1 (Zhang 2002)	N/R	N/R	OR 1.01 (0.9-1.12)	N/C	Very low	Cohort study	Very serious[3,9,10]	No serious	Serious[4]	Serious[5]	None
T2DM (any duration) (fasting glucose≥7.0 mmol/L or 2 hour glucose ≥11.1 mmol/litre) (average 4 years follow-up)											
1 (Zhang 2002)	N/R	N/R	OR 1.10 (1.01-1.18)	N/C	Very low	Cohort study	Very serious[3,9,10]	No serious	Serious[4]	No serious	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Current HRT	No HRT	Hazard ratio, odds ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
T2DM (any duration) (2 hour glucose ≥11.1 mmol/litre) (average 4 years follow-up)											
1 (Zhang 2002)	N/R	N/R	OR 1.10 (1.01-1.19)	N/C	Very low	Cohort study	Very serious[3,9,10]	No serious	Serious[4]	No serious	None

1. Evidence was downgraded due to attrition bias
2. Evidence was downgraded due to selection bias
3. Evidence was downgraded due to detection bias
4. Majority of evidence had only 1 indirect aspect of PICO (population)
5. Evidence was downgraded by 1 due to confidence interval crossing 1 default MID (0.75 to 1.25)
6. Evidence was downgraded by 2 due to confidence interval crossing 2 default MIDs (0.75 to 1.25)
7. Adjusted for age and BMI
8. Adjusted for BMI, waist-to-hip ratio, American Indian Heritage
9. Adjusted for BMI and hysterectomy status
10. Number of participants in treatment groups was not reported for each outcome

Table 55: GRADE profile: past HRT use versus no HRT use for the outcome of type 2 diabetes

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Past HRT	No HRT	Hazard ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
T2DM (12-year follow-up)											
1 (Manson 1992)	342/106,063 person-years	747/225,248 person-years	RR 1.07 (0.93-1.23)	0 more per 1000 (from 0 fewer to 1 more)	Very low	Cohort study	Very serious[2,5]	No serious	No serious	Serious[4]	None
T2DM (12-year follow-up) (duration of past use of HRT < 1 year) (subgroup analysis)											
1 (Manson 1992)	79/27,670 person-years	747/225,248 person-years	RR 0.86 (0.67-1.12)	0 fewer per 1000 (from 1 fewer to 0 more)	Very low	Cohort study	Very serious[2,5]	No serious	No serious	Serious[4]	None
T2DM (12-year follow-up) (duration of past use of HRT 1-3 years) (subgroup analysis)											
1 (Manson 1992)	133/39,914 person-years	747/225,248 person-years	RR 1.05 (0.85-1.29)	0 more per 1000 (from 0 fewer to 1 more)	Very low	Cohort study	Very serious[2,5]	No serious	No serious	Serious[4]	None
T2DM (12-year follow-up) (duration of past use of HRT 4-6 years) (subgroup analysis)											

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Past HRT	No HRT	Hazard ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
1 (Manson 1992)	57/17,277 person-years	747/225,248 person-years	RR 1.29 (0.97-1.71)	1 more per 1000 (from 0 fewer to 2 more)	Very low	Cohort study	Very serious[2,5]	No serious	No serious	Serious[4]	None
T2DM (12-year follow-up) (duration of past use of HRT more than 7 years) (subgroup analysis)											
1 (Manson 1992)	55/16,355 person-years	747/225,248 person-years	RR 1.13(0.84-1.52)	0 more per 1000 (from 1 fewer to 2 more)	Very low	Cohort study	Very serious[2,5]	No serious	No serious	Serious[4]	None
T2DM (duration of past HRT use>1 year) (14-year follow-up)											
1 (de Lauzon-Guillain 2009)	244/,35,384	518/18,230	HR 0.90 (0.76-1.07)	3 fewer per 1000 (from 7 fewer to 2 more)	Very low	Cohort study	Very serious[1,6]	No serious	Serious[3]	Serious[4]	None

1. Evidence was downgraded due to attrition bias
2. Evidence was downgraded due to selection bias
3. Majority of evidence had only 1 indirect aspect of PICO (population)
4. Evidence was downgraded by 1 due to serious imprecision as 95% CI crossed 1 default MID (0.75 to 1.25)
5. Adjusted for only age and BMI
6. Adjusted for age, age at menarche, parity, breast feeding, age at menopause, type of menopause, family history of diabetes, physical activity, alcohol intake, total energy intake exclusive of alcohol intake, education, baseline cholesterol level, hypertension, smoking and BMI during follow-up as time dependent variable

Table 56: GRADE profile: HRT ever use (current and past) versus no HRT use for the outcome of type 2 diabetes, (subgroup analyses on route of administration)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Ever HRT	No HRT	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
T2DM (14-year follow-up) (Oral) (subgroup analysis)											
1 (de Lauzon-Guillain 2009)	121/11,263	518/18,230	HR 0.61 (0.50-0.76)	11 fewer per 1000 (from 7 fewer to 14 fewer)	Very low	Cohort study	Serious[1,5]	No serious	Serious[3]	Serious[4]	None
T2DM (14-year follow-up) (cutaneous) (subgroup analysis)											
1	425/25,740	518/18,230	HR	6 fewer per 1000	Very low	Cohort study	Very serious[1,5]	No serious	Serious[3]	Serious[4]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Ever HRT	No HRT	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
(de Lauzon-Guillain 2009)			0.78 (0.67-0.90)	(from 3 fewer to 9 fewer)							
T2DM (14-year follow-up) (Other route of administration) (subgroup analysis)											
1 (de Lauzon-Guillain 2009)	49/2,533	518/18,230	HR 0.76 (0.56-1.04)	7 fewer per 1000 (from 12 fewer to 1 more)	Very low	Cohort study	Very serious[1,5]	No serious	Serious[3]	Serious[4]	None

N/R: not reported;
1. Evidence was downgraded due to attrition bias
2. Evidence was downgraded due to selection bias
3. Majority of evidence had only 1 indirect aspect of PICO (population)
4. Evidence was downgraded by 1 due to serious imprecision as 95% CI crossed 1 default MID (0.75 to 1.25)
5. Adjusted for age, age at menarche, parity, breast feeding, age at menopause, type of menopause, family history of diabetes, physical activity, alcohol intake, total energy intake exclusive of alcohol intake, education, baseline cholesterol level, hypertension, smoking and BMI during follow-up as time dependent variable

I.5.4 Management of type 2 diabetes – control of blood sugar

Table 57: GRADE profile: sequential combined HRT versus placebo for the outcome of HbA1c at 3 months (RCTs)

Quality assessment							No of patients		Effect			
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	HRT	Placebo	Relative (95% CI)	Absolute	Quality	Importance
HbA1c (oral 2mg 17β oestradiol/2mg 17-β oestradiol plus 1mg norethisterone) (follow-up mean 12 weeks; measured with:%; Better indicated by lower values)												
1 (Darko2001)	Randomised trials	Serious[1]	No serious inconsistency	No serious indirectness	Serious[2]	None	11	13	-	MD 0.6 lower (1.72 lower to 0.52 higher)	Low	CRITICAL
HbA1c (transdermal patch 50µg per 24 hrs 17-β oestradiol/50µg 17-β oestradiol plus 170µg norethisterone) (follow-up mean 12 weeks; measured with: %; Better indicated by lower values)												
1 (Darko 2001)	Randomised trials	Serious[1]	No serious inconsistency	No serious indirectness	Very serious[3]	None	9	13	-	MD 0.4 higher (1.06 lower to 1.86 higher)	Very low	CRITICAL
HbA1c (oral 1mg 17-β oestradiol/0.5mg norethisterone) (follow-up mean 3 months; measured with: %; Better indicated by lower values)												
1 (Kernohan 2007)	Randomised trials	Serious[4]	No serious inconsistency	No serious indirectness	Serious[2]	None	14	14	-	MD 0.7 lower (1.59 lower to 0.19 higher)	Low	CRITICAL

1. Risk due to selection, performance and detection bias
2. Evidence was downgraded by 1 due to serious imprecision as 95% CI crossed 1 default MID (-/+0.5 times SD)
3. Evidence was downgraded by 2 due to very serious imprecision as 95% CI crossed 2 default MIDs (-/+0.5 times SD)
4. Risk due to selection and detection bias

Table 58: GRADE profile: continuous combined HRT (oral or transdermal) versus placebo for the outcome of blood glucose at 3 months (RCTs)

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	HRT	Placebo	Relative (95% CI)	Absolute		
Fasting plasma glucose (oral 2mg 17-β oestradiol/2mg 17-β oestradiol plus 1mg norethisterone) (follow-up mean 12 weeks; measured with: mmol/litre; Better indicated by lower values)												
1 (Darko 2001)	Randomised trials	Serious[1]	No serious inconsistency	No serious indirectness	Very serious[2]	None	11	13	-	MD 0.8 lower (3.49 lower to 1.89 higher)	Very low	CRITICAL
Fasting plasma glucose (transdermal patch 50µg per 24 hrs 17-β oestradiol/50µg 17-β oestradiol plus 170µg norethisterone) (follow-up mean 12 weeks; measured with: mmol/litre; Better indicated by lower values)												
1 (Darko 2001)	Randomised trials	Serious[1]	No serious inconsistency	No serious indirectness	Serious[3]	None	9	13	-	MD 1.5 higher (1.51 lower to 4.51 higher)	Low	CRITICAL
Fasting plasma glucose (oral 1mg 17-β oestradiol/0.5mg norethisterone) (follow-up mean 3 months; measured with: mmol/litre; Better indicated by lower values)												
1 (Kernohan 2007)	Randomised trials	Serious[4]	No serious inconsistency	No serious indirectness	Serious[3]	None	14	14	-	MD 1.7 lower (3 to 0.4 lower)	Low	CRITICAL

1. Risk due to selection, performance and detection bias
2. Evidence was downgraded by 2 due to very serious imprecision as 95% CI crossed 2 default MIDs (-/+0.5 times SD)
3. Evidence was downgraded by 1 due to serious imprecision as 95% CI crossed 1 default MID (-/+0.5 times SD)
4. Risk due to selection and detection bias

Table 59: GRADE profile: Conjugated equine oestrogen versus placebo for the outcome of HbA1c at 6 months (RCTs)

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	HRT versus placebo	Control	Relative (95% CI)	Absolute		
HbA1c (1mg oestradiol/0.5mg norethisterone versus placebo for glycaemic control (follow-up mean 6 months; measured with: %; Better indicated by lower values)												
1 (McKenzie 2003)	Randomised trials	Very serious[1]	No serious inconsistency	No serious indirectness	Serious[2]	None	22	23	-	MD 0.59 lower (1.45 lower to 0.27 higher)	Very low	CRITICAL

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	HRT versus placebo	Control	Relative (95% CI)	Absolute		
HbA1c (transdermal 80mcg oestradiol patch/1mg oral norethisterone) versus placebo for glycaemic control (follow-up mean 6 months; measured with: %; Better indicated by lower values)												
1 (Perera 2000)	Randomised trials	Very serious[3]	No serious inconsistency	No serious indirectness	Serious[2]	None	22	21	-	MD 0.2 lower (1.05 lower to 0.65 higher)	Very low	CRITICAL
HbA1c (oral 0.625mg conjugated equine oestrogen/2.5mg medroxyprogesterone acetate) versus placebo (follow-up mean 6 months; measured with: %; Better indicated by lower values)												
1 (Sutherland 2001)	Randomised trials	Very serious[3]	No serious inconsistency	No serious indirectness	Serious[2]	None	28	19	-	MD 0.6 lower (1.71 lower to 0.51 higher)	Very low	CRITICAL

1. Very high risk due to performance and detection bias
2. Evidence was downgraded by 1 due to serious imprecision as 95% CI crossed 1 default MID (-/+0.5 times SD)
3. Very high risk due to selection, performance, attrition and detection bias

Table 60: GRADE profile: HRT (oral or transdermal) versus placebo for the outcome of blood glucose at 6 months (RCTs)

Quality assessment							No of patients		Effect		Quality	Importance
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	HRT	Placebo	Relative (95% CI)	Absolute		
Blood glucose (oral HRT 1mg oestradiol/0.5mg norethisterone) versus placebo (follow-up mean 6 months; measured with: mmol/litre; Better indicated by lower values)												
1 (McKenzie 2003)	Randomised trials	Very serious[1]	No serious inconsistency	No serious indirectness	Serious[2]	None	22	23	-	MD 2.16 lower (4.06 to 0.26 lower)	Very low	CRITICAL
Blood glucose (transdermal 80mcg oestradiol patch/1mg oral norethisterone) versus placebo (follow-up mean 6 months; measured with: mmol/litre; Better indicated by lower values)												
1 (Perera 2000)	Randomised trials	Very serious[3]	No serious inconsistency	No serious indirectness	Very serious[4]	None	22	21	-	MD 0 higher (1.53 lower to 1.53 higher)	Very low	CRITICAL
Blood glucose (oral 0.625mg conjugated equine oestrogen/2.5mg medroxyprogesterone acetate) versus placebo (follow-up mean 6 months; measured with: mmol/litre; Better indicated by lower values)												
1 (Sutherl	Randomised trials	Very serious[5]	No serious inconsistency	No serious indirectness	Serious[2]	none	28	19	-	MD 2.01 lower (4.01 to 0.01 lower)	Very low	CRITICAL

Quality assessment							No of patients		Effect			
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	HRT	Placebo	Relative (95% CI)	Absolute	Quality	Importance
and 2001)												

1. Very high risk due to performance and detection bias
2. Evidence was downgraded by 1 due to serious imprecision as 95% CI crossed 1 default MID (-/+0.5 times SD)
3. Very high risk due to selection, performance, attrition and detection bias
4. Evidence was downgraded by 2 due to very serious imprecision as 95% CI crossed 2 default MIDs (-/+0.5 times SD)
5. Very high risk due to performance, attrition and detection bias

Table 61: GRADE profile: HRT versus no HRT use for the outcome of HbA1c during 2 year (cross sectional study)

Quality assessment							No of patients		Effect			
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	HRT	No HRT	Relative (95% CI)	Absolute	Quality	Importance
HbA1c during 2 years (follow-up 2 years[1]; measured with: %; Better indicated by lower values)												
1 (Ferrara 2001)	Observational studies	serious risk of bias[1,2]	No serious inconsistency	No serious indirectness	No serious imprecision	None	3406	11583	-	MD 0.6 lower (0.67 to 0.53 lower)	Low	CRITICAL

1. Due to the study design, data in the study was reported at a time point during the 2 year study
2. Data has been adjusted for age in both treatment groups

I.5.5 Breast cancer

Table 62: GRADE profile: HRT user versus no placebo for the outcome of breast cancer (RCTs)

	Number of participants		Effect				Quality assessment				
Number of studies	HRT	No HRT	Hazard ratio or risk ratio (95% CI)	Absolute (95% CI)	Quality	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Breast cancer											
1 (Schierbeck 2012)	17/502	10/504	HR (95%CI): 0.59 (0.27-1.30)	9 fewer per 1000 (from 16 fewer to 7 more)	Very low	Randomised trials	Serious[1]	No serious	No serious	Very serious[2]	None
Breast cancer (Oestrogen plus progesterone)											

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
1 (Manson 2013)	55/2837	42/2683	HR (95%CI): 1.21 (0.81 -1.80)	5 more per 1000 (from 4 fewer to 36 more	Low	Randomised trials	Serious[3,4,5,6]	No serious	No serious	Serious[7]	None
1 (Vickers 2007)	5/2196	7/2189	RR (95%CI): 0.71 (0.18-2.61)	7 fewer per 1000 (from 18 fewer to 36 more)	Very low	Randomised trials	Serious[8]	No serious	No serious	Very serious[2]	None
Breast cancer (Oestrogen only)											
1 (Manson 2013)	29/1639	36/1674	HR (95%CI): 0.82 (0.50-1.34)	4 fewer per 1000 (from 11 fewer to 8 more)	Low	Randomised trials	Serious[9,10,11,12]	No serious	No serious	Serious[7]	None
Breast cancer (Oestrogen plus Progesterone versus oestrogen)											
1 (Vickers 2007)	815	826	RR (95%CI): 1.52 (0.17-18.24)	12 more per 1000 (from 19 fewer to 388 more)	Very low	Randomised trials	Serious[8]	No serious	No serious	Very serious[2]	None
Randomised controlled trials with post-intervention follow-up											
Breast cancer (current HRT user, 10-year follow-up)											
1 (Schierbeck 2012)	502	504	HR (95%CI): 0.92 (0.52-1.62)	2 fewer per 1000 (from 11 fewer to 14 more)	Very low	Randomised trial with post-intervention)	Serious[1]	N/A	No serious	Very serious[2]	None
Breast cancer (Oestrogen plus progesterone, 8.2 years post-intervention follow-up)											
1 (Manson 2013)	132/2,837	93/2,683	HR (95%CI): 1.34 (1.03-1.75)	8 more per 1000 (from 1 fewer to 17 more)	Low	Randomised trial with post-intervention	Serious[3,5,6]	N/A	No serious	Serious[7]	None
Breast cancer (Oestrogen, 6.6 years post-intervention follow-up)											

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
1 (Manson 2013)	46/	61/	HR (95%CI): 0.76 (0.52-1.11)	5 fewer per 1000 (from 11 fewer to 2 more)	Low	Randomised trial with post-intervention	Serious[9, 10,11,12]	No serious	No serious	Serious[7]	None
Breast cancer (Oestrogen, 10.6 years post-intervention follow-up)											
1 (Cherry 2014)	2/162	5/134	RR (95%CI): 0.33 (0.06-1.68)	15 fewer per 1000 (from 21 fewer to 15 more)	Very low	Randomised trialwith post-intervention	Serious[13]	No serious	Serious	Very serious[2]	None

1. Evidence was downgraded due to lack of blinding (open-label RCT) and relatively high attrition (at 5-year follow-up, about 25% of women dropped-out)
2. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)
3. Overall breaking of blinding was relatively high because of the need to unmask 40.5% of the gynaecologists in the treatment group (due to vaginal bleeding after taking HRT) compared with 6.8% in the placebo group, though the distribution of that in the age group this review is interested is unclear
4. High attrition bias among all participants; about 42% of women in the oestrogen plus progesterone and 38% of women in the placebo stopped taking the study drugs during follow-up. The drop-in rates were 6.2% in the intervention group and 10.7% in the placebo group, respectively. However, the distribution of that in the age group this review is interested is unclear
5. In the WHI CEE plus progesterone trial, about 26% participants in each group had used HRT prior the enrolment of the trial in their lifetime; and about 6% in each group were current HRT users. The distribution of that in the age group this review is interested is unclear
6. Among all participants in the WHI, about 34% of women had a BMI higher than 30 kg/m2 in each group. The distribution of that in the age group this review is interested is unclear
7. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
8. Evidence was downgraded by 1 due to high and un-proportional drop-out rates in the arms (20% in the intervention arm and 9% in the placebo arm)
9. An average follow-up of 6.8 years, the study was terminated earlier than expected
10. Relatively high drop-out and drop- in rates in both the CEE and placebo groups in the WHI CEE trial. When the CEE trial was terminated, earlier than expected, overall about 54% of women had already stopped taking study medication. About 5.7% in the CEE group and 9.1% in the placebo group initiated HRT use through their own clinicians. However, the distribution of that in the age group this review is interested is unclear
11. In the WHI CEE trial, about 36% participants in each group had used HRT prior the enrolment of the trial in their lifetime; and about 13% in each group were current HRT users. However, the distribution of that in the age group this review is interested is unclear
12. BMI was high in both groups at baseline (mean 30.1 ± 6.1 in CEE group and 30.1 ± 6.2 in the placebo group, respectively), not really the "healthy" women at baseline as the authors stated. However, the distribution of that in the age group this review is interested is unclear
13. Evidence was downgraded by 1 due to selection (participants were originally recruited from an RCT and have had a myocardial infarction). HRT use or not during post-study intervention phase which this study examined was not followed up or ascertained

Table 63: GRADE profile: HRT use versus never use for the outcome of breast cancer (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Breast cancer (ever users)											
15 (Bakken, 2004; Beral, 2003; Ewertz, 2005; Folsom, 1995; Fournier, 2005; Hedblad, 2002; Lando, 1999; Lund, 2007; Manjer, 2001; Mills, 1989; Saxena, 2010; Schuurman, 1995; Stahlberg, 2004; Stahlberg, 2005; Tjonneland, 2004)	606,002	594,962	RR 1.46 (1.34-1.60)	10 more per 1000 (from 8 more to 13 more)	Very low	Prospective Cohort	Very serious[1]	Very serious[2]	No serious	No serious	None
Breast cancer (current users)											
9 (Beral, 2003; Ewertz, 2005; Lund, 200; Mills, 1989; Stahlberg, 2004; Stahlberg, 2005; Tjonneland, 2004; Bakken, 2004; Grodstein, 1997)	511,647	515,517	RR 1.79 (1.52-2.11)	18 more per 1000 (from 12 more to 25 more)	very low	Prospective Cohort	Very serious[1]	Very serious[2]	No serious	No serious	None
Breast cancer (Past users)											
9 (Beral, 2003; Ewertz, 2005; Lund, 2007; Mills, 1989; Stahlberg, 2004; Stahlberg, 2005; Tjonnela nd, 2004; Bakken, 2004; Grodstein, 1997)	511,647	515,517	RR 1.02 (0.96-1.08)	0 fewer per 1000 (from 1 fewer to 2 more)	Low	Prospective cohort	Very serious[1]	No serious	No serious	No serious	None
Breast cancer (Ever users of oestrogen)											
3 (Willis, 1996; Lund, 2007; Sourander, 1998)	199,955	250,373	RR 1.01 (0.76-1.36)	0 fewer per 1000 (from 5 fewer to 8 more)	very low	Prospective cohort	Very serious[1]	Very serious[2]	No serious	Serious[3]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Breast cancer (Current users of oestrogen)											
4 (Bakken, 2011; Lund, 2007; Saxena, 2010; Sourander, 1998)	115,379	101,296	RR 1.25 (1.03-1.52)	6 more per 1000 (from 1 fewer to 12 more)	Very low	Prospective cohort	Very serious[1]	Very serious[2]	No serious	Serious[3]	None
Breast cancer (Past users of oestrogen)											
4 (Willis, 1996; Lund, 2007; Saxena, 2010; Sourander, 1998)	242,600	262,704	RR 1.02 (0.76-1.37)	0 fewer per 1000 (from 5 fewer to 8 more)	Very low	Prospective cohort	Very serious[1]	Very serious[2]	No serious	Serious[3]	None
Breast cancer (Ever users of oestrogen plus Progesterone)											
2 (Jernstrom, 2003; Lund, 2007)	7,442	11,305	RR 2.29 (1.24-4.24)	29 more per 1000 (from 5 more to 73 more)	Very low	Prospective cohort	Very serious[1]	Very serious[2]	No serious	Serious[3]	None
Breast cancer (Current users of oestrogen plus Progesterone)											
3 (Bakken, 2011; Lund, 2007; Saxena, 2010)	113,634	95,724	RR 1.75 (1.64-1.88)	17 more per 1000 (from 14 more to 20 more)	Low	Prospective cohort	Very serious[1]	No serious	No serious	No serious	None
Breast cancer (Past users of oestrogen plus Progesterone)											
2 (Lund, 2007; Saxena, 2010)	51,978	23,636	RR 0.88 (0.50-1.54)	3 fewer per 1000 (from 11 fewer to 12 more)	Very low	Prospective cohort	Very serious[1]	Serious[4]	No serious	Very serious5	None
Breast cancer (Up to 2 years HRT use)											
4 (Fournier, 2005; Stahlberg, 2004; Mills, 1989; Bakken, 2004)	68,537	53,338	RR 1.63 (1.17-2.28)	14 more per 1000 (from 4 more to 29 more)	Very low	Prospective cohort	Very serious[1]	Very serious[2]	No serious	Serious[3]	None
Breast cancer (Up to 4 years HRT use)											
4 (Fournier, 2005; Lando, 1999; Stahlberg, 2004; Bakken, 2004)	64,893	54,450	RR 1.35 (0.91-1.99)	8 more per 1000 (from 2 fewer to 22 more)	Very low	Prospective cohort	Very serious[1]	Very serious[2]	No serious	Serious[3]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Breast cancer (Up to 5 years HRT use)											
2 (Mills, 1989; Folsom, 1995)	23,375	29,163	RR 1.49 (1.12-1.97)	11 more per 1000 (from 3 more to 22 more)	Very low	Prospective cohort	Very serious[1]	No serious	No serious	Serious[3]	None
Breast cancer (More than or equal to 4 years HRT use)											
2 (Fournier, 2005; Folsom, 1995)	45,215	50,403	RR 1.21 (0.99-1.47)	5 more per 1000 (from 0 fewer to 11 more)	Very low	Prospective cohort	Very serious[1]	No serious	No serious	Serious[3]	None
Breast cancer (3 to 9 years HRT use)											
1 (Lando, 1999)	2,197	3,564	RR 0.50 (0.29-0.87)	11 fewer per 1000 (from 16 fewer to 3 fewer)	Very low	Prospective cohort	Very serious[1]	N/A	No serious	Serious[3]	None
Breast cancer (5 to 10 years HRT use)											
3 (Stahlberg, 2004; Mills, 1989; Bakken, 2004)	40,856	29,646	RR 2.03 (1.37-3.02)	23 more per 1000 (from 8 more to 45 more)	Very low	Prospective cohort	Very serious[1]	Very serious[2]	No serious	No serious	None
Breast cancer (More than or equal to 10 years HRT use)											
3 (Lando, 1999; Mills, 1989; Bakken, 2004)	38,745	26,644	RR 1.19 (0.64-2.22)	4 more per 1000 (from 8 fewer to 27 more)	Very low	Prospective cohort	Very serious[1]	Very serious[2]	No serious	Very serious[5]	None
Breast cancer (10 to 14 years HRT use)											
1 (Stahlberg, 2004)	4,308	6,566	RR 3.08 (1.87-5.07)	47 more per 1000 (from 20 more to 91 more)	Moderate	Prospective cohort	Serious[6]	N/A	No serious	No serious	None
Breast cancer (More than or equal to 15 years HRT use)											
1 (Stahlberg, 2004)	4,308	6,566	RR 3.08 (1.87-5.07)	47 more per 1000 (from 20 more to 91 more)	Moderate	Prospective cohort	Serious[6]	N/A	No serious	No serious	None
Breast cancer (Up to 2 years of oestrogen use)											
5 (Beral, 2003; Bakken, 2011; Colditz, 1992; Willis, 1996; Fournier, 2008)	748,816	740,566	RR 0.93 (0.77-1.13)	2 fewer per 1000 (from 5 fewer to 3 more)	Low	Prospective cohort	Very serious[1]	No serious	No serious	No serious	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Breast cancer (Up to 5 years of oestrogen use)											
7 (Beral, 2003; Bakken, 2011; Colditz, 1992; Willis, 1996; Fournier, 2008; Saxena, 2010; Bakken, 2004)	820,581	770,653	RR 1.16 (0.95-1.42)	4 more per 1000 (from 1 fewer to 9 more)	Very low	Prospective cohort	Very serious[1]	Very serious[2]	No serious	Serious[3]	None
Breast cancer (4 to 10 years of oestrogen use)											
5 (Beral, 2003; Bakken, 2011; Colditz, 1992; Willis, 1996; Fournier, 2008)	748,816	740,566	RR 1.23 (0.94-1.61)	5 more per 1000 (from 1 fewer to 14 more)	Very low	Prospective cohort	Very serious[1]	Very serious[2]	No serious	Serious[3]	None
Breast cancer (More than or equal to 5 years of oestrogen use)											
3 (Colditz, 1992; Fournier, 2008; Bakken, 2004)	89,346	59,981	RR 1.42 (1.10-1.82)	9 more per 1000 (from 2 more to 18 more)	Very low	Prospective cohort	Very serious[1]	No serious	No serious	Serious[3]	None
Breast cancer (Up to 7 years of oestrogen use)											
1 (Schairer, 2000)	27,075	19,280	RR 1.00 (0.83-1.21)	0 fewer per 1000 (from 4 fewer to 5 more)	Moderate	Prospective cohort	No serious	N/A	No serious	No serious	None
Breast cancer (More than or equal to 10 years of oestrogen use)											
3 (Beral, 2003; Bakken, 2011; Willis, 1996)	686,699	698,341	RR 1.10 (0.77-1.55)	2 more per 1000 (from 5 fewer to 12 more)	Very low	Prospective cohort	Very serious[1]	Very serious[2]	No serious	Serious[3]	None
Breast cancer (6 to 15 years of oestrogen use)											
2 (Schairer, 2000; Saxena, 2010)	71,611	31,611	RR 1.14 (0.91-1.43)	3 more per 1000 (from 2 fewer to 10 more)	Very low	Prospective cohort	No serious	Very serious[2]	No serious	Serious[3]	None
Breast cancer (More than or equal to 15 years of oestrogen use)											
2 (Schairer, 2000; Saxena, 2010)	71,611	31,611	RR 1.20 (1.06-1.36)	4 more per 1000 (from 1 more to 8 more)	Very low	Prospective cohort	No serious	Very serious[2]	No serious	Serious[3]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Breast cancer (Up to 2 years of oestrogen plus progesterone use)											
4 (Beral, 2003; Bakken, 2011; Fournier, 2008; Schairer, 2000)	581,571	507,828	RR 1.14 (0.84-1.55)	3 more per 1000 (from 4 fewer to 12 more)	Very low	Prospective cohort	Very serious[1]	Serious[4]	No serious	Serious[3]	None
Breast cancer (Up to 5 years oestrogen plus progesterone use)											
6 (Beral, 2003; Bakken, 2011; Fournier, 2008; Saxena, 2010; Schairer, 2000; Bakken, 2004)	653,336	537,915	RR 1.52 (1.25-1.85)	12 more per 1000 (from 6 more to 19 more)	Very low	Prospective cohort	Very serious[1]	Very serious[2]	No serious	No serious	None
Breast cancer (4 to 10 years oestrogen plus progesterone use)											
3 (Beral, 2003; Bakken, 2011; Fournier, 2008)	554,496	488,548	RR 1.94 (1.41-2.66)	21 more per 1000 (from 9 more to 37 more)	Very low	Prospective cohort	Very serious[1]	Very serious[2]	No serious	No serious	None
Breast cancer (More than or equal to 4 years oestrogen plus progesterone use)											
3 (Bakken, 2004; Schairer, 2000; Fournier, 2008)	110,978	60,739	RR 1.81 (1.12-2.91)	18 more per 1000 (from 3 more to 43 more)	Very low	Prospective cohort	Very serious[1]	Very serious[2]	No serious	Serious[3]	None
Breast cancer (More than or equal to 10 years oestrogen plus progesterone use)											
2 (Beral, 2003; Bakken, 2011)	497,822	464,845	RR 2.30 (2.07-2.55)	29 more per 1000 (from 24 more to 35 more)	Low	Prospective cohort	Very serious[1]	No serious	No serious	No serious	None
Breast cancer (6 to 15 years oestrogen plus progesterone use)											
1 (Saxena, 2010)	44,536	12,331	RR 1.57 (1.40-1.76)	13 more per 1000 (from 9 more to 17 more)	Moderate	Prospective cohort	No serious	N/A	No serious	No serious	None
Breast cancer (More than or equal to 15 years oestrogen plus progesterone use)											
1 (Saxena, 2010)	44,536	12,331	RR 1.83 (1.48-2.26)	19 more per 1000 (from 11 more to 28 more)	Moderate	Prospective cohort	No serious	N/A	No serious	No serious	None
Breast cancer (Up to 5 years since last use of HRT)											
2 (Beral, 2003)	437,905	394,193	RR 1.05 (0.96-1.13)	1 more per 1000 (from 1 fewer to 3 more)	Low	Prospective cohort	Very serious[1]	No serious	No serious	No serious	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Breast cancer (4 to 10 years since last use of HRT use)											
2 (Beral, 2003)	437,905	394,193	RR 1.01 (0.88-1.16)	0 fewer per 1000 (from 3 fewer to 4 more)	Low	Prospective cohort	Very serious[1]	No serious	No serious	No serious	None
Breast cancer (More than or equal to 10 years since last use of HRT)											
1 (Beral, 2003)	436,166	392,757	RR 0.90 (0.72-1.12)	2 fewer per 1000 (from 6 fewer to 3 more)	Moderate	Prospective cohort	Serious[6]	N/A	No serious	No serious	None
Breast cancer (Up to 5 years since last use of oestrogen)											
3 (Willis, 1996; Fournier, 2008; Schairer, 2000)	272,626	276,479	RR 1.20 (0.90-1.60)	4 more per 1000 (from 2 fewer to 13 more)	Very low	Prospective cohort	Serious[6]	Very serious[2]	No serious	Serious[3]	None
Breast cancer (5 to 10 years last use of oestrogen)											
2 (Willis, 1996; Schairer, 2000)	215,952	252,776	RR 0.76 (0.60-0.95)	5 fewer per 1000 (from 9 fewer to 1 fewer)	Low	Prospective cohort	Serious[6]	No serious	No serious	Serious[3]	None
Breast cancer (More than or equal to 5 years since last use of oestrogen)											
2 (Fournier, 2008; Schairer, 2000)	83,749	42,983	RR 1.10 (0.96-1.27)	2 more per 1000 (from 1 fewer to 6 more)	Low	Prospective cohort	Serious[6]	No serious	No serious	Serious[3]	None
Breast cancer (More than or equal to 11 years since last use of oestrogen use)											
1 (Willis, 1996)	188,877	233,496	RR 0.84 (0.70-1.01)	4 fewer per 1000 (from 7 fewer to 0 fewer)	Moderate	Prospective cohort	No serious	N/A	No serious	Serious[3]	None
Breast cancer (Up to 5 years since last use of oestrogen plus progesterone)											
2 (Fournier, 2008; Schairer, 2000)	83,749	42,983	RR 1.13 (0.90-1.41)	3 more per 1000 (from 2 fewer to 9 more)	Low	Prospective cohort	Serious[6]	No serious	No serious	Serious[3]	None
Breast cancer (4 to 10 years since last use of oestrogen plus progesterone)											
1 (Schairer, 2000)	27,075	19,280	RR 0.60 (0.17-2.16)	9 fewer per 1000 (from 19 fewer to 26 more)	Low	Prospective cohort	No serious	N/A	No serious	Very serious[5]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Breast cancer (More than or equal to 6 years since last use of oestrogen plus progesterone)											
1 (Schairer, 2000)	27,075	19,280	RR 0.60 (0.30-1.60)	9 fewer per 1000 (from 16 fewer to 13 more)	Low	Prospective cohort	No serious	N/A	No serious	Very serious[5]	None
Breast cancer (Oestrogen only, timing of use not specified)											
7 (Fournier, 2005; Ewertz, 2005; Stahlberg, 2004; Colditz, 1992; Saxena, 2010; Schairer, 2000; Bakken, 2004)	153,733	166,939	RR 1.27 (1.13-1.43)	6 more per 1000 (from 3 more to 10 more)	Very low	Prospective cohort	Very serious[1]	Serious[4]	No serious	Serious[3]	None
Breast cancer (Oestrogen plus Progesterone, timing of use not specified)											
6 (Fournier, 2005; Stahlberg, 2004; Colditz, 1992; Saxena, 2010; Schairer, 2000; Bakken, 2004)	140,704	101,839	RR 1.64 (1.33-2.01)	14 more per 1000 (from 7 more to 23 more)	Very low	Prospective cohort	Very serious[1]	Very serious[2]	No serious	No serious	None
Breast cancer (Progesterone only, timing of use not specified)											
3 (Ewertz, 2005; Saxena, 2010; Schairer, 2000)	84,408	96,489	RR 1.19 (0.92-1.54)	4 more per 1000 (from 2 fewer to 12 more)	Very low	Prospective cohort	Very serious[1]	No serious	No serious	Serious[3]	None
Breast cancer Incident cases (Ever users of HRT)											
7 (Beral, 2003; Lando, 1999; Tjonneland, 2004; Stahlberg, 2005; Folsom, 1995; Bakken, 2004; Hedblad, 2002)	144,519	162,646	RR 1.47 (1.24-1.75)	11 more per 1000 (from 5 more to 17 more)	Very low	Prospective cohort	Very serious[1]	Very serious[2]	No serious	Serious[3]	None
Breast cancer mortality, (Ever users of HRT)											
2 (Beral, 2003; Stahlberg, 2005)	440,474	399,323	RR 1.31 (0.94-1.84)	1 more per 1000 (from 0 fewer to 2 more)	Very low	Prospective cohort	Very serious[1]	Very serious[2]	No serious	Serious[3]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Breast cancer incidence (Current users of HRT)											
4 (Beral, 2003; Tjonneland, 2004; Stahlberg, 2005; Bakken, 2004)	479,520	428,880	RR 2.03 (1.65-2.50)	10 more per 1000 (from 6 more to 14 more)	Very low	Prospective cohort	Very serious[2]	Very serious[2]	No serious	No serious	None
Breast cancer mortality (Current users of HRT)											
3 (Beral, 2003; Stahlberg, 2005; Grodstein, 1997)	440,474	399,323	RR 1.16 (0.76-1.77)	0 fewer per 1000 (from 0 fewer to 1 more)	Very low	Prospective cohort	Very serious[1]	Very serious[2]	No serious	Serious[3]	None
Breast cancer incidence (Past users of HRT)											
4 (Beral, 2003; Tjonneland, 2004; Stahlberg, 2005; Bakken, 2004)	479,520	428,880	RR 1.02 (0.96-1.09)	0 fewer per 1000 (from 1 fewer to 2 more)	Low	Prospective cohort	Very serious[1]	No serious	No serious	No serious	None
Breast cancer mortality (Past users of HRT)											
3 (Beral, 2003; Stahlberg, 2005; Grodstein, 1997)	440,474	399,323	RR 0.98 (0.84-1.15)	0 fewer per 1000 (from 0 fewer to 0 fewer)	Low	Prospective cohort	Very serious[2]	No serious	No serious	No serious	None

1. High risk of performance, attrition, detection, and attrition biases in included studies;
2. Evidence was downgraded by 2 due to very serious heterogeneity (chi-squared p<0.1, I-squared inconsistency statistic > 75%) and no plausible explanation was found with sensitivity or subgroup analysis;
3. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25);
4. Evidence was downgraded by 1 due to serious heterogeneity (chi-squared p<0.1, I-squared inconsistency statistic of 50%-74.99% and no plausible explanation was found with sensitivity or subgroup analysis;
5. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)
6. High risk of performance and/or attrition biases

I.5.6 Osteoporosis

Table 64: GRADE profile: current use of HRT versus no current use of HRT for the outcomes of any fracture, any non- vertebral fracture, hip fracture, vertebral fracture, wrist fracture (RCTs)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	Placebo	Relative risk (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Any fracture											
5 (Cherry 2002, Mosekilde 2000, PEPI 1996 Ravn 1999, Veerus 2006))	119/2724	178/2564	RR 0.67 (0.53 to 0.85)	23 fewer per 1000 (from 10 fewer to 33 fewer)	Low	Randomised trials	Serious[3]	No serious	No serious	Serious[1]	None
Any non-vertebral fracture											
9 (Bjarnson and Christiansen 2000, Delmas et al., 2000, Genant 1997, Hosking 1998, Komulainen 1998, Mosekilde 2000, Lees and Stevenson 2001, Weiss 1999, Wimalawansa 1998)	65/1962	90/1603	RR 0.65 (0.47 to 0.90)	20 fewer per 1000 (from 6 fewer to 30 fewer)	Low	Randomised trials	Serious[3]	No serious	No serious	Serious[1]	None
Hip fracture											
2 (Mosekilde 2000, Vickers 2007)	3/2698	3/2693	RR 1.00 (0.23 to 4.39)	0 fewer per 1000 (from 1 fewer to 4 more)	Very low	Randomised trials	Serious[3]	No serious	No serious	Serious[2]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	Placebo	Relative risk (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Vertebral fracture											
5 (Delmas 2000, Lufkin 1992, Mosekilde 2000, Reid 2004, Wimalawansa 1998)	18/804	24/758	RR 0.75 (0.43 to 1.30)	8 fewer per 1000 (from 18 fewer to 9 more)	Very low	Randomised trials	Very serious[4]	No serious	No serious	Very Serious[2]	None
Wrist fracture											
2 (Komulainene 1998, Mosekilde 2000)	10/618	32/620	RR 0.31 (0.16 to 0.63)	36 fewer per 1000 (from 19 fewer to 43 fewer)	Moderate	Randomised trials	Serious[3]	No serious	No serious	No serious	None
(Any non-vertebral fracture (duration up to 2 years)											
5 (Delmas 2000, Genant 1997, Hosking 1998, Lees and Stevenson 2001, Weiss 1999)	20/ 1098	22/ 808	RR 0.74 (0.37 to 1.49)	1 fewer per 1000. (from 2 fewer to 1 more)	Very low	Randomised trials	Very serious[4]	No serious	No serious	Very serious[2]	None
Hip fracture (duration up to 2 years)											
1 (Vickers 2007)	2/ 2196	3/ 2189	RR 0.66 (0.11 to 3.97)	0 fewer per 1000. (from 1 fewer to 4 more)	Very low	Randomised trials	Serious[3]	No serious	No serious	Very serious[2]	None
Vertebral fracture (duration up to 2 years)											
2 (Delmas 2000, Lufkin 1992)	7/ 126	14/ 84	RR 0.51 (0.24 to 1.10)	82 fewer per 1000. (From 127 fewer to 17 more)	Very low	Randomised trials	Very serious[4]	No serious	No serious	Serious[2]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	Placebo	Relative risk (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Any fracture (duration2 to 5 years)											
4 (Mosekilde 2000, PEPI 1996 Ravn 1999, Veerus 2006)	108/ 2211	160/ 2060	RR 0.68 (0.53 to 0.88)	25 fewer per 1000. (from 9 fewer to 37 fewer)	Low	Randomised trials	Serious[3]	No serious	No serious	Serious[1]	None
Any non-vertebral fracture (duration 2 to 5 years)											
3 (Komulainen 1998, Mosekilde 2000, Wimalawansa 1998)	34/ 636	59/ 638	RR 0.58 (0.38 to 0.87)	39 fewer per 1000. (from 12 fewer to 57 fewer)	Moderate	Randomised trials	No serious	No serious	No serious	Serious[1]	None
Hip fracture (duration 2 to 5 years)											
1 (Mosekilde 2000)	1/ 502	0/ 504	RR 3.01 (0.12 to 73.76)	Unable to calculate as no events in control group	Low	Randomised trials	No serious	No serious	No serious	Very serious[2]	None
Vertebral fracture (duration 2 to 5 years)											
3 (Mosekilde 2000, Reid 2004, Wimalawansa 1998)	11/ 678	10/ 674	RR 1.10 (0.48 to 2.52)	1 more per 1000. (from 8 fewer to 23 more)	Very low	Randomised trials	Serious[3]	No serious	No serious	Very serious[2]	None
Wrist fracture (duration 2 to 5 years)											
2 (Komulainen 1998, Mosekilde 2000)	8/ 618	22/ 620	RR 0.36 (0.16 to 0.81)	23 fewer per 1000. (from 7 fewer to 30 fewer)	Moderate	Randomised trials	No serious	No serious	No serious	Serious[1]	None

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MID (0.75 to 1.25)
3. Evidence was downgraded by 1 taking into account the weight from studies with high risk due to selection, performance, attrition or detection bias
4. Evidence was downgraded by 2 taking into account weight from studies with very high risk due to selection, performance, attrition or detection bias

Table 65: GRADE profile: current use of oestrogen plus progestogen versus no current use of HRT for the outcomes of any fracture, any osteoporotic fracture, hip fracture, vertebral fracture, wrist fracture (RCTs)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	Placebo	Hazard ratio or relative risk (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Any fracture											
2 (Ravn 1999, Veerus 2006,)	49/1008	108/1382	RR 0.62 (0.44 to 0.87)	30 fewer per 1000 (from 10 fewer to 44 fewer)	Low	Randomised trials	Serious[7]	No serious	No serious	Serious[1]	None
Any fracture											
1 (Manson 2013)	741/8506	903/8102	HR 0.76 (0.69 to 0.83)	26 fewer per 1000 (From 18 fewer to 33 fewer)	Low	Randomised trial with post-intervention	Serious[3, 4,5,6]	No serious	No serious	Serious[1]	None
Any non-vertebral fracture											
5 (Delmas 2000, Hosking 1998, Komulainen 1998, Lees and Stevenson 2001Wimalawansa 1998)	28/916	47/910	RR 0.58 (0..36 to 0.94)	22 fewer per 1000 (from 3 fewer to 33 fewer)	Moderate	Randomised trials	No serious	No serious	No serious	Serious[1]	None
Hip fracture											
1 (Vickers 2007)	2/2196	3/2189	RR 0.66 (0.11 to 3.97)	0 fewer per 1000 (from 1 fewer to 4 more)	very low	Randomised trials	Serious[8]	No serious	No serious	Very serious[2]	None
Hip fracture											
1 (Manson 2013)	232/8506	270/8102	HR 0.81 (0.68 to 0.97)	6 fewer per 1000 (from 1 fewer to 11 fewer)	Low	Randomised trial with post-intervention	Serious[3, 4,5,6]	No serious	No serious	Serious[1]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	Placebo	Hazard ratio or relative risk (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Vertebral fracture											
3 (Delmas 2000, Lufkin 1992, Wimalawansa 1998)	9/144	19/102	RR 0.48 (0.25 to 0.96	97 fewer per 1000 (from 7 fewer to 140 fewer)	Very low	Randomised trials	serious[7]	No serious	No serious	Serious[1]	None
Vertebral fracture											
1 (Manson 2013)	56/8506	78/8102	HR 0.68 (0.48 to 0.96)	3 fewer per 1000 (from 0 fewer to 5 fewer)	Low	Randomised trials with post-intervention	Serious[3, 4,5,6]	No serious	No serious	Serious[1]	None
Wrist fracture											
1 (Komulainen 1998)	2/116	7/116	RR 0.29 (0.06 to 1.35)	43 fewer per 1000 (from 57 fewer to 21 more)	Very low	Randomised trials	Serious[8]	No serious	No serious	Very Serious[2]	None
Osteoporotic fracture											
1 (Vickers 2007)	40/ 2196	58/ 2189	RR 0.69 (0.46 to 1.03)	8 fewer per 1000 (from 3 fewer to 12 fewer)	Low	Randomised trials	Serious[8]	No serious	No serious	Serious[1]	None

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)
3. Overall breaking of blinding was relatively high because of the need to unmask 40.5% of the gynaecologists in the treatment group (due to vaginal bleeding after taking HRT) compared with 6.8% in the placebo group, though the distribution of that in the age group this review is interested is unclear;
4. High attrition bias among all participants; about 42% of women in the oestrogen plus progesterone and 38% of women in the placebo stopped taking the study drugs during follow-up. The drop-in rates were 6.2% in the intervention group and 10.7% in the placebo group, respectively. However, the distribution of that in the age group this review is interested is unclear;
5. In the WHI CEE plus progesterone trial, about 26% participants in each group had used HRT prior the enrolment of the trial in their lifetime; and about 6% in each group were current HRT users. The distribution of that in the age group this review is interested is unclear.
6. Among all participants in the WHI, about 34% of women had a BMI higher than 30 kg/m2 in each group. The distribution of that in the age group this review is interested is unclear;
7. Evidence was downgraded by 1 taking into account the weight from studies with high risk due to selection, performance, attrition or detection bias
8. Evidence was downgraded by 1 due to selection, performance, attrition or detection bias

Table 66: GRADE profile: current use of oestrogen plus progestogen versus no current HRT use (subgroup analysis age) for the outcomes of any fracture or hip fracture (RCTs)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	Placebo	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Any fracture (age 50-54 years)											
1 (Cauley 2003)	67/ 1139	90/ 1050	HR 0.68 (0.49 to 0.93)	27 fewer per 1000. (from 6 fewer to 43 fewer)	Low	Randomised trials	Serious[3,4,5,6]	No serious	No serious	Serious[1]	None
Hip fracture (age 50-59 years)											
1 (Manson 2013)	1/ 2839	5/ 2683	HR 0.17 (0.02 to 1.43)	2 fewer per 1000. (from 2 fewer to 1 more)	Very low	Randomised trial with post-intervention	Serious[3,4,5,6]	No serious	No serious	Very serious[2]	None
Hip fracture (age 50 to 59 years)											
1 (Manson 2013)	17/8506	28/8102	HR 0.57 (0.31 to 1.04)	1 fewer per 1000 (from 2 fewer to 0 more)	Low	Randomised trial with post-intervention	Serious[3,4,5,6]	No serious	No serious	Serious[1]	None
Any fracture (age 55 to 59 years)											
1 (Cauley 2003)	124/ 1877	126/ 1744	HR 0.91 (0.71 to 1.16)	6 fewer per 1000. (from 20 fewer to 11 more)	Low	Randomised trials with post-intervention	Serious[3,4,5,6]	No serious	No serious	Serious[1]	None
Any fracture (age 60 to 64 years)											
1 (Cauley 2003)	168/ 1961	184/ 1776	HR 0.80 (0.65 to 0.98)	20 fewer per 1000. (from 2 fewer to 35 fewer)	Low	Randomised trials	Serious [3,4,5,6]	No serious	No serious	Serious[1]	None
Any fracture (age 65 to 69 years)											
1 (Cauley 2003)	161/ 1879	238/ 1809	HR 0.68 (0.49 to 0.93)	40 fewer per 1000. (from 9 fewer to 65 fewer)	Low	Randomised trials	Serious[3,4,5,6]	No serious	No serious	Serious[1]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	Placebo	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Hip fracture age (60 to 69 years)											
1 (Cauley 2003)	19/ 3853	23/ 3657	HR 0.76 (0.41 to 1.39)	2 fewer per 1000. (from 4 fewer to 2 more)	Very Low	Randomised trials	Serious[3,4,5,6]	No serious	No serious	Very serious[2]	None
Hip fracture (age 60 to 59 years)											
1 (Manson 2013)	103/8506	100/8102	HR 0.94 (0.71 to 1.24)	1 fewer per 1000 (from 4 fewer to 3 more)	Low	Randomised trial with post-intervention	Serious[3,4,5,6]	No serious	No serious	Serious[1]	None

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Evidence was downgraded by 2 due to serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)
3. Overall breaking of blinding was relatively high because of the need to unmask 40.5% of the gynaecologists in the treatment group (due to vaginal bleeding after taking HRT) compared with 6.8% in the placebo group, though the distribution of that in the age group this review is interested is unclear;
4. High attrition bias among all participants; about 42% of women in the oestrogen plus progesterone and 38% of women in the placebo stopped taking the study drugs during follow-up. The drop-in rates were 6.2% in the intervention group and 10.7% in the placebo group, respectively. However, the distribution of that in the age group this review is interested is unclear;
5. In the WHI CEE plus progesterone trial, about 26% participants in each group had used HRT prior the enrolment of the trial in their lifetime; and about 6% in each group were current HRT users. The distribution of that in the age group this review is interested is unclear.
6. Among all participants in the WHI, about 34% of women had a BMI higher than 30 kg/m2 in each group. The distribution of that in the age group this review is interested is unclear;
7. Evidence was downgraded by 1 taking into account the weight from studies with high risk due to selection, performance, attrition or detection bias

Table 67: GRADE profile: current use of oestrogen plus progestogen versus placebo (subgroup analysis duration) for the outcome of any osteoporotic fracture (RCTs)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	Placebo	Hazard ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Osteoporotic fracture (up to 2 years)											
1 (Vickers 2007)	40/ 2196	58/ 2189	RR 0.69 (0.46 to 1.03)	8 fewer per 1000. (from 14 fewer to 1 more)	Low	Randomised trials	Serious[2]	No serious	No serious	Serious[1]	None

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Evidence was downgraded by 1 due to selection, performance, attrition or detection bias

Table 68: GRADE profile: current use of oestrogen versus no current use of HRT for the outcomes of any fracture, non- vertebral fracture, hip fracture, vertebral fracture, wrist fracture (RCTs)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	Placebo	Hazard ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Any fracture											
1 (Cherry 2002)	11/513	18/504	RR 0.60 (0.29 to 1.26)	14 fewer per 1000 (from 25 fewer to 9 more)	Moderate	Randomised trials	Serious[7]	No serious	No serious	Very Serious[2]	None
Any fracture (intervention phase)											
1 (Manson 2013)	544/5310	767/5429	HR 0.72 (0.64 to 0.80)	37 fewer per 1000 (from 27 fewer to 98 fewer)	Low	Randomised trial with post-intervention	Serious[3, 4,5,6]	No serious	No serious	Serious[1]	None
Any non-vertebral fracture											
2 (Aitken 1973, Weiss 1999)	3/197	3/112	RR 0.52 (0.10 to 2.73)	13 fewer per 1000 (from 24 fewer to 46 more)	Low	Randomised trials	Serious[8]	No serious	No serious	Very Serious[2]	None
Hip fracture											
1 (Jackson 2006)	46/ 5310	73/ 5429	HR 0.64 (0.45 to 0.93)	5 fewer per 1000. (from 1 fewer to 7 fewer)	Low	Randomised trials	Serious[3, 4,5,6]	No serious	No serious	Serious[1]	None
Hip fracture (during and post intervention)											
1 (Manson 2013)	134/5310	148/5429	HR 0.91 (0.72 to 1.15)	2 fewer per 1000 (from 8 fewer to 4 more)	Low	Randomised trial with post-intervention	Serious[3, 4,5,6]	No serious	No serious	Serious[1]	None
Vertebral fracture											
1 (Reid 2004)	1/158	1/152	RR 0.96 (0.06 to 15.24)	0 fewer per 1000 (from 6 fewer to 94 more)	Moderate	Randomised trials	Serious[7]	No serious	No serious	Very Serious[2]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	Placebo	Hazard ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Vertebral fracture											
1 (Manson 2013)	44/5310	70/5429	HR 0.64 (0.44 to 0.93)	5 fewer per 1000 (from 1 fewer to 7 fewer)	Low	Randomised trial with post-intervention	Serious[3, 4,5,6]	No serious	No serious	Serious[1]	None
Wrist fracture											
1 (Jackson 2006)	130/ 5310	227/ 5429	HR 0.58 (0.47 to 0.72)	17 fewer per 1000. (from 12 fewer to 22 fewer)	Moderate	Randomised trials	Serious[3, 4,5,6]	No serious	No serious	No serious	None

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)
3. Overall breaking of blinding was relatively high because of the need to unmask 40.5% of the gynaecologists in the treatment group (due to vaginal bleeding after taking HRT) compared with 6.8% in the placebo group, though the distribution of that in the age group this review is interested is unclear;
4. High attrition bias among all participants; about 42% of women in the oestrogen plus progesterone and 38% of women in the placebo stopped taking the study drugs during follow-up. The drop-in rates were 6.2% in the intervention group and 10.7% in the placebo group, respectively. However, the distribution of that in the age group this review is interested is unclear;
5. In the WHI CEE plus progesterone trial, about 26% participants in each group had used HRT prior the enrolment of the trial in their lifetime; and about 6% in each group were current HRT users. The distribution of that in the age group this review is interested is unclear.
6. Among all participants in the WHI, about 34% of women had a BMI higher than 30 kg/m2 in each group. The distribution of that in the age group this review is interested is unclear;
7. Evidence was downgraded by 1 due to selection, performance, attrition or detection bias
8. Evidence was downgraded by 1 taking into account the weight from studies with high risk due to selection, performance, attrition or detection bias

Table 69: GRADE profile: current use of oestrogen versus placebo (subgroup analysis age) for the outcome of any fracture or hip fracture (RCTs)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	Placebo	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Hip fracture (age 50 to 59 years)											
1 (Jackson 2006)	5/ 1637	1/ 1673	HR 5.02 (0.59 to 43.02)	2 more per 1000. (from 0 fewer to 25 more)	Very low	Randomised trials	Serious[3, 4,5,6]	No serious	No serious	Very serious[2]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	Placebo	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Hip fracture (age 50 to 59 years)											
1 (Manson 2013)	5/5310	1/5429	HR 5.01 (0.59 to 42.91)	1 more per 1000 (from 0 fewer to 8 more)	Very low	Randomised trial with post-intervention	Serious[3,4,5,6]	No serious	No serious	Very serious[2]	None
Hip fracture (age 50 to 59 years)											
1 (Manson 2013)	9/5310	10/5429	HR 0.88 (0.36 to 2.17)	0 fewer per 1000 (from 1 fewer to 2 more)	Very low	Randomised trials with post-intervention	Serious[3,4,5,6]	No serious	No serious	Very serious[2]	None
Any fracture (age 50 to 59 years)											
1 (Jackson 2006)	153/ 1637	173/ 1673	HR 0.90 (0.72 to 1.12)	10 fewer per 1000. (from 28 fewer to 12 more)	Low	Randomised trials	Serious[3,4,5,6]	No serious	No serious	Serious[1]	None
Any fracture (age 60 to 69 years)											
1 (Jackson 2006)	220/ 2387	348/ 2465	HR 0.63 (0.53 to 0.75)	50 fewer per 1000. (from 33 fewer to 64 fewer)	Moderate	Randomised trials	Serious[3,4,5,6]	No serious	No serious	No serious	None
Hip fracture (age 60 to 69 years)											
1 (Jackson 2006)	9/ 2387	20/ 2465	HR 0.47 (0.22 to 1.04)	4 fewer per 1000. (from 6 fewer to 0 more)	Very Low	Randomised trials	Serious[3,4,5,6]	No serious	No serious	Very serious[2]	None
Hip fracture (age 60 to 69 years)											
1 (Manson 2013)	9/5310	20/5429	HR 0.47 (0.22 to 1.04)	2 fewer per 1000 (from 3 fewer to 0 more)	Low	Randomised trial with post-intervention	Serious[3,4,5,6]	No serious	No serious	Serious[1]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	Placebo	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Hip fracture (age 60 to 69 years)											
1 (Manson 2013)	46/5310	49/5429	HR 0.95 (0.64 to 1.43)	0 fewer per 1000 (from 3 fewer to 4 more)	Very low	Randomised trial with post-intervention	Serious[3, 4,5,6]	No serious	No serious	Very serious[2]	None

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Evidence was downgraded by 2 due to serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)
3. Overall breaking of blinding was relatively high because of the need to unmask 40.5% of the gynaecologists in the treatment group (due to vaginal bleeding after taking HRT) compared with 6.8% in the placebo group, though the distribution of that in the age group this review is interested is unclear;
4. High attrition bias among all participants; about 42% of women in the oestrogen plus progesterone and 38% of women in the placebo stopped taking the study drugs during follow-up. The drop-in rates were 6.2% in the intervention group and 10.7% in the placebo group, respectively. However, the distribution of that in the age group this review is interested is unclear;
5. In the WHI CEE plus progesterone trial, about 26% participants in each group had used HRT prior the enrolment of the trial in their lifetime; and about 6% in each group were current HRT users. The distribution of that in the age group this review is interested is unclear.
6. Among all participants in the WHI, about 34% of women had a BMI higher than 30 kg/m2 in each group. The distribution of that in the age group this review is interested is unclear; a Stratified by age, prior disease and randomisation status in the WHI dietary intervention trial

Table 70: GRADE profile: current use of progestogen versus no current use of HRT for the outcome of vertebral fracture (RCTs)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	Placebo	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Vertebral fracture											
1 (Liu 2005)	0/ 65	0/ 23	unable to calculate as no events in either group	-	Moderate	Randomised trials	No serious	No serious	No serious	N/C[1]	None

1. Imprecision was not calculable

Table 71: GRADE profile: current use of HRT versus no current use or never use of HRT for the outcome of any fracture, any non-vertebral fracture, osteoporotic fracture, hip fracture, vertebral fracture, wrist fracture (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio, odds ratio, or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Any fracture											
1 (Huopio 2000)	48/ 799	209/ 2269	RR 0.65 (0.47 to 0.88)	32 fewer per 1000. (from 11 fewer to 49 fewer)	Very low	Prospective cohort	Serious[2,6]	No serious	No serious	Serious[1]	None
Any fracture											
1 (Banks 2004)	1179/ 46122	3010 70297	RR 0.62 (0.58 to 0.66)	16 fewer per 1000. (from 15 fewer to 18 fewer)	Low	Prospective cohort	Serious[2,7]	No serious	No serious	No serious	None
Any fracture											
1 (Lafferty 1994)	3/ 81	16/ 76	RR 0.28 (0.09 to 0.89)	152 fewer per 1000. (from 23 fewer to 192 fewer)	Very low	Prospective cohort	Serious[4,8]	No serious	No serious	Serious[1]	None
Any fracture											
1 (Randell 2002)	94/ 1335	352/ 3335	RR 0.62 (0.48 to 0.79)	40 fewer per 1000. (from 22 fewer to 55 fewer)	Low	Prospective cohort	Serious[9]	No serious	No serious	Serious1	None
Any non-vertebral fracture											
1 (Hundrup 2004)	50/ 1936	215/ 4019	HR 0.50 (0.35 to 0.71)	26 fewer per 1000. (from 15 fewer to 34 fewer)	Low	Prospective cohort	Serious[1,10]	No serious	No serious	No serious	None
Any non-vertebral fracture											
1 (Lafferty 1994)	2/ 81	11/ 76	RR 0.23 (0.06 to 0.97)	111 fewer per 1000. (from 4 fewer to 136 fewer)	Very low	Prospective cohort	Serious[3,8]	No serious	No serious	Serious[4]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio, odds ratio, or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Hip fracture											
1 (Høidrup 1999)	37/ 1314	326/ 4832	RR 0.71 (0.50 to 1.01)	20 fewer per 1000. (from 34 fewer to 1 more)	Very low	Prospective cohort	Serious[1,11]	No serious	No serious	Serious[1]	None
Hip fracture											
1 (Yates 2004)	66/ 67973	149/ 53723	OR 0.60 (0.44 to 0.82)	1 fewer per 1000. (from 0 fewer to 2 fewer)	Low	Prospective cohort	Serious[1,2]	No serious	No serious	Serious[1]	None
Vertebral fracture											
1 (Lafferty 1994)	1/ 81	6/ 76	RR 0.27 (0.12 to 0.60)	58 fewer per 1000. (from 32 fewer to 69 fewer)	Low	Prospective cohort	Serious[5,8]	No serious	No serious	No serious	None
Wrist fracture											
1 (Randell 2002)	22/ 1335	145/ 3335	RR 0.41 (0.26 to 0.67)	26 fewer per 1000. (from 14 fewer to 32 fewer)	Moderate	Prospective cohort	Serious[9]	No serious	No serious	No serious	None
Wrist fracture											
1 (Honkanen 2000)	110/ 4842	258/ 6956	HR 0.37 (0.23 to 0.61)	23 fewer per 1000. (from 14 fewer to 28 fewer)	Moderate	Prospective cohort	Serious[1,3]	No serious	No serious	No serious	None
Osteoporotic fracture											
1 (Engel 2011)	not reported	1981/ 18651	HR 0.78 (0.73 to 0.83)	22 fewer per 1000. (from 17 fewer to 28 fewer)	Low	Prospective cohort	serious[1,4]	No serious	No serious	Serious[1]	None

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Data on use of HRT only collected at baseline, not at follow up. Therefore "current users" and "non-users" at baseline may have changed HRT status by follow up
3. Data from a practice of a single individual

4. Subjects identified through private practice of a single individual
5. Data from individual private practice of one clinician
6. Adjusted for age, height, weight, menopausal status, BMD, previous fractures, maternal hip fracture, calcium intake, smoking, multiple chronic health disease
7. Adjusted for age, region, socioeconomic status, time since menopause, BMI, physical activity
8. Adjusted for age
9. Adjusted for age, time since menopause, BMI, number of chronic health diseases, history of previous fractures
10. Adjusted for age at menopause, BMI and family history
11. Adjusted for age, BMI, physical activity, smoking, alcohol, cohabitation, marital status, education, age at menopause, parity
12. Adjusted for age, BMI, previous fracture, health status, maternal history of fractures, cortisone use
13 Adjusted for age, menopausal status, BMI, calcium intake, previous wrist fracture, parity
14 Adjusted for BMI, physical activity, age at menopause, parity, previous use of oral contraceptive, previous use of calcium supplement, education

Table 72: GRADE profile: current HRT use versus no HRT use (subgroup analysis duration) for the outcome of any fracture (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Relative risk (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Any fracture (< 1 year duration)											
1 (Banks 2004)	81/ 2801	3010/ 70297	RR 0.75 (0.60 to 0.93)	11 fewer per 1000. (from 3 fewer to 17 fewer)	Very low	Prospective cohort	Serious[2,3]	No serious	No serious	Serious[1]	None
Any fracture (1 to 4 years)											
1 (Banks 2004)	405/ 15707	3010/ 70297	RR 0.66 (0.60 to 0.74)	15 fewer per 1000. (from 11 fewer to 17 fewer)	Low	Prospective cohort	Serious[2,3]	No serious	No serious	No serious	None
Any fracture (5 to 9 years duration)											
1 (Banks 2004)	458/ 18604	3010/ 70297	RR 0.58 (0.53 to 0.65)	18 fewer per 1000. (from 15 fewer to 20 fewer)	Low	Prospective cohort	Serious[2,3]	No serious	No serious	Serious[1]	None
Any fracture (≥ 10 years)											
1 (Banks 2004)	206/ 7956	3010/ 70297	RR 0.57(0.50 to 0.66)	18 fewer per 1000. (from 15 fewer to 21 fewer)	Low	Prospective cohort	Serious[2,3]	No serious	No serious	No serious	None

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Data on use of HRT only collected at baseline, not at follow up. Therefore "current users" and "non-users" at baseline may have changed HRT status by follow up

3. Adjusted for age, region, socioeconomic status, time since menopause, BMI, physical activity

Table 73: GRADE profile: current use of HRT versus no HRT use (subgroup analysis duration) for the outcome of osteoporotic fracture (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio or odds ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Osteoporotic fracture (< 2 years duration)											
1 (Engel 2011)	N/R	1981/ 18651	HR 0.89 (0.80 to 1.00)	11 fewer per 1000. (from 20 fewer to 0 more)	Low	Prospective cohort	Serious[2]	No serious	No serious	Serious[1]	None
Osteoporotic fracture(2 to 4.9 years duration)											
1 (Engel 2011)	N/R	1981/ 18651	HR 0.71 (0.64 to 0.79)	30 fewer per 1000. (from 21 fewer to 37 fewer)	Low	Prospective cohort	Serious[2]	No serious	No serious	Serious[1]	None
Osteoporotic fracture (≤ 5 years duration)											
1 (Barrett-Connor 2003)	220/ 23295	974/ 53737	OR 0.75 (0.65 to 0.88)	4 fewer per 1000. (from 2 fewer to 6 fewer)	Low	Prospective cohort	Serious[3]	No serious	No serious	Serious[1]	None
(Subgroup analysis- duration Osteoporotic fracture(duration ≥ 5 years)											
1 (Engel 2011)	N/R	1981/ 18651	HR 0.77 (0.71 to 0.84)	23 fewer per 1000. (from 16 fewer to 30 fewer)	Low	Prospective cohort	Serious[2]	No serious	No serious	Serious[1]	None
Osteoporotic fracture (6 to 10 years duration)											
1 (Barrett-Connor 2003)	152/ 16737	974/ 53737	OR 0.71 (0.59 to 0.84)	5 fewer per 1000. (from 3 fewer to 7 fewer)	Low	Prospective cohort	Serious[3]	No serious	No serious	Serious[1]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio or odds ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Subgroup analysis- duration Osteoporotic fracture (> 10 years duration)											
1 (Barrett-Connor 2003)	333/ 27941	974/ 53737	OR 0.75 (0.66 to 0.85)	4 fewer per 1000. (from 3 fewer to 6 fewer)	Low	Prospective cohort	Serious[3]	No serious	No serious	Serious[1]	None

N/R: not reported
1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Adjusted for BMI, physical activity, age at menopause, parity, previous use of oral contraceptive, previous use of calcium supplements, education
3. Adjusted for age, prior fracture, health status, maternal history of fracture, cortisone use

Table 74: GRADE profile: current use of HRT versus no HRT use (subgroup analysis duration) for the outcome of any non- vertebral fracture and hip fracture (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio or odds ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Any non-vertebral fracture (< 5 years duration)											
1 (Hundrup 2004)	20/ 723	215/ 4019	HR 0.65 (0.37 to 1.14)	18 fewer per 1000. (from 33 fewer to 7 more)	Very low	Prospective cohort	Serious[2]	No serious	No serious	Serious[1]	None
Hip fracture (≤ 5 years duration)											
1 (Yates 2004)	11/ 23282	149/ 53737	OR 0.35 (0.18 to 0.67)	2 fewer per 1000. (from 1 fewer to 2 fewer)	Moderate	Prospective cohort	Serious[3]	No serious	No serious	No serious	None
Any non-vertebral fracture (5 to 10 years duration)											
1 (Hundrup 2004)	20/ 566	215/ 4019	HR 0.62 (0.36 to 1.07)	20 fewer per 1000. (from 34 fewer to 4 more)	Very low	Prospective cohort	Serious[2]	No serious	No serious	Serious[1]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio or odds ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Hip fracture (6 to 10 years duration)											
1 (Yates 2004)	15/ 16722	149/ 53737	OR 0.71 (0.41 to 1.23)	1 fewer per 1000. (from 2 fewer to 1 more)	Low	Prospective cohort	Serious[3]	No serious	No serious	Serious[1]	None
Any non-vertebral fracture (≥ 10 years duration)											
1 (Hundrup 2004)	10/ 570	215/ 4019	HR 0.32 (0.16 to 0.64)	36 fewer per 1000. (from 1 fewer to 35 fewer)	Low	Prospective cohort	Serious[2]	No serious	No serious	No serious	None
Hip fracture (> 10 years duration)											
1 (Yates 2004)	40/ 27901	149/ 53737	OR 0.66 (0.46 to 0.95)	1 fewer per 1000. (from 0 fewer to 1 fewer)	Low	Prospective cohort	Serious[3]	No serious	No serious	Serious[1]	None

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Adjusted for age at menopause, BMI, family history
3. Adjusted for age, BMI, previous fracture, health status, maternal history of fracture, parity

Table 75: GRADE profile: ever use of HRT versus never use of HRT for the outcome of hip fracture, vertebral fracture, wrist fracture (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio, odds ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Any fracture											
1 (Tuppurainen 1995)	N/R	N/R	OR 0.70 (0.50 to 0.96)	N/C	Low	Prospective cohort	Serious[3]	No serious	No serious	Serious[3]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio, odds ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Hip fracture											
1 (Maxim 1995)	14/ 245	15/ 245	RR 1.31 (0.55 to 3.12)	19 more per 1000. (from 28 fewer to 130 more)	Very low	Prospective cohort	Serious[4]	No serious	No serious	Very serious[2]	None
Hip fracture											
1 (Melton III 1996)	N/R	N/R	RR 0.8 (0.2 to 2.6)	N/C	Very low	Prospective cohort	Serious[5]	No serious	No serious	Very serious[2]	None
1 (Paganini-Hill 1991)	163/ 4866	166/ 3708	RR 1.02 (0.81 to 1.27)	1 more per 1000. (from 9 fewer from 12 more)	Low	Prospective cohort	Serious[6]	No serious	No serious	Serious[1]	None
Vertebral fracture											
1 (Melton III 1996)	N/R	N/R	RR 0.8 (0.4 to 1.9)	N/C	Very low	Prospective cohort	Serious[5]	No serious	No serious	Very serious[2]	None
Vertebral fracture											
1 (Maxim 1995)	59/ 245	98/ 245	RR 0.60 (0.36 to 0.99)	160 fewer per 1000. (From 4 fewer to 256 fewer)	Low	Prospective cohort	Serious[4]	No serious	No serious	Serious[1]	None
Vertebral fracture											
1 (Paganini-Hill 2005)	342/ 4987	268/ 3863	HR 0.95 (CI not reported, but NS)	3 fewer per 1000 (Unable to calculate CI)	Low	Prospective cohort	Serious[6]	No serious	No serious	N/R	None
Wrist fracture											
1 (Melton III 1996)	N/R	N/R	RR 1.6 (0.8 to 3.2)	Unable to calculate	Low	Prospective cohort	Serious[5]	No serious	No serious	Serious[1]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio, odds ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Wrist fracture											
1 (Maxim 1995)	23/ 245	41/ 245	RR 0.44 (0.23 to 0.84)	94 fewer per 1000. From 27 fewer to 129 fewer	Low	Prospective cohort	Serious[4]	No serious	No serious	Serious[1]	None
Wrist fracture											
1 (Paganini-Hill 2005)	276/ 4987	186/ 3863	HR 0.93 (CI not reported but NS)	3 fewer per 1000. (Unable to calculate CI)	Low	Prospective cohort	Serious[6]	No serious	No serious	N/R	None
Osteoporotic fracture											
1 (Engel 2011)	3608/ 51531	1981/ 18651	HR 0.85 (0.81 to 0.91)	15 fewer per 1000. (From 9 fewer to 19 fewer)	Moderate	Prospective cohort	Serious[7]	No serious	No serious	No serious	None

N/C: not calculable; N/R: not reported; NS: not significant

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)

2. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)

3. Adjusted for age

4. Adjusted for age at menopause, BMI, smoking history

5. Adjusted for age

6. Adjusted for smoking, exercise, and attitude

7. Adjusted for BMI, physical activity, age at menopause, parity, previous use of calcium supplements, previous use of oral contraceptive, education

Table 76: GRADE profile: ever use of HRT versus never use of HRT (subgroup analysis duration) for the outcomes of hip, vertebral fracture, wrist fracture (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Hip fracture (≤ 3 years duration)											
1 (Paganini-Hill 1991)	63/ 1449	166/ 3708	RR 1.19 (0.89 to 1.60)	9 more per 1000. (From 5 fewer to 27 more)	Low	Prospective cohort	Serious[2]	No serious	No serious	Serious[1]	None
Vertebral fracture (< 3 years duration)											
1 (Paganini-Hill 2005)	75/ 1065	268/ 3863	HR 0.79 (CI not reported, but NS)	N/C	Low	Prospective cohort	Serious[3]	No serious	No serious	N/C	None
Wrist fracture(< 3 years duration)											
1 (Paganini-Hill 2005)	78/ 1065	186/ 3863	HR 1.15 (CI not reported, but NS)	N/C	Low	Prospective cohort	Serious[3]	No serious	No serious	N/C	None
Vertebral fracture (3 to 14 years duration)											
1 (Paganini-Hill 2005)	142/ 2037	268/ 3863	HR 1.01 (CI not reported, but NS)	N/C	Low	Prospective cohort	Serious[3]	No serious	No serious	N/C	None
Wrist fracture (3 to 14 years duration)											
1 (Paganini-Hill 2005)	111/ 2037	186/ 3863	HR 0.85 (CI not reported, but NS)	N/C	Low	Prospective cohort	Serious[3]	No serious	No serious	N/C	None
Hip fracture (4 to 14 years duration)											
1 (Paganini-Hill 1991)	46/ 1769	166/ 3708	RR 0.89 (0.63 to 1.23)	5 fewer per 1000. (From 17 fewer to 10 more)	Low	Prospective cohort	Serious[2]	No serious	No serious	Serious[1]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Hip fracture (≥ 15 years duration)											
1 (Paganini-Hill 1991)	43/ 1513	166/ 3708	RR 0.88 (0.63 to 1.24)	5 fewer per 1000. (From 1 fewer to 11 more)	Low	Prospective cohort	Serious[2]	No serious	No serious	Serious[1]	None
Vertebral fracture (≥ 15 years duration)											
1 (Paganini-Hill 2005)	106/ 1537	268/ 3863	HR 0.93 (CI not reported, but NS)	N/C	Low	Prospective cohort	Serious[3]	No serious	No serious	N/C	None
Wrist fracture (≥ 15 years duration)											
1 (Paganini-Hill 2005)	77/ 1537	186/ 3863	HR 0.85 (CI not reported, but NS)	N/C	Low	Prospective cohort	Serious[3]	No serious	No serious	N/C	None

N/C: not calculable; NS: not significant

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Adjusted for age
3. Adjusted for age, history of fracture, BMI, blood pressure medication, non-prescription pain medication, smoking, exercise, attitude

Table 77: GRADE profile: previous use of HRT versus never use of HRT for the outcomes of any fracture, osteoporotic fracture (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No treatment	Hazard ratio, odds ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Any fracture											
1 (Banks 2004)	841/ 18939	3010/ 70297	RR 0.98 (0.71 to 1.34)	1 fewer per 1000. (From 12 fewer to 15 more)	Very low	Prospective cohort	Serious[4]	No serious	No serious	Very serious[2]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No treatment	Hazard ratio, odds ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Any fracture											
1 (Bagger 2004)	27/ 155	36/ 108	OR 0.48 (0.26 to 0.88)	140 fewer per 1000. (From 28 fewer to 218 fewer)	Low	Prospective cohort	Serious[5]	No serious	No serious	Serious[1]	None
Any non-vertebral fracture											
1 (Hundrup 2004)	62/ 922	215/ 4019	HR 1.23 (0.89 to 1.70)	12 more per 1000. (From 6 fewer to 36 more)	Very low	Prospective cohort	Serious[3,6]	No serious	No serious	Serious[1]	None
Any non-vertebral fracture											
1 (Bagger 2004)	12/ 155	13/ 108	OR 0.68 (0.30 to 1.60)	35 fewer per 1000. (From 81 fewer to 59 more)	Very low	Prospective cohort	Serious[5]	No serious	No serious	Very serious[2]	None

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)
3. Evidence was downgraded by 1 due to performance and attrition bias
4. Adjusted for age, region, socioeconomic status, time since menopause, BMI, physical activity
5. Adjusted for age at baseline, BMC, spine BMD
6. Adjusted for age at menopause, BMI, family history

Table 78: GRADE profile: previous use of HRT versus never HRT use (subgroup analysis duration) for the outcome of any non - vertebral fracture (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Any non-vertebral fracture (less than 5 years duration)											
1 (Hundrup 2004)	43/ 577	215/ 4019	HR 1.41 (0.97 to 2.05)	21 more per 1000. (From 2 fewer to 53 more)	Very low	Prospective cohort	Serious[3,4]	No serious	No serious	Serious[1]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Any non-vertebral fracture 5 years or more (duration)											
1 (Hundrup 2004)	17/ 313	215/ 4019	HR 0.94 (0.54 to 1.64)	3 fewer per 1000. (From 24 fewer to 33 more)	Very low	Prospective cohort	Serious[3,4]	No serious	No serious	Very serious[2]	None

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)
3. Evidence was downgraded by 1 due to performance and attrition bias
4. Adjusted for age at menopause, BMI, family history

Table 79: GRADE profile: previous use of HRT versus never HRT use (subgroup analysis duration) for the outcome of osteoporotic fracture (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Osteoporotic fracture (< 2 years duration)											
1 (Engel 2011)	N/R	1981/ 18651	HR 1.04 (0.94 to 1.15)	4 more per 1000. (From 6 fewer to 15 more)	Moderate	Prospective cohort	Serious[1]	No serious	No serious	No serious	None
Osteoporotic fracture (2 to 4.9 years duration)											
1 (Engel 2011)	N/R	1981/ 18651	HR 0.99 (0.88 to 1.11)	1 fewer per 1000. (From 12 fewer to 1 more)	Moderate	Prospective cohort	Serious[1]	No Serious	No serious	No serious	None
Osteoporotic fracture(≥ 5 years duration)											
1 (Engel 2011)	N/R	1981/ 18651	HR 0.89 (0.80 to 0.99)	11 fewer per 1000. (From 1 fewer to 20 fewer)	Moderate	Prospective cohort	Serious[1]	No serious	No serious	No serious	None

N/R: not reported
1. Adjusted for BMI, physical activity, age at menopause, parity, previous use of oral contraceptive, previous use of calcium supplements, education

Table 80: GRADE profile: previous use of HRT versus never HRT use (subgroup analysis duration) for the outcome of hip fracture (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Fractures HRT	No HRT	Odds ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Hip fracture (≤ 5 years duration)											
1 (Yates 2004)	32/ 13592	149/ 53737	OR 1.00 (0.68 to 1.48)	0 fewer per 1000. (From 1 fewer to 1 more)	Very low	Prospective cohort	Serious[3]	No serious	No serious	Very serious[2]	None
Hip fracture(6 to 10 years duration)											
1 (Yates 2004)	11/ 2616	149/ 53737	OR 1.69 (0.91 to 3.12)	2 more per 1000. (From 0 fewer to 6 more)	Low	Prospective cohort	Serious[3]	No serious	No serious	Serious[1]	None
Hip fracture (> 10 years duration)											
1 (Yates 2004)	11/ 2608	149/ 53737	OR 1.24 (0.67 to 2.30)	1 more per 1000. (From 1 fewer to 4 more)	Very low	Prospective cohort	Serious[3]	No serious	No serious	Very serious[2]	None

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)
3. Adjusted for age, BMI, previous fracture, health status, maternal history of fractures, cortisone use

Table 81: GRADE profile: previous use of HRT versus never use for the outcome of vertebral fracture (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Odds ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Vertebral fracture											
1 (Bagger 2004)	18/ 155	26/ 108	OR 0.47 (0.24 to 0.93)	111 fewer per 1000. (From 13 fewer to 170 fewer)	Low	Prospective cohort	Serious[2]	No serious	No serious	Serious[1]	None

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Adjusted for age, baseline BMC, spine BMD

Table 82: GRADE profile: previous use of HRT versus never HRT use (subgroup analysis of time of discontinuation) for the outcome of vertebral fracture, wrist fracture (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Vertebral fracture (discontinued 0 to 1 years ago)											
1 (Paganini-Hill 2005)	85/ 1444	268/ 3863	HR 0.82 (CI not reported, but NS)	N/C	Low	Prospective cohort	Serious[1]	No serious	No serious	N/C	None
Wrist fracture (discontinued 0 to 1 years ago)											
1 (Paganini-Hill 2005)	58/ 1444	186/ 3863	HR 0.60 (CI not reported, p < 0.05)	N/C	Low	Prospective cohort	Serious[1]	No serious	No serious	N/C	None
Vertebral fracture (discontinued 2 to 14 years ago)											
1 (Paganini-Hill 2005)	134/ 1876	268/ 3863	HR 1.05 (CI not reported, but NS)	N/C	Low	Prospective cohort	Serious[1]	No serious	No serious	N/C	None
Wrist fracture (discontinued 2 to 14 years ago)											
1 (Paganini-Hill 2005)	117/ 1876	186/ 3863	HR 0.90 (CI not reported, but NS)	N/C	Low	Prospective cohort	Serious[1]	No serious	No serious	N/C	None
Wrist fracture (discontinued ≥ 5 years ago)											
1 (Randell 2002)	65/ 1212	145/ 3335	RR 1.44 (1.06 to 1.95)	N/C	Moderate	Prospective cohort	Serious[2]	No serious	No serious	No serious	None
Vertebral fracture (discontinued ≥ 15 years ago)											
1 (Paganini-Hill 2005)	106/ 1553	268/ 3863	HR 0.82 (CI not reported, but NS)	N/C	Low	Prospective cohort	Serious[1]	No serious	No serious	N/C	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Wrist fracture (discontinued ≥ 15 years ago)											
1 (Paganini-Hill 2005)	96/ 1553	186/ 3863	HR 1.30 (CI not reported, but NS)	N/C	Low	Prospective cohort	Serious[1]	No serious	No serious	N/C	None

N/C: not calculable; NS: not significant
1. Adjusted for age, history of fractures, BMI, blood pressure medication, non-prescription pain medication, smoking, exercise, attitude
2. Adjusted for age, time since menopause, BMI, number of chronic health diseases, history of previous fractures

Table 83: GRADE profile: previous use of HRT versus never HRT use (subgroup analysis for time of discontinuation) for the outcome of only any fracture, vertebral fracture (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Any fracture (discontinued < 1 year ago)											
1 (Banks 2004)	130/ 2904	3010/ 70297	RR 1.09 (0.91 to 1.30)	4 more per 1000. (From 4 fewer to 13 more)	Very low	Prospective cohort	Very serious[3, 4]	No serious	No serious	Serious[1]	None
Any fracture(discontinued 1 to 2 years ago)											
1 (Banks 2004)	250/ 6263	3010/ 70297	RR 0.96 (0.85 to 1.10)	2 fewer per 1000. (From 6 fewer to 4 more)	Low	Prospective cohort	Very serious[3, 4]	No serious	No serious	No serious	None
Any fracture (discontinued 3 to 4 years ago)											
1 (Banks 2004)	160/ 3525	3010/ 70297	RR 1.09 (0.93 to 1.28)	4 more per 1000. (From 3 fewer to 12 more)	Very low	Prospective cohort	Very serious[3, 4]	No serious	No serious	Serious[1]	None
Any fracture discontinued ≥ 5 years ago)											
1 (Banks 2004)	301/ 6247	3010/ 70297	RR 1.10 (0.97 to 1.23)	4 more per 1000. (From 1 fewer to 10 more)	Low	Prospective cohort	Very Serious[3 ,4]	No serious	No serious	No serious	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Any fracture(discontinued ≥ 5 years ago)											
1 (Randell 2002)	130/ 1212	352/ 3335	RR 1.02 (0.82 to 1.26)	2 more per 1000. (From 19 fewer to 27 more)	Low	Prospective cohort	Serious[5]	No serious	No serious	Serious[1]	None
Any fracture (previous use 2 to 4 years, discontinued ≥ 4 years ago)											
1 (Middleton and Steel, 2007)	6/ 60	54/ 340	RR 0.46 (0.14 to 1.57)	86 fewer per 1000. (From 137 fewer to 91 more)	Very low	Prospective cohort	Serious[6]	No serious	No serious	Very serious[2]	None

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)
3. Evidence was downgraded by 1 due to performance and attrition bias
4. Adjusted for age, region, socioeconomic status, time since menopause, BMI, physical exercise
5. Adjusted for age, time since menopause, BMI, number of chronic health diseases, history of previous fractures
6. Adjusted for baseline BMD

Table 84: GRADE profile: previous use of HRT versus never HRT use (subgroup analysis of time of discontinuation) for the outcome of only non -vertebral fracture (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Any non-vertebral fracture (discontinued < 5 years ago)											
1 (Hundrup 2004)	22/ 418	215/ 4019	HR 1.05 (0.63 to 1.73)	3 more per 1000. (From 19 fewer to 37 more)	Very low	Prospective cohort	Serious[3]	No serious	No serious	Very serious[2]	None
Any non-vertebral fracture (discontinued 5 to 10 years ago)											
1 (Hundrup 2004)	16/ 251	215/ 4019	HR 0.85 (0.45 to 1.61)	8 fewer per 1000. (From 29 fewer to 31 more)	Very low	Prospective cohort	Very serious[3, 4]	No serious	No serious	Very serious[2]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Any non-vertebral fracture (discontinued 10 or more years ago)											
1 (Hundrup 2004)	23/ 229	215/ 4019	HR 2.03 (1.25 to 3.29)	52 more per 1000. (From 13 more to 112 more)	Low	Prospective cohort	Very serious[3, 4]	No serious	No serious	No serious	None
Any non-vertebral fracture (Previous use less than 5 years discontinued < 5 years ago)											
1 (Hundrup 2004)	12/ 246	215/ 4019	HR 1.03 (0.52 to 2.04)	2 more per 1000. (From 25 fewer to 53 more)	Very low	Prospective cohort	Very serious[3, 4]	No serious	No serious	Very serious[2]	None
Any non-vertebral fracture (Previous use less than 5 years, discontinued > 5 years ago)											
1 (Hundrup 2004)	31/ 327	215/ 4019	HR 1.65 (1.07 to 2.53)	33 more per 1000. (From 4 more to 76 more)	Very low	Prospective cohort	Very serious[3, 4]	No serious	No serious	Serious[1]	None
Any non-vertebral fracture (Previous more than 5 years discontinued < 5 years ago)											
1 (Hundrup 2004)	10/ 166	215/ 4019	HR 1.11 (0.54 to 2.27)	6 more per 1000. (From 24 fewer to 64 more)	Very low	Prospective cohort	Very serious[3, 4]	No serious	No serious	Very serious[2]	None
Any non-vertebral fracture (Previous use more than 5 years discontinued > 5 years ago)											
1 (Hundrup 2004)	7/ 146	215/ 4019	HR 0.84 (0.36 to 1.92)	8 fewer per 1000. (From 34 fewer to 47 more)	Very low	Prospective cohort	Very serious[3, 4]	No serious	No serious	Very serious[2]	None

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)
3. Evidence was downgraded by 1 due to performance and attrition bias
4. Adjusted for age at menopause, BMI, family history

Table 85: GRADE profile: previous use of HRT versus never HRT use (subgroup analysis of time of discontinuation) for the outcome of osteoporotic fracture (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio or odds ratio (95% CI)	Absolute (95% CI)			HRT	No treatment	Relative (95% CI)	Absolute (95% CI)	Other considerations
Osteoporotic fracture (discontinued < 5 years ago)											
1 (Engel 2011)	N/R	1981/ 18651	HR 0.92 (0.83 to 1.01)	8 fewer per 1000. (From 17 fewer to 1 more)	Moderate	Prospective cohort	Serious[3]	No serious	No serious	No serious	None
Osteoporotic fracture (discontinued ≥ 5 years ago)											
1 (Engel 2011)	N/R	1981/ 18651	HR 1.05 (0.96 to 1.14)	5 more per 1000. (From 5 fewer to 14 more)	Moderate	Prospective cohort	Serious[3]	No serious	No serious	No serious	None
Osteoporotic fracture (Previous use < 2 years, discontinued < 5 years ago)											
1 (Engel 2011)	N/R	1981/ 18651	HR 0.95 (0.83 to 1.09)	5 fewer per 1000. (From 17 fewer to 9 more)	Moderate	Prospective cohort	Serious[3]	No serious	No serious	No serious	None
Osteoporotic fracture (Previous use < 2 years discontinued ≥ 5 years ago)											
1 (Engel 2011)	N/R	1981/ 18651	HR 1.14 (1.00 to 1.30)	14 more per 1000. (From 0 to 30 more)	Low	Prospective cohort	Serious[3]	No serious	No serious	Serious[1]	None
Osteoporotic fracture (Previous use 2 to 4.9 years discontinued < 5 years ago)											
1 (Engel 2011)	N/R	1981/ 18651	HR 0.93 (0.79 to 1.09)	7 fewer per 1000. (From 21 fewer to 9 more)	Moderate	Prospective cohort	Serious[3]	No serious	No serious	No serious	None
Osteoporotic fracture (Previous use ≤ 5 years discontinued ≤ 5 years ago)											
1 (Barrett-Connor 2003)	75/ 5981	974/ 53737	OR 0.90 (0.71 to 1.15)	2 fewer per 1000. (From 5 fewer to 3 more)	Low	Prospective cohort	Serious[4]	No serious	No serious	Serious[1]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio or odds ratio (95% CI)	Absolute (95% CI)			HRT	No treatment	Relative (95% CI)	Absolute (95% CI)	Other considerations
Osteoporotic fracture (Previous use 2 to 4.9 years discontinued ≥ 5 years ago)											
1 (Engel 2011)	N/R	1981/ 18651	HR 1.06 (0.91 to 1.24)	6 more per 1000. (From 9 fewer to 24 more)	Moderate	Prospective cohort	Serious[3]	No serious	No serious	No serious	None
Osteoporotic fracture (Previous use ≤ 5 years discontinued > 5 years ago)											
1 (Barrett-Connor 2003)	160/ 7643	974/ 53737	OR 1.09 (0.92 to 1.29)	2 more per 1000. (From 1 fewer to 5 more)	Low	Prospective cohort	Serious[4]	No serious	No serious	Serious[1]	None
Osteoporotic fracture (Previous use ≥ 5 years discontinued < 5 years ago)											
1 (Engel 2011)	N/R	1981/ 18651	HR 0.79 (0.66 to 0.95)	21 fewer per 1000. (From 5 fewer to 35 fewer)	Low	Prospective cohort	Serious[3]	No serious	No serious	Serious[1]	None
Osteoporotic fracture (Previous use 6 to 10 years discontinued ≤ 5 years ago)											
1 (Barrett-Connor 2003)	18/ 1297	974/ 53737	OR 0.98 (0.61 to 1.57)	0 fewer per 1000. (From 7 fewer to 10 more)	Very low	Prospective cohort	Serious[4]	No serious	No serious	Very serious[2]	None
Osteoporotic fracture (Previous use ≥ 5 years, discontinued > 5 years ago)											
1 (Engel 2011)	N/R	1981/ 18651	HR 0.95 (0.85 to 1.07)	5 fewer per 1000. (From 15 fewer to 7 more)	Moderate	Prospective cohort	Serious[3]	No serious	No serious	No serious	None
Osteoporotic fracture (Previous use 6 to 10 years discontinued > 5 years ago)											
1 (Barrett-Connor et al., 2003)	37/ 1332	974/ 53737	OR 1.39 (0.99 to 1.94)	7 more per 1000. (From 0 fewer to 16 more)	Low	Prospective cohort	Serious[4]	No serious	No serious	Serious[1]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio or odds ratio (95% CI)	Absolute (95% CI)			HRT	No treatment	Relative (95% CI)	Absolute (95% CI)	Other considerations
Osteoporotic fracture (Previous use > 10 years discontinued ≤ 5 years ago)											
1 (Barrett-Connor 2003)	34/ 1445	974/ 53737	OR 1.32 (0.93 to 1.87)	6 more per 1000. (From 1 fewer to 15 more)	Low	Prospective cohort	Serious[4]	No serious	No serious	Serious[1]	None
Osteoporotic fracture (Previous use > 10 years discontinued > 5 years ago)											
1 (Barrett-Connor 2003)	28/ 1176	974/ 53737	OR 1.06 (0.72 to 1.56)	1 more per 1000. (From 5 fewer to 10 more)	Very low	Prospective cohort	Serious[4]	No serious	No serious	Very serious[2]	None

N/R: not reported
1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)
3. Adjusted for BMI, physical activity, age at menopause, parity, previous use of oral contraceptive, previous use of calcium supplements, education
4. Adjusted for age, prior fracture, health status, maternal history, of fracture, cortisone use

Table 86: GRADE profile: previous HRT use versus never HRT use (subgroup analysis of time of discontinuation) for the outcome of hip fracture (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Odds ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Hip fracture (discontinued 0 to 1 years ago)											
1 (Paganini-Hill 1991)	28/ 1422	166/ 3708	RR 0.80 (0.53 to 1.21)	9 fewer per 1000. (From 21 fewer to 9 more)	Low	Prospective cohort	Serious[3]	No serious	No serious	Serious[1]	None
Hip fracture (discontinued ≤ 5 years ago)											
1 (Yates 2004)	23/ 8723	149/ 53737	OR 1.65 (1.05 to 2.59)	2 more per 1000. (From 0 more to 4 more)	Low	Prospective cohort	Serious[4]	No serious	No serious	Serious[1]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Odds ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Hip fracture (discontinued > 5 years ago)											
1 (Yates 2004)	31/ 10151	149/ 53737	OR 0.93 (0.63 to 1.38)	0 fewer per 1000. (From 1 fewer to 1 more)	Very low	Prospective cohort	Serious[4]	No serious	No serious	Very serious[2]	None
Hip fracture (discontinued 2 to 14 years ago)											
1 (Paganini-Hill 1991)	47/ 1836	166/ 3708	RR 0.88 (0.63 to 1.23)	5 fewer per 1000. (From 17 fewer to 10 more)	Low	Prospective cohort	Serious[3]	No serious	No serious	Serious[1]	None
Hip fracture (discontinued ≥ 15 years ago)											
1 (Paganini-Hill 1991)	78/ 1499	166/ 3708	RR 1.15 (0.88 to 1.50)	7 more per 1000. (From 5 fewer to 22 more)	Low	Prospective cohort	Serious[3]	No serious	No serious	Serious[1]	None
Hip fracture (Previous use of HRT for ≤ 3 years) discontinued 0 to 1 year ago)											
1 (Paganini-Hill 1991)	3/ 148	166/ 3708	RR 0.87 (0.28 to 2.73)	6 fewer per 1000. (From 32 fewer to 77 more)	Very low	Prospective cohort	Serious[3]	No serious	No serious	Very serious[2]	None
Hip fracture (Previous use of HRT for ≤ 3 years discontinued 2 to 14 years ago)											
1 (Paganini-Hill 1991)	8/ 378	166/ 3708	RR 0.79 (0.38 to 1.60)	9 fewer per 1000. (From 28 fewer to 27 more)	Very low	Prospective cohort	Serious[3]	No serious	No serious	Very serious[2]	None
Hip fracture (Previous use of HRT for ≤ 3 years discontinued ≥ 15 years ago)											
1 (Paganini-Hill 1991)	52/ 916	166/ 3708	RR 1.33 (0.97 to 1.82)	15 more per 1000. (From 1 fewer to 37 more)	Low	Prospective cohort	Serious[3]	No serious	No serious	Serious[1]	None
Hip fracture (Previous use of HRT for 4 to 14 years discontinued 0 to 1 year ago)											
1 (Paganini-Hill 1991)	22/ 481	166/ 3708	RR 0.72 (0.31 to 1.64)	13 fewer per 1000. (From 31 fewer to 29 more)	Very low	Prospective cohort	Serious[3]	No serious	No serious	Very serious[2]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Odds ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Hip fracture (Previous use of HRT for 4 to 14 years discontinued 2 to 14 years ago)											
1 (Paganini-Hill 1991)	18/ 846	166/ 3708	RR 0.86 (0.52 to 1.42)	13 fewer per 1000. (From 31 fewer to 29 more)	Very low	Prospective cohort	Serious[3]	No serious	No serious	Very serious[2]	None
Hip fracture (Previous use of HRT for ≥ 15 years discontinued 0 to 1 year ago)											
1 (Paganini-Hill 1991)	3/ 89	166/ 3708	RR 0.85 (0.53 to 1.38)	7 fewer per 1000. (From 21 fewer to 17 more)	Very low	Prospective cohort	Serious[3]	No serious	No serious	Very serious[2]	None
Hip fracture (Previous use of HRT for ≥ 15 years discontinued 2 to 14 years ago)											
1 (Paganini-Hill 1991)	21/ 605	166/ 3708	RR 0.97 (0.61 to 1.53)	1 fewer per 1000. (From 17 fewer to 24 more)	Very low	Prospective cohort	Serious[3]	No serious	No serious	Very serious[2]	None
Hip fracture (Previous use of HRT for ≥ 15 years discontinued ≥ 15 years ago)											
1 (Paganini-Hill 1991)	19/ 819	166/ 3708	RR 0.57 (0.18 to 1.79)	19 fewer per 1000. (From 37 fewer to 35 more)	Very low	Prospective cohort	Serious[3]	No serious	No serious	Very serious[2]	None

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)
3. Adjusted for age
4. Adjusted for age, BMI, previous fracture, health status, maternal history of fracture, cortisone use

Table 87: GRADE profile: current use of oestrogen plus progestogen versus no current use of HRT for the outcome of any fracture, non- vertebral fracture, hip fracture (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Any non-vertebral fracture											
1 (Hundrup 2004)	27/ 1214	215/ 4019	HR 0.48 (0.32 to 0.74)	27 fewer per 1000. (From 14 fewer to 36 fewer)	Low	Prospective cohort	Serious[1,2]	No serious	No serious	No serious	None

1. Evidence was downgraded by 1 due to performance and attrition bias
2. Adjusted for age at menopause, BMI, family history

Table 88: GRADE profile: current use of oestrogen plus progestogen versus no current HRT use (subgroup analysis of HRT initiation years since menopause) for the outcome of hip fracture - combined analysis of WHI trial (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Hip fracture (within 2 years (< 2 years) since menopause)											
1 (Prentice 2009)	N/R	N/R	HR 0.35 (0.1 to 1.17)	N/C	Very low	Prospective cohort	Serious[1,3]	No serious	no serious indirectness	Serious[2]	None
Hip fracture (within 2-4 years since menopause)											
1 (Prentice 2009)	N/R	N/R	HR 0.33 (0.1 to 1.1)	N/C	Very low	Prospective cohort	Serious[1,3]	No serious	no serious indirectness	Serious[2]	None

N/C: not calculable; N/R: not reported
1. Evidence was downgraded by 1 due to selection and attrition bias
2. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
3. Adjusted for age, BMI, education, smoking, alcohol, physical activity, family history of fracture, personal history of fracture, duration of prior HRT use, number of falls, calcium intake, waist to hip ratio, height, history of treated diabetes, NSAID use, history of hypertension, history of high cholesterol requiring medication, history of breast cancer, personal history of non-melanoma skin cancer, prior oral contraceptive use

Table 89: GRADE profile: current use of oestrogen plus progestogen for 5.2 years duration versus no current HRT use (subgroup analysis time of discontinuation) for the outcome of hip fracture (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Hip fracture (Previous use 5.2 years discontinued 2.4 years ago)											
1 (Heiss 2008)	107/ 8506	132/ 8102	HR 0.78 (0.60 to 1.00)	4 fewer per 1000. (From 6 fewer to 0 more)	Low	Prospective cohort	Serious[2]	No serious	No serious	Serious[1]	None
Hip fracture (Previous use 5.2 years discontinued 8.2 years ago)											
1 (Manson 2013)	232/ 8506	270/ 8102	HR 0.81 (0.68 to 0.97)	6 fewer per 1000. (From 1 fewer to 11 fewer)	Low	Prospective cohort	Serious[3]	No serious	No serious	Serious[1]	None
Hip fracture (Previous use 5.2 years discontinued 8.2 years ago, age 50 to 59)											
1 (Manson 2013)	17/ 2837	28/ 2683	HR 0.57 (0.31 to 1.04)	4 fewer per 1000. (From 7 fewer to 0 more)	Low	Prospective cohort	Serious[3]	No serious	No serious	Serious[1]	None
Hip fracture (Previous use 5.2 years discontinued 8.2 years ago, age 60 to 69)											
1 (Manson 2013)	103/ 3854	100/ 3655	HR 0.94 (0.71 to 1.24)	2 fewer per 1000. From 8 fewer to 6 more.	Low	Prospective cohort	Serious[3]	No serious	No serious	Serious[1]	None

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Adjusted for age, ethnicity, education, body mass index, smoking, self-reported general health, night sweats, hot flashes, breast tenderness, and treatment assignment, and at year 1, breast tenderness, night sweats, and hot flashes
3. Adjusted for age, BMI, education, smoking, alcohol, physical activity, family history of fracture, personal history of fracture, duration of prior HRT use, number of falls, calcium intake, waist to hip ratio, height, history of treated diabetes, NSAID use, history of hypertension, history of high cholesterol requiring medication, history of breast cancer, personal history of non-melanoma skin cancer, prior oral contraceptive use

Table 90: GRADE profile: current use of oestrogen plus progestogen versus no HRT use (subgroup analysis of HRT duration) for the outcomes of any fracture, vertebral fracture (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Any fracture (Previous use 5.2 years discontinued 2.4 years ago)											
1 (Heiss 2008)	1078/ 8506	1249/ 8102	HR 0.80 (0.73 to 0.86)	29 fewer per 1000. From 20 fewer to 39 fewer.	Low	Prospective cohort	Serious[2]	No serious	No serious	Serious[1]	None
Vertebral fracture (Previous use5.2 years discontinued 2.4 years ago)											
1 (Heiss 2008)	102/ 8506	125/ 8102	HR 0.78 (0.60 to 1.01)	3 fewer per 1000. From 6 fewer to 0 more.	Low	Prospective cohort	Serious[2]	No serious	No serious	Serious[1]	None

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Adjusted for age, ethnicity, education, body mass index, smoking, self-reported general health, night sweats, hot flashes, breast tenderness, and treatment assignment, and at year 1, breast tenderness, night sweats, and hot flashes

Table 91: GRADE profile: current use of oestrogen plus progestogen versus no HRT use (subgroup analysis of previous use years from current HRT episode) for the outcome of hip fracture-combined analysis of WHI (comparative cohort studies

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Hip fracture (within 2 years (< 2 years) since menopause)											
1 (Prentice 2009)	N/R	N/R	HR 0.94 (0.19 to 4.58)	N/C	Very low	Prospective cohort	Serious[1,3]	No serious	No serious indirectness	Very serious[2]	None
Hip fracture (within 2-4 years since menopause)											
1 (Prentice 2009)	N/R	N/R	HR 0.26 (0.05 to 1.25)	N/C	Very low	Prospective cohort	Serious[1,3]	No serious	No serious indirectness	Very serious[2]	None

N/C: not calculable; N/R: not reported
1. Evidence was downgraded by 1 due to selection and attrition bias
2. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)
3. Adjusted for age, BMI, education, smoking, alcohol, physical activity, family history of fracture, personal history of fracture, duration of prior HRT use, number of falls, calcium intake, waist to hip ratio, height, history of treated diabetes, NSAID use, history of hypertension, history of high cholesterol requiring medication, history of breast cancer, personal history of non-melanoma skin cancer, prior oral contraceptive use

Table 92: GRADE profile: current use of oestrogen versus no current use of HRT for the outcomes of any fracture, non- vertebral fracture (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Any non-vertebral fracture											
1 (Hundrup 2004)	23/ 722	215/ 4019	HR 0.53 (0.30 to 0.96)	25 fewer per 1000. From 2 fewer to 37 fewer.	Very low	Prospective cohort	Very serious[2,3]	No serious	No serious	Serious[1]	None

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Evidence was downgraded by 1 due to performance and attrition bias
3. Adjusted for age at menopause, BMI, family history

Table 93: GRADE profile: current use of oestrogen versus no current use of HRT (subgroup analysis of Initiation, years from menopause) for the outcome of hip fracture- combined analysis of WHI trial (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No treatment	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Hip fracture (within 2 years (< 2 years) since menopause)											
1 (Prentice 2009)	N/R	N/R	HR 0.46 (0.04 to 4.88)	N/C	Very low	Prospective cohort	Serious[1,3]	No serious	No serious indirectness	Very serious[2]	None
Hip fracture (within 2-4 years since menopause)											
1 (Prentice 2009)	N/R	N/R	HR 0.53 (0.11 to 2.51)	N/C	Very low	Prospective cohort	Serious[1,3]	No serious	No serious indirectness	Very serious[2]	None

N/C: not calculable; N/R: not reported
1. Evidence was downgraded by 1 due to selection and attrition bias
2. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)
3. Adjusted for age, BMI, education, smoking, alcohol, physical activity, family history of fracture, personal history of fracture, duration of prior HRT use, number of falls, calcium intake, waist to hip ratio, height, history of treated diabetes, NSAID use, history of hypertension, history of high cholesterol requiring medication, history of breast cancer, personal history of non-melanoma skin cancer, prior oral contraceptive use

Table 94: GRADE profile: current use of oestrogen versus no current HRT use (subgroup analysis initiation, years since menopause) for the outcome of hip fracture- combined analysis of WHI (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	HRT	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Hip fracture (within 2 years (< 2 years) since menopause)											
1 (Prentice 2009)	N/R	N/R	HR 0.60 (0.11 to 3.24)	N/C	Very low	Prospective cohort	Serious[1,4]	No serious	No serious indirectness	Very serious[2]	None
Hip fracture (within 2-4 years since menopause)											
1 (Prentice 2009)	N/R	N/R	HR 0.13 (0.02 to 1.08)	N/C	Very low	Prospective cohort	Serious[1,4]	No serious	No serious indirectness	Serious[3]	None

N/C: not calculable
1. Evidence was downgraded by 1 due to selection and attrition bias
2. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)
3. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
4. Adjusted for age, BMI, education, smoking, alcohol, physical activity, family history of fracture, personal history of fracture, duration of prior HRT use, number of falls, calcium intake, waist to hip ratio, height, history of treated diabetes, NSAID use, history of hypertension, history of high cholesterol requiring medication, history of breast cancer, personal history of non-melanoma skin cancer, prior oral contraceptive use

Table 95: GRADE profile: current use of HRT versus no current HRT use (subgroup analysis time of discontinuation 7.2 years) for the outcome of hip fracture) (comparative cohort studies)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Hip fracture (discontinued 6.6 years ago)											
1 (Manson 2013)	134/ 5310	148/ 5429	HR 0.91 (0.72 to 1.15)	2 fewer per 1000. (From 8 fewer to 4 more)	Low	Prospective cohort	Serious[3]	No serious	No serious	Serious[1]	None
Hip fracture (discontinued 3.9 years ago)											
1 (LaCroix 2011)	114/ 5310	127/ 5429	HR 0.92 (0.71 to 1.18)	3 fewer per 1000. (From 9 fewer to 6 more)	Low	Prospective cohort	Serious[4]	No serious	No serious	Serious[1]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Hip fracture (discontinued 3.9 years ago age 60 to 69)											
1 (LaCroix 2011)	38/ 1740	45/ 1799	HR 0.87 (0.57 to 1.35)	3 fewer per 1000. (From 11 fewer to 9 more)	Very low	Prospective cohort	Serious[4]	No serious	No serious	Very serious[2]	None
Hip fracture (discontinued 6.6 years ago age 60 to 69)											
1 (Manson 2013)	46/ 2386	49/ 2465	HR 0.95 (0.64 to 1.43)	1 fewer per 1000. (From 7 fewer to 8 more)	Very low	Prospective cohort	Serious[3]	No serious	No serious	Very serious[2]	None

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)
3. Adjusted for age, BMI, education, smoking, alcohol, physical activity, family history of fracture, personal history of fracture, duration of prior HRT use, number of falls, calcium intake, waist to hip ratio, height, history of treated diabetes, NSAID use, history of hypertension, history of high cholesterol requiring medication, history of breast cancer, personal history of non-melanoma skin cancer, prior oral contraceptive use
4. Stratified by age, prior disease (if appropriate), and randomization status in the WHI Dietary Modification Trial

I.5.7 Dementia

Table 96: GRADE profile: HRT versus no HRT for the outcome of cerebral metabolism change

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Continued conjugated equine oestrogen	Discontinued 17β oestradiol	Relative risk (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Dementia (cerebral metabolism change) (2 year follow-up)											
1 (Rasgon 2014	28/28	17/17	RR 1.00 (0.91 to 1.10)	0 fewer per 1000 (from 90 fewer to 100 more)	Low	Randomised trials	Serious[4]	No serious	Serious[3]	No serious	None
Dementia (medial cortical area decline) (2 year follow-up)											
1 (Rasgon 2014)	16/28	13/17	RR 0.75 (0.49 to 1.13)	191 fewer per 1000 (from 390 fewer to 99 more)	Very low	Randomised trials	Serious[4]	No serious	Serious[3]	Serious[1]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Continued conjugated equine oestrogen	Discontinued 17β oestradiol	Relative risk (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Dementia (posterior cingulate decline (2 year follow-up)											
1 (Rasgon 2014)	7/28	6/17	RR 0.71 (0.29 to 1.76)	102 fewer per 1000 (from 251 fewer to 268 more)	Very low	Randomised trials	Serious[4]	No serious	Serious[3]	Very serious[2]	None

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)
3. Majority of evidence had only one indirect PICO (outcome)
4. Evidence was downgraded by 1 due to selection and performance bias

Table 97: GRADE profile: HRT versus no HRT for the outcome of dementia

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Dementia											
1 (Shao 2012)	87/1105	89/663	HR 0.80 (0.58 to 1.09)	25 fewer per 1000 (from 54 fewer to 11 more)	Very low	Prospective cohort	Serious[7]	No serious	Serious[5]	Serious[1]	None
Dementia (9 years follow-up)											
1 (Whitmer 2011)	1384/5504	2454/5504	HR 0.74 (0.58 to 0.94)	92 fewer per 1000 (from 20 fewer to 156 fewer)	Very low	Retrospective cohort	Very serious[4,9]	No serious	No serious	Serious[1]	None
Cognitive decline (TICs score) 6 ≥ 5 points (2 year follow-up)											
1 (Kang 2004)	196/3814	169/3615	RR 1.10 (0.88 to 1.38)	5 more per 1000 (from 6 fewer to 18 more)	Very low	Retrospective cohort	Serious[3,8]	No serious	Serious[3]	Serious[1]	None
Cognitive decline											
1 (Fillenbaum 2001)	Not reported	Not reported	OR 1.17 (0.76 to 1.79)	NC	Low	Prospective cohort	Serious[10]	No serious	No serious	Serious[1]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Cognitive decline (by ≥5 points (TICs)6											
1 (Kang 2004)	249/4611	202/4258	RR 1.07 (0.87 to 1.30)	3 more per 1000 (from 6 fewer to 14 more)	Very low	Retrospective Cohort study	Serious[4,8]	No serious	Serious[3]	Serious[1]	None
Cognitive impairment (5 years follow-up)											
1 (Mitchell 2003)	1420/1462	1420/1462	OR 1.0 (0.6 to 1.8)	0 fewer per 1000 (from 18 fewer to 13 more)	Very low	Prospective cohort	Serious[11]	No serious	No serious	Very serious[2]	None
Cognitive impairment (5 years follow-up)											
1 (Mitchell 2003)	1303/1462	1303/1462	OR 0.7 (0.3 to 1.8)	40 fewer per 1000 (from 180 fewer to 45 more)	Very low	Prospective cohort	Serious[11]	No serious	No serious	Very serious[2]	None
Cognitive decline											
1 (Fillenbaum 2001)	Not reported	Not reported	OR 0.94 (0.42 to 2.15)	NC	Low	Prospective cohort	Serious[10]	No serious	No serious	Serious[1]	None
Cognitive decline (intermittent use of HRT)											
1 (Fillenbaum 2001)	Not reported	Not reported	OR 1.16 (0.76 to 1.75)	NC	Low	Prospective cohort	Serious[10]	No serious	No serious	Serious[1]	None
Cognitive decline (continuous use of HRT)											
1 (Fillenbaum 2001)	Not reported	Not reported	OR 0.68 (0.23 to 1.99)	N/C	Very low	Prospective cohort	Very serious[10]	No serious	No serious	Very serious[2]	None
Dementia (age <80.4 years for "mid-life")											
1 (Whitmer 2011)	121/579	253/1167	RR 0.96 (0.80 to 1.17))	9 fewer per 1000 (from 43 fewer to 37 more)	Very low	Retrospective cohort study	Very serious[4,9]	No serious	No serious	No serious	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Dementia (age <80.4 years for "late-life")											
1 (Whitmer 2011	99/132	253/1167	RR 0.79 (0.64 to 0.97)	46 fewer per 1000 (from 7 fewer to 78 fewer)	Very low	Retrospective cohort	Very serious[4,9]	No serious	No serious	Serious[1]	None
Cognitive decline (≥5 years use) (5 years follow-up)											
1 (Mitchell 2003)	1402/1462	1402/1462	OR 0.7 (0.4 to 1.4)	17 fewer per 1000 (from 56 fewer to 11 more)	Very low	Prospective cohort	Serious[11]	No serious	No serious	Very serious[2]	None
Dementia (10 years or more versus <10 years) (surgical menopause)											
1 (Bove 2014)	592/607	Not reported	HR 0.917 (0.7 to 1.1)	N/C	Very low	Retrospective Cohort study	Serious[12]	No serious	No serious	Very serious[2]	None
Dementia (initiation within 5 years of menopause) (7 years follow-up)											
1 (Shao 2012)	52/727	89/663	HR 0.70 (0.49 to 0.99)	38 fewer per 1000 (from 1 more to 865 fewer)	Very low	Retrospective cohort	Serious[7]	No serious	Serious[3]	Serious[1]	None
Dementia (initiation within 10 years of menopause) (5 years follow-up)											
1 (Petitti 2008)	91/957	95/977	HR 0.95 (0.71 to 1.28)	5 fewer per 1000 (from 27 fewer to 25 more)	Low	Prospective cohort	Serious[13]	No serious	No serious	Serious[1]	None

N/C: not calculable
1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)
3. Evidence was downgraded by 1 due to selection bias
4. Evidence was downgraded by 2 due to selection, performance, attrition and detection bias
5. Majority of evidence had only 1 indirect PICO (population)
6. TICs: Telephone interview for cognitive status- validated scale for (0-50) detecting cognitive impairment, with any score greater than 27 points indicating severe impairment
7. Adjusted for education, alcohol use, smoking, body mass index, history of hypertension, high cholesterol, diabetes, stroke, heart attack, coronary artery bypass graft surgery, physical activity, regular social activity, dietary scores reflecting adherence to Mediterranean or Dietary Approaches to Stop Hypertension diets
8. Adjusted for factors (age, education, diabetes, blood pressure, vitamin E supplements, body mass index, smoking, physical activity, socioeconomic status, antidepressant use, alcohol intake, aspirin use, other NSAID use, baseline cognitive score, mental health index, energy fatigue index)
9. Adjusted for age, education, race, mid-life body mass index, diabetes, hypertension, hyperlipidaemia, stroke, hysterectomy status

10. Adjusted for age, education, race, marital status, number of natural children, body mass index, smoking, alcohol consumption, medications that may influence cognitive impairment (thyroid, benzodiazepine, NSAIDs), stroke, diabetes, hip fracture, other broken bones, arthritis, heart attack, hypertension, incontinence, self-rated health, health status (as measured by Rosow-Breslau physical health scale)
11. Adjusted for age, body mass index, education, exercise, marital status, employment status, income, self-reported health status, smoking, alcohol use
12. Adjusted for age at enrolment, education, smoking, and study (ROS vs MAP)
13. Adjusted for age, education, myocardial infarction, stroke, Parkinson's disease, diabetes mellitus, and hypertension

Table 98: GRADE profile: oestrogen or progestogen use versus no HRT use for the outcome of dementia

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No treatment	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Dementia (by prescription) (5 years follow-up)											
1 (Petitti 2008)	15/340	80/879	HR 1.64 (0.94 to 2.88)2	54 more per 1000 (from 5 fewer to 149 more)	Low	Prospective cohort	Serious[2]	No serious	No serious	Serious[1]	None

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Adjusted for age, education, myocardial infarction, stroke, Parkinson's disease, diabetes mellitus, and hypertension

Table 99: GRADE profile: oestrogen use versus no HRT use (including subgroup analysis of timing and duration) for the outcome of dementia

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio, odds ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Dementia											
1 (Tang 1996)	156/303	968/1778	RR 0.95 (0.84 to 1.06)	27 fewer per 1000 (from 87 fewer to 33 more)	Low	Retrospective Cohort study	Serious[3]	No serious	Serious[5]	No serious	
Dementia risk (by prescription and self-report) (5 years follow-up)											
1 (Petitti 2008)	80/879	99/1011	HR 1.07 (0.79 to 1.44)	6 more per 1000 (from 20 fewer to 20 more)	Low	Prospective cohort	Serious[11]	No serious	No serious	Serious[1]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio, odds ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Cognitive decline (by <2 points (MMSE)8) (4 year follow-up)											
1 (Ryan, 2008)	N/R	N/R	OR 1.08 (0.66 to 1.76)	N/C	Very low	Prospective cohort	Serious[12]	No serious	No serious	Very serious[2]	None
Cognitive decline (by ≥5 points (TICs)7)											
1 (Kang 2004)	181/3580	202/4258	RR 1.06 (0.85 to 1.32)	3 more per 1000 (from 7 fewer to 15 more)	Very low	Retrospective cohort	Serious[3, 10]	No serious	Serious[5]	Serious[1]	None
Cognitive decline (by <2 points (MMSE) (4 year follow-up)											
1 (Ryan, 2008)	N/R	N/R	OR 0.93 (0.61 to 1.43)	N/C	Very low	Prospective cohort	Serious[12]	No serious	No serious	Very serious[2]	None
Dementia (>0.5 years versus 0 years duration)											
1 (Kawas 1997)	N/R	N/R	RR 0.443 (0.13 to 1.51)	N/C	Very low	Prospective cohort	Serious[6]	No serious	Serious[5]	Serious[1]	None
Cognitive decline (by <2 points (MMSE)8), (0-9 years duration)											
1 (Ryan, 2008)	N/R	N/R	OR 0.75 (0.28 to 2.02)	N/C	Very low	Prospective cohort	Serious[12]	No serious	No serious	Very serious[2]	None
Cognitive decline (by <2 points (MMSE)8) (≥ 10 years duration)											
1 (Ryan, 2008)	N/R	N/R	OR 1.20 (0.70 to 2.06)	N/C	Very low	Prospective cohort	Serious[12]	No serious	No serious	Very serious[2]	None
Dementia (5-10 years versus 0 years duration)											
1 (Kawas 1997)	N/R	N/R	RR 0.338 (0.05 to 2.5)	N/C	Very low	Prospective cohort	Serious[6]	No serious	Serious[5]	Very serious[2]	None
Dementia (>10 years versus 0 years duration)											
1 (Kawas 1997)	N/R	N/R	RR 0.5 (0.5 to 0.170)	N/C	Very low	Prospective cohort	Serious[6]	No serious	Serious[5]	No serious	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio, odds ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Cognitive decline (by ≥5 points (TICs7) (20+ years duration)											
1 (Kang 2004)	55/1134	202/4258	RR 0.95 (0.69 to 1.32)	2 fewer per 1000 (from 15 fewer to 15 more)	Very low	Retrospective cohort	Serious[3]	No serious	Serious[5]	Serious[1]	None
Cognitive decline (by <2 points (MMSE)8) (0-9 years past duration)											
1 (Ryan 2008)	N/R	N/R	OR 0.70 (0.40 to 1.22)	N/C	Low	Prospective cohort	Serious[12]	No serious	No serious	Serious[1]	None
Cognitive decline (by <2 points (MMSE)8) (≥10 years past duration)											
1 (Ryan 2008)	N/R	N/R	OR 1.37 (0.77 to 2.45)	N/C	Low	Prospective cohort	Serious[12]	No serious	No serious	Serious[1]	None
Cognitive decline (by ≥10% decrease (MMSE)8) (early initiation)											
1 (Khoo 2010)	68/158	0/213	HR 0.28 (0.08 to 0.97)	N/C	Moderate	Prospective cohort	No serious	No serious	No serious	Serious[1]	None
Cognitive decline (≥ 5 points (TICs)7) (recent initiation) (2 years follow-up)											
1 (Kang 2004)	22/282	169/3615	RR 1.74 (1.08 to 2.81)	35 more per 1000 (from 4 more to 85 more)	Very low	Retrospective cohort	Serious[3]	No serious	Serious[5]	Serious[1]	None
Cognitive decline (by ≥10% decrease (MMSE)8), (late initiation)											
1 (Khoo 2012)	14/39	213	HR 1.28 (0.31 to 5.25)	N/C	Very low	Prospective cohort	Serious[9]	No serious	No serious	Very serious[2]	None

N/C: not calculable; N/R: not reported
1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossing 1 default MID (0.75 to 1.25)
2. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossing 2 default MIDs (0.75 to 1.25)
3. Evidence was downgraded by 1 due to selection bias
4. Evidence was downgraded by 2 due to selection, performance, attrition and detection bias
5. Majority of evidence had only 1 indirect PICO (population)
6. Evidence was downgraded by 1 due to selection or detection bias
7. TICs: Telephone interview for cognitive status- validated scale for (0-50) detecting cognitive impairment, with any score greater than 27 points indicating severe impairment.
8. MMSE: Mini Mental State Examination- questionnaire that measures cognitive status(0-30), with any score greater than or equal to 27/30 points indicating normal cognition. Scores below 9 indicate severe cognitive impairment, or mild (19-24 points).

9. Logistic regression model controlling for confounding factors (age, body mass index, physical activity, smoking, and alcohol intake)
10. Adjusted for confounding factors (age, education, diabetes, blood pressure, vitamin E supplements, body mass index, smoking, physical activity, socioeconomic status, antidepressant use, alcohol intake, aspirin use, other NSAID use, baseline cognitive score, mental health index, energy fatigue index)
11. Adjusted for age, education, myocardial infarction, stroke, Parkinson's disease, diabetes mellitus, and hypertension
12. Adjusted for age, education, and baseline cognitive test score

Table 100: GRADE profile: progestogen use versus no HRT use for the outcome of dementia

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Dementia (by prescription) (5 years follow-up)											
1 (Petitti 2008)	38/493	80/879	HR 0.80 (0.54 to 1.19)	18 fewer per 1000 (from 41 fewer to 16 more)	Low	Prospective cohort	Serious[2]	No serious	No serious	Serious[1]	None

1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossing 1 default MID (0.75 to 1.25)
2. Adjusted for age, education, myocardial infarction, stroke, Parkinson's disease, diabetes mellitus, and hypertension

Table 101: GRADE profile: oestrogen plus progestogen versus no HRT use (including subgroup analysis for timing and duration) for the outcome of dementia

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Dementia (by prescription and self-report) (5 years follow-up)											
1 (Petitti 2008)	48/410	80/879	HR 1.32 (0.91 to 1.91)	27 more per 1000 (from 8 fewer to 76 more)	Low	Prospective cohort	Serious[10]	No serious	No serious	Serious[1]	None
Cognitive decline (by ≥5 points (TICs)[6])											
1 (Kang 2004)	82/1358	202/4258	RR 1.27 (0.97 to 1.68)	13 more per 1000 (from 1 fewer to 32 more)	Low	Retrospective cohort	Serious[9]	No serious	No serious	Serious[1]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Hazard ratio or risk ratio (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Cognitive decline (by ≥5 points (TICs)[6]) (10+ years duration)											
1 (Kang 2004)	48/732	202/4258	RR 1.36 (0.97 to 1.92)	17 more per 1000 (from 1 fewer to 44 more)	Very low	Retrospective cohort	Serious[3,9]	No serious	Serious[5]	Serious[1]	None
Cognitive decline ((MMSE)[7] by ≥10%) (early initiation)											
1 (Khoo 2010)	90/158	213	HR 0.85 (0.38 to 1.88)	N/C	Very low	Prospective cohort	Serious[4,8]	No serious	No serious	Very Serious[2]	None
Cognitive decline ((MMSE)[7] by ≥10%) (late initiation)											
1 (Khoo 2012)	25/39	213	HR 1.43 (0.53 to 3.89)	N/C	Very low	Prospective cohort	Serious[4,8]	No serious	No serious	Very Serious[2]	None

N/C: not calculable
1. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID (0.75 to 1.25)
2. Evidence was downgraded by 2 due to very serious imprecision as 95% confidence interval crossed 2 default MIDs (0.75 to 1.25)
3. Evidence was downgraded by 1 due to selection bias
4. Evidence was downgraded by 2 due to selection, performance, attrition and detection bias
5. Majority of evidence had only 1 indirect PICO (population)
6. TICs: Telephone interview for cognitive status- validated scale for (0-50) detecting cognitive impairment, with any score greater than 27 points indicating severe impairment.
7. MMSE: Mini Mental State Examination- questionnaire that measures cognitive status(0-30), with any score greater than or equal to 27/30 points indicating normal cognition. Scores below 9 indicate severe cognitive impairment, or mild (19-24 points).
8. Adjusted for confounding factors (age, body mass index, physical activity, smoking, and alcohol intake)
9. Adjusted for confounding factors (age, education, diabetes, blood pressure, vitamin E supplements, body mass index, smoking, physical activity, socioeconomic status, antidepressant use, alcohol intake, aspirin use, other NSAID use, baseline cognitive score, mental health index, energy fatigue index)
10. Adjusted for age, education, myocardial infarction, stroke, Parkinson's disease, diabetes mellitus, and hypertension

I.5.8 Loss of muscle mass (sarcopenia)

Table 102: GRADE profile: HRT versus no HRT use for the outcomes of change in muscular strength and change in muscle mass

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Change in muscular strength											
Knee extension torque (isometric)											
2 (Sipila, 2001 and Taaffe, 2005	40	40	-	MD 11.40 higher (1.79 to 21.01 higher)	Low	Randomised trials	Serious[1]	No serious	No serious	Serious[2]	None
Knee extension strength (isokinetic)											
1 (Ribom, 2002)	20	20	-	MD.95 higher (3.87 lower to 13.77 higher)	Low	Randomised trials	Serious[3]	N/A	No serious	Serious[2]	None
Knee flexion strength (isokinetic)											
1 (Ribom, 2002)	20	20	-	MD 2.80 higher (4.02 lower to 9.62 higher)	Low	Randomised trials	Serious[1]	N/A	No serious	Serious[2]	None
Handgrip strength											
2 (Armstrong, 1996; Ribom, 2002)	77	79	-	MD 0.01 higher (0.92 lower to 0.94 higher)	Low	Randomised trials	Serious[1]	No serious	No serious	Serious[2]	None
Adductor pollicis muscle strength											
1 (Skelton, 1999)	50	52	-	Mean percentage difference 15.4 higher (12.9 higher to 17.9 higher)	Low	Randomised trials	Very serious[3]	N/A	No serious	No serious	None
Change in muscle mass											
Quadriceps muscle CSA											
2 (Sipila, 2001; Taaffe, 2005)	40	40	-	MD (95%CI): 2.35 higher (0.28 higher to 4.42 higher)	Low	Randomised trials	Serious[1]	No serious	No serious	Serious[2]	None
Quadriceps muscle LCSA											
1 (Sipila, 2001)	20	20	-	MD (95%CI): 2.40 higher (0.48 lower to 5.28 higher)	Low	Randomised trials	Serious[1]	N/A	No serious	Serious[2]	None
Lower leg muscle CSA											
1 (Sipila, 2001)	20	20	-	MD (95%CI): 1.60 higher (1.54 lower to 4.74 higher)	Low	Randomised trials	Serious[1]	N/A	No serious	Serious[2]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	No HRT	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Lower leg muscle LCSA											
1 (Sipila, 2001)	20	20	-	MD (95%CI): 1.50 higher (1.51 lower to 4.51 higher)	Low	Randomised trials	Serious[1]	N/A	No serious	Serious[2]	None
Appendicular skeletal mass											
1 (Kenny, 2005)	83	84	-	MD (95%CI): 0.20 higher (0.16 higher to 0.24 higher)	Moderate	Randomised trials	Serious[1]	N/A	No serious	No serious	None
Posterior muscle CSA											
1 (Taaffe, 2005)	20	20	-	MD (95%CI): 2.00 higher (0.32 lower to 4.32 higher)	Low	Randomised trials	Serious[1]	N/A	No serious	Serious[2]	None

1. Unclear allocation concealment and randomization method in one trial
2. Evidence was downgraded by 1 due to serious imprecision as 95% CI crossed 1 default MID (-/+0.5 times SD)
3. Unblinded trial with no information on randomization and allocation concealment

Table 103: GRADE profile: HRT versus placebo for the outcome of change in muscle strength (total muscle strength)

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	HRT	Placebo	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Change in muscular strength											
Total muscle strength (composite)											
1 (Maddalozzo, 2004)	67	59	-	MD (95%CI): 0.52 lower (3.91 lower to 2.87 higher)	Very low	Prospective cohort	Very serious[1,2,3]	N/A	No serious	Serious[4]	None

1. High risk of selection bias
2. High risk of performance bias
3. High risk of detection bias
4. Evidence was downgraded by 1 due to serious imprecision as 95% CI crossed one default MID (-/+0.5 times SD)

I.6 Premature ovarian insufficiency

I.6.1 Diagnosis of premature ovarian insufficiency

Table 104: GRADE profile: diagnosis of premature ovarian insufficiency by the outcomes of AMH levels, inhibin B, oestradiol, FSH, antral follicle count, combination of FSH and AMH, combination of antral follicle count and inhibin B, and combination of antral follicle count and AMH

Number of studies	Number of participants	Measure of diagnostic accuracy				Quality	Quality assessment					
		Sensitivity, %	Specificity, %	Positive likelihood ratio	Negative likelihood ratio		Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
AMH level for the diagnosis of POI in high risk women. Cut-off <2 pmol/litre												
1 (Giuseppe 2007)	29	73 (35 to 91)	77 (58 to 92)	3.17 (1.30 to 7.72)	0.35 (0.11 to 1.12)	Low	Prospective case series	Serious[1]	No serious	No serious	Serious[2]	None
AMH level for the diagnosis of POI in high risk women. Cut-off < 8pmol/litre												
2 (Hagen 2010, Jadoul 2011)	98	97 (90 to 100)	62 (41 to 80)	2.99 (0.34 to 26.39)	0.05 (0.01 to 0.17)	Very low	Retrospective/prospective case series	Serious[1]	Very serious[3]	No serious	Very serious[4]	None
Inhibin B level for the diagnosis of POI in high risk women. Cut-off < 60 pg/mL												
1 (Giuseppe 2007)	29	59 (24 to 84)	77 (58 to 92)	2.47 (0.92 to 6.65)	0.56 (0.24 to 1.28)	Low	Prospective case series	Serious[1]	No serious	No serious	Serious[2]	None
Oestradiol level for the diagnosis of POI in high risk women. Cut-off < 50 pg/mL												
1 (Jadoul 2011)	31	52 (30 to 74)	33 (10 to 65)	0.79 (0.44 to 1.39)	1.43 (0.57 to 3.58)	Moderate	Prospective case series	Serious[1]	No serious	No serious	No serious	None
FSH level for the diagnosis of POI in high risk women. Cut-off ≥ 10 mIU/mL												
1 (Giuseppe 2007)	29	55 (24 to 84)	85 (64 to 95)	3.66 (1.11 to 12.12)	0.53 (0.24 to 1.16)	Very low	Prospective case series	Serious[1]	No serious	No serious	Very serious[4]	None
FSH level for the diagnosis of POI in high risk women. Cut-off > 30 mIU/mL												
1 (Jadoul 2011)	31	38 (18 to 62)	100 (74 to 100)	N/C	0.62 (0.44 to 0.87)	Moderate	Prospective case series	Serious[1]	No serious	No serious	Serious[5]	None
FSH level for the diagnosis of POI in high risk women. Cut-off > 30 mIU/mL (taken prior to hormonal treatment)												
1 (Jadoul 2011)	30	100 (84 to 100)	100 (69 to 100)	N/C	0.00 (N/C)	Very low	Prospective case series	Very serious[6]	No serious	No serious	Serious[5]	None
Antral follicle count for the diagnosis of POI in high risk women												
1 (Giuseppe 2007)	29	83 (47 to 97)	74 (53 to 89)	3.13 (1.44 to 6.86)	0.23 (0.05 to 1.09)	Low	Prospective case series	Serious[1]	No serious	No serious	Serious[2]	None

Number of studies	Number of participants	Measure of diagnostic accuracy				Quality	Quality assessment					
		Sensitivity, %	Specificity, %	Positive likelihood ratio	Negative likelihood ratio		Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Combination of FSH level and AMH level for the diagnosis of POI in high risk women												
1 (Giuseppe 2007)	29	55 (24 to 84)	89 (70 to 97)	4.91 (1.26 to 19.09)	0.51 (0.23 to 1.11)	Very low	Prospective case series	Serious[1]	No serious	No serious	Very serious[4]	None
Combination of antral follicle count and inhibin B level for the diagnosis of POI in high risk women												
1 (Giuseppe 2007)	29	83 (47 to 97)	87 (70 to 97)	6.38 (2.02 to 20.16)	0.20 (0.04 to 0.91)	Very low	Prospective case series	Serious[1]	No serious	No serious	Very serious[4]	None
Combination of antral follicle count and AMH level for the diagnosis of POI in high risk women												
1 (Giuseppe 2007)	29	83 (47 to 97)	88 (70 to 97)	7.03 (2.10 to 23.60)	0.19 (0.04 to 0.90)	Very low	Prospective case series	Serious[1]	No serious	No serious	Very serious[4]	None

N/C: not calculable
1. Selection bias as no clear methods are described in the recruitment of sample.
2. Evidence was downgraded by 1 due to 95% confidence interval for positive likelihood ratio ranges from not useful (<5) to moderately useful (5 to 10).
3. Evidence was downgraded by 2 due to very serious heterogeneity (chi-squared p<0.1, I-squared inconsistency statistic of >75%).
4. Evidence was downgraded by 2 due to 95% confidence interval for positive likelihood ratio ranges from not useful (<5) to very useful (>10).
5 .Confidence interval for positive likelihood ratio not calculable.
6. FSH level used as part of diagnostic criteria for POI.

I.6.2 Management of premature ovarian insufficiency

Table 105: GRADE profile: hormone replacement therapy versus combined oral contraceptives for management of premature ovarian insufficiency

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Intervention (HRT)	Comparator (OCP)	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Cardio/metabolic markers											
24 hour mean systolic blood pressure (mmHg, at 12 months)											
1 (Langrish 2009)	17	17	-	MD 7.3 lower (2.5 lower to 12.0 lower)	Low	Randomised controlled cross-over trial	Very serious[1]	No serious	No serious	No serious	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Intervention (HRT)	Comparator (OCP)	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
24 hour mean diastolic blood pressure (mmHg, at 12 months)											
1 (Langrish 2009)	17	17	-	MD 7.4 lower (3.9 lower to 11.0 lower)	Low	Randomised controlled cross-over trial	Very serious[1]	No serious	No serious	No serious	None
Triglyceride level (mmol/L at 6 months)											
1 (Guttmann 2001)	25	25	-	MD 0.10 lower (0.50 lower to 0.30 higher)	Very low	Randomised controlled cross-over trial	Very serious[1]	No serious	No serious	Serious[2]	None
HDL cholesterol level (mmol/litre at 6 months)											
1 (Guttmann 2001)	25	25	-	MD 0.03 higher (0.38 lower to 0.44 higher)	Very low	Randomised controlled cross-over trial	Very serious[1]	No serious	No serious	Serious[2]	None
LDL cholesterol level (mmol/litre at 6 months)											
1 (Guttmann 2001)	25	25	-	MD 0.55 lower (1.12 lower to 0.02 higher)	Very low	Randomised controlled cross-over trial	Very serious[1]	No serious	No serious	Serious[2]	None
Discontinuation rate											
1 (Langrish 2009)	10/29 (34.5%)	5/24 (20.8%)	RR 1.66 (0.65 - 4.18)	137 more per 1000 (from 73 fewer to 662 more)	Very low	Randomised controlled cross-over trial	Very serious[1]	No serious	No serious	Serious[2]	None
Discontinuation due to adverse effects											
1 (Langrish 2009)	8/29 (27.6%)	1/24 (4.2%)	RR 6.62 (0.89 - 49.28)	234 more per 1000 (from 5 fewer to 1000 more)	Very low	Randomised controlled cross-over trial	Very serious[1]	No serious	No serious	Serious[2]	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Intervention (HRT)	Comparator (OCP)	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Bone density											
Lumbar spine BMD (z-score)											
1 (Langrish 2009)b	18	18	-	MD (95% CI): 0.09 higher (0.06 lower to 0.25 higher)	Very low	Randomised controlled cross-over trial	Very serious[1]	No serious	No serious	Serious[2]	None
ALP (Absolute value in U/litre at 6 months)											
1 (Guttmann 2001)	17	17	-	MD 35 higher (11.13 higher to 58.87 higher)	Low	Randomised controlled cross-over trial	Very serious[1]	No serious	No serious	No serious	None
25 OH Vitamin D (nmol/litre at 6 months)											
1 (Guttmann 2001)	17	17	-	MD 9.98 lower (31.93 lower to 11.93 higher)	Very low	Randomised controlled cross-over trial	Very serious[1]	No serious	No serious	Serious[2]	None
1, 25 (OH)2 Vitamin D3 (pmol/litre at 6 months)											
1 (Guttmann et al. 2001)	17	17	-	MD 7.21 lower (28.29 lower to 13.87 higher)	Very low	Randomised controlled cross-over trial	Very serious[1]	No serious	No serious	Serious[2]	None
Osteocalcin (µg/litre at 6 months)											
1 (Guttmann 2001)	17	17	-	MD 4.50 higher (1.81 higher to 7.19 higher)	Low	Randomised controlled cross-over trial	Very serious[1]	No serious	No serious	No serious	None

Number of studies	Number of participants		Effect		Quality	Design	Quality assessment				
	Intervention (HRT)	Comparator (OCP)	Relative (95% CI)	Absolute (95% CI)			Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations
Urinary deoxypyridinoline cross-links (DPD) (nmol/mmol Cr at 6 months)											
1 (Guttmann 2001)	17	17	-	MD 1.40 higher (1.96 lower to 4.76 higher)	Very low	Randomised controlled cross-over trial	Very serious[1]	No serious	No serious	Serious[2]	None

1. Evidence was downgraded by 2 due to selection and lack of blinding bias
2. Evidence was downgraded by 1 due to serious imprecision as 95% confidence interval crossed 1 default MID

I.7 Abbreviations used in GRADE tables

Abreviation	Definition
AFC	Antral follicle count
ALP	Alkaline phosphate
AMH	Anti-Müllerian
ARD	Absolute risk difference
AUC	Area under the curve
BKMI	Blatt-Kupperman Menopausal Index
BMD	Bone mineral density
BMI	Body mass index
BNF	British National Formulary
BP	Blood pressure
CBT	Cognitive behavioural therapy
CEE	Conjugated equine estrogens
CEO	Combined equine oestrogens
CHD	Coronary heart disease
CI	Confidence interval
CNS	Central nervous system
CrI	Credible interval
CVD	Cardiovascular disease
CVA	Stroke or Cerebral Vascular Accident
DEXA	Dual energy X-ray absorptiometry
DIC	Deviance information criteria
DVT	Deep vein thrombosis
ECG	Electrocardiogram
EPT	Oestrogen and progestogen therapy
ESCIT	Escitalopram
FRAX	Fracture risk assessment tool
FSH	Follicle-stimulating hormone
GRADE	Grading of Recommendations Assessment, Development and Evaluation
HbA1c	Glycated haemoglobin
HCP	Healthcare professional
HDL	High density lipoprotein
HRQoL	Health related quality of life
HRT	Hormone replacement therapy
HT	Hormone therapy
HTA	Health technology assessment
ICER	Incremental cost-effectiveness ratio
IFG	Impaired fasting glycaemia
IHD	Ischaemic heart disease
LDL	Low density lipoprotein
LETR	Linking evidence to recommendations
LH	Luteinizing Hormone
LMP	Last menstrual period

Abreviation	Definition
LNG-IUS	Levonorgestrel-releasing intra-uterine system
MDD	Major depressive disorder
MHRA	Medicines and healthcare product regulatory authority
MHT	Menopausal hormone therapy
MI	Myocardial infarction
MID	Minimally important difference
MPA	Medroxyprogesterone acetate
MR	Means ratio
NCC-WCH	National Collaborating Centre for Women's and Children's Health
NETA	Norethisterone acetate
NHANES	National Health and Nutrition Examination Survey
NICE	National Institute for Health and Care Excellence
NIHR	National Institute for Health Research
NMA	Network meta-analysis
NPV	Negative predictive value
N/A	Not applicable
N/C	Not calculable
N/R	Not reported
OCP	Oral contraceptive pill
ONS	Office of National Statistics
OR	Odds ratio
PE	Pulmonary embolism
PET	Positron emission tomography
PICO	Population, intervention, comparison, outcome
POI	Premature ovarian insufficiency
PPV	Positive predictive value
QALY	Quality adjusted life year
QUADAS	Quality assessment tool for diagnostic accuracy studies
RCOG	Royal College of Obstetricians and Gynaecologists
RCT	Randomised control trial
ReSTAGE	Staging of reproductive aging
ROM	Ratio of means
RR	Risk ratio/relative risk
SD	Standard deviation
SE	Standard error
SMD	Standardised mean difference
SNRI	Norepinephrine reuptake inhibitors
SSRI	Selective serotonin reuptake inhibitors
STRAW	The Stages of Reproductive Aging Workshop
SWAN	Study of Women Across the Nation
T2DM	Type 2 diabetes mellitus
TIA	Transient ischemic attack
TS	Turner Syndrome
UA	Urinary atrophy
USD	US dollars

Abreviation	Definition
UTI	Urinary tract infections
VMS	Vasomotor symptoms
VTE	Venous thromboembolism
VVA	Vulvovaginal atrophy
WHI	Women's Health Initiative
WHO	World Health Organization

Appendix J: Forest plots

J.1 Diagnosis of perimenopause and menopause

Figure 1: Diagnosis of menopause from perimenopausal women

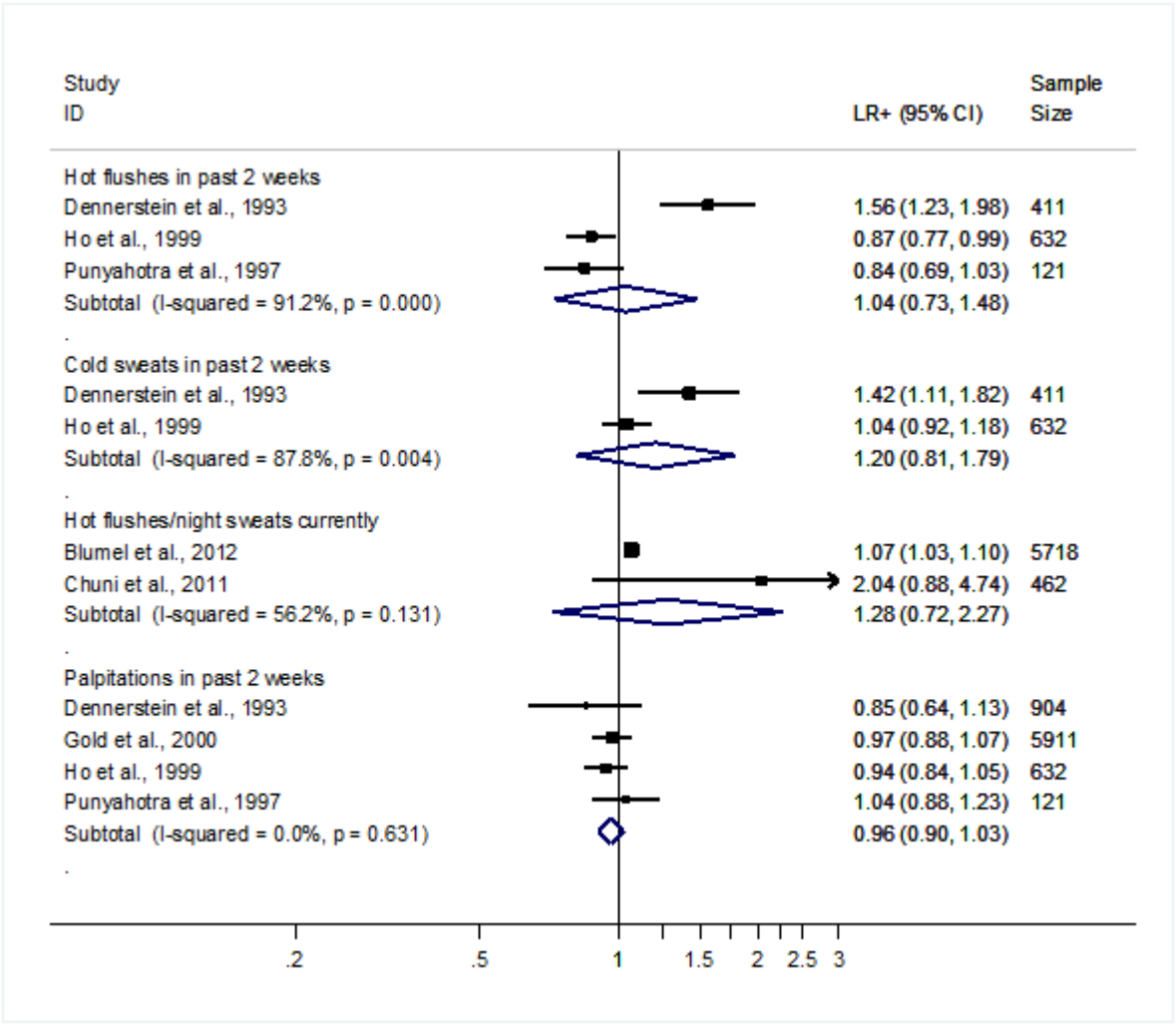


Figure 2: Diagnosis of menopause from premenopausal women

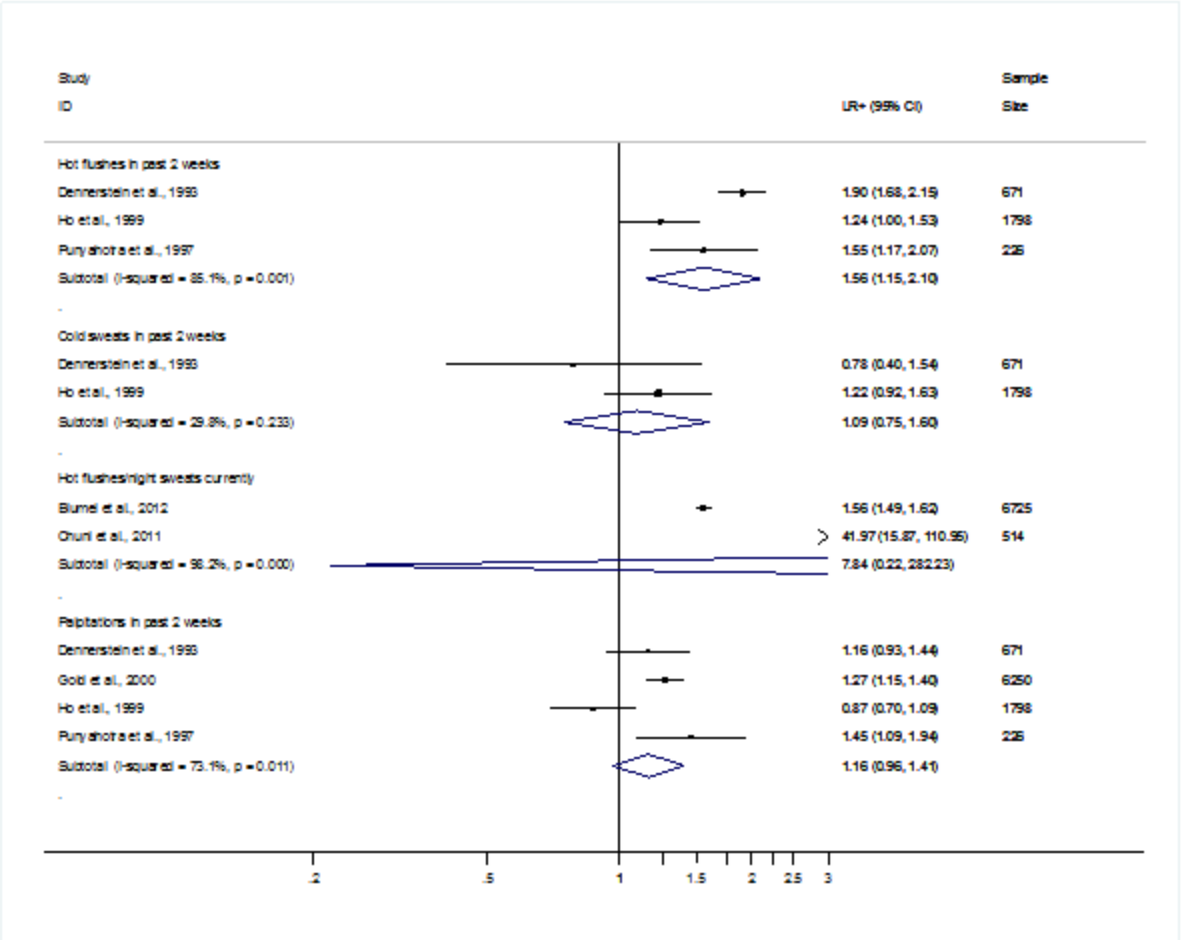


Figure 3: Diagnosis of postmenopause from all other women

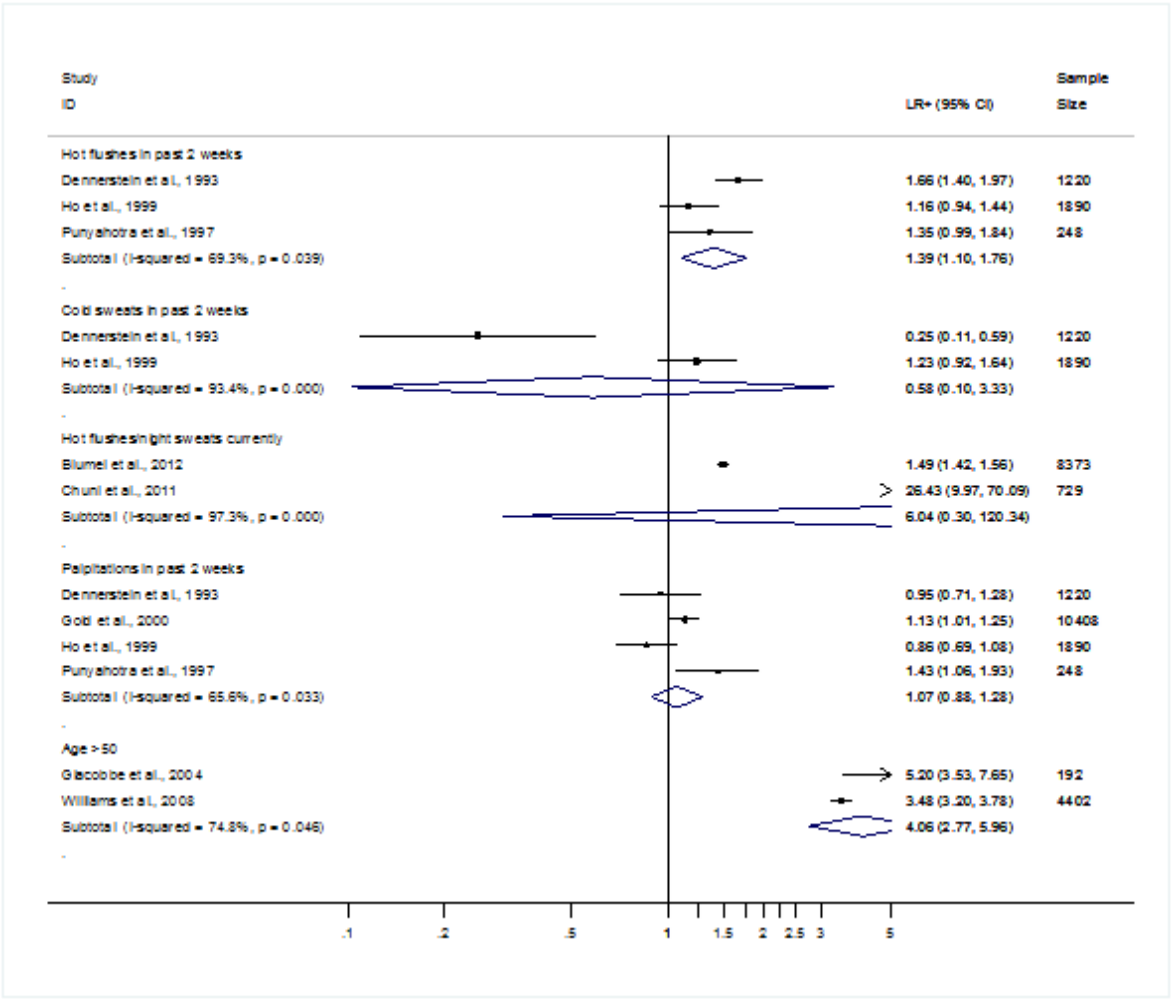


Figure 4: Diagnosis of perimenopause from postmenopausal women

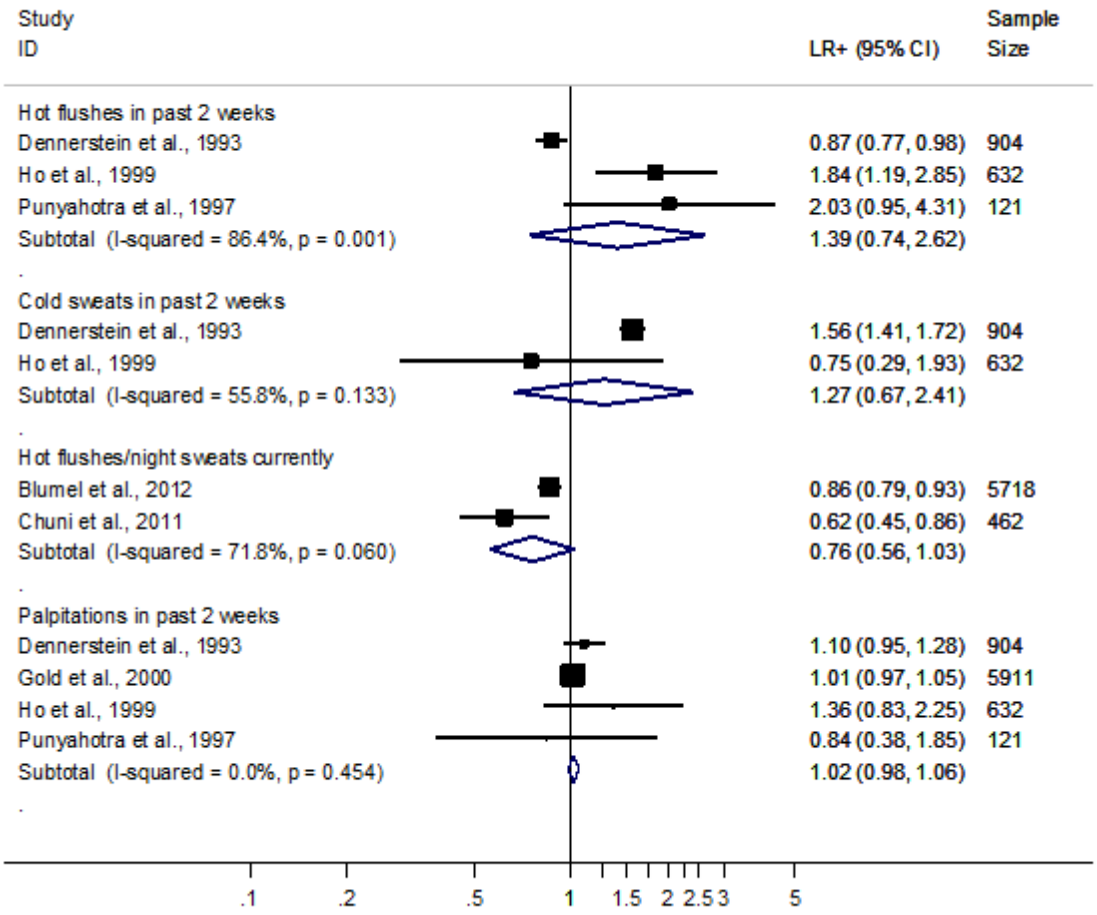


Figure 5: Diagnosis of perimenopause from premenopausal women

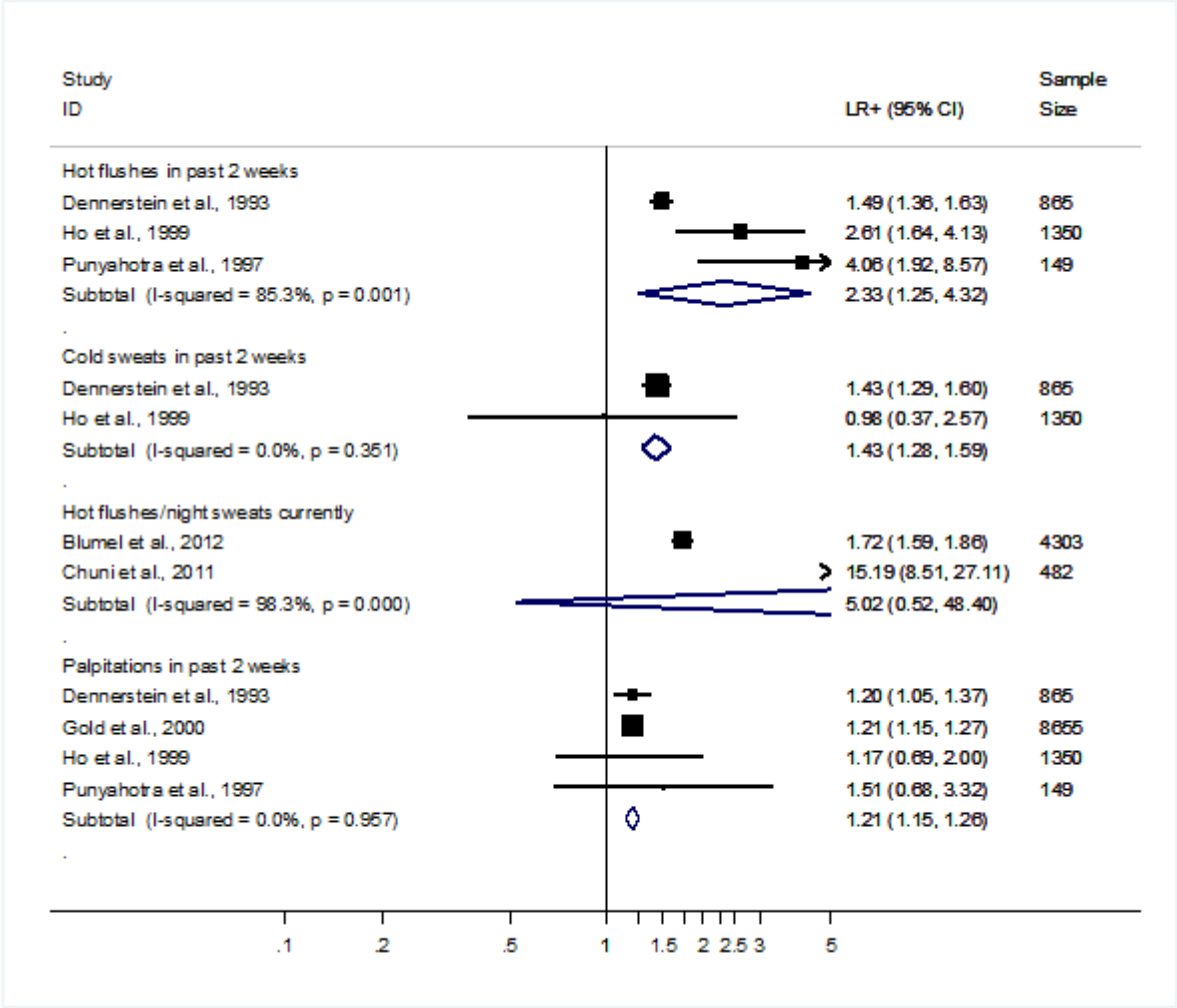
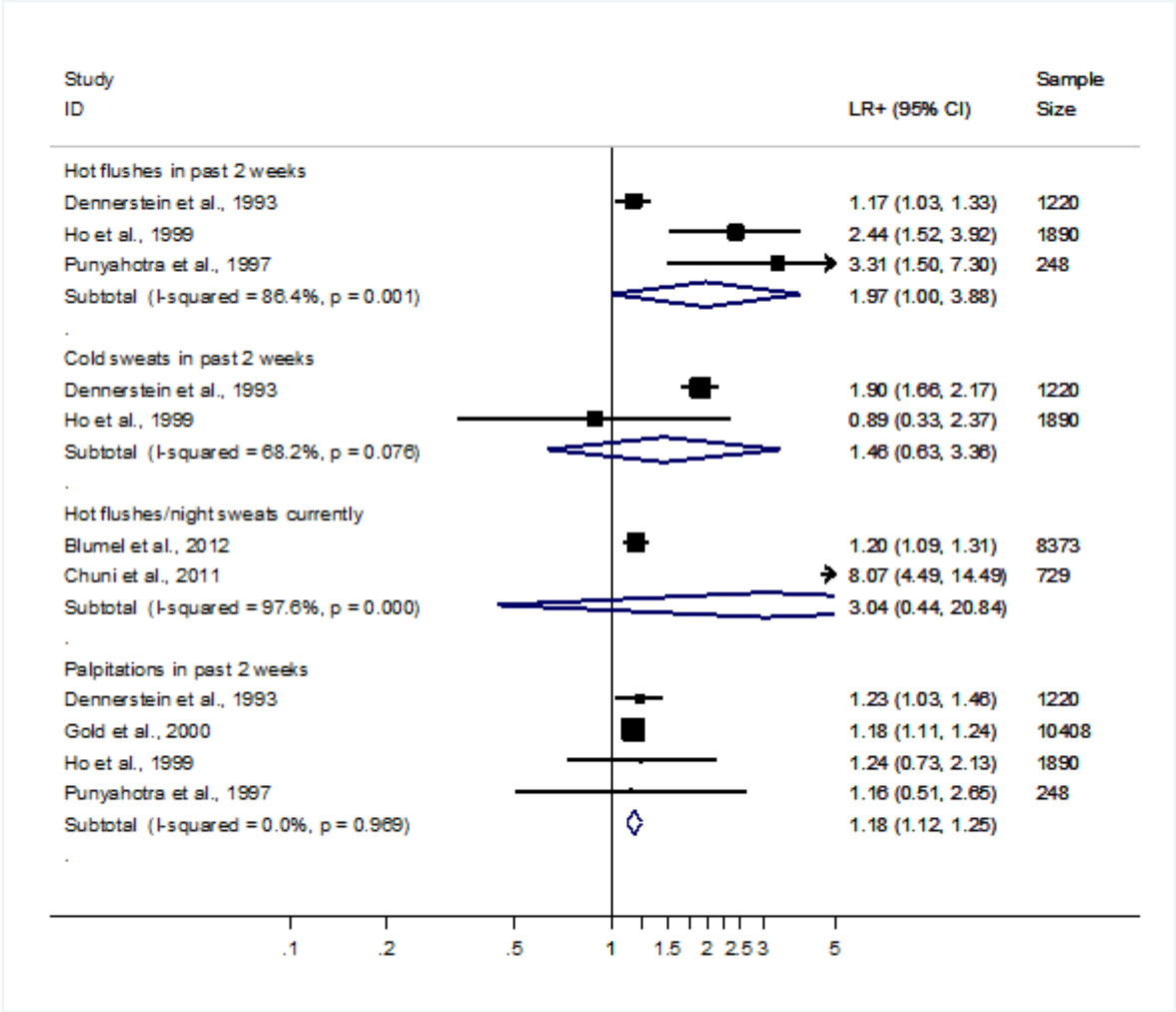


Figure 6: Diagnosis of perimenopause from all other women



J.2 Classification systems for the diagnosis of menopause

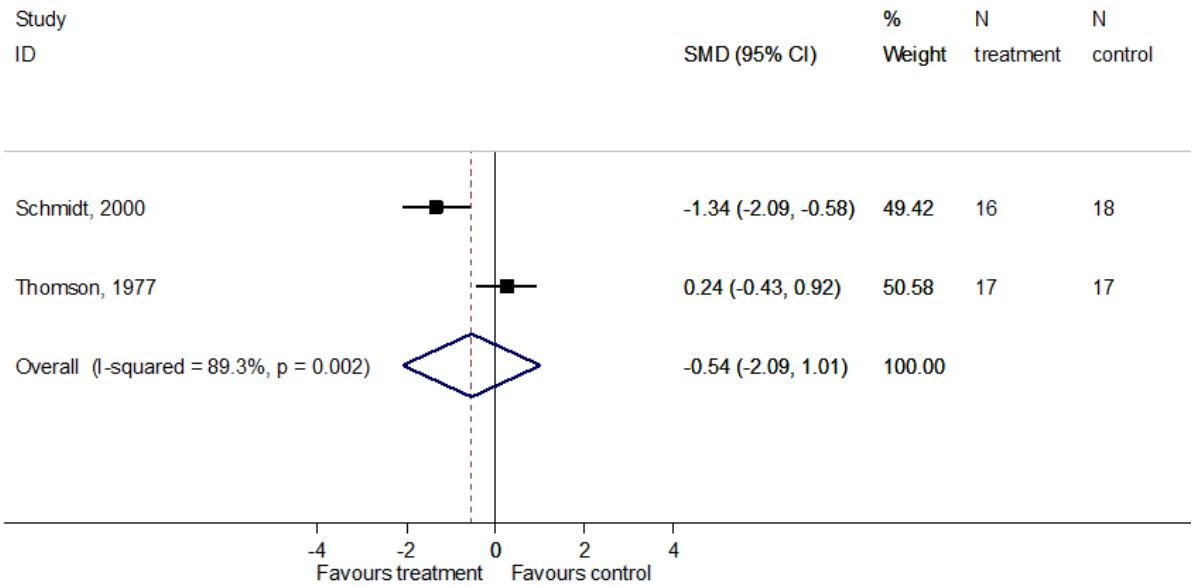
There were no forest plots for this review.

J.3 Information and advice

There were no forest plots for this review.

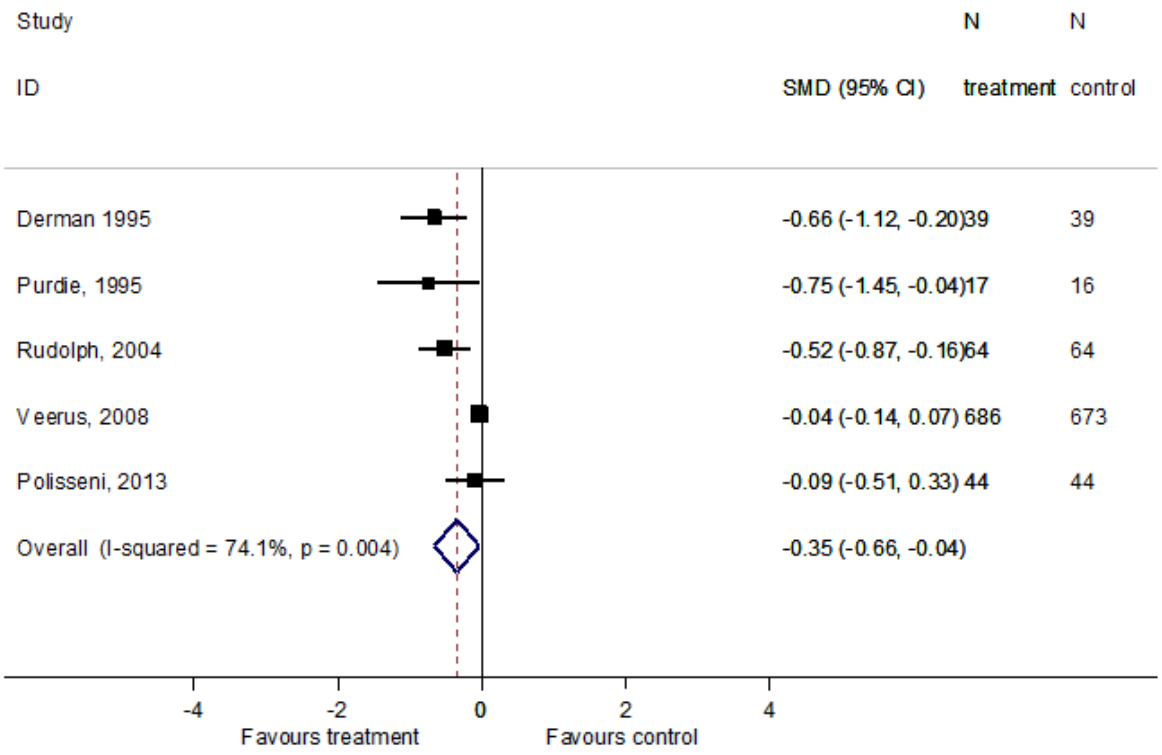
J.4 Managing short-term symptoms

Figure 7: Oestrogen versus no treatment/placebo for low mood measured by various scales



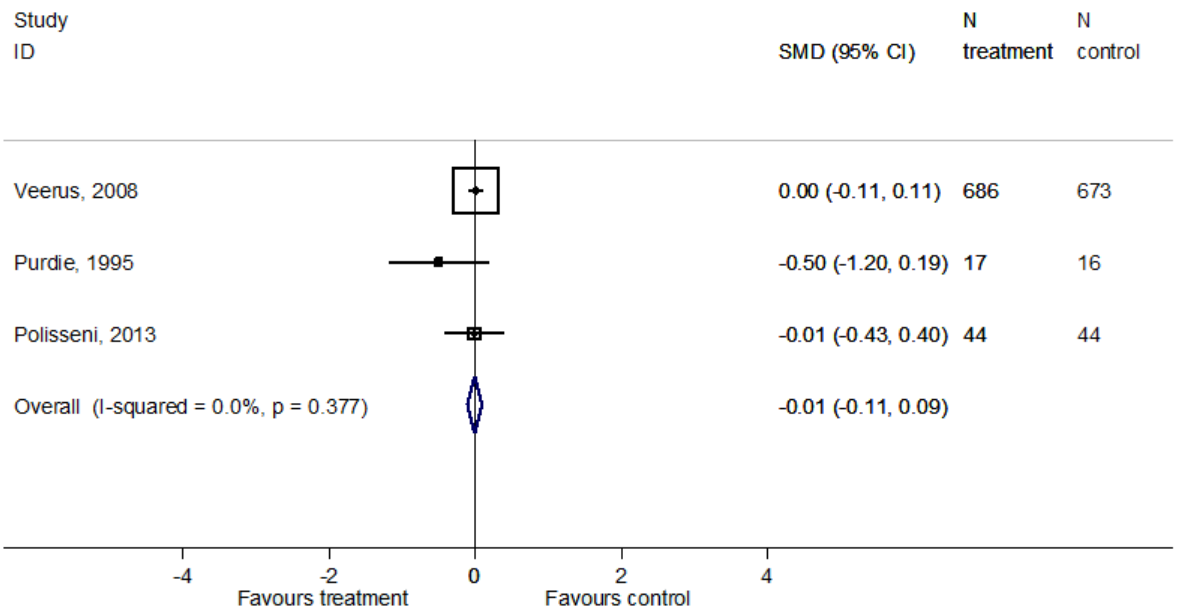
Study-level estimates pooled using standardised mean differences

Figure 8: Oestrogen versus no treatment/placebo for low mood measured by various scales



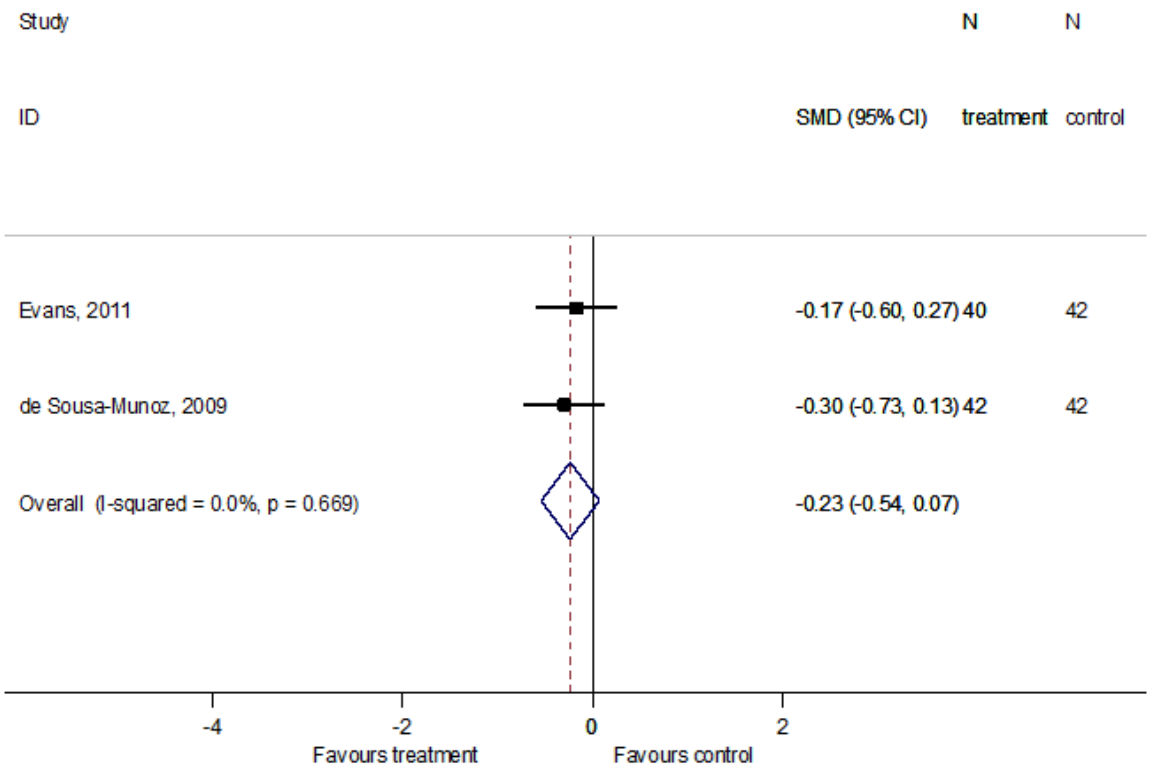
Study-level estimates pooled using standardised mean differences

Figure 9: Oestrogen versus no treatment/placebo for anxiety measured by various scales



Study-level estimates pooled using standardised mean differences

Figure 10: Phytoestrogen versus no treatment/placebo for low mood measured by various scales



Study-level estimates pooled using standardised mean differences

J.4.1 Urogenital atrophy

Figure 11: Percentage change in Parabasal cells after treatment of Ospemifene (60mg) for less than one year compared to placebo

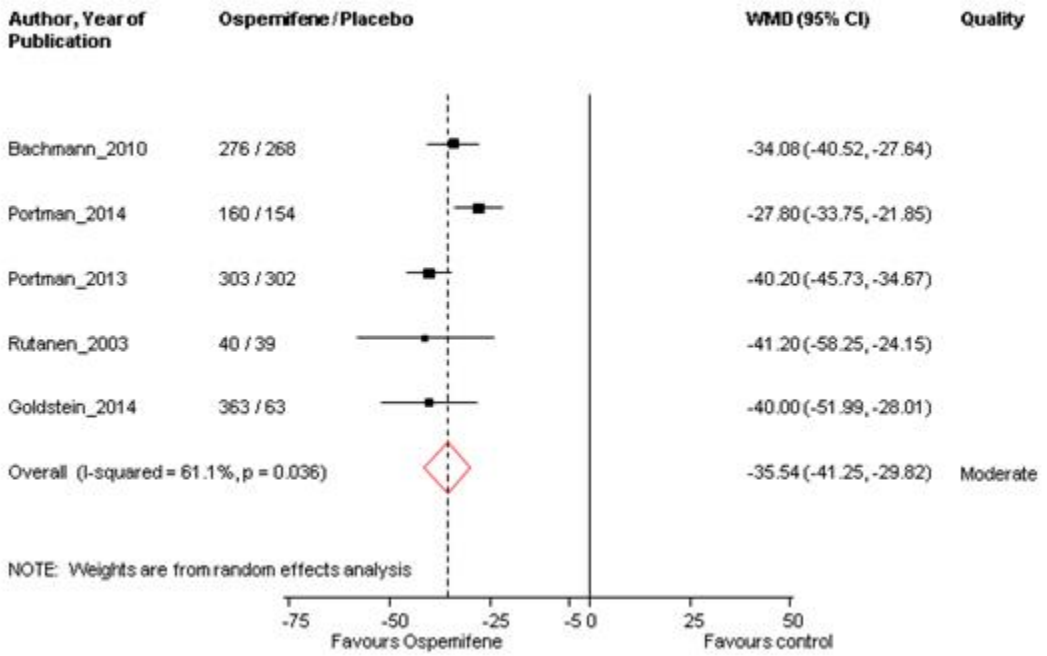


Figure 12: Percentage change in Superficial cells after treatment of Ospemifene (60mg) for less than one year compared to placebo

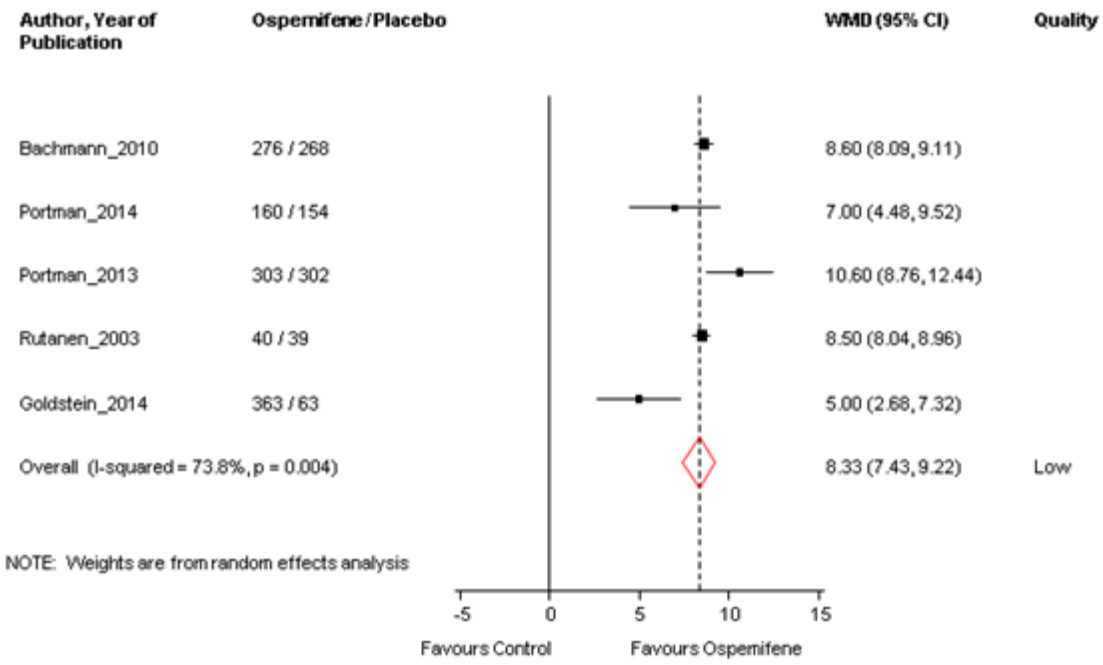


Figure 13: Change in dyspareunia severity score after treatment of Ospemifene for less than one year compared to placebo

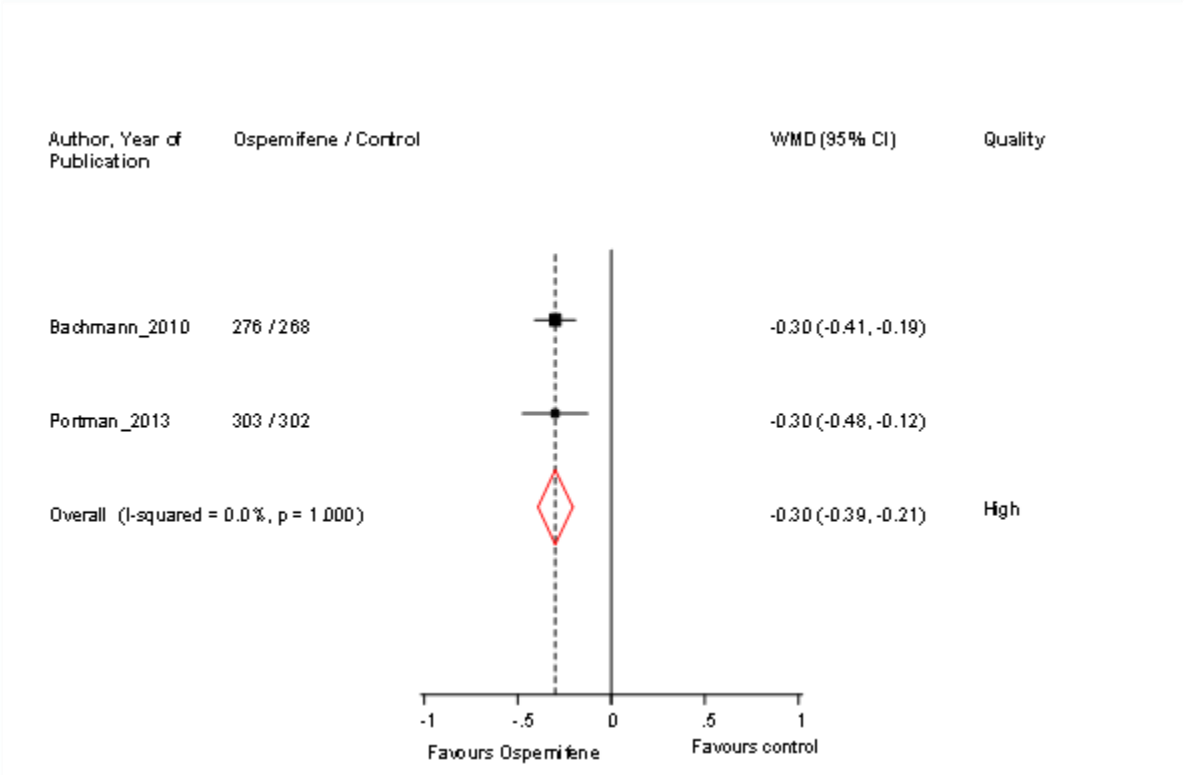


Figure 14: Change in vaginal pH after treatment of Ospemifene for less than one year compared to placebo

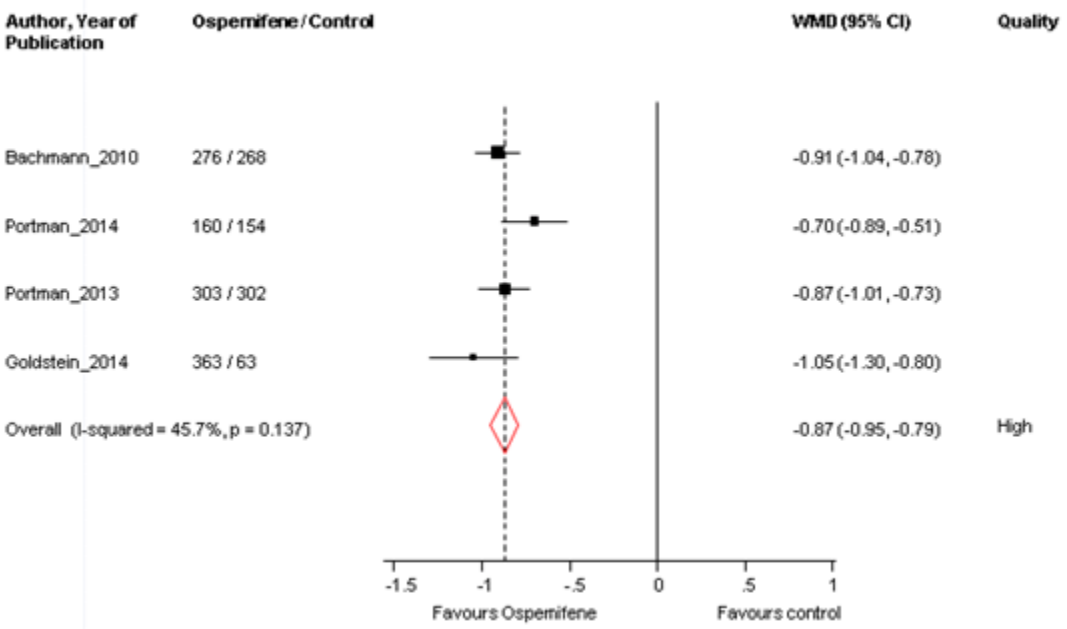


Figure 15: Change in endometrial thickness after treatment with different doses of Ospemifene for less than one year compared to placebo

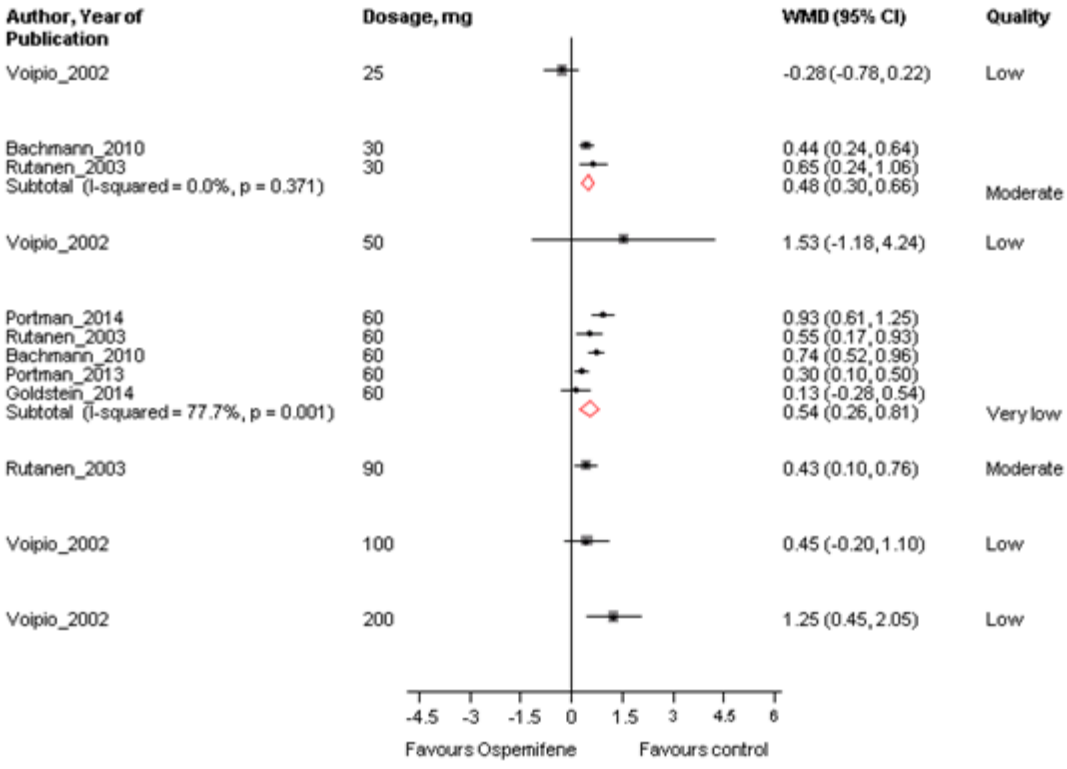


Figure 16: Frequency of adverse events relating to treatment with different doses of Ospemifene for less than one year compared to placebo

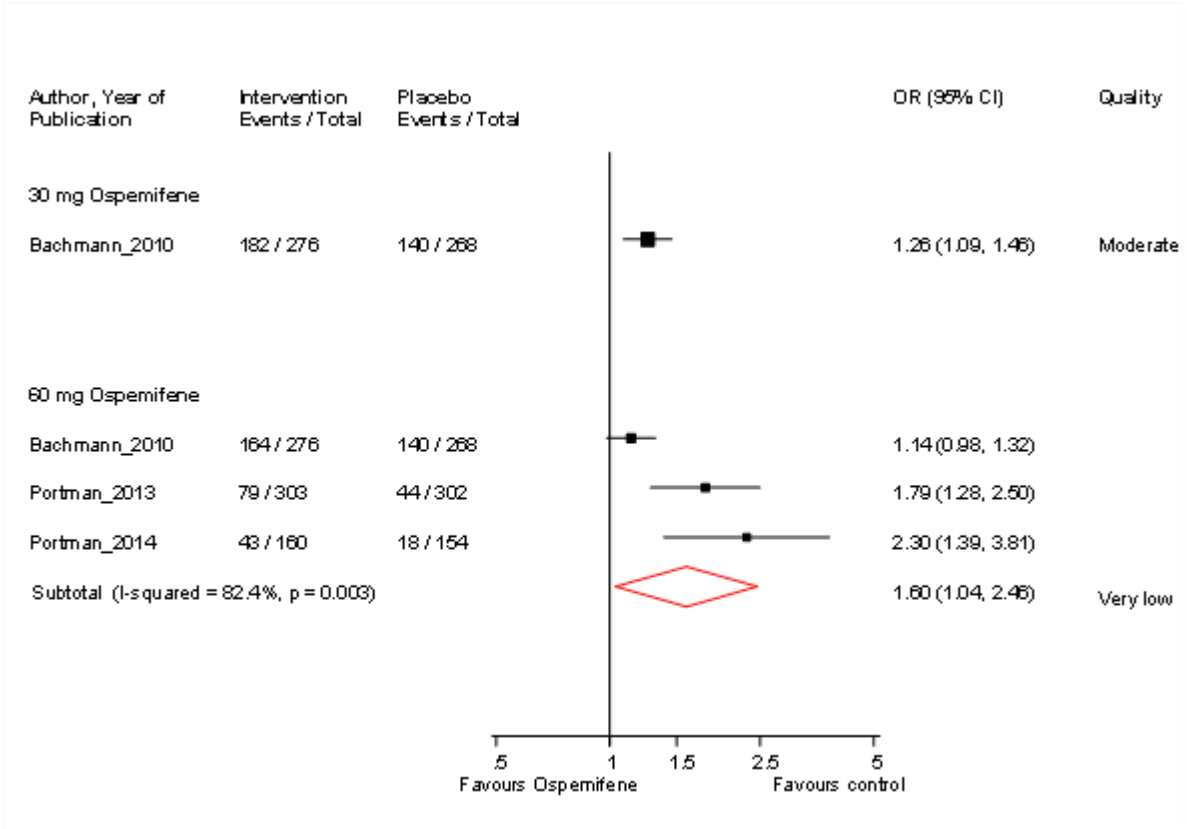


Figure 17: Withdrawal due to treatment-related adverse events with different doses of Ospemifene for less than one year compared to placebo

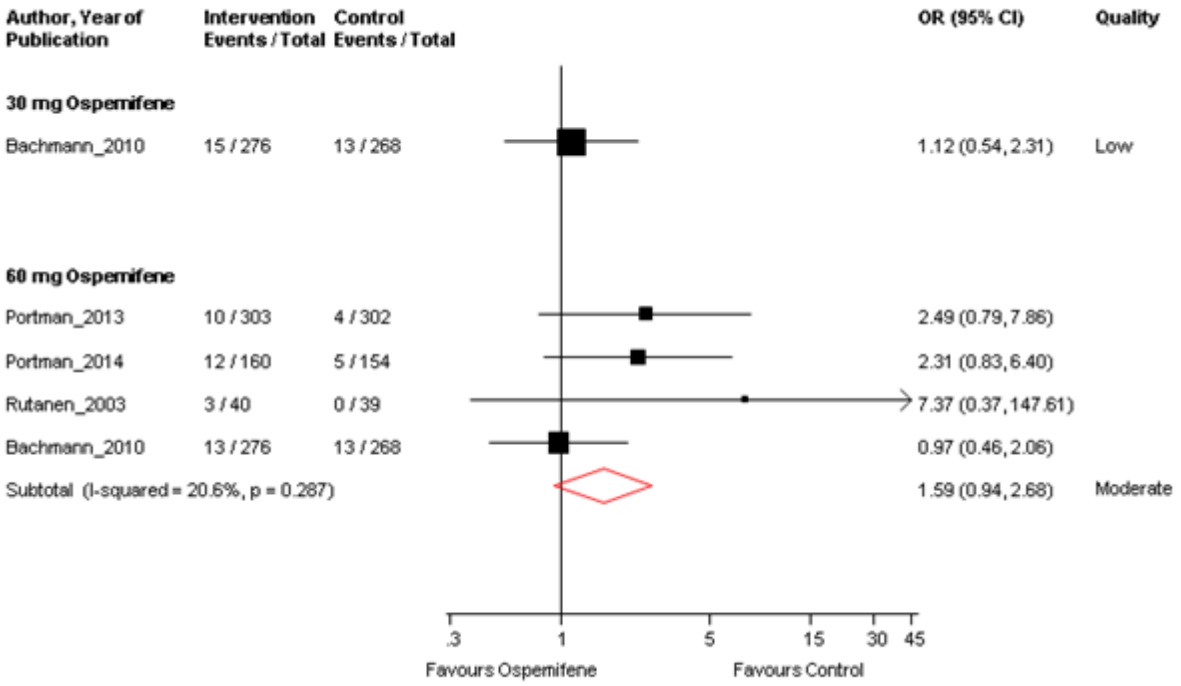


Figure 18: Change in endometrial thickness after treatment with Ospemifene for more than one year compared to placebo

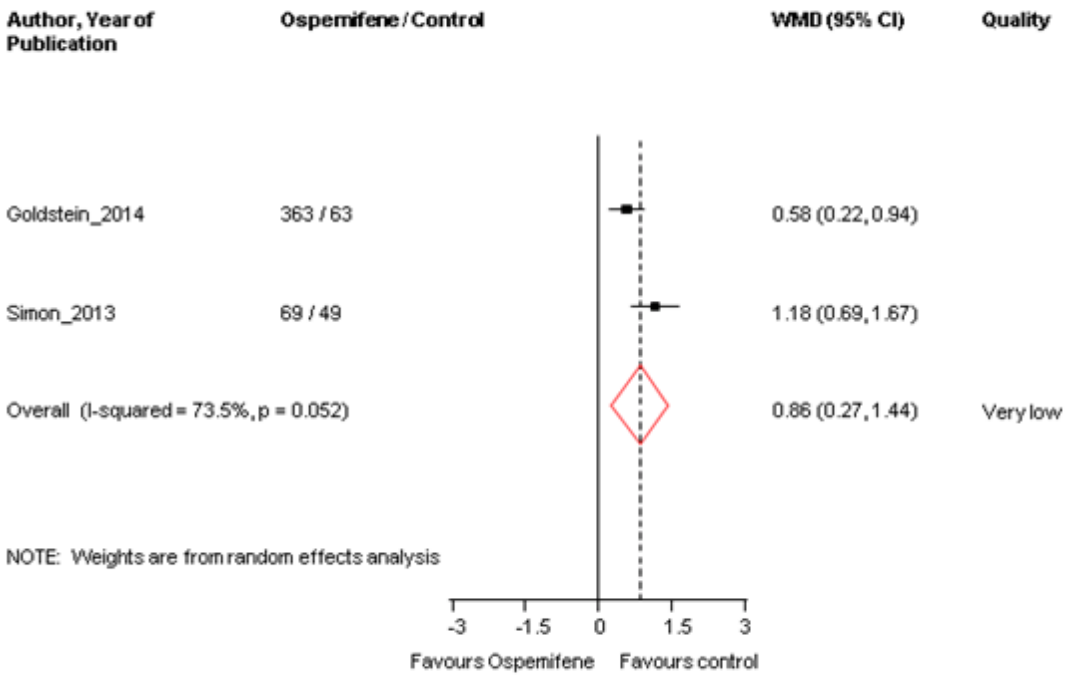


Figure 19: Frequency of adverse events relating to treatment with different doses of Ospemifene for moire than one year compared to placebo

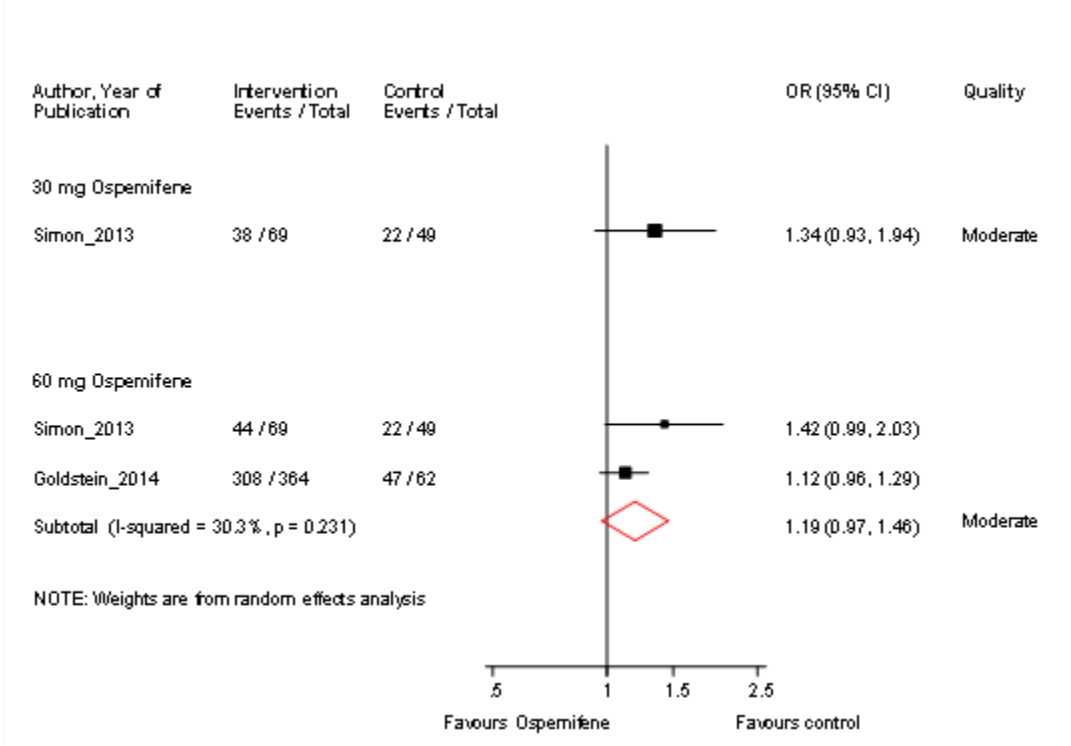
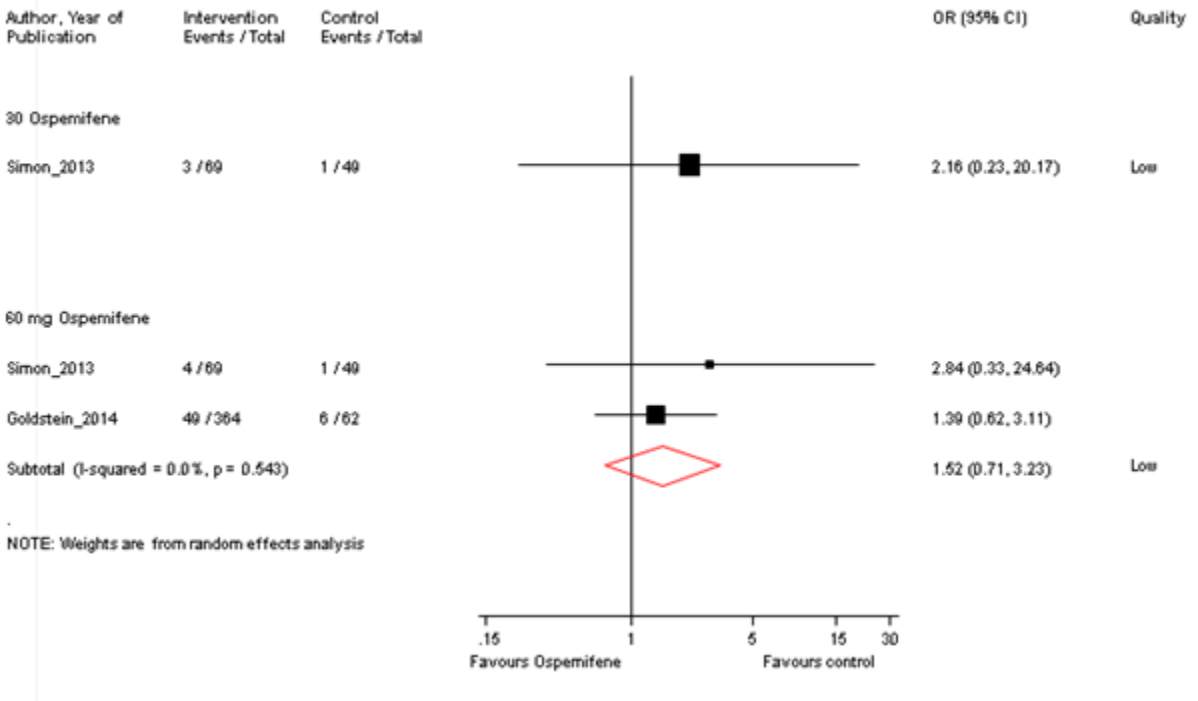


Figure 20: Withdrawal due to treatment-related adverse events with different doses of Ospemifene for more than one year compared to placebo



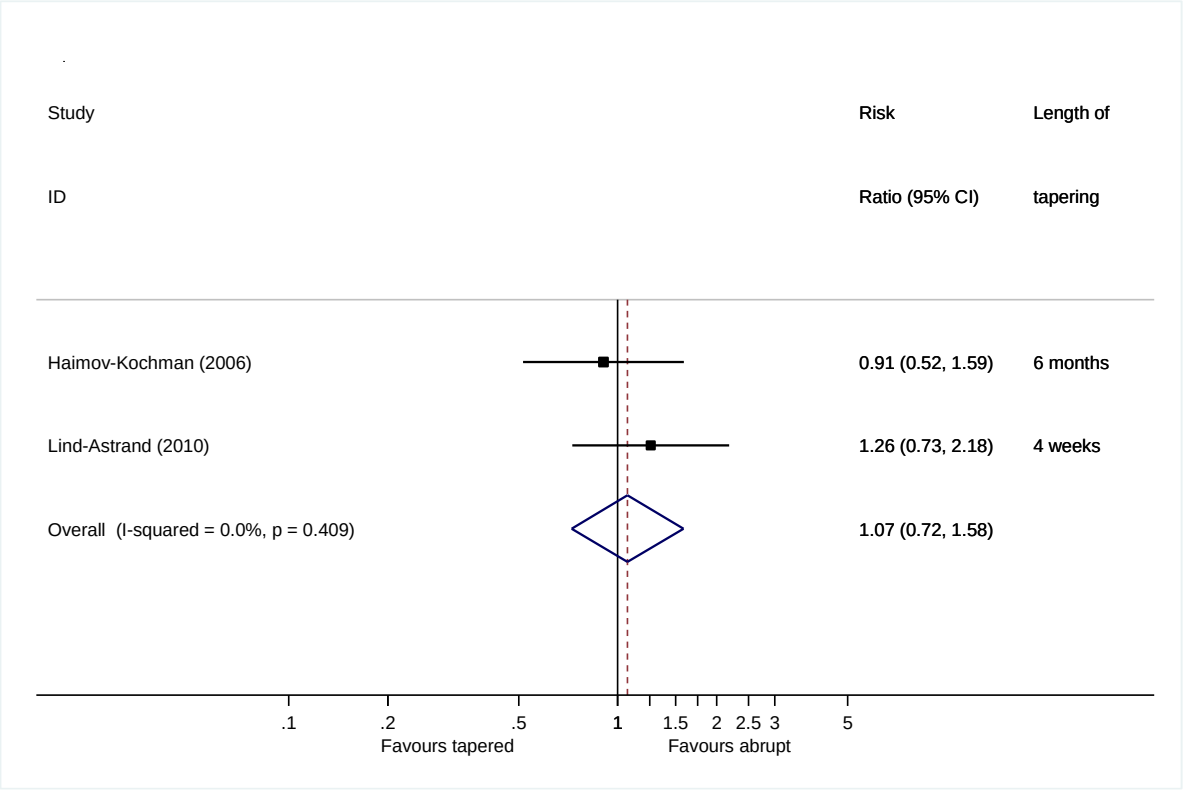
J.5 Review and referral

There are no forest plots for this review.

J.6 Starting and stopping HRT

J.6.1 Recommencing HRT

Figure 21: Recommencing HRT treatment by 12 months after tapering over 4 weeks or 6 months, versus abrupt discontinuation



J.7 Long-term benefits and risks of HRT

J.7.1 Venous thromboembolism

Figure 22: Relative risk of VTE in participants using HRT versus participants treated with placebo

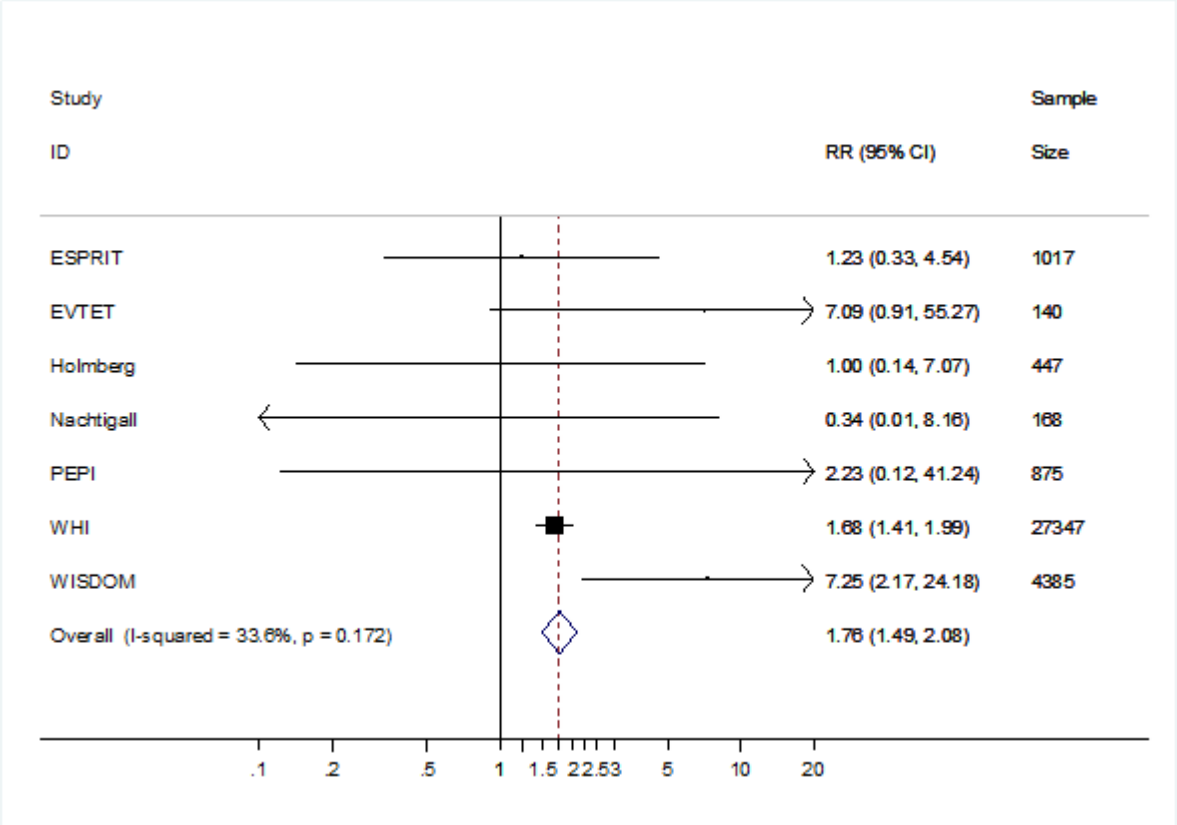


Figure 23: Relative risk of VTE in participants using oestrogen alone (HRT) versus participants treated with placebo

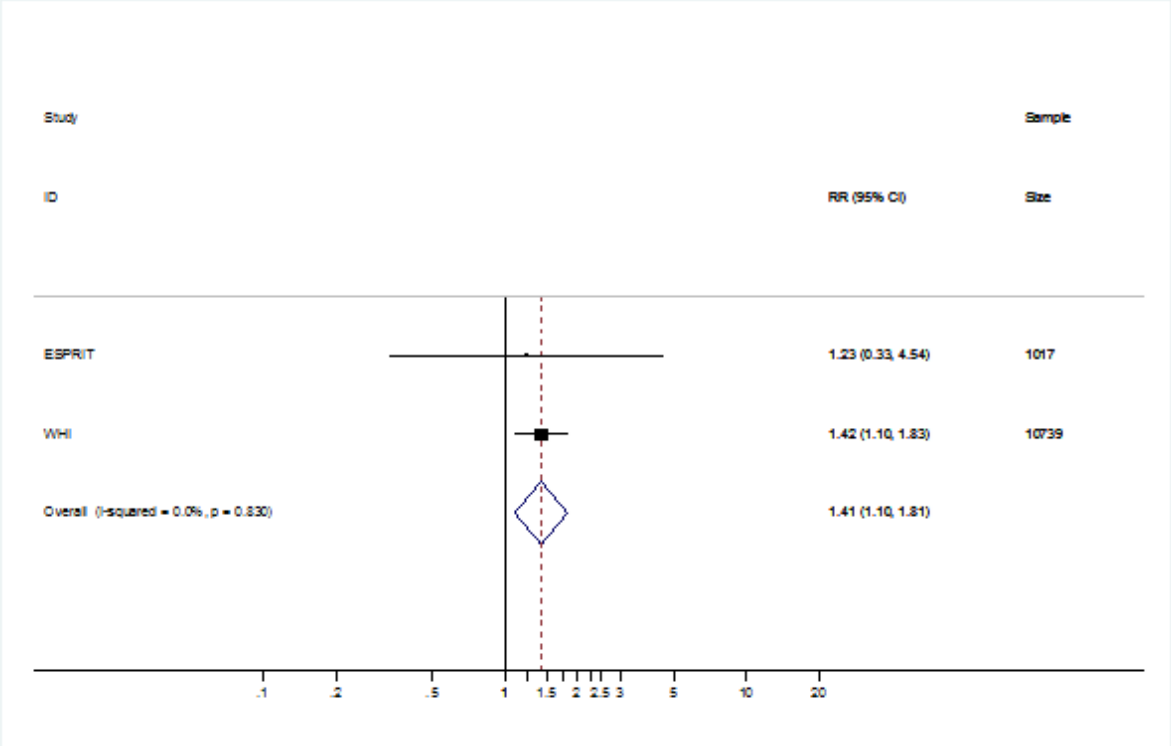


Figure 24: Relative risk of VTE in participants using oestrogen plus progestogen (HRT) versus participants treated with placebo

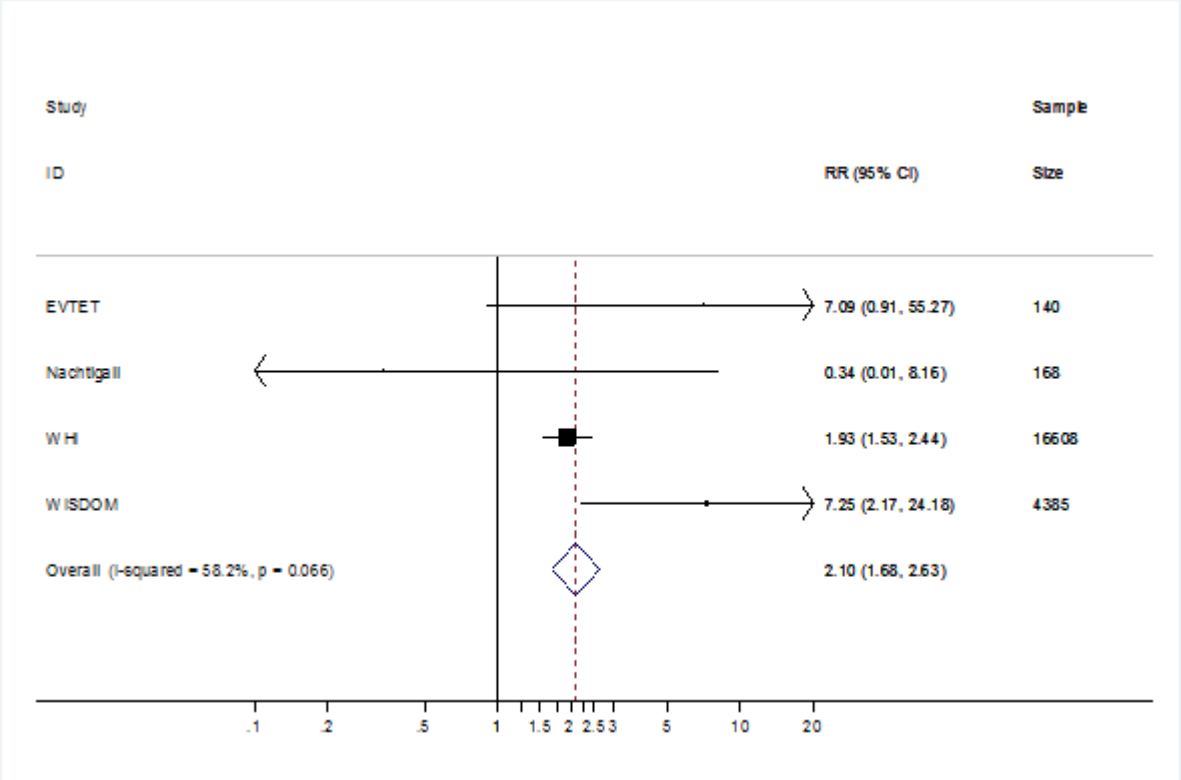


Figure 25: Relative risk of VTE in participants using HRT for between 1 and 5 years versus participants treated with placebo

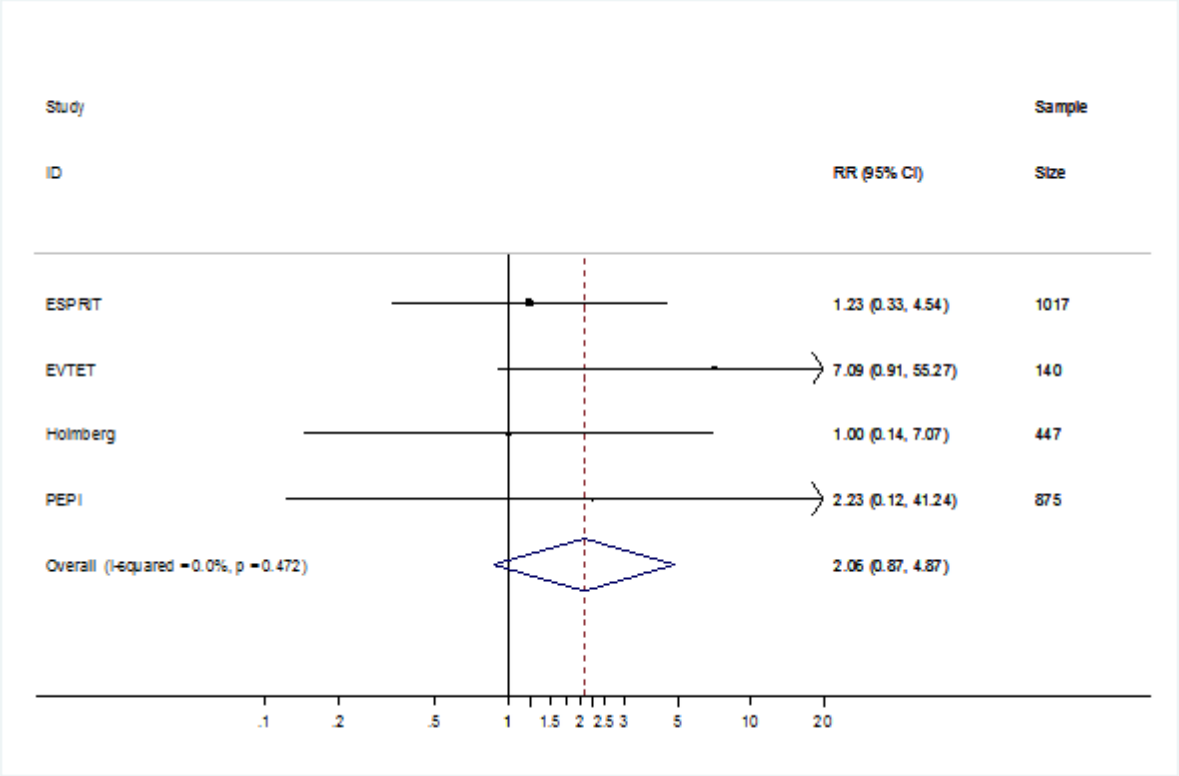
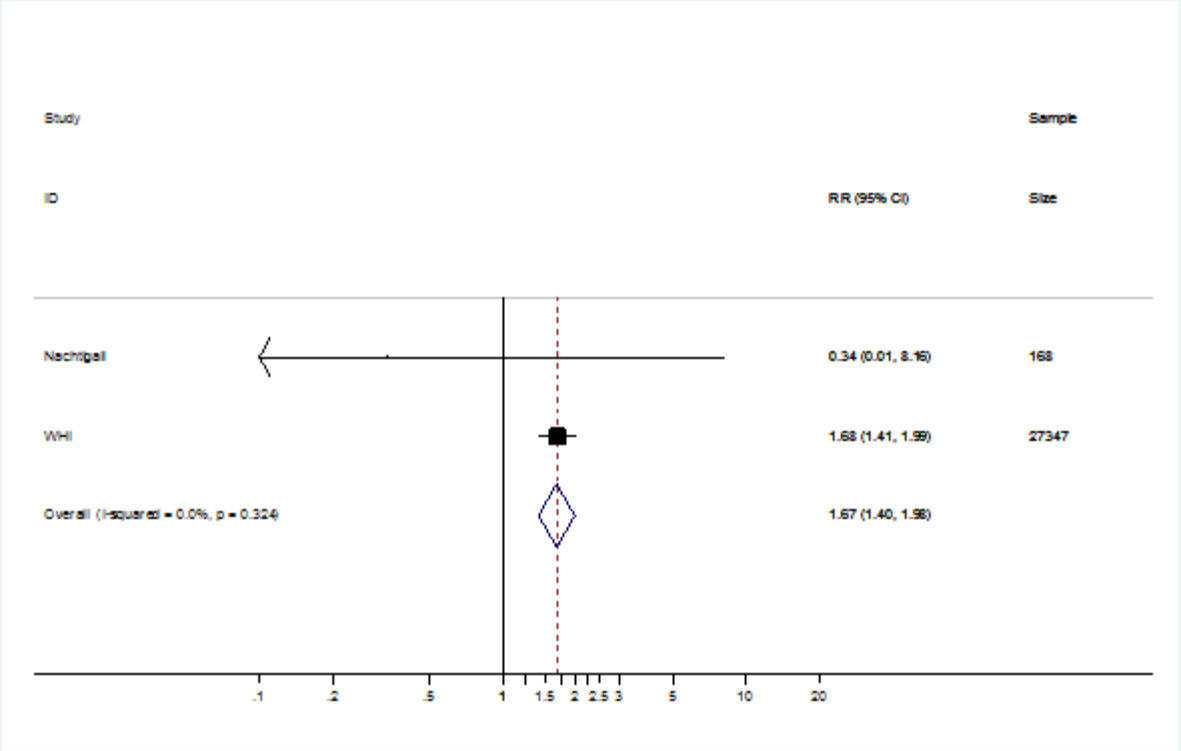


Figure 26: Relative risk of VTE in participants using HRT for more than 5 years versus participants treated with placebo



J.7.2 Cardiovascular disease

Figure 27: Forest plot showing the association between HRT use and CHD mortality

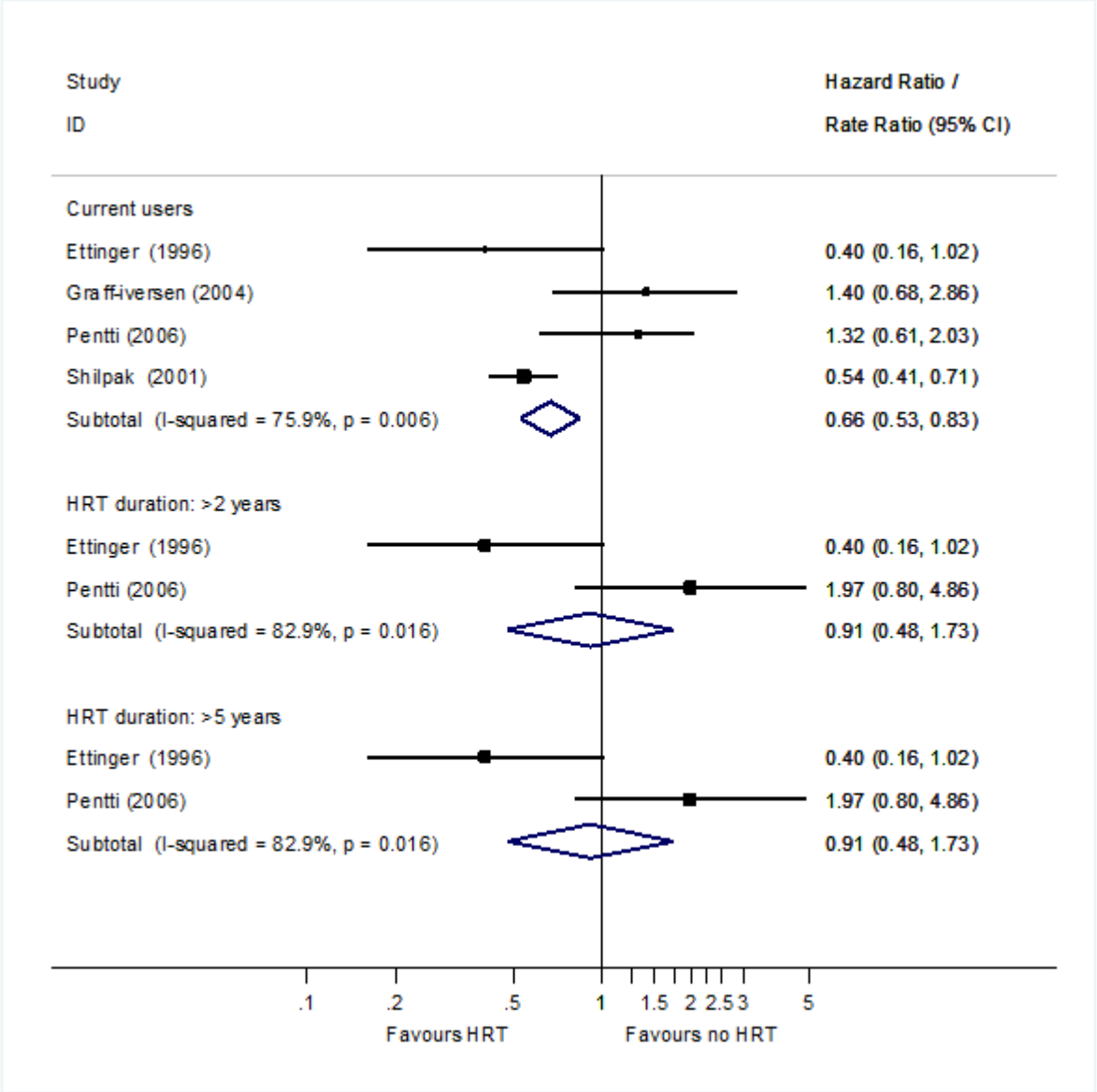


Figure 28: Forest plot showing the association between HRT use and CHD in different populations

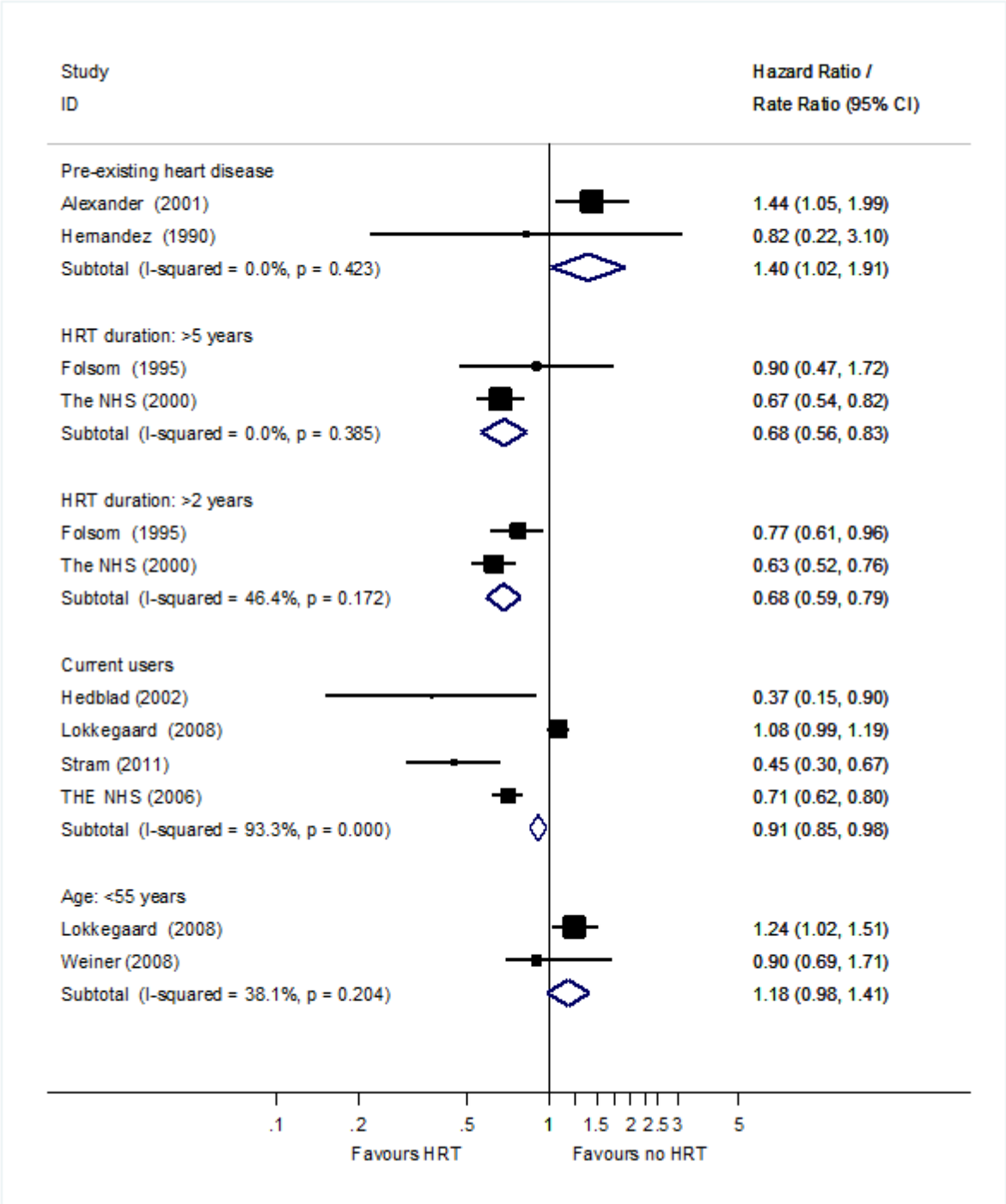


Figure 29: Forest plot showing the association between HRT use and CVD mortality

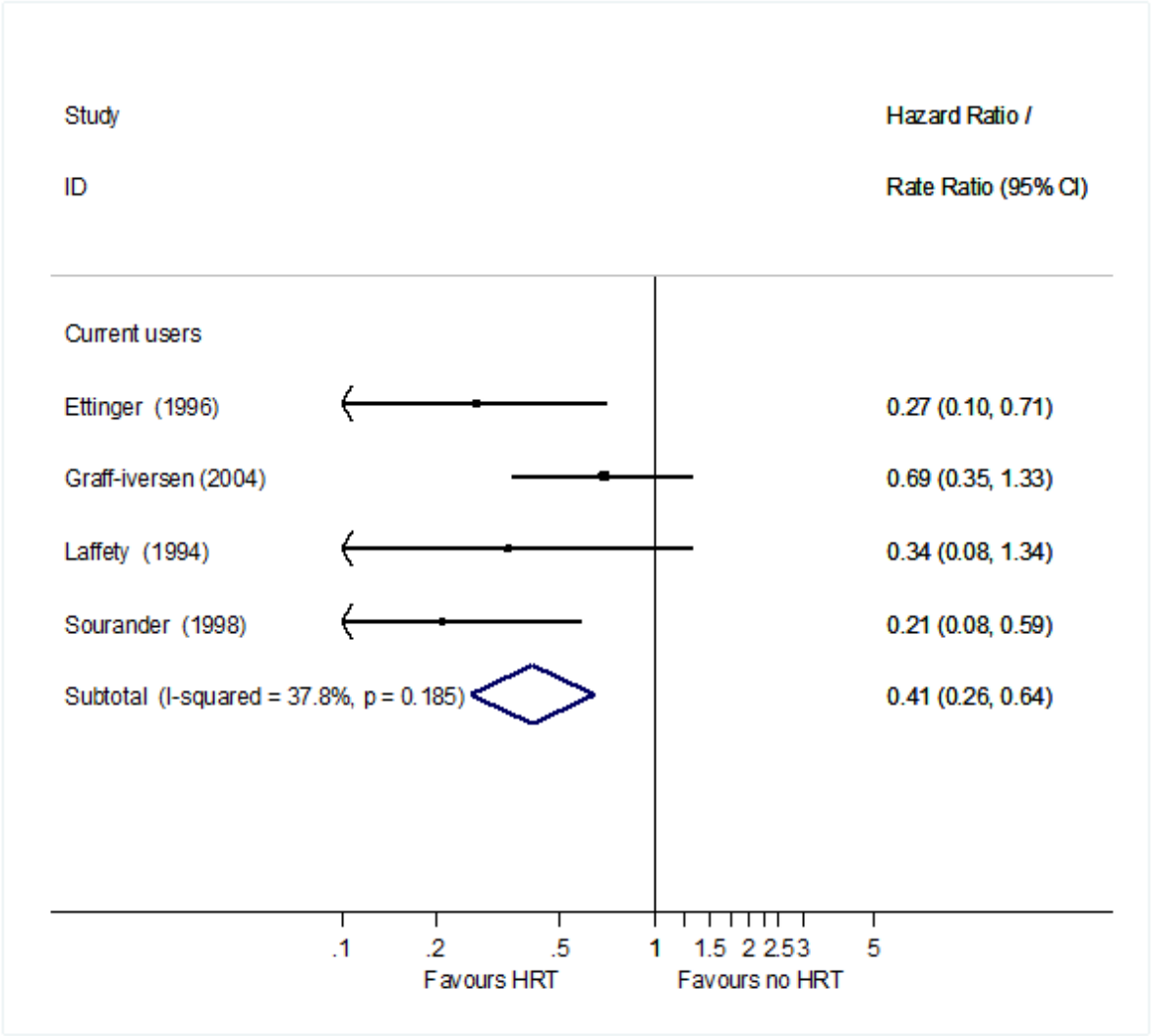
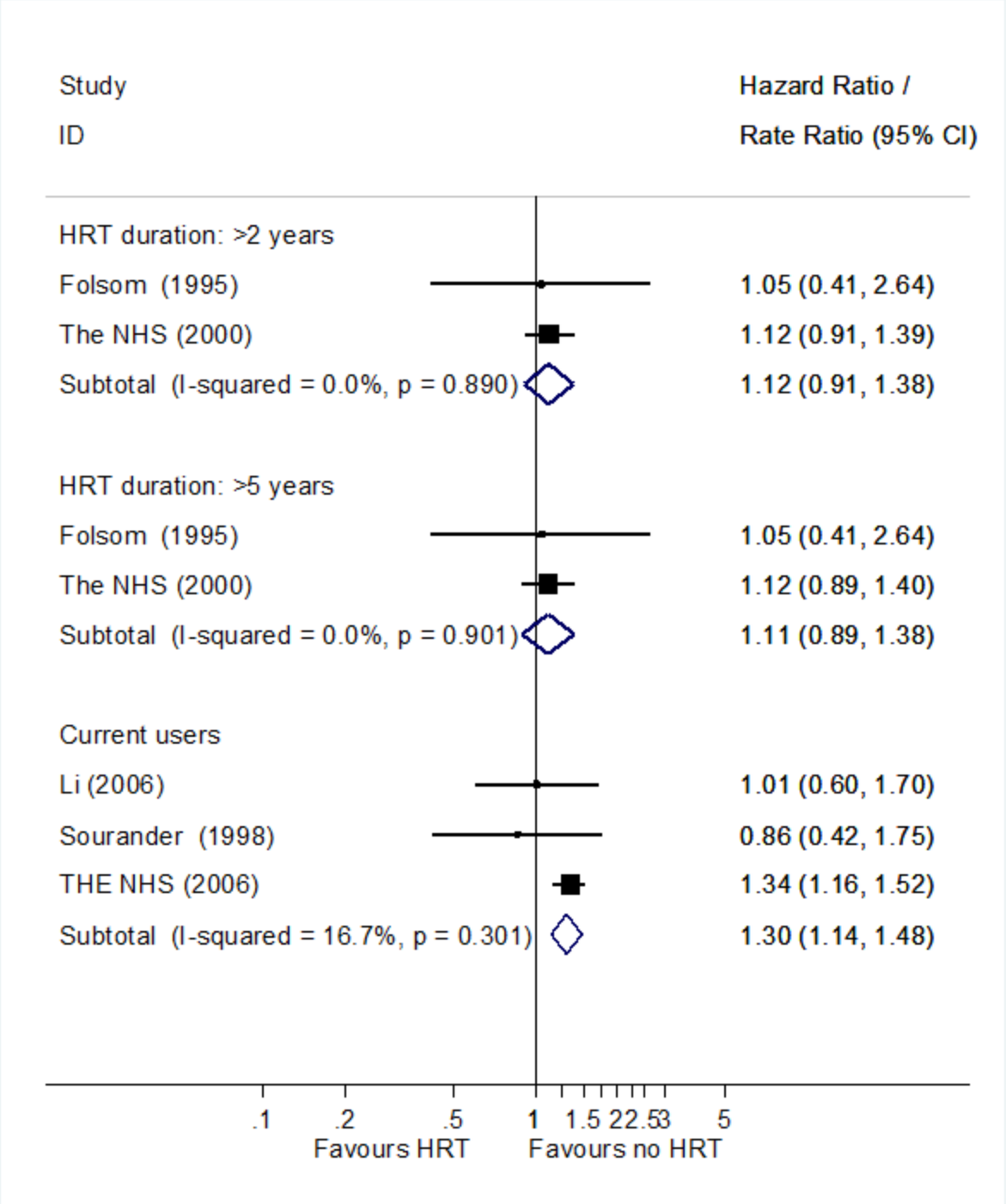


Figure 30: Forest plot showing the association between HRT use and the occurrence of stroke



J.7.3 Development of type 2 diabetes

There are no forest plots for this review

J.7.4 Management of type 2 diabetes – control of blood sugar

There are no forest plots for this review

J.7.5 Breast Cancer

Figure 31: Cohort studies: ever use versus never use of HRT

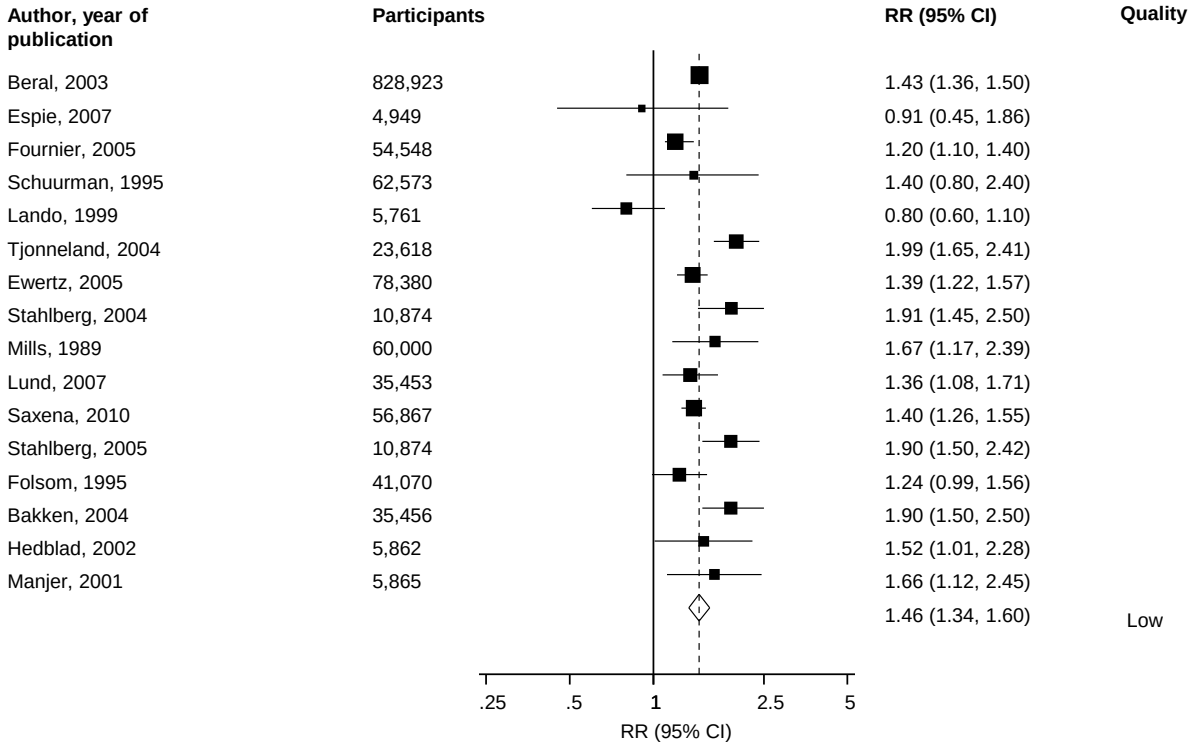


Figure 32: Cohort studies: current use versus never use of HRT

Author, year of publication	Participants	RR (95% CI)	Quality
Beral, 2003	828923	1.66 (1.60, 1.72)	
Tjonneland, 2004	23618	2.22 (1.80, 2.75)	
Ewertz, 2005	78380	1.61 (1.38, 1.88)	
Stahlberg, 2004	10874	2.42 (1.81, 3.26)	
Grodstein, 1997	34625	0.76 (0.56, 1.02)	
Mills, 1989	60000	2.53 (1.62, 3.98)	
Lund, 2007	35453	1.53 (1.18, 1.98)	
Stahlberg, 2005	10874	2.42 (1.81, 3.26)	
Bakken, 2004	31451	2.10 (1.70, 2.70)	
		1.79 (1.52, 2.11)	Low

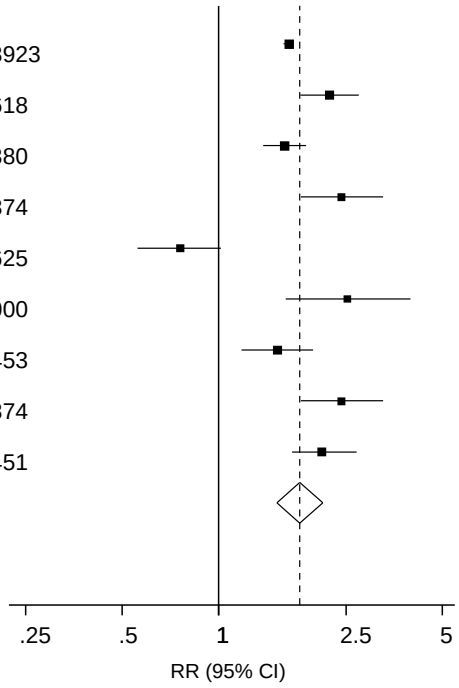


Figure 33: Cohort studies: past use versus never use of HRT

Author, year of publication	Participants	RR (95% CI)	Quality
Beral, 2003	828923	1.01 (0.95, 1.08)	
Tjonneland, 2004	23618	1.35 (0.90, 2.02)	
Ewertz, 2005	78380	1.03 (0.82, 1.28)	
Stahlberg, 2004	10874	1.16 (0.76, 1.77)	
Grodstein, 1997	34625	0.83 (0.63, 1.09)	
Mills, 1989	60000	1.44 (0.95, 2.17)	
Lund, 2007	35453	0.87 (0.53, 1.44)	
Stahlberg, 2005	10874	1.16 (0.76, 1.77)	
Bakken, 2004	35456	1.00 (0.60, 1.60)	
		1.02 (0.96, 1.08)	Low

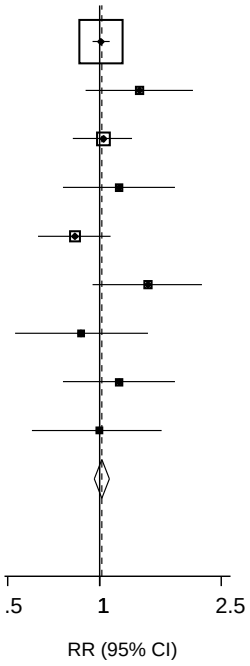


Figure 34: Cohort studies: use of oestrogen

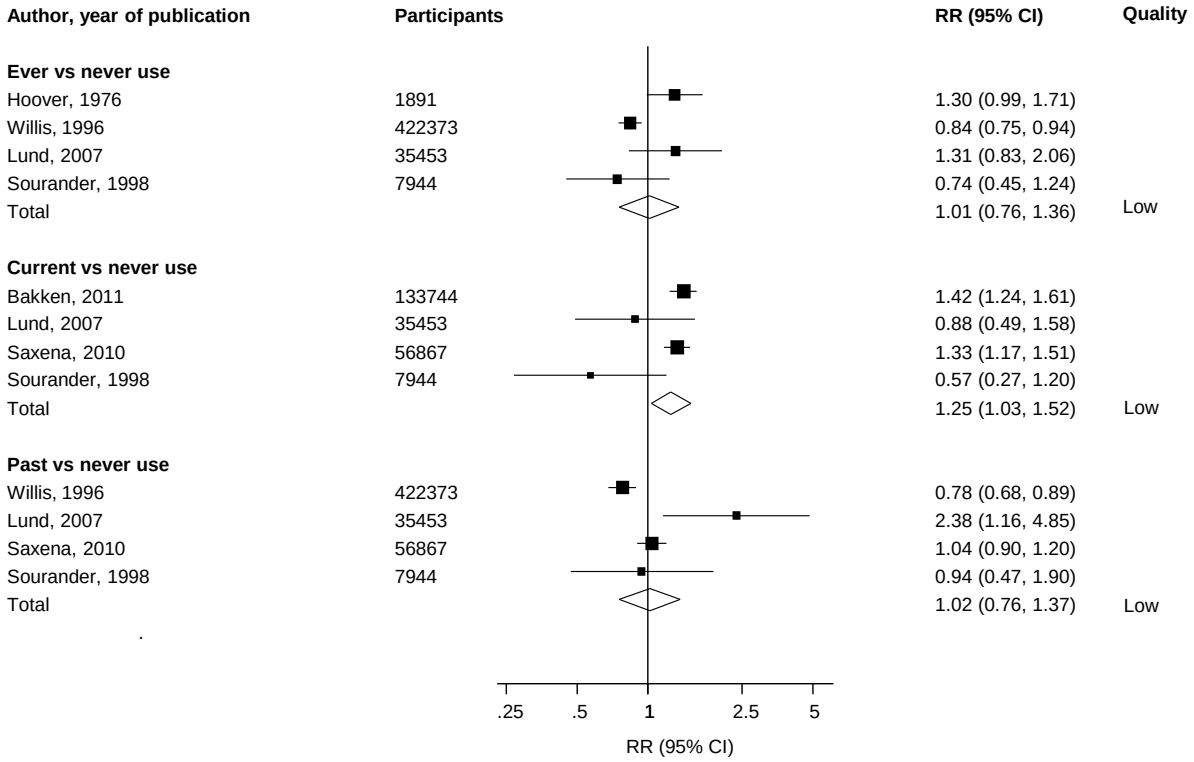


Figure 35: Cohort studies: use of oestrogen plus progestogen

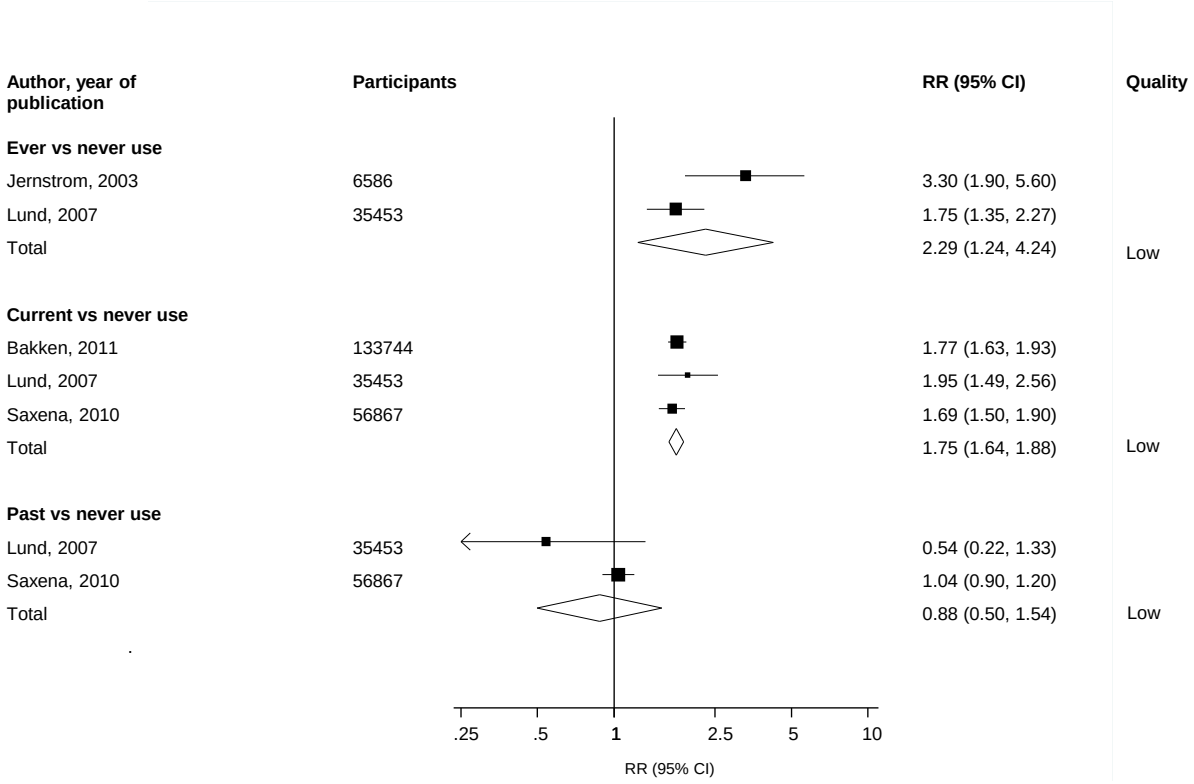


Figure 36: Cohort studies: duration of HRT use

Duration of use	No. of studies	HRT users / non-HRT users	RR (95% CI)	Quality
Up to 2 years	4	68,537 / 53,338	1.63 (1.17, 2.28)	Low
Up to 4 years	5	64,893 / 54,450	1.35 (0.91, 1.99)	Low
Up to 5 years	2	23,375 / 29,163	1.49 (1.12, 1.97)	Low
More than or equal to 4 years	2	45,215 / 50,403	1.21 (0.99, 1.47)	Low
3 to 9 years	1	2,197 / 3,564	0.50 (0.29, 0.87)	Low
5 to 10 years	4	40,856 / 29,646	2.03 (1.37, 3.02)	Moderate
More than or equal to 10 years	4	38,745 / 26,644	1.19 (0.64, 2.22)	Low
10 to 14 years	1	4,308 / 6,566	3.08 (1.87, 5.07)	Moderate
More than or equal to 15 years	1	4,308 / 6,566	3.08 (1.87, 5.07)	Moderate

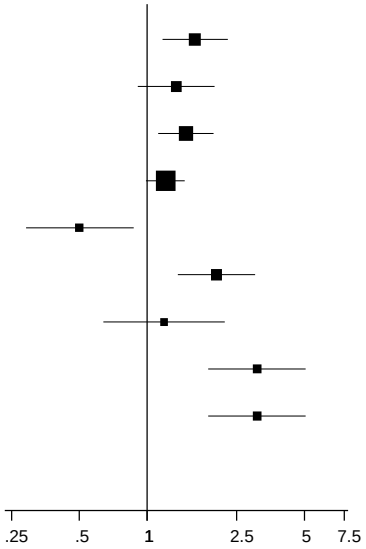


Figure 37: Cohort studies: duration of oestrogen use

Duration of use	No. of studies	O users / Non-HRT users	RR (95% CI)	Quality
Up to 2 years	5	748,816 / 740,566	0.93 (0.77, 1.13)	Low
Up to 5 years	7	820,581 / 770,653	1.16 (0.95, 1.42)	Low
4 to 10 years	5	748,816 / 740,566	1.23 (0.94, 1.61)	Low
More than or equal to 5 years	3	89,346 / 59,981	1.42 (1.10, 1.82)	Low
Up to 7 years	1	27,075 / 19,280	1.00 (0.83, 1.21)	Moderate
More than or equal to 10 years	4	686,699 / 698,341	1.10 (0.77, 1.55)	Low
6 to 15 years	2	71,611 / 31,611	1.14 (0.91, 1.43)	Low
More than or equal to 15 years	2	71,611 / 31,611	1.20 (1.06, 1.36)	Low

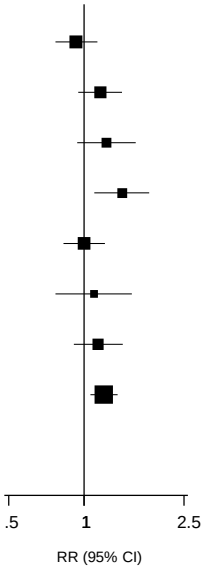


Figure 38: Cohort studies: duration of oestrogen plus progestogen use

Duration of use	No. of studies	OP users / Non-HRT users	RR (95% CI)	Quality
Up to 2 years	4	581,571 / 507,828	1.14 (0.84, 1.55)	Low
Up to 5 years	6	653,336 / 537,915	1.52 (1.25, 1.85)	Low
4 to 10 years	3	554,496 / 488,548	1.94 (1.41, 2.66)	Low
More than or equal to 4 years	3	110,978 / 60,739	1.81 (1.12, 2.91)	Low
More than or equal to 10 years	2	497,822 / 464,845	2.30 (2.07, 2.55)	Moderate
6 to 15 years	1	44,536 / 12,331	1.57 (1.40, 1.76)	Moderate
More than or equal to 15 years	1	44,536 / 12,331	1.83 (1.48, 2.26)	Moderate

Figure 39: Cohort studies: time since last use of HRT

Time since last use	No. of studies	HRT users / Non-HRT users	RR (95% CI)	Quality
Up to 5 years	2	437,905 / 394,193	1.05 (0.96, 1.13)	Low
4 to 10 years	2	437,905 / 394,193	1.02 (0.89, 1.16)	Low
More than or equal to 10 years	1	436,166 / 392,757	0.90 (0.72, 1.12)	Low

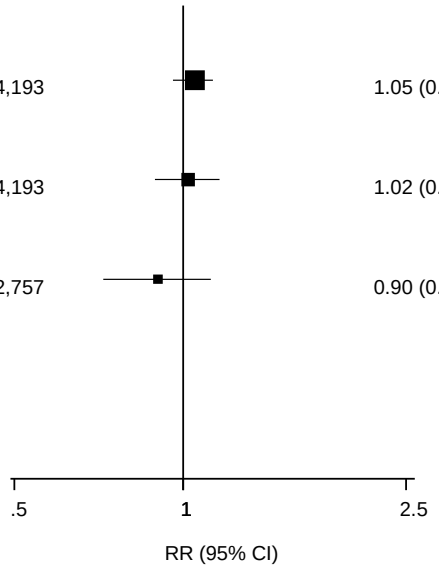


Figure 40: Cohort studies: time since last use of oestrogen

Time since last use	No. of studies	O users / Non-HRT users	RR (95% CI)	Quality
Up to 5 years	3	272,626 / 276,479	1.20 (0.90, 1.60)	Low
5 to 10 years	2	215,952 / 252,776	0.76 (0.60, 0.95)	Low
More than or equal to 5 years	2	83,749 / 42,983	1.10 (0.96, 1.27)	Low
More than or equal to 11 years	1	188,877 / 233,496	0.84 (0.70, 1.01)	Low

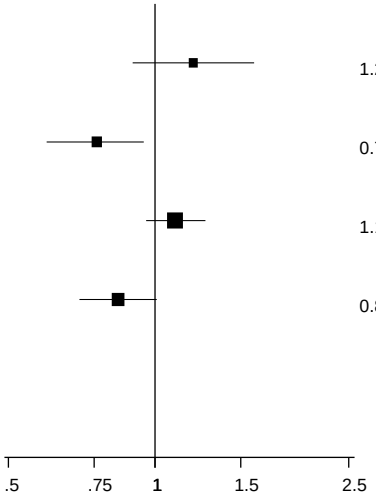


Figure 41: Cohort studies: time since last use of oestrogen plus progestogen

Time since last use	No. of studies	OP users / Non-HRT users	RR (95% CI)	Quality
Up to 5 years	2	83,749 / 42,983	1.11 (0.90, 1.41)	Low
4 to 10 years	1	27,075 / 19,280	0.60 (0.17, 2.16)	Low
More than or equal to 6 years	1	27,075 / 19,280	0.60 (0.30, 1.60)	Low

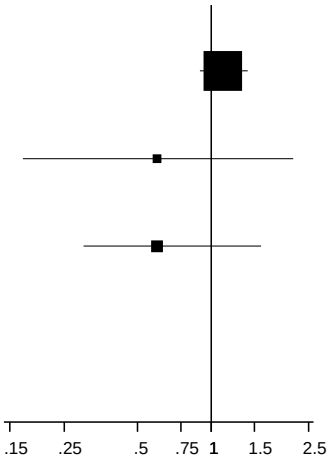


Figure 42: Cohort studies: type of HRT (timing of use not specified)

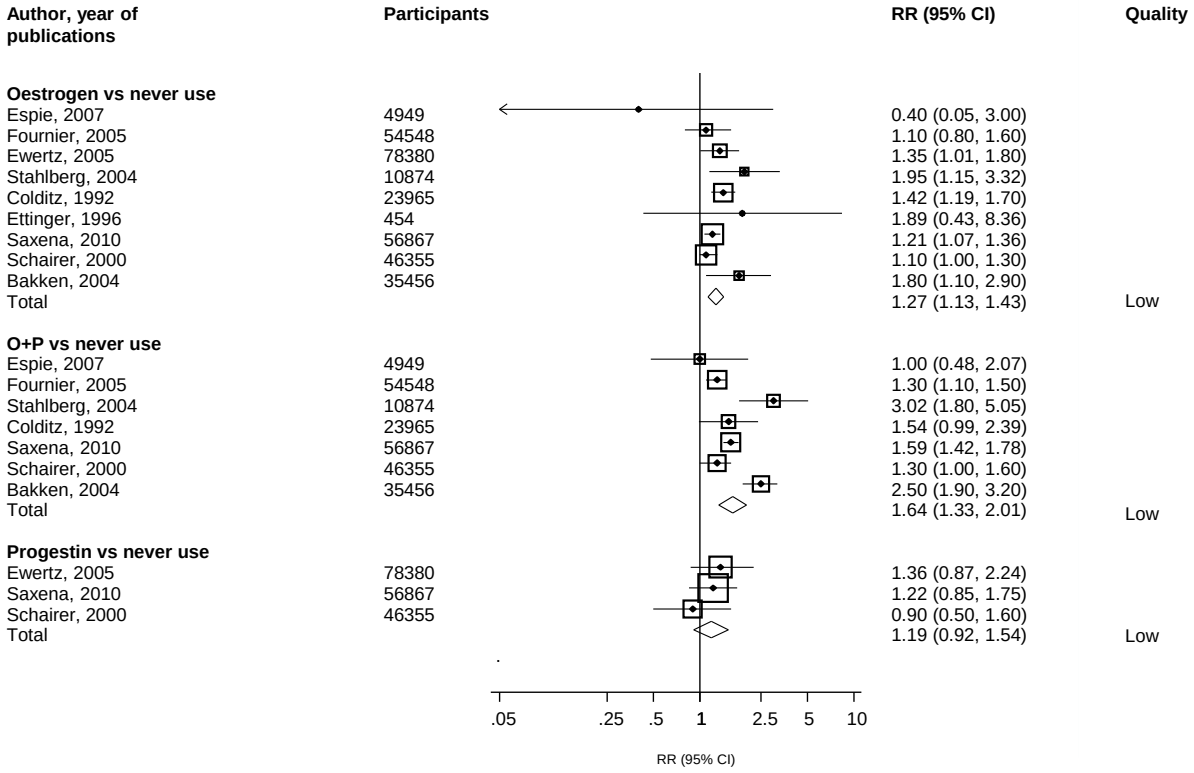


Figure 43: Cohort studies: breast cancer incidence and mortality (ever use versus never use of HRT)

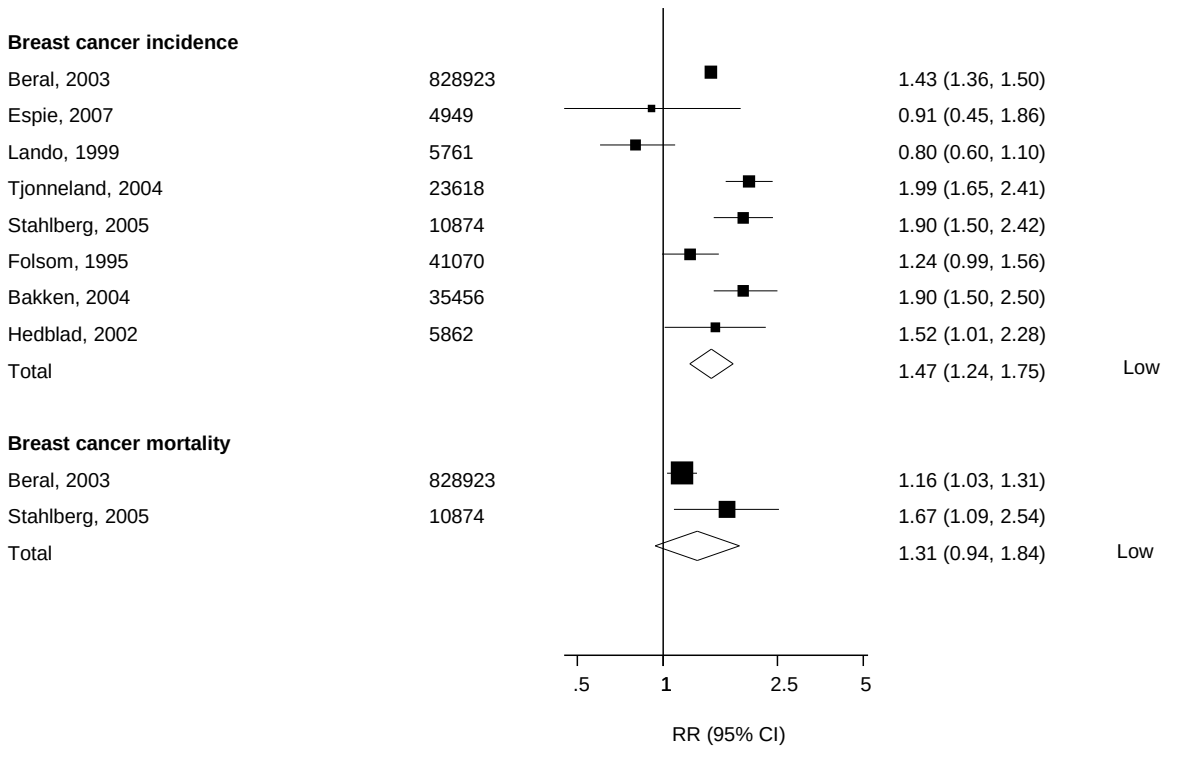


Figure 44: Cohort studies: breast cancer incidence and mortality (current use versus never use of HRT)

Author, year of publication	Participants	RR (95% CI)	Quality
Breast cancer incidence			
Beral, 2003	285987	1.66 (1.60, 1.72)	
Tjonneland, 2004	23618	2.22 (1.80, 2.75)	
Stahlberg, 2005	10874	2.42 (1.81, 3.26)	
Bakken, 2004	31451	2.10 (1.70, 2.70)	
Total		2.03 (1.65, 2.50)	Moderate
.			
Breast cancer mortality			
Beral, 2003	285987	1.22 (1.05, 1.41)	
Grodstein, 1997	34625	0.76 (0.56, 1.02)	
Stahlberg, 2005	10874	1.97 (1.14, 3.42)	
Total		1.16 (0.76, 1.77)	Low
.			

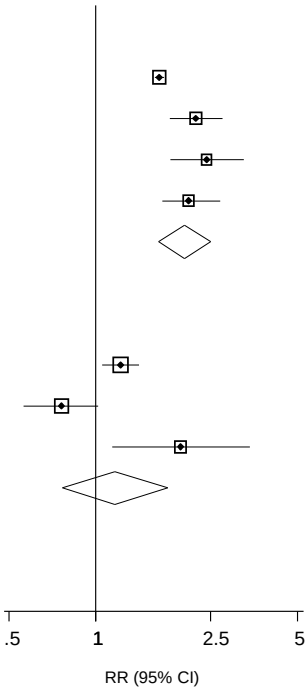
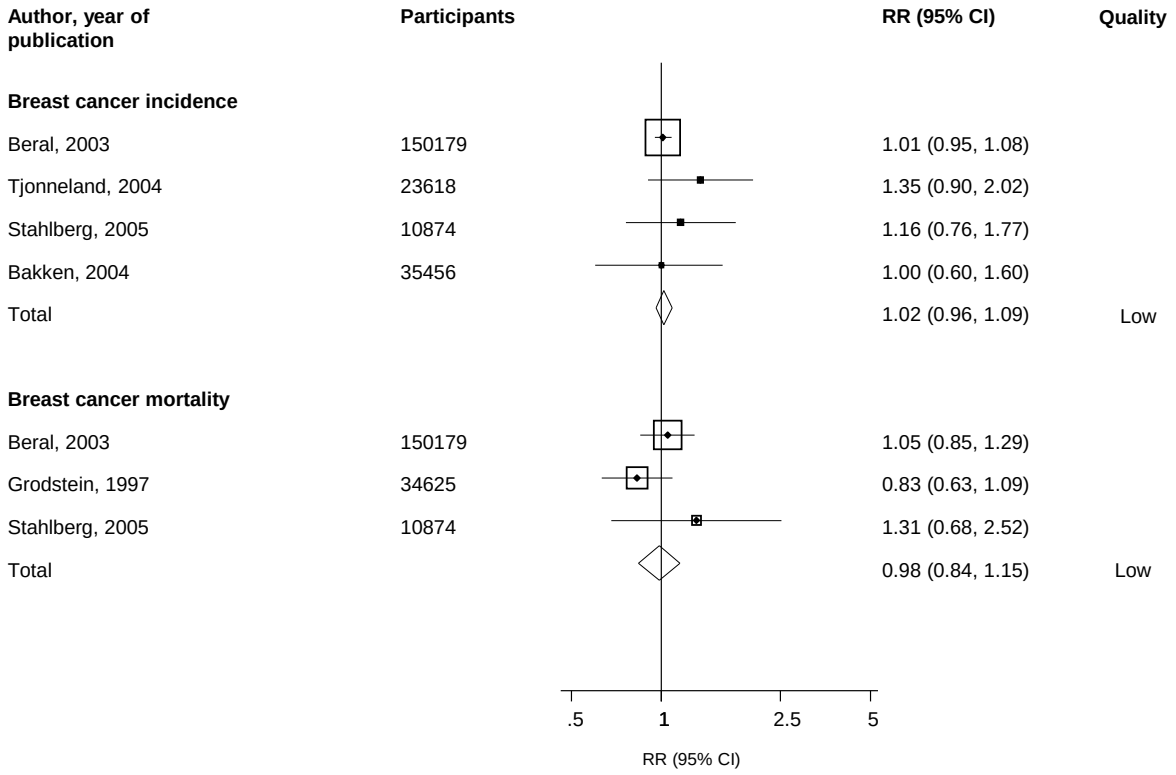


Figure 45: Cohort studies: breast cancer incidence and mortality (past use versus never use of HRT)



J.7.6 Osteoporosis

Figure 46: Risk of any fracture with current use of HRT compared to no HRT

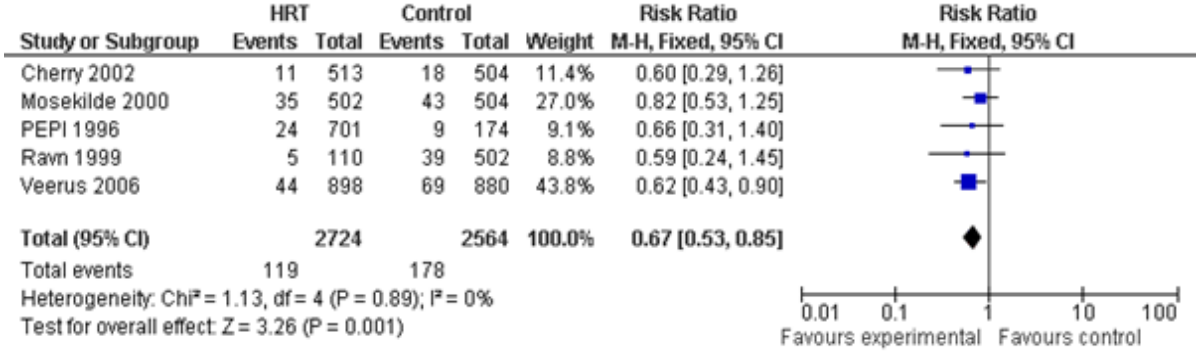


Figure 47: Risk of any non-vertebral fracture with current use of HRT compared to no HRT

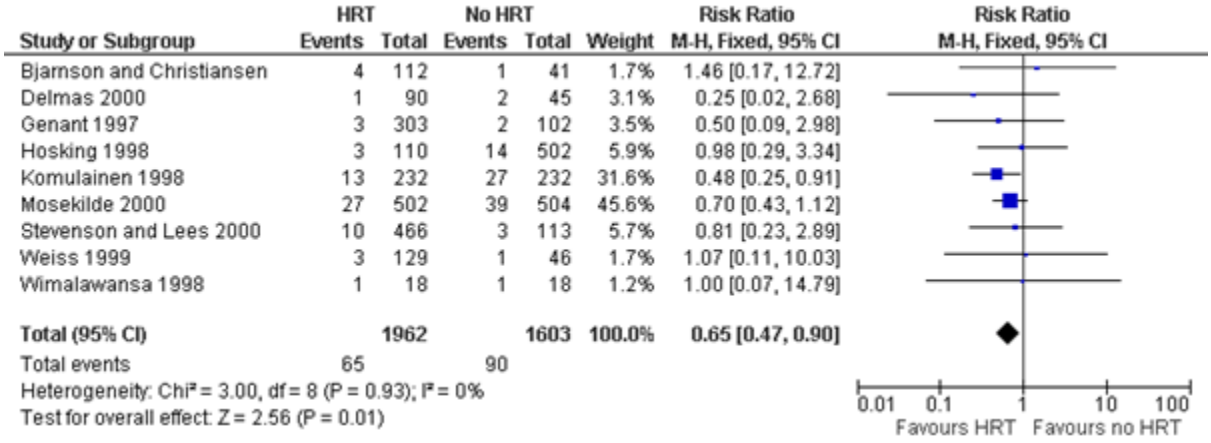


Figure 48: Risk of hip fracture with current use of HRT compared to no HRT

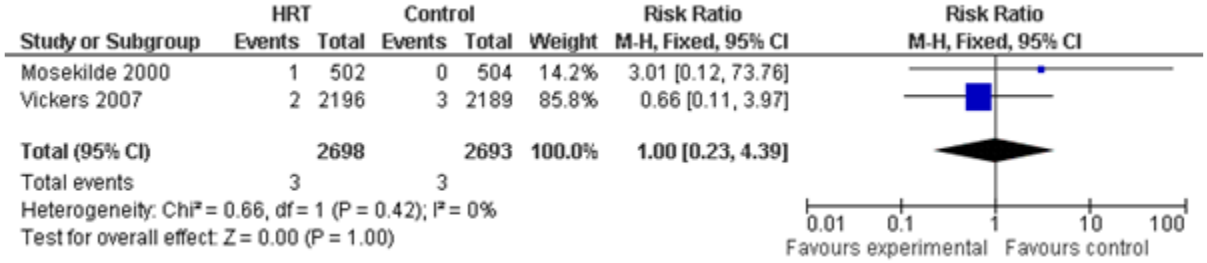
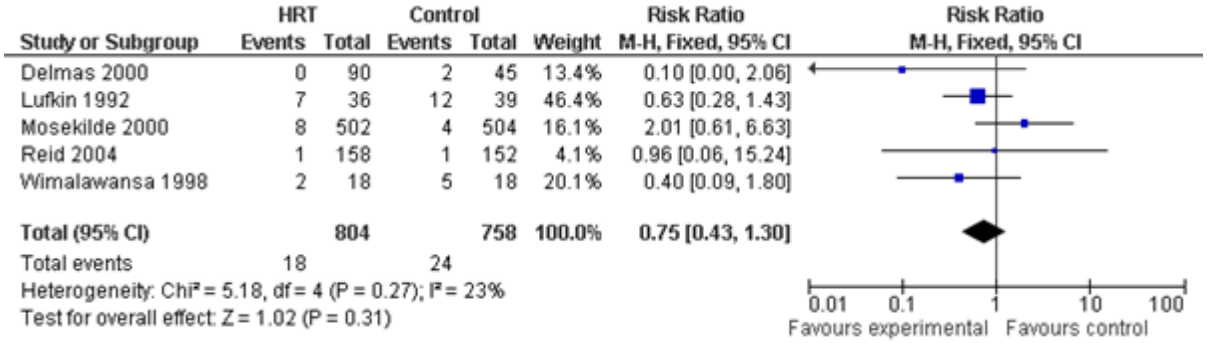


Figure 49: Risk of vertebral fracture with current use of HRT compared to no HRT



<Insert Note here>

Figure 50: Risk of wrist fracture with current use of HRT compared to no HRT

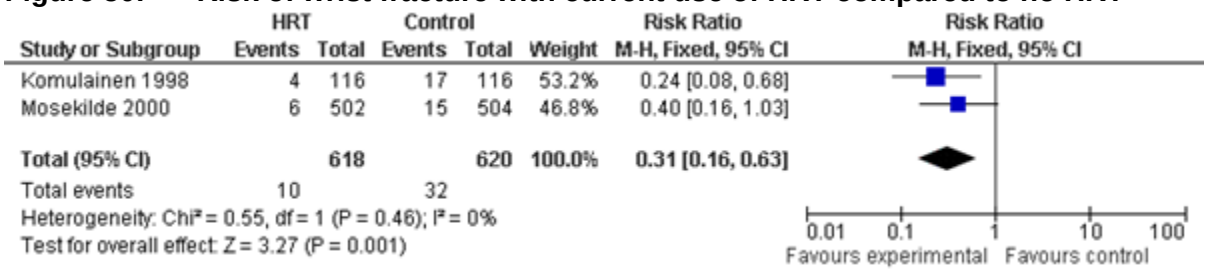


Figure 51: Risk of any non-vertebral fracture with HRT use for up to 2 years duration compared to no HRT

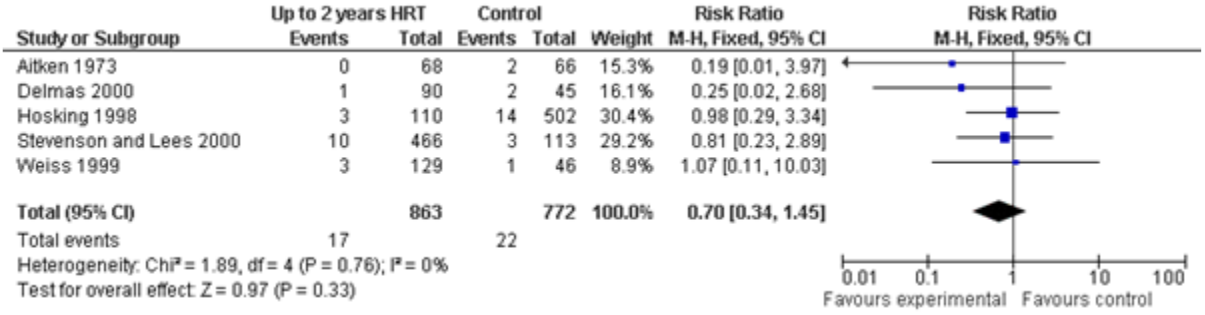


Figure 52: Risk of any vertebral fracture with HRT use for up to 2 years duration compared to no HRT

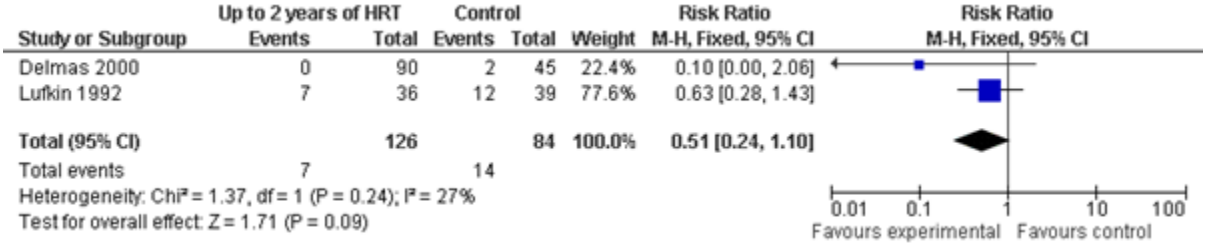


Figure 53: Risk of any fracture with HRT use for 2 to 5 years duration compared to no HRT

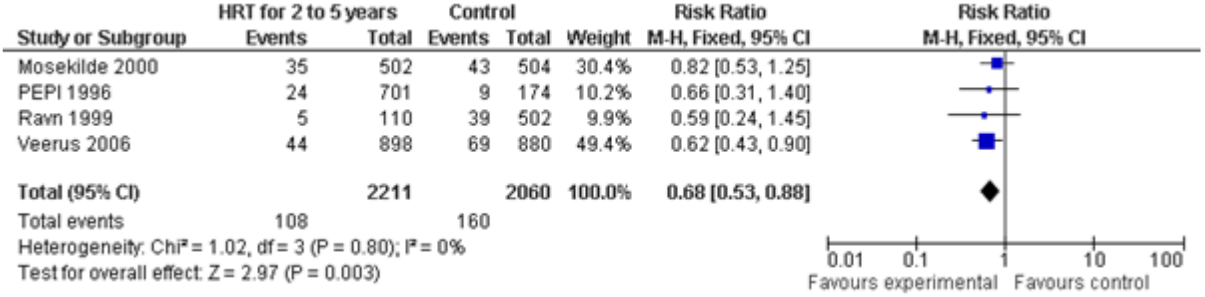


Figure 54: Risk of any non-vertebral fracture with HRT use for 2 to 5 years duration compared to no HRT

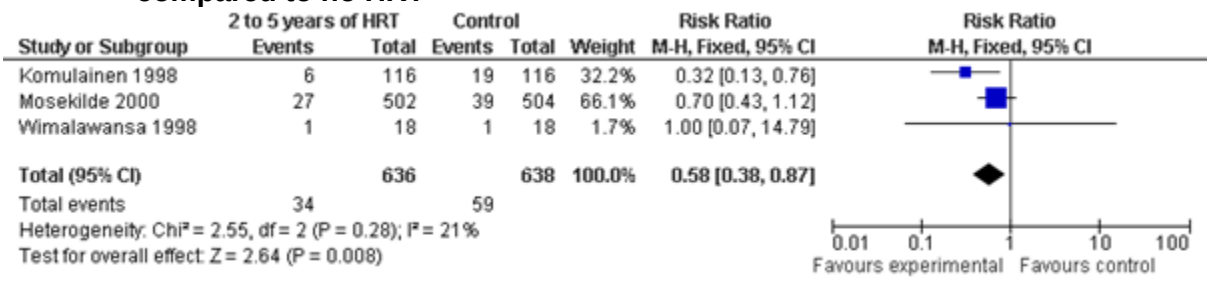


Figure 55: Risk of any vertebral fracture with HRT use for 2 to 5 years duration compared to no HRT

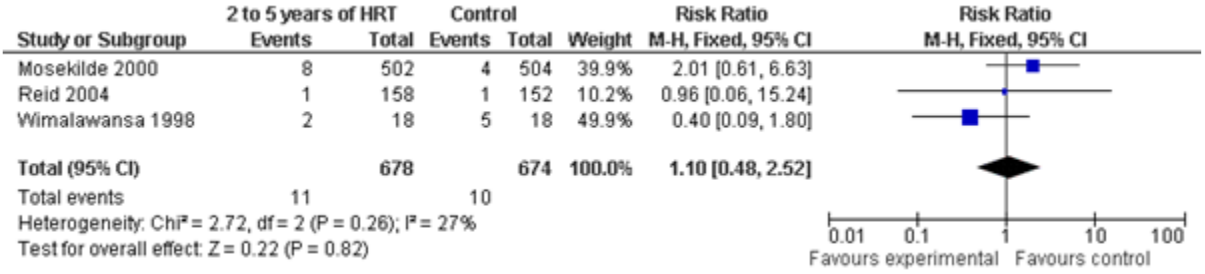


Figure 56: Risk of any wrist fracture with HRT use for 2 to 5 years duration compared to no HRT

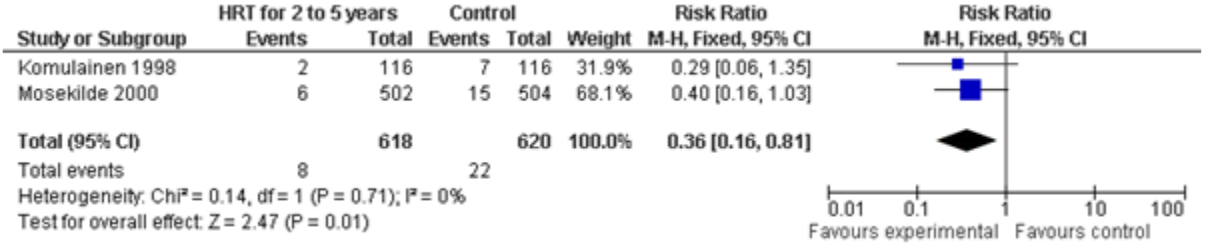


Figure 57: Risk of any fracture with current use of oestrogen plus progestogen compared to no current use of HRT

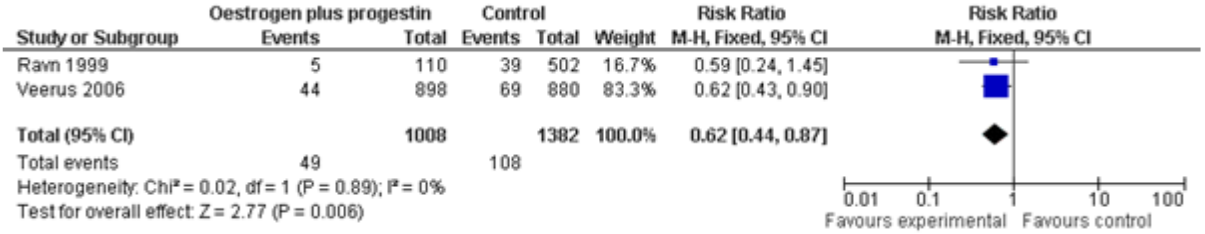


Figure 58: Risk of any non-vertebral fracture with current use of oestrogen plus progestogen compared to no current use of HRT

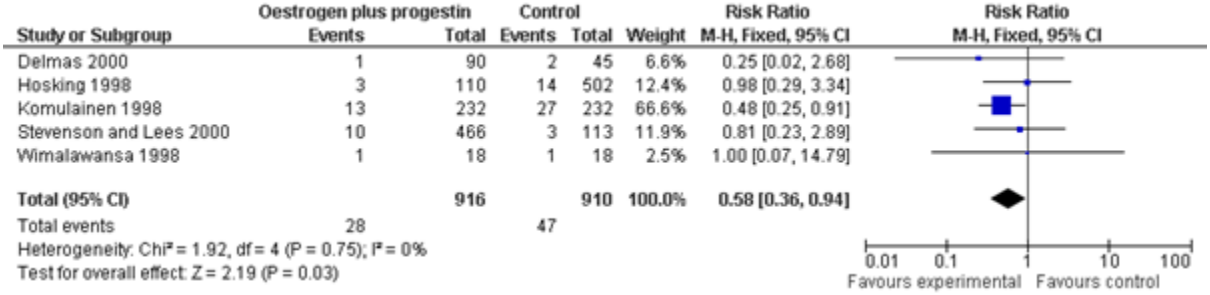


Figure 59: Risk of vertebral fracture with current use of oestrogen plus progestogen compared to no current use of HRT

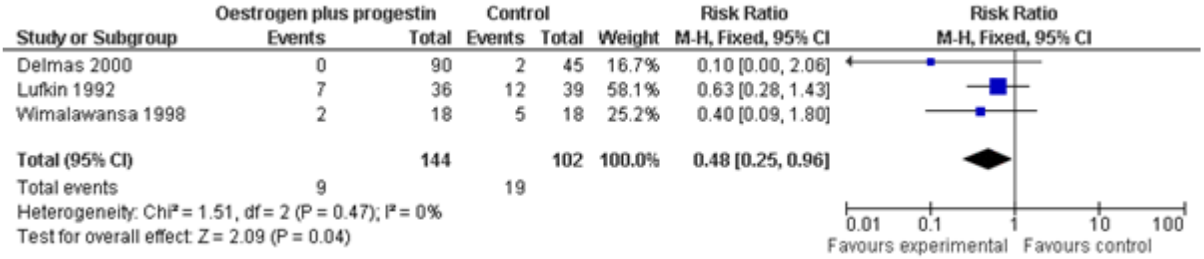
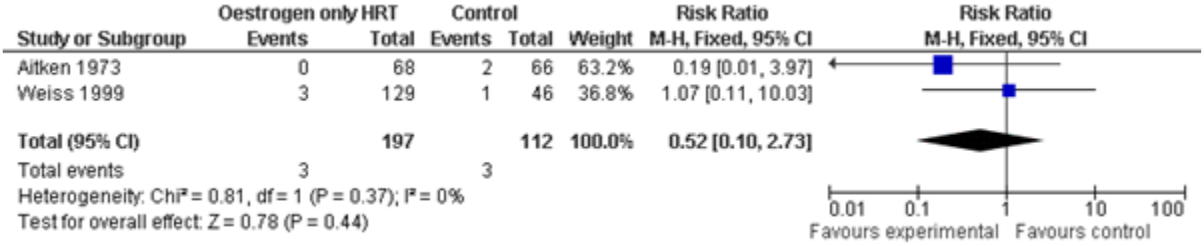


Figure 60: Risk of non-vertebral fracture with current use of oestrogen alone compared to no current use of HRT



J.7.7 Dementia

There are no forest plots for this review.

J.7.8 Loss of muscle mass (sarcopenia)

Figure 61: Change in knee extension torque (isometric) after treatment with HRT compared to no HRT

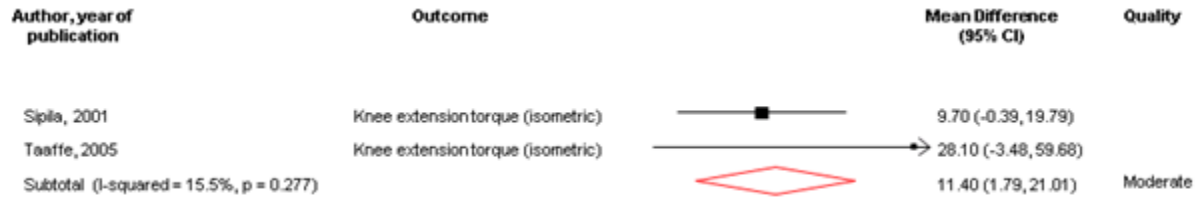


Figure 62: Change in handgrip strength after treatment with HRT compared to no HRT

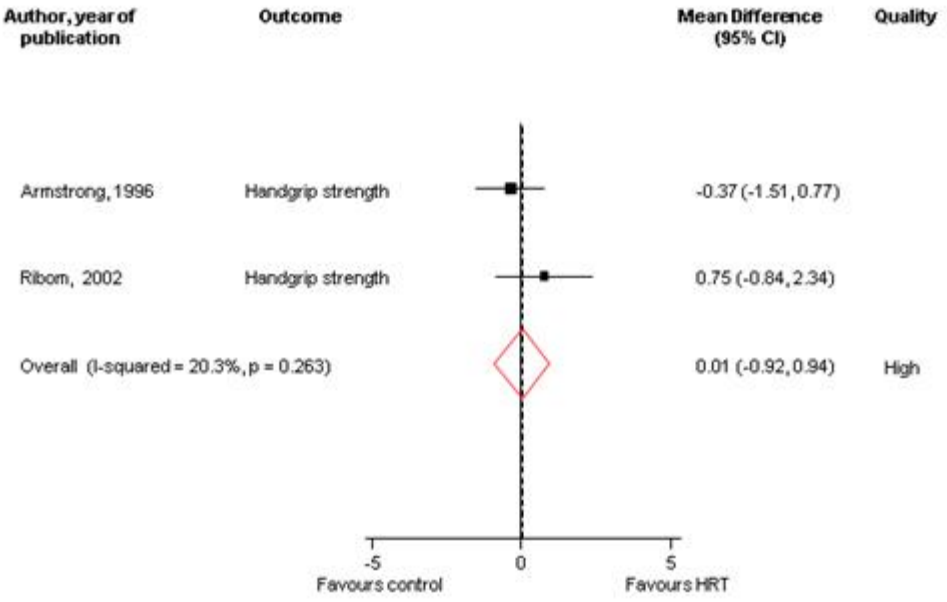
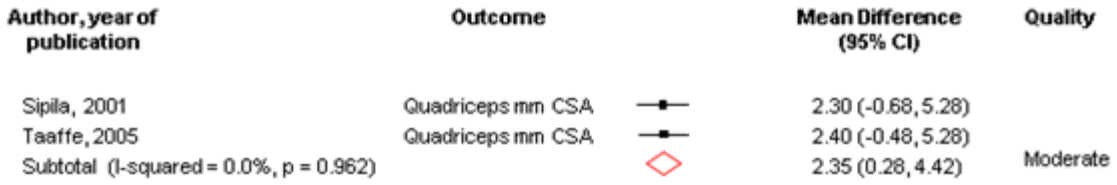


Figure 63: Change in quadriceps muscle mass after treatment with HRT compared to no HRT



J.8 Premature ovarian insufficiency

There are no forest plots for this review.

J.8.1 Diagnosis of premature ovarian insufficiency

There are no forest plots for this review.

J.8.2 Management of premature ovarian insufficiency

There are no forest plots for this review.

Appendix K: Network meta-analysis of interventions in the pharmacological and non-pharmacological treatment of short-term symptoms for women in menopause

K.1 Introduction

The results of conventional pairwise meta-analyses of direct evidence alone for the review question: "What is the most clinically effective treatment for the relief of individual menopause-related symptoms for women at menopause?" (as presented in Chapter 7 and forest plots in Appendix J) do not help to fully inform which intervention is most effective in the treatment of short term symptoms for women in menopause. The challenge of interpretation has arisen for two main reasons:

- In isolation, each pairwise comparison does not fully inform the choice between the different treatments (pharmacological and non-pharmacological) and having a series of discrete pair wise comparisons can be disjoint and difficult to interpret.
- Direct comparison of treatments of clinical interest is not available, for example comparison between different types of HRT which makes choice difficult unless based on patient preference or price.

To overcome these issues, a hierarchical Bayesian network meta-analysis (NMA) was performed. Advantages of performing this type of analysis are:

- It allows the synthesis of data from direct and indirect comparisons without breaking randomisation, to produce measures of treatment effect and ranking of different interventions. If treatment A has never been compared against treatment B head to head, but these two interventions have been compared to a common comparator, then an indirect treatment comparison can use the relative effects of the two treatments versus the common comparator. This is also the case whenever there is a path linking two treatments through a set of common comparators. All the randomised evidence is considered within the same model.
- For every intervention in a connected network, a relative effect estimate (with its 95% credible intervals) can be estimated versus any other intervention. These estimates provide a useful clinical summary of the results and facilitate the formation of recommendations based on all of the best available evidence, whilst appropriately accounting for uncertainty. Furthermore, these estimates will be used to parameterise treatment effectiveness in the de novo cost-effectiveness modelling.

Conventional fixed effects meta-analysis assumes that the relative effect of one treatment compared to another is the same across an entire set of trials. In a random effects model, it is assumed that the relative effects are different in each trial but that they are from a single common distribution and that this distribution is common across all sets of trials.

NMA requires an additional assumption over conventional meta-analysis. The additional assumption is that intervention A has the same effect on people in trials of intervention A compared to intervention B as it does for people in trials of intervention A versus intervention C, and so on. Thus, in a random effects network meta-analysis, the assumption is that intervention A has the same effect across trials of A versus B, A versus C and so on.

The terms indirect treatment comparisons, mixed treatment comparisons, and network meta-analysis are used interchangeably. We use the term NMA as the network consists of both indirect treatment comparisons (some trials have a common comparator and some do not)

and mixed treatment comparisons (with at least one closed loop, combination of direct and indirect evidence).

K.2 Methods

K.2.1 Study selection and data collection

To estimate the relative efficacy of different interventions, a NMA was conducted using all the relevant RCT evidence identified in the clinical evidence review (conventional meta-analysis). As with conventional meta-analyses, this type of analysis does not break the randomisation of the evidence, nor does it make any assumptions about the additive effects of combination interventions. The effectiveness of a particular intervention (pharmacological or non-pharmacological) was derived only from RCTs that included one of the selected treatments in a trial arm.

From the outset, we sought to minimise any clinical or methodological heterogeneity by focusing the analysis on selected studies that matched the pre specified NMA protocol (Table 1).

Table 1: Protocol of the NMA

Item	Details
Review question	**What is the most clinically effective treatment for the relief of individual menopause-related symptoms for women at menopause?**
Objective	The aim for this review will be to assess the relative effectiveness of all the main treatments used to treat short term menopause-related symptoms in five clinical categories: • vasomotor • Adverse events (discontinuation, bleeding)
Population	All women with menopause Exclusion criterion: pre-menopausal women
Stratified analyses	• Peri or postmenopausal women with uterus • Peri or postmenopausal women without uterus (hysterectomized) • Women with a history/history of breast cancer.
Interventions	***Hormonal pharmaceutical treatments:*** • oestrogen combined with progestogen/ progesterone (oral) • oestrogen combined with progestogen/ progesterone (topical – patch, cream) • oestrogen (oral) • oestrogen (topical – patch, cream) • oestrogen (depot) • progestogen alone • tissue-selective oestrogen complexes • testosterone • tibolone • bio-identical hormones licensed for use in the UK • selective oestrogen-receptor modulators (oral) ***Non-hormonal pharmaceutical treatments:*** • selective serotonin reuptake inhibitors • serotonin–noradrenaline reuptake inhibitors • gabapentin

Item	Details
	• clonidine ***Non-pharmaceutical treatments:*** • phytoestrogens (including red clover) • herbal preparations (including black cohosh) • acupuncture • lifestyle advice • relaxation therapies (including yoga) • psychological therapies including cognitive behavioural therapy
Comparisons	• All interventions listed above • Placebo
Outcomes	• The following outcomes at the end of treatment (unless end of treatment is after 26 weeks follow-up) will be included: • Frequency of vasomotor symptoms (modelled as a rate). We will not consider severity of symptoms as part of this outcome due to the variation in scores used to measure them. ○ Hot flushes and night sweats will be included. Where a study reports frequency of both hot flushes and night sweats they will be added together (by treating them as independent outcomes) to give an overall frequency of vasomotor symptoms. • Discontinuation (modelled as OR) – assuming constant probability of discontinuation after 4 weeks of treatment • Vaginal bleeding (modelled as OR) – assuming constant probability of bleeding after 4 weeks of treatment (only for women with uterus and women with a history of breast cancer) ○ Only studies which report the number of women with bleeding will be included in the analysis. Studies reporting only the total number of bleeds will not be included, as we cannot ascertain the number of bleeds per woman nor the number of women with bleeding. • For HRT study arms we will take the latest time point possible that is longer than 12 weeks and less than 26 weeks follow-up • For non-HRT study arms we will take the latest time point possible that is longer than 4 weeks and less than 26 weeks follow-up
Study design	Only RCTS will be considered for inclusion. Cross over RCTs will be only considered if provided separated data on the first period or data are reported in a linear mixed model that adjusts for treatment period and reports the coefficient for the effect of treatment versus placebo. Exclusion criteria: studies with a duration of less than 4 weeks, studies including non UK license drugs.
Population size and directness	Studies with indirect populations will be considered under the following assumptions: Mixed population studies: we will only include mixed population studies if more than 2/3 of the sample falls within the pre specified strata. For the non HRT trials: if population not specified with regards to hysterectomy status we will include studies in NMAs of women with a uterus and women without a uterus because we would assume that the efficacy of different non-HRT interventions would be exchangeable across the two populations For HRT trials: if trials have not explicitly stated history of breast cancer as an exclusion criterion, but have excluded current breast cancer as an exclusion reason, then we would assume that the trials would have excluded both types of breast cancer.

Item	Details
	If a trial does not explicitly state that women with breast cancer, a history of breast cancer, or those who had contraindications to HRT, were included/excluded, we will assume that the authors did not include these patients. If a trial including breast cancer patients has specified that premenopausal women as assessed before breast cancer diagnosis were included, then this trial would still be included, as breast cancer treatment can induce menopausal symptoms in some women. Within each population, treatment efficacy will be independent of the cause of menopause (i.e. surgical vs natural).
Search strategy	See separate document
Review strategy	• Synthesis of data • Network meta-analysis will be conducted using Winbugs codes (TSU Bristol Unit) • NMA will be based on final scores • If final scores are not reported but trials have reported changes from baseline scores, these will only be used if they also report baseline values. • We will exclude trials which reported change from baseline as a percentage • We will use the ratio of means in reporting the frequency of VSM symptoms (95% CI) • We will use the RRs (95% CI) for reporting the results of bleeding, discontinuation • We will exclude trials which reported outcomes in mean changes without measure of variation (SD, SE, 95% CI)

Therefore, 7 networks were formulated for NMA, defined by population and outcome measure:

For women in menopause with uterus:

1. Network 1: Frequency of vasomotor symptoms at the end of treatment (up to 26 weeks)
2. Network 2: Proportion of women in menopause who discontinued treatment (up to 26 weeks)
3. Network 3: Proportion of women in menopause with vaginal bleeding episodes under treatment (up to 26 weeks)

For women in menopause without uterus:

4. Network 1: Frequency of vasomotor symptoms at the end of treatment (up to 26 weeks)
5. Network 2: Proportion of women in menopause who discontinued treatment (pharmacological and non-pharmacological) (up to 26 weeks)

For women with breast cancer/history of breast cancer:

6. Network 1: Frequency of vasomotor symptoms at the end of treatment (up to 26 weeks)
7. Network 2: Proportion of women in menopause who discontinued treatment (pharmacological and non-pharmacological) (up to 26 weeks)

Limited data did not allow the formulation of a network for the outcome of vaginal bleeding for women with breast cancer or a history of breast cancer. For women at high risk of breast cancer please see specific NICE guideline: Familial breast cancer: Classification and care of people at risk of familial breast cancer and management of breast cancer and related risks in people with a family history of breast cancer: (http://www.nice.org.uk/guidance/cg164/chapter/recommendations)

K.2.2 Outcome measures

The Guideline Development Group considered the following outcomes as the most important in assessing the effectiveness of interventions (pharmacological and non-pharmacological treatments) for the relief of short term menopause related symptoms in order to inform the health economic analysis and furthermore the decision making about the most appropriate treatment for women in menopause.

- The frequency of vasomotor symptoms at the end of treatment (up to 26 weeks) was selected as an important outcome as it is the most common symptom experienced by women in menopause and is the main reason for initial consultation with the health professionals
- Discontinuation of treatment and vaginal bleeding during treatment as the most common adverse events that can lead to change of treatment plan. Vaginal bleeding was not considered as an outcome for women without uterus.

Outcome measures were calculated on an intention to treat analysis if reported by the authors unless specified (the available case analysis would be preferred compared to intention to treat analysis with imputation).

K.2.3 Methods

The GDG decided at the protocol stage to investigate a class effect for the included interventions for the prediction of short term symptom relief for menopausal women. However, due to the complexity of different HRT treatments and insufficient data available, it was decided for the case of oestrogen and progestogen, that the route of administration (oral, non-oral (transdermal) should be considered as a different level. Placebo was selected as the baseline comparator (treatment "1") for all networks. Details about the categorization of different interventions in classes used in the NMAs are given in Table 2.

Table 2: Categorization of interventions into classes for the NMAs

Classes in the NMAs	Interventions in the included trials
Placebo	Placebo
Sham acupuncture	Sham acupuncture
Acupuncture	Acupuncture
Normal living/Usual care/Attention	Waiting List
	Normal living/Usual care/Attention
Non oral oestrogen alone	Oestrogen alone transdermal Low dose
	Oestrogen alone transdermal Ave dose
	Oestrogen alone transdermal High dose
	Oestrogen vaginal Ave dose
	Oestrogen vaginal High dose
	Oestrogen nasal spray Ave dose
	Oestrogen nasal spray High dose
Oral oestrogen alone	Oestrogen alone oral Low dose
	Oestrogen alone oral Ave dose
	Oestrogen alone oral High dose
	Conjugated equine estrogen (CEE) Low dose
	Conjugated equine estrogen (CEE) Ave dose
	Conjugated equine estrogen (CEE) High dose
	Conjugated equine estrogen (CEE)
	Oestrogen valerate Ave dose

Classes in the NMAs	Interventions in the included trials
	Oestrogen valerate High dose
Non oral oestrogen plus progestogen	Oestrogen transdermal + progestogen transdermal Low dose
	Oestrogen transdermal + progestogen transdermal Ave dose
	Oestrogen transdermal + progestogen transdermal High dose
	Oestrogen transdermal + progestogen oral Ave dose
Oral oestrogen plus progestogen	Oestrogen oral + progestogen oral Low dose
	Oestrogen oral + progestogen oral Ave dose
	Oestrogen oral + progestogen oral High dose
	Oestrogen oral + progestogen oral
	Oestrogen valerate + oral progestogen Ave dose
	Conjugated equine estrogen and progestogen High dose
Progestogen alone	Progestogen alone
Conjugated oestrogens plus bazedoxifene	Conjugated oestrogens plus bazedoxifene
Tibolone	Tibolone Low dose
	Tibolone Ave dose
	Tibolone High dose
Raloxifine	Raloxifine
SSRIs	Venlafaxine
	Desvenlafaxine
	Fluoxetine
	Paroxetine
	Sertraline
	Citalopram
	5-HTP
Gabapentin	Gabapentin
Clonidine	Clonidine
Isoflanoves	Isoflavones/Genistein/soy
	Lignans
	Red clover
Chinese herbal medicine	Chinese herbal medicine
Black cohosh	Black cohosh
St John's Wort	St John's Wort
Dong Quoi	Dong Quoi
Multibotanicals	Multibotanicals
Acupuncture	Acupuncture
Cognitive behavioural therapy	Cognitive Behavioural Therapy
Relaxation	Relaxation
Hypnosis	Hypnosis
Vitamin E	Vitamin E
Evening primrose oil	Evening primrose oil
Valerian root	Valerian root

A hierarchical Bayesian network meta-analysis (NMA) was performed using the software WinBugs version 1.4.3. This is a method which preserves randomisation within trials.

Data were available on dosing for many treatments, but the sparseness of the networks meant that it was necessary to borrow strength on dosing within treatments using a multi-level model, with each dose of a treatment at the first level and the class/treatment (see Table 2) itself at the second level. Common class variance was also assessed to check if it improved model fit and reduced heterogeneity but no significant improvement of the models was ever observed. Therefore, two models for this were explored: an exchangeable dose effects model, where the pooled relative effects of different treatment doses were assumed to be randomly distributed within each treatment with a common variance; and a fixed dose effects model, where the pooled relative dose effects are assumed equal for all doses of a treatment. For treatments where dosing information was not available, the relative effect at the dose level was assumed to be equal to the treatment effect in both models.

As it is the case for ordinary pairwise meta-analysis, NMA may be conducted using either fixed or random treatment effects models. A fixed effects model typically assumes that there is no variation in relative effects across trials for a particular pairwise comparison and any observed differences are solely due to chance. For a random effects model, it is assumed that the relative effects are different in each trial but that they are from a single common distribution. The variance reflecting heterogeneity is often assumed to be constant across trials. For all the networks set up in our NMA, both models (fixed and random effect) were performed and then these models were compared based on residual deviance and deviance information criteria (DIC). The model with the smallest DIC is estimated to be the model that would best predict a replicate dataset which has the same structure as that currently observed. A small difference in DIC between the fixed and random effects models (3-5 points) implies that the better fit obtained by adding random effects does not justify the additional complexity. However, if the difference in DIC between a fixed and random effect model was less than 5 points, and the models made very similar inferences, then we would report the results from a fixed effects model results as it contains fewer parameters and is easier for clinical interpretation than the random effects model.

In a Bayesian analysis, for each parameter the evidence distribution is weighted by a distribution of prior beliefs. Markov Chain Monte Carlo (MCMC) algorithm was used to generate a sequence of samples from a joint posterior distribution of two or more random variables and is particularly well adapted to sampling the treatment effects (known as posterior distribution) of a Bayesian network. A non-informative prior distribution was used to maximise the weighting given to the data and to generate the posterior distribution for each log odds ratio (OR) or log mean ratio (MR) of interest in the networks. We used the median of the distribution as our point estimate and the centiles provided the 95% credible interval (95% CrI).

Non-informative priors were selected which were normally distributed with a mean of 0 and standard deviation of 100. However, for networks where data were sparse, informative priors generated from empirical data were used to give a more stable between-study variance. The priors for between-study variances in these instances were that it was log normally distributed with mean equal to -4.06 and precision equal to 0.4756. This allowed for more precise estimation of random effects (Turner 2012 - http://ije.oxfordjournals.org/content/41/3/818.long#T4).

One of the main advantages of the Bayesian approach is that the method leads to a decision framework that supports decision making. The Bayesian approach also allows the probability that each intervention is best for achieving a particular outcome, as well as its ranking, to be calculated.

We adapted a random effects model template for continuous and dichotomous data available from NICE DSU technical support document number 2:

http://www.nicedsu.org.uk/Evidence-Synthesis-TSD-series(2391675).htm. This model accounts for the within-study correlation between treatment effects induced by multi-arm trials.

For the analyses, a series of 40,000 burn-in simulations were run to allow the posterior distributions to convergence and then a further 60,000 simulations were run to produce the outputs. Convergence was assessed by examining the history, autocorrelation and Brooks-Gelman-Rubin plots.

Goodness of fit of the model was also estimated by using the posterior mean of the sum of the deviance contributions for each item by calculating the residual deviance and deviance information criteria (DIC). If the residual deviance was close to the number of unconstrained data points (the number of trial arms in the analysis) then the model was explaining the data at a satisfactory level. The choice of a fixed or random effects model can be made by comparing their goodness-of-fit to the data.

The outputs of the NMA were:

- Treatment specific log odds ratios (ORs) and log mean ratios (MRs) with their 95% credible intervals (CI) were generated for every possible pairs of comparisons by combining direct and indirect evidence in each network.
- The probability that each treatment is ranked best, 2nd best etc, based on the proportion of Markov chain iterations in which the log OR for an intervention is ranked best, 2nd best, etc.
- The ranking of treatments compared to placebo (presented as median rank and its 95% credible intervals)
- The assessment of probability that each intervention was the best by calculating the log OR of each drug compared to placebo, and counting the proportion of simulations of the Markov chain in which each intervention had the highest log OR, the overall ranking of interventions was also calculated according to their log ORs compared to placebo (baseline comparator).

The baseline probabilities for vasomotor and vaginal bleeding outcomes were taken from high quality observational studies. For discontinuation, the baseline probability was calculated by performing a fixed-effects meta-analysis of studies that reported placebo-arm data. Once the treatment specific probabilities for response were calculated, they were divided by the baseline probability to get treatment specific relative risk.

Differences between treatments were considered statistically significant at the 0.05 level if the 95% credible interval for the OR or the mean ratio did not cross 1.

There are two key assumptions behind a NMA, *similarity* and *consistency.*

Similarity across trials is the critical rationale for the consistency assumption to be valid as by ensuring the clinical characteristics of the trials are similar we ensure consistency in the data analysis.

More specifically, randomisation holds only within individual trials, not across the trials. Therefore, if the trials differ in terms of patient characteristics, measurement and/or definition of outcome, length of follow up across the direct comparisons (e.g. tibolone versus placebo trial differ from oestrodial alone versus. placebo trial), the similarity assumption is violated and this would bias the analysis. Potential sources of heterogeneity arising from trials of interventions for short term relief of menopause related symptoms are:

- Different population, for example, mixed populations of women with and without uterus and different duration or dosages of interventions. As described in the NMA protocol, a sensitivity analysis was performed to test the validity of the assumption of similarity of effect for HRT treatments between women with and without uterus.

- Different dosages of pharmacological treatment (categorized as low, medium and high) were grouped under the same class
- Different routes of treatment's interventions (oral, non-oral) were grouped under the same class with the exception of oestrogen and oestrogen plus progesterone that they fitted in the network as separate classes.

Consistency assumption - it is important that for a network that contains loops, the indirect comparisons are consistent with the direct comparisons. Discrepancies between direct and indirect estimates of effect may result from several possible causes. One possible cause is 'chance', and if this is the case then the NMA results are likely to be more precise as they pool together more data than conventional meta-analysis estimates alone. However, a second possible cause could be due to differences between the trials included in terms of their clinical or methodological characteristics, which would therefore raise concerns about the validity of the network.

We aimed to explore network inconsistency of direct and indirect treatment comparisons by checking whether the estimates (MR or OR) of the direct treatment comparisons (reported by the study) were within the confidence intervals of the estimates generated from the NMA, for the same treatment comparison. If the estimate (MR or OR) of a direct treatment comparison is outside the confidence intervals of the estimate generated from the NMA, it indicates inconsistency for that specific treatment comparison.

K.2.4 Studies excluded from the NMA

The studies presented in Table 3 were excluded from the networks built up for the purposes of this NMA. Detailed exclusion reasons are given per study. The main exclusion reasons were lack of information on variation of vasomotor symptoms (for example SE, or SD) (that would preclude even a pair-wise meta-analysis) and lack of information on baseline scores when only change from baseline was reported, thus preventing estimation of final scores, which was the selected way of analysing vasomotor symptoms.

Table 3: Excluded studies – reason for exclusion from NMAs

Study name	Reason for exclusion	Interventions (sample size)	Outcomes	Populations
Aguirre 2010	HRT study includes women with and without uterus but does not report separately	Oestrogen alone transdermal Low (N=22); Gaberpentin (N=23)	VMS, Discontinuation, Bleeding	Uterus, No uterus,
Al-azzawi 1997	Study only reports outcome for HRT at >26 weeks follow-up	Oestrogen alone transdermal Ave (N=134); Oestrogen alone transdermal High (N=131)	VMS, Discontinuation,	No uterus,
Allameh 2013	Study reports mean number of flushes without reporting a measure of uncertainty (SE or SD)	Conjugated equine estrogen (CEE) Ave (N=40); Gaberpentin (N=30); Gaberpentin (N=30)	VMS, Discontinuation,	Uterus, No uterus,
Archer 2003	HRT study includes women with and without uterus but does not report separately	Placebo (N=73); Oestrogen alone transdermal Low (N=75); Oestrogen alone transdermal Ave (N=73)	VMS, Discontinuation,	Uterus, No uterus,
Archer 1992	Study reports mean number of flushes without reporting a measure of uncertainty (SE or SD)	Placebo (N=25); Oestrogen alone oral Ave (N=27); Oestrogen alone oral High (N=25); Conjugated equine estrogen (CEE) Ave (N=25); Conjugated equine estrogen (CEE) High (N=26)	VMS	Uterus, No uterus,
Archer 2012	Study reports mean number of flushes without reporting a measure of uncertainty (SE or SD)	Placebo (N=73); Oestrogen alone transdermal Ave (N=75); Oestrogen alone transdermal High (N=73)	VMS, Discontinuation, Bleeding	Uterus, No uterus,
Bacchi-Modena 1997	HRT study includes women with and without uterus but does not report separately	Placebo (N=56); Oestrogen alone transdermal Ave (N=53)	VMS, Discontinuation,	Uterus, No uterus,
Bachmann 2007	Study reports mean number of flushes without reporting a measure of uncertainty (SE or SD)	Placebo (N=133); Oestrogen alone transdermal Low (N=147); Oestrogen transdermal + progestogen transdermal Low (N=145)	VMS, Discontinuation,	Uterus, No uterus,

Study name	Reason for exclusion	Interventions (sample size)	Outcomes	Populations
Barton 2010	Study reports mean number of flushes without reporting a measure of uncertainty (SE or SD)	Placebo (N=83); Citalopram (N=56)	VMS	Breast cancer/history
Bertelli 2002	Study only reports outcome for HRT at 6 weeks follow-up	Oestrogen valerate + oral progestogen Ave (N=37); Oestrogen valerate + oral progestogen Ave (N=)	VMS, Discontinuation,	Breast cancer/history
Buster 2008	Study only reports final values adjusted for baseline - unadjusted final values cannot be calculated from this	Placebo (N=76); Oestrogen alone transdermal Low (N=77); Oestrogen alone transdermal Ave (N=76); Oestrogen alone transdermal High (N=76)	VMS, Discontinuation,	Uterus, No uterus,
Carranza-Lira 2001	Median and range reported	Placebo (N=15); Conjugated equine estrogen (CEE) Ave (N=15); Clonidene (N=15)	VMS,	Uterus, No uterus,
Cohen 1999	Study reports mean number of flushes without reporting a measure of uncertainty (SE or SD)	Placebo (N=130); Oestrogen alone transdermal Ave (N=127)	VMS, Discontinuation,	Uterus, No uterus,
Crisafulli 2004	Study only reports relative effects	Placebo (N=30); Oestrogen oral + progestogen oral Ave (N=30); Isoflavones/Genistein/soy (N=30)	VMS,	Uterus, No uterus,
D'Anna 2009	Study reports mean number of flushes without reporting a measure of uncertainty (SE or SD)	Placebo (N=191); Isoflavones/Genistein/soy (N=198)	VMS,	Uterus, No uterus,
Davis 2001	Study reports % change from baseline so SE for treatment group final scores cannot be calculated	Placebo (N=27); Chinese herbal medicine (N=28)	VMS,	Uterus, No uterus,
De Aloysio 2000	Study reports mean number of flushes without reporting a measure of uncertainty (SE or SD)	Placebo (N=52); Oestrogen alone transdermal Low (N=52)	VMS,	Uterus, No uterus,
de Vrijer 2000	HRT study includes women with and without uterus but does not report separately	Placebo (N=86); Oestrogen alone transdermal Ave (N=82); Oestrogen alone transdermal High (N=86)	Discontinuation,	Uterus, No uterus,
Derman 1995	Study reports mean number of flushes without reporting a measure of uncertainty (SE or SD)	Placebo (N=42); Oestrogen oral + progestogen oral High (N=40)	VMS,	Uterus, No uterus,
Ettinger 2004	Study reports % change from baseline so SE for treatment group final scores cannot be calculated	Placebo (N=85); Red clover (N=84)	VMS,	Uterus, No uterus,
Farzaneh 2013	Numbers of participants not reported	Placebo (N=?); Evening primrose oil (N=?)	VMS,	Uterus, No uterus,
Frisk 2012	Median and range reported	Oestrogen oral + progestogen oral (N=18); Acupuncture (N=27)	VMS,	Breast cancer/history
Geller 2009	Study only reports relative effects	Placebo (N=22); Conjugated equine estrogen and progestogen High (N=23); Red clover (N=22); Black cohosh (N=22)	VMS,	Uterus, No uterus,
Good 1996	HRT study includes women with and without uterus but does not report separately	Placebo (N=91); Oestrogen alone transdermal Ave (N=88); Oestrogen alone transdermal High (N=94)	Discontinuation,	Uterus, No uterus,
Haines 2009	Study reports mean number of flushes without reporting a measure of uncertainty (SE or SD)	Placebo (N=84); Oestrogen alone oral Low (N=81)	VMS, Bleeding	Uterus, No uterus,
Hedrick 2009	HRT study includes women with and without uterus but does not report separately	Placebo (N=125); Oestrogen alone transdermal Low (N=123); Oestrogen alone transdermal Ave (N=125)	Discontinuation,	Uterus, No uterus,
Hitchcock 2012	SE for final values could not be calculated from change from baseline due to mathematical complications (square-root of negative number)	Placebo (N=58); Progestogen alone (N=75)	VMS, Bleeding	Uterus, No uterus,

Study name	Reason for exclusion	Interventions (sample size)	Outcomes	Populations
Huber 2002	Study only reports outcome for HRT at >26 weeks follow-up	Conjugated equine estrogen (CEE) Low (N=251); Tibolone High (N=250)	Discontinuation, Bleeding	Uterus,
Kim 2011	BL not reported for change from baseline so final values could not be calculated	Sham acupuncture (N=27); Acupuncture (N=27)	VMS,	Uterus, No uterus,
Lee 2007	Study reports mean number of flushes without reporting a measure of uncertainty (SE or SD)	Placebo (N=45); Oestrogen oral + progestogen oral Ave (N=45)	VMS,	Uterus, No uterus,
Lindh-Astrand 2002	Study only reports discontinuation and bleeding in one trial arm	Oestrogen alone oral High (N=15); Excercise (N=15)	VMS,	Uterus, No uterus,
Loibl 2007	Median and range reported	Venlafaxine (N=40); Clonidene (N=40)	VMS,	Uterus, No uterus,
Loprinzi 1994	Study only reports outcome for HRT at 4 weeks follow-up	Placebo (N=NA); Progestogen alone (N=NA)	Bleeding	Breast cancer/history
Loprinzi 2000	Study reports median change from baseline	Placebo (N=72); Venlafaxine (N=78)	VMS,	Uterus, No uterus,
Loprinzi 2002	Median and range reported	Placebo (N=62); Venlafaxine (N=66); Fluoxitene (N=)	VMS,	Breast cancer/history
Loprinzi 2009	Study reports median change from baseline	Placebo (N=320); Gaberpentin (N=314)	VMS,	Breast cancer/history
Meuwissen 2001	Study reports number of bleeds rather than number of women with bleeds	Oestrogen oral + progestogen oral High (N=40); Oestrogen oral + progestogen oral High (N=40)	Bleeding	Uterus,
Nahas 2007	Study only reports outcome for HRT at >26 weeks follow-up	Placebo (N=66); Isoflavones/Genistein/soy (N=68)	VMS, Discontinuation,	Uterus, No uterus,
Notelovitz 2000	HRT study includes women with and without uterus but does not report separately	Placebo (N=80); Oestrogen alone oral Low (N=80); Oestrogen alone oral Ave (N=77); Oestrogen alone oral High (N=74)	Discontinuation, Bleeding	Uterus, No uterus,
Rovati 2000	HRT study includes women with and without uterus but does not report separately	Placebo (N=57); Oestrogen alone transdermal Low (N=54); Oestrogen alone transdermal Ave (N=54); Oestrogen alone transdermal High (N=)	Discontinuation, Bleeding	Uterus, No uterus,
Rozenbaum 2002	Study reports mean number of flushes without reporting a measure of uncertainty (SE or SD)	Placebo (N=10); Oestrogen nasal spray Ave (N=9); Oestrogen nasal spray High (N=)	VMS, Discontinuation, Bleeding	Uterus, No uterus,
Scharf 2007	Study reports mean number of flushes without reporting a measure of uncertainty (SE or SD)	Placebo (N=18); Conjugated equine estrogen (CEE) Low (N=20)	VMS, Discontinuation,	Uterus, No uterus,
Simbalista 2010	SE not reported / SE units not reported	Placebo (N=48); Lignans (N=72)	VMS,	Uterus, No uterus,
Simon 2001	HRT study includes women with and without uterus but does not report separately	Placebo (N=137); Conjugated equine estrogen (CEE) Ave (N=147)	VMS, Discontinuation, Bleeding	Uterus, No uterus,
Simon 2007	Study reports mean number of flushes without reporting a measure of uncertainty (SE or SD)	Placebo (N=54); Oestrogen alone transdermal Low (N=54); Oestrogen alone transdermal Ave (N=)	VMS, Discontinuation,	Uterus, No uterus,
Speroff 1996	Study reports mean number of flushes without reporting a measure of uncertainty (SE or SD)	Placebo (N=54); Oestrogen alone transdermal Low (N=54)	VMS,	Uterus, No uterus,
Speroff 1996	Study reports mean number of flushes without reporting a measure of uncertainty (SE or SD)	Placebo (N=108); Oestrogen alone transdermal Low (N=113)	VMS, Discontinuation,	Uterus, No uterus,
Speroff 2003	Study reports mean number of flushes without reporting a measure of uncertainty (SE or SD)	Placebo (N=108); Oestrogen vaginal Ave (N=113); Oestrogen vaginal High (N=112)	VMS,	Uterus, No uterus,
Speroff 2004	Study reports mean number of flushes without reporting a measure of uncertainty (SE or SD)	Placebo (N=54); Oestrogen vaginal Ave (N=54); Oestrogen vaginal High (N=)	VMS,	Uterus, No uterus,

Study name	Reason for exclusion	Interventions (sample size)	Outcomes	Populations
Speroff 2006 Stevens 2000	Study reports mean number of flushes without reporting a measure of uncertainty (SE or SD)	Placebo (N=48); Oestrogen alone transdermal Low (N=72)	VMS, Bleeding	Uterus, No uterus,
	HRT study includes women with and without uterus but does not report separately	Placebo (N=16); Conjugated equine estrogen (CEE) Ave (N=16)	VMS, Bleeding	Uterus, No uterus,
Studd 1996	Study compares O alone in women with uterus	Oestrogen alone transdermal Ave (N=17); Conjugated equine estrogen (CEE) Ave (N=17)	VMS, Discontinuation,	Uterus,
Thomson 1977	HRT study includes women with and without uterus but does not report separately	Placebo (N=87); Oestrogen alone oral High (N=90)	VMS,	Uterus, No uterus,
Upmalis 2000	Study reports mean number of flushes without reporting a measure of uncertainty (SE or SD)	Placebo (N=72); Isoflavones/Genistein/soy (N=68)	VMS,	Uterus, No uterus,
Utian 2004	HRT study includes women with and without uterus but does not report separately	Placebo (N=87); Conjugated equine estrogen (CEE) Low (N=56); Conjugated equine estrogen (CEE) High (N=87)	VMS, Discontinuation,	Uterus, No uterus,
Utian 2004	Study reports mean number of flushes without reporting a measure of uncertainty (SE or SD)	Placebo (N=93); Conjugated equine estrogen (CEE) Low (N=93); Conjugated equine estrogen (CEE) Low (N=); Conjugated equine estrogen (CEE) High (N=)	VMS,	Uterus, No uterus,
Von Holst 2000	Study reports mean number of flushes without reporting a measure of uncertainty (SE or SD)	Placebo (N=51); Oestrogen alone transdermal Ave (N=51)	VMS, Discontinuation,	Uterus, No uterus,
Washburn 1999	Study only reports outcome for HRT at 6 weeks follow-up	Placebo (N=NA); Isoflavones/Genistein/soy (N=NA)	VMS,	Uterus, No uterus,
Wren 1986	Study reports mean number of flushes without reporting a measure of uncertainty (SE or SD)	Placebo (N=56); Clonidene (N=54)	VMS,	Uterus, No uterus,

K.2.5 Content of networks

The following section describes the composition of networks for each outcome per population. In order to be included in the analysis, a fundamental requirement is that each treatment is connected directly or indirectly to every other intervention in the network. By that meaning there is a path connecting each treatment to every other. For each outcome for each population subgroup, a diagram of the evidence network was produced in Figure 64-70 and presented in the next section .

The thickness of the line connecting two interventions in the graphs indicates the number of included studies in which the interventions connected by the line were compared directly (the thicker the line the more trials were included for this comparison). The size of the circle under each intervention in the graphs reflects the number of participants included in the trials who received the specific intervention (the bigger the circle the more participants were included for this comparison).

K.2.5.1 Women with and without uterus

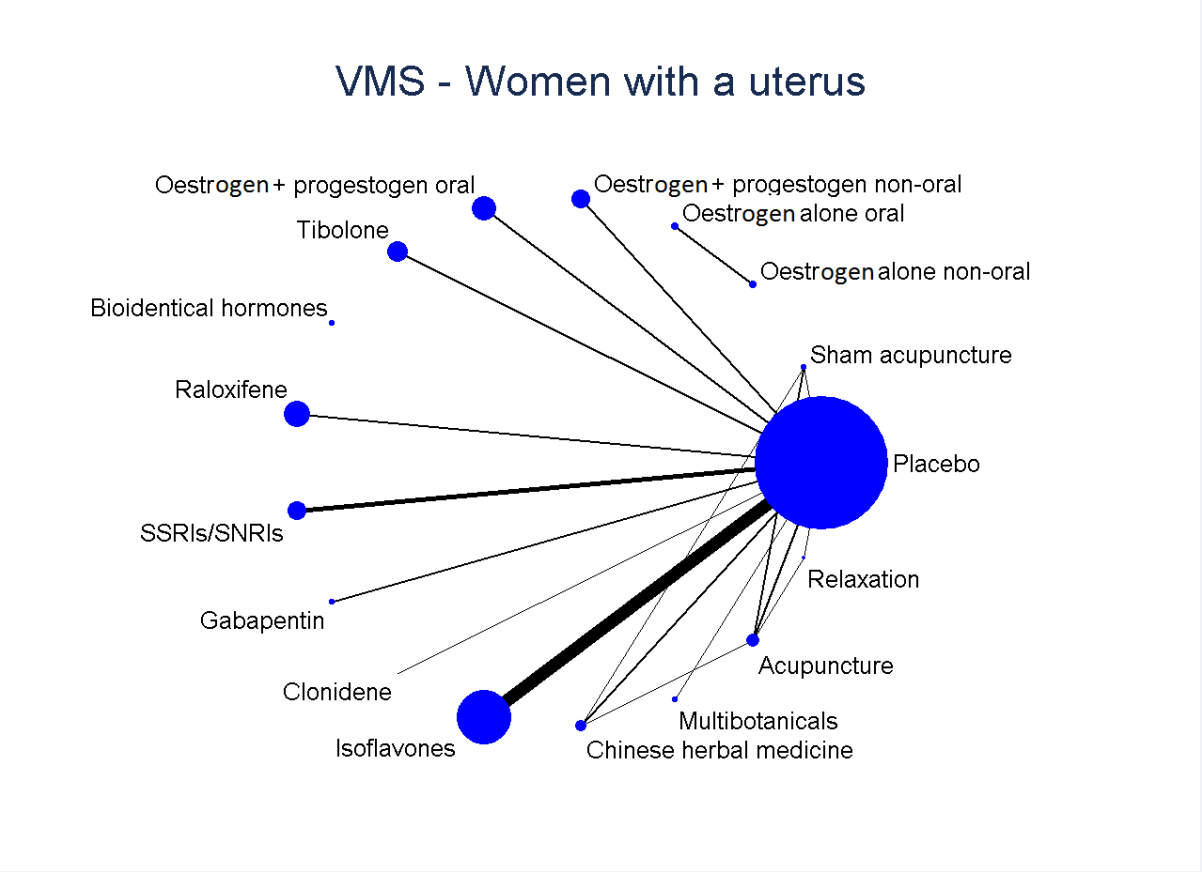
Vasomotor symptoms

As a first step we built up the networks for the outcome of vasomotor symptoms separately for the population of women with uterus and for women without uterus.

The network for the population of women with uterus included all the clinically relevant interventions for the relief of short term menopause symptoms that would be helpful for the group's decision making (Figure 64). On the other hand, for the network of women without uterus, the treatment of oestrogen alone, which is the most common treatment for women

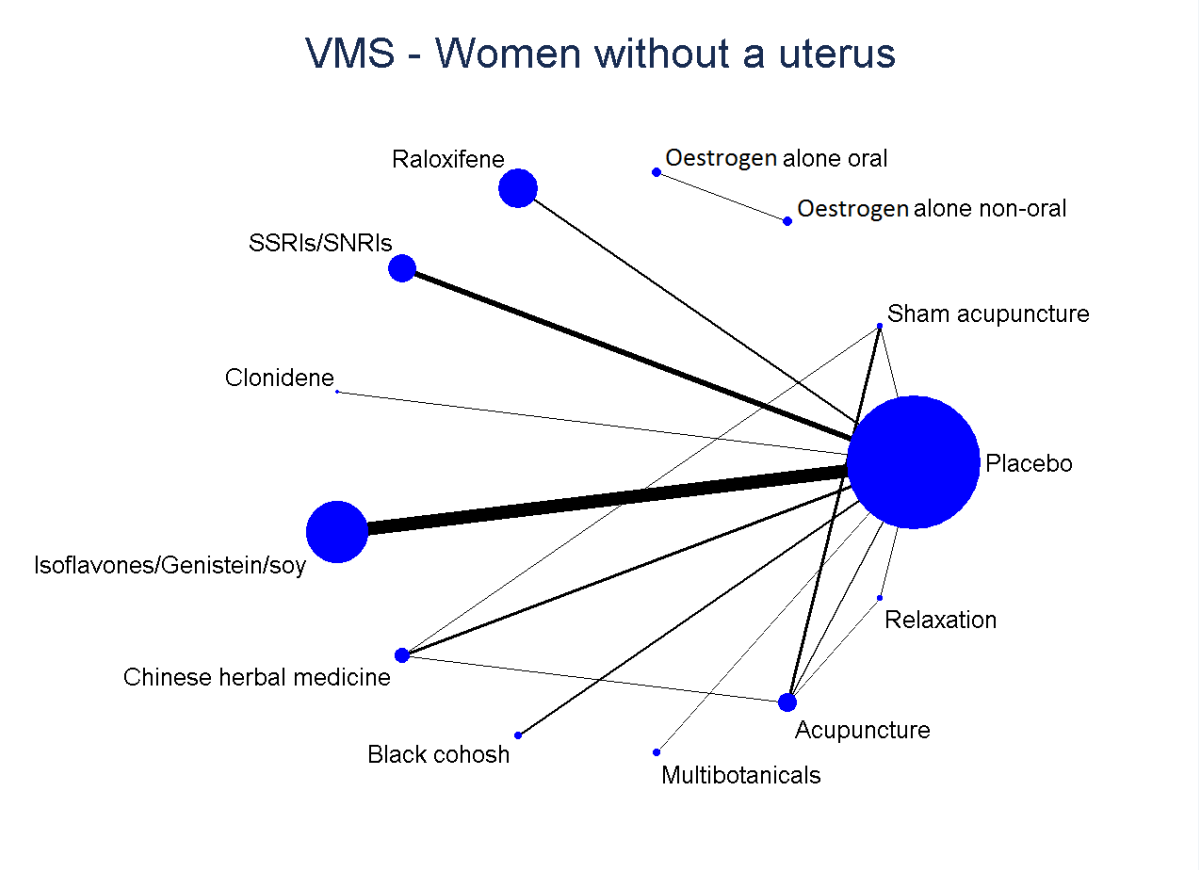
without uterus in the UK, did not connect with other interventions in the network, therefore it was excluded (Figure 65).

Figure 64: Network for the outcome of vasomotor symptoms for the population of women with uterus only



2 interventions are not connected (oestrogen alone non-oral, oestrogen alone oral) and are therefore not compared in the NMA.

Figure 65: Network for the outcome of vasomotor symptoms for the population of women without uterus only



2 interventions are not connected (oestrogen alone non-oral, oestrogen alone oral) and are therefore not compared in the NMA.

After discussion with the GDG about the potential limitations of interpretation of results from the network of women without uterus due to exclusion of oestrogen alone, we attempted to fit all the data from both networks (including also mixed population studies of both women with and without uterus) in one general network for the outcome of vasomotor symptoms. However, the model failed to converge and the main reason for this was the wide variability of studies and the heterogeneity of populations included. Two main conclusions were made:

- This limitation of the data analysis of including all populations further confirmed our prior decision to separate the networks for the populations of women with and without uterus.
- Further assumptions will be made in the HE modelling to address the weakness of the results from the network of women without uterus to include the intervention of oestrogen alone.

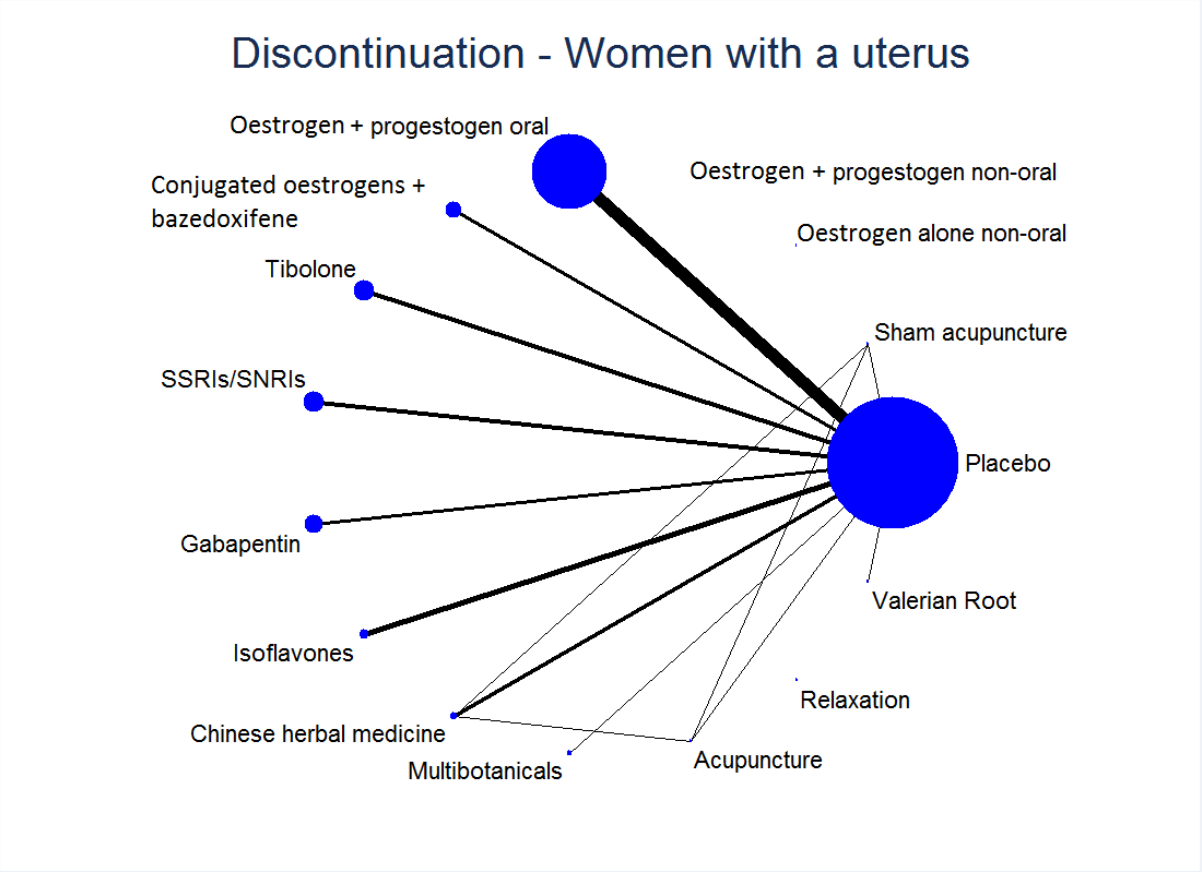
K.2.5.2 Women with uterus

Discontinuation of treatment

The network for the outcome of treatment's discontinuation for women with uterus is presented in the following graph (Figure 66). 11 classes of interventions (oral oestrogen plus progesterone, conjugated oestrogens plus bazedoxifene, tibolone, SSRIs/SNRIs, gabapentin, isoflavones, chinese herbal medicines, multibotanicals, acupuncture, valerian root and sham acupuncture) were connected to the network. Most were compared directly to placebo and not within each other. Most of the evidence fitted in this network came from the trials comparing oral oestradial plus progesterone versus placebo.

After exclusion of studies that could not be included in the network, no potential for inconsistency was possible as no “indirect” evidence was available for any comparison.

Figure 66: Network of women with uterus for the outcome of discontinuation of treatment



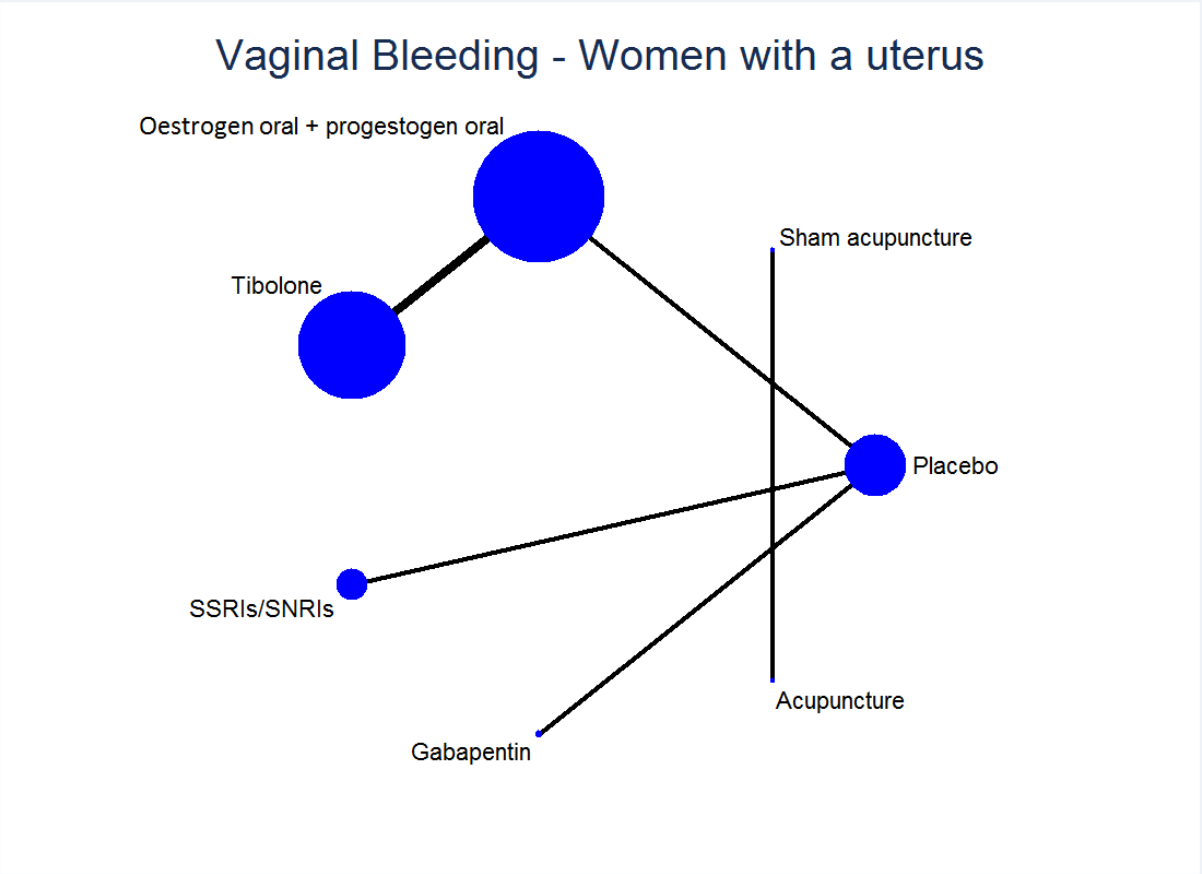
3 interventions are not connected (oestrogen alone non-oral, oestrogen alone oral, relaxation) and is therefore not compared in the NMA.

Vaginal bleeding

The network for the outcome of bleeding for women with uterus is presented in the following graph (Figure 67). 4 classes of interventions (oral oestrogen plus progesterone, tibolone, gabapentin) were connected to the network. Tibolone was not compared directly to placebo, but was connected to the network through oral oestrogen plus progesterone. Most of the evidence fitted in this network came from the trials comparing oral oestradial plus progesterone versus tibolone.

There was no potential to assess inconsistency as no direct evidence between treatments was available to compare with “indirect” evidence.

Figure 67: Network of women with uterus for the outcome of bleeding



2 interventions are not connected (sham acupuncture, acupuncture) and are therefore not compared in the NMA.

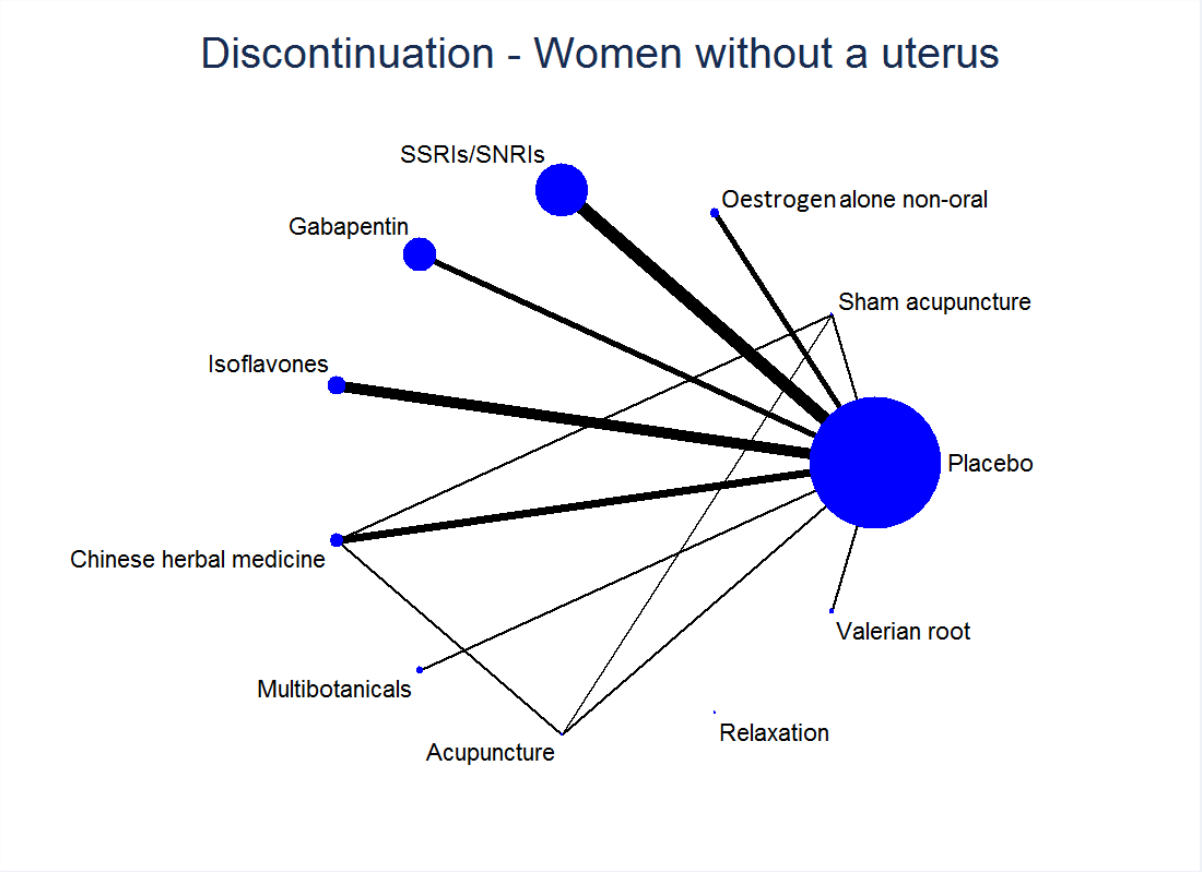
K.2.5.3 Women without uterus

Discontinuation of treatment

The network for the outcome of treatment's discontinuation for women without uterus is presented in the following graph (Figure 68). 9 interventions (SSRIs, non-oral oestrogen alone, sham acupuncture, acupuncture, valerian root, isoflavones, gabapentin, Chinese herbal medicines, multibotanicals) were connected in the network and all of other treatments were compared directly to placebo and not with each other, except in one study which compared sham acupuncture, acupuncture, Chinese herbal medicine and placebo.

After exclusion of studies that could not be included in the network, no potential for inconsistency was possible as no "indirect" evidence was available for any comparison.

Figure 68: Network of women without uterus for the outcome of discontinuation of treatment



1 intervention was not connected (relaxation) and therefore is not compared in the NMA.

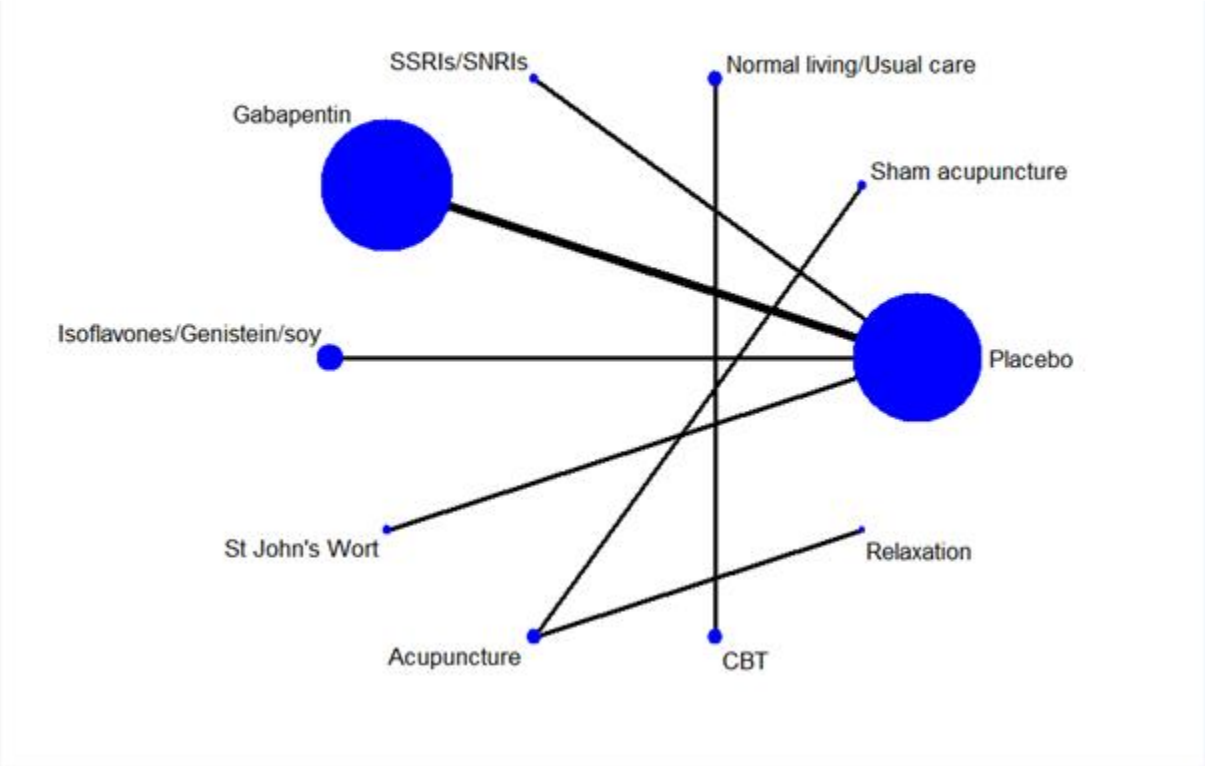
K.2.5.4 Women with breast cancer/history of breast cancer

Vasomotor symptoms

The network for the outcome of vasomotor symptoms for women with breast cancer/history of breast cancer is presented in the following graph (Figure 69). 4 classes of interventions (gabapentin, isoflavones, St John's Wort, SSRIs/SNRIs) were connected to the network. 5 other interventions were not connected to the network so could not be included in the NMA (CBT, normal living/usual care, sham acupuncture, acupuncture, relaxation). Most of the evidence fitted in this network came from the trials comparing gabapentin versus placebo.

There was no potential to assess inconsistency as no direct evidence between treatments was available to compare with "indirect" evidence.

Figure 69: Network for the outcome of vasomotor symptoms for the population of women with breast cancer/history of breast cancer

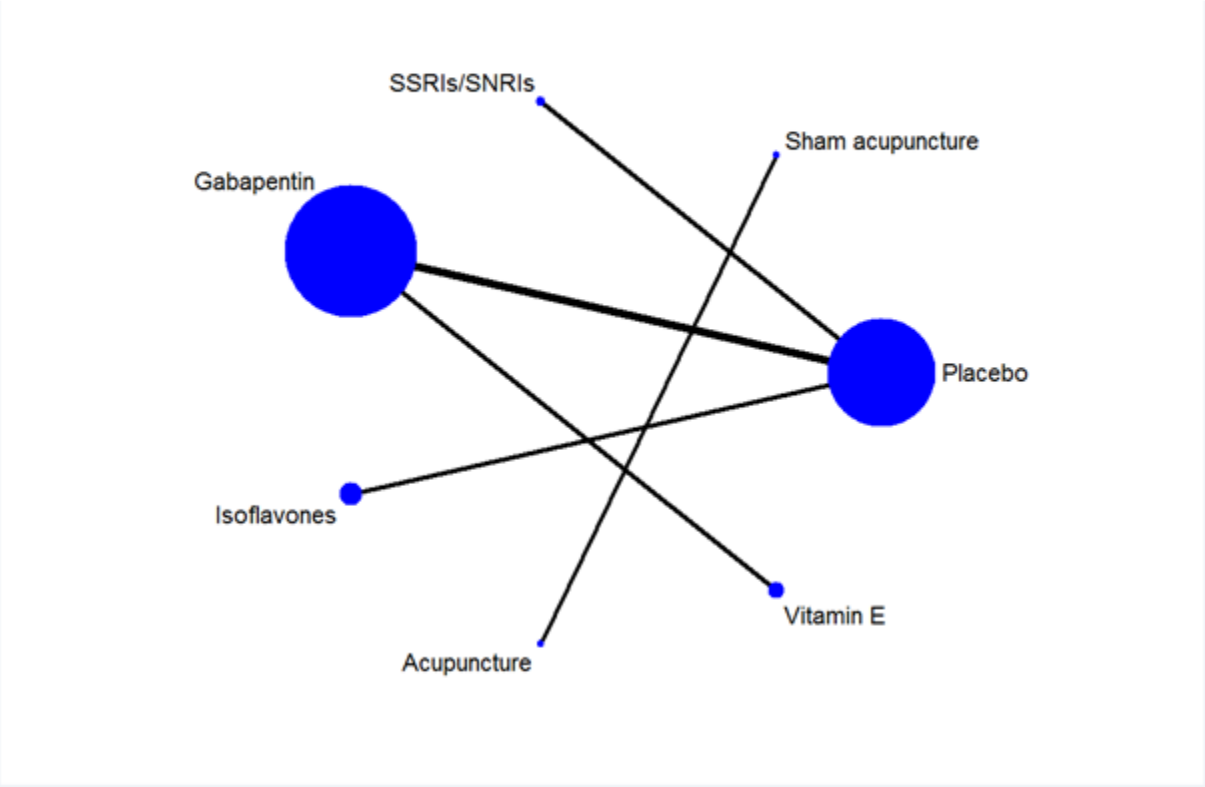


5 interventions are not connected (CBT, normal living/usual care, sham acupuncture, acupuncture, relaxation) and are therefore not compared in the NMA.

Discontinuation of treatment

The network for the outcome of discontinuation of treatment women with breast cancer/history of breast cancer is presented in the following graph (Figure 70). 4 classes of interventions (SSRIs/SNRIs, gabapentin, isoflavones, vitamin E) were connected to the network. Vitamin E was connected to the network through gabapentin. Most of the evidence fitted in this network came from the trials comparing gabapentin versus placebo.

Figure 70: Network for the outcome of discontinuation of treatment for the population of women with breast cancer/history of breast cancer



2 interventions are not connected (sham acupuncture, acupuncture) and are therefore not compared in the NMA.

K.3 NMA Results

K.3.1.1 Studies included in the NMA

The studies presented in Table 4 were included in the NMA networks. Risk of bias, time that the outcome was reported for use in the NMA, baseline age of participants, dose and frequency of intervention, and number of participants are shown.

Table 4: Included studies – Characteristics, outcomes and populations

Study name	Risk of bias	Time of outcome reported (weeks)	Age (range or mean (SD))	Sample size per group	Description of treatment	Outcomes	Populations
Al-Akoum 2009	Low	12.9	Placebo: 53.4 (4.8); St John's Wort: 54.0 (5.8)	Placebo (N=25); St John's Wort (N=22)	Placebo (TID); Ethanolic St John's wort extract, 900mg (300mg TID)	VMS,	Breast cancer/history
Al-Azzawi 1999	High	25.8	O+P oral: 53.4 (5.0); Tibolone: 54.2 (4.7)	Oestrogen oral + progestogen oral High (N=116); Tibolone High (N=191)	2mg micronized oestrogen valerate and 0.7 mg norethisterone; 2.5mg/day tibolone	Bleeding	Uterus,
Albertazzi 1998	Moderate	12	48-61	Placebo (N=53); Isoflavones/Genistein/soy (N=51)	60g of placebo (casein) daily: 40g of proteins but no isoflavones: powder form in sachets of 30g each; 60g of isolated soy protein daily: contains 40g of proteins and 76mg of isoflavones (aglycone units) - powder form in sachets of 30g each	Discontinuation,	Uterus, No uterus,
Baber 1999	High	12	45-65	Placebo (N=26); Isoflavones/Genistein/soy (N=25)	Placebo; 40mg/day phytoestrogen	VMS,	Uterus, No uterus,
Burke 2003	High	25.8	45-55	Placebo (N=70); Isoflavones/Genistein/soy (N=76); Isoflavones/Genistein/soy (N=65)	25 g of soy protein, alcohol washed to remove isoflavones (≤ 4 mg/day) (placebo); 25 g of soy protein with a medium dose of isoflavones (42 mg/day); 25 g of soy protein with a higher dose of isoflavones (58 mg/day)	VMS,	Uterus, No uterus,
D'Anna 2009	High	25.8	50-70	Placebo (N=191); Isoflavones/Genistein/soy (N=198)	Placebo; 54mg/day genestein	VMS,	Uterus, No uterus,
Endrikat 2007	Moderate	12	52-65	Placebo (N=162); Oestrogen valerate + oral progestogen Ave (N=162)	Placebo; 2mg dienogest/1mg estradiol valerate	Discontinuation,	Uterus,

Study name	Risk of bias	Time of outcome reported (weeks)	Age (range or mean (SD))	Sample size per group	Description of treatment	Outcomes	Populations
Evans 2010	Low	12	Placebo: 53.39 (5.05); Genestein: 53.50 (4.44)	Placebo (N=42); Isoflavones/Genistein/soy (N=42)	Placebo; 30mg/d genistein	Discontinuation,	Uterus, No uterus,
Faure 2002	V high	16	53-54	Placebo (N=36); Isoflavones/Genistein/soy (N=39)	2x2 capsules of placebo (cellulose microcrystalline/sodium magnesium stearic) per day; 2x2 capsules of soy isoflavone extract per day	VMS,	Uterus, No uterus,
Ferrari 2009	High	12	40-65	Placebo (N=95); Isoflavones/Genistein/soy (N=85)	Placebo; 80mg/day phytoestogen (corresponding to 60mg of genistein)	VMS, Discontinuation,	Uterus, No uterus,
Freedman 2010	Low	4	50-52	Placebo (N=12); 5-HTP (N=12)	Placebo; 150 mg of 5-hydroxytroptophan given daily	VMS,	Uterus, No uterus,
Freedman 2011	Low	8	52-53	Placebo (N=14); Citalopram (N=12)	Placebo; 10-20mg/day Escalitopram	VMS,	Uterus, No uterus,
Freeman 2011	Low	8	42-56	Placebo (N=101); Citalopram (N=104)	Placebo; 10 to 20 mg of escitalopram daily	VMS, Discontinuation,	Uterus, No uterus,
Garcia 2010	Moderate	12	45-60	Placebo (N=39); Multibotanicals (N=120)	Placebo; Mung legume extract combined with Eucommia ulmoides	VMS, Discontinuation,	Uterus, No uterus,
Gordon 2006	Low	4	40-65	Placebo (N=41); Sertraline (N=46)	Placebo; 50mg/day Setraline	VMS,	Uterus, No uterus,
Grady 2007	Moderate	6	50	Placebo (N=49); Sertraline (N=50)	Placebo; 50mg/day Setraline	VMS,	Uterus, No uterus,
Guttuso 2003	Moderate	17	53	Placebo (N=29); Gaberpentin (N=54)	Identically appearing placebo capsules; 900mg capsules of gabapentin/day	Discontinuation, Bleeding	Uterus, No uterus,
Hachul 2011	Moderate	17.2		Placebo (N=19); Isoflavones/Genistein/soy (N=19)	Placebo; 80mg/day isoflavone	VMS,	Uterus, No uterus,
Hammar 2007	Moderate	25.8	45-65	Tibolone High (N=285); Oestrogen oral + progestogen oral Ave (N=284)	2.5 mg tibolone; 1 mg 17b oestrogen plus 0.5 mg norethisterone acetate daily for 48 weeks	Bleeding	Uterus,
Joffe 2014	Low	8	Placebo: 54.3 (3.8); Venlafaxine: 54.9 (4.1)	Placebo (N=146); Oestrogen oral + progestogen oral Low (N=96); Venlafaxine (N=97)	Placebo; Oestrogen oral + progestogen oral Low (0.5mg per day O + 10mg/day medroxyprogesterone if women had uterus); Venlafaxine (37.5mg/day for 1 week then 75mg/day for 7 weeks)	VMS, Discontinuation, Bleeding	Uterus, No uterus,
Kimmick 2006	Low	12	52	Placebo (N=29); Sertraline (N=33)	Placebo; 50mg/day sertraline	VMS, Discontinuation,	Breast cancer/history
Knight 1999	Low	12	40-65	Placebo (N=12); Isoflavones/Genistein/soy (N=12);	Placebo; 1 tablet (40 mg) of Promensil daily; 4 tablets (160 mg) of Promensil daily	VMS,	Uterus, No uterus,

Study name	Risk of bias	Time of outcome reported (weeks)	Age (range or mean (SD))	Sample size per group	Description of treatment	Outcomes	Populations
				Isoflavones/Genistein/soy (N=12)			
Knight 2001	Moderate	12	40-65	Placebo (N=12); Isoflavones/Genistein/soy (N=12)	Isoflavone-free, isocaloric casein-based beverage; Dietary beverage in the form of soy powder containing isoflavones, daily dose of 4 scoops or 60g	VMS,	Uterus, No uterus,
Landgren 2005	Moderate	12	51-53	Placebo (N=58); Tibolone Low (N=73); Tibolone Ave (N=68); Tibolone High (N=57)	Placebo; Daily oral 1.25mg tibolone; Daily oral 2.5mg tibolone; Daily oral 5.0mg tibolone	VMS, Discontinuation,	Uterus,
Lin 2011	Moderate	16	52	Placebo (N=62); Oestrogen oral + progestogen oral Ave (N=187)	Oral placebo once daily; Oral 2mg drospirenone/1mg estradiol (DRSP/E2) once daily	VMS, Discontinuation,	Uterus,
Lipovac 2011	Moderate	12.9	40 and over	Placebo (N=60); Red clover (N=53)	Placebo; 40mg red clover	VMS,	Uterus, No uterus,
Mirabi 2013	Moderate	8	45-55	Placebo (N=38); Valerian root (N=38)	Placebo; Valerian root (225mg, 3 times per day)	Discontinuation,	Uterus, No uterus,
Nedeljkovic 2013	Low	24	51-54	Placebo (N=10); Sham acupuncture (N=10); Chinese herbal medicine (N=10); Acupuncture (N=10)	Placebo; Sham acupunture; Chinese herbal medicine (Zhi Mu 14 3g/d); Acupuncture	VMS,	Uterus, No uterus,
Nir 2007	Moderate	7	57	Sham acupuncture (N=17); Acupuncture (N=12)	Placebo acupuncture, 9 sessions twice weekly during the first 2 weeks, once weekly for the remaining 5 weeks ; Active acupuncture, 9 sessions twice weekly during the first 2 weeks, once weekly for the remaining 5 weeks	VMS,	Uterus, No uterus,
Notelovitz 2000	High	12	40-70	Placebo (N=53); Oestrogen transdermal + progestogen transdermal Low (N=55); Oestrogen transdermal + progestogen transdermal Ave (N=59); Oestrogen transdermal + progestogen	Transdermal placebo patch; Transdermal patch 50mcg/d estradiol plus combination patch 50mcg/d estradiol plus 140 mcg/d of norethindrone acetate; Transdermal patch 50mcg/d estradiol plus combination patch 50mcg/d estradiol plus 250 mcg/d of norethindrone acetate; Transdermal patch 50mcg/d estradiol plus combination patch 50mcg/d estradiol plus 400 mcg/d of norethindrone acetate	VMS,	Uterus,

Study name	Risk of bias	Time of outcome reported (weeks)	Age (range or mean (SD))	Sample size per group	Description of treatment	Outcomes	Populations
				transdermal High (N=53)			
Palacios 2004	High	8.6	58	Placebo (N=159); Raloxifene (N=161); Raloxefine (N=167)	Placebo; 60mg/day raloxifene (RLX); 60mg/day raloxifene every other day for 1st 2 months, followed by 60mg/d for remainder of study (SDE)	VMS,	Uterus, No uterus,
Panay 2009	Low	12	55	Placebo (N=201); Oestrogen oral + progestogen oral Low (N=194); Oestrogen oral + progestogen oral Low (N=182)	Placebo; 0.5mg NETA + 0.1mg oestrogen; 0.5mg NETA + 0.25mg oestrogen	Discontinuation,	Uterus,
Pandya 2005	Moderate	8	54	Placebo (N=137); Gaberpentin (N=144); Gaberpentin (N=139)	Placebo; 300mg/day gabapentin; 900mg/day gabapentin	VMS, Discontinuation,	Breast cancer/history
Penotti 2003	High	25.8	45-60	Placebo (N=34); Isoflavones/Genistein/soy (N=28)	Two 0.5g of talc and 0.5g of microcrystalline cellulose placebo tablets per day (placebo); Two 72 mg of soy-derived isoflavones tablets per day	VMS,	Uterus, No uterus,
Pinkerton 2009	Moderate	12	40-65	Placebo (N=66); Bazadoxifene + oestrogen (N=133); Bazadoxifene + oestrogen (N=133)	Placebo; Bazedoxifene 20mg with conjugated estrogen 0.45mg once daily; Bazedoxifene 20mg with conjugated estrogen 0.625mg once daily	Discontinuation,	Uterus,
Pinkerton 2012	Moderate	12	45 and over	Placebo (N=190); Desvenlafaxine (N=200)	Placebo; Desvenlafaxine 100mg/d	Discontinuation,	Uterus, No uterus,
Pinkerton 2013	Moderate	24	54	Placebo (N=294); Gaberpentin (N=299)	Placebo; Gabapentin (600mg am/1200 mg pm)	Discontinuation,	Uterus, No uterus,
Rotem 2007	Moderate	12.9	55	Placebo (N=25); Black cohosh (N=25)	Placebo; Phyto-Female Complex (standardized extracts of black cohosh, dong quai, milk thistle, red clover, American ginseng, chaste-tree berry) daily	VMS,	Uterus, No uterus,
Schurmann 2004	High	16	45-65	Placebo (N=61); Oestrogen oral + progestogen oral Ave (N=57); Oestrogen oral + progestogen oral Ave (N=55); Oestrogen oral +	Placebo; 1mg estradiol and 1mg drospirenone; oral tablet once daily; 1mg estradiol and 2mg drospirenone; oral tablet once daily; 1mg estradiol and 3mg drospirenone; oral tablet once daily	Discontinuation,	Uterus,

Study name	Risk of bias	Time of outcome reported (weeks)	Age (range or mean (SD))	Sample size per group	Description of treatment	Outcomes	Populations
				progestogen oral Ave (N=52)			
Shahnazi 2013	Low	8	45-60	Placebo (N=42); Black cohosh (N=42)	Placebo; Black cohosh	VMS,	Uterus, No uterus,
Speroff 1996	Moderate	12	49	Placebo (N=52); Oestrogen alone transdermal Low (N=54); Oestrogen alone transdermal Low (N=53)	One placebo transdermal system applied weekly; Two placebo transdermal system applied weekly; One 7-day transdermal system which delivered 0.02mg of 17beta-estradiol/day applied every week	Discontinuation,	No uterus,
Stearns 2013	High	6	35-64	Placebo (N=56); Paroxitene (N=58); Paroxitene (N=51)	Placebo; 12.5mg/d paroxetine; 25mg/d paroxetine	Discontinuation,	Uterus, No uterus,
Stevenson 2010	Moderate	13	54	Placebo (N=127); Oestrogen oral + progestogen oral Low (N=124); Oestrogen oral + progestogen oral Ave (N=62)	Placebo; 0.5mg/2.5mg CEE daily; 1mg/5mg CEE daily	VMS, Discontinuation, Bleeding	Uterus,
van de Weijer 2002	Moderate	12	49-65	Placebo (N=16); Isoflavones/Genistein/soy (N=16)	Placebo; 80 mg isoflavones	VMS, Discontinuation,	Uterus, No uterus,
Van Patten 2002	Moderate	12	Placebo: 54.9 (6.5); Isoflavones: 55.5 96.3)	Placebo (N=79); Isoflavones/Genistein/soy (N=78)	Rice beverage; 0.90mg isoflavones beverage	VMS, Discontinuation,	Breast cancer/history
Verhoeven 2005	High	12	45-65	Placebo (N=64); Isoflavones/Genistein/soy (N=60)	2,000 mg/day olive oil (placebo); 50mg/day isoflavone	VMS,	Uterus, No uterus,
Wyon 2004	Low	12	48-63	Sham acupuncture (N=13); Acupuncture (N=15)	14 half-hour sham acupuncture treatments; 14 half-hour active acupuncture treatments	VMS,	Uterus, No uterus,
Xia 2012	High	8	50	Placebo (N=36); Chinese herbal medicine (N=36)	Cornstarch and maltodextrin placebo daily; 3.5g of Chinese herbal medication daily	VMS, Discontinuation,	Uterus, No uterus,
Zaborowska 2007	High	12	not reported	Placebo (N=21); Acupuncture (N=30); Relaxation (N=15)	Placebo; 14 acupuncture sessions; 12 60 min training sessions	VMS,	Uterus, No uterus,
Zhong 2013	Low	12	50	Placebo (N=54); Chinese herbal medicine (N=54)	Placebo; Chinese herbal medicine (Er-Xian decoction)	VMS, Discontinuation,	Uterus, No uterus,

K.3.1.2 Studies excluded from the NMA (due to no connectedness to the networks)

The studies presented in Table 5 could not be included in the NMAs due to technical reasons identified after network plots had been drawn. Detailed exclusion reasons are given per study. The main exclusion reasons were studies not being connected to the networks through any treatment comparison, or studies making comparisons that were coded as being within the same class (e.g. using different frequencies of dosing of the same treatment).

Table 5: Excluded studies

Study name	Reason for exclusion	Interventions (sample size)	Outcomes	Populations
Kim 2010	Study makes within-treatment comparison only	Normal living/Usual care/Attention (N=59); Acupuncture (N=116)	VMS,	Uterus, No uterus,
Wang 2013	Study makes within-treatment comparison only	Chinese herbal medicine (N=20); Chinese herbal medicine (N=20)	VMS, Discontinuation, Bleeding	Uterus, No uterus,
Nagamani 1987	Study prevents convergence of model and provides no indirect evidence	Placebo (N=15); Clonidene (N=15)	VMS,	Uterus, No uterus,
Ozsoy 2002	Treatments in study are not connected to network	Oestrogen nasal spray High (N=101); Oestrogen alone oral High (N=100)	VMS,	Uterus,
Utian 2005	Treatments in study are not connected to network	Oestrogen alone oral Ave (N=84); Oestrogen valerate Ave (N=79); Conjugated equine estrogen (CEE) Ave (N=85)	VMS,	Uterus,
Parsey 2000	Treatments in study are not connected to network	Oestrogen alone transdermal Low (N=95); Conjugated equine estrogen (CEE) Low (N=98)	VMS,	No uterus,
Hervik & Mjaland 2009	Treatments in study are not connected to network	Sham acupuncture (N=29); Acupuncture (N=30)	VMS,	Breast cancer/history
Nedstrand 2006	Treatments in study are not connected to network	Acupuncture (N=19); Relaxation (N=19)	VMS,	Breast cancer/history
Elkins 2013	Study not connected to network	Normal living/Usual care/Attention (N=94); Hypnosis (N=93)	VMS,	Uterus,
Ayers 2012	Treatments in study are not connected to network	Normal living/Usual care/Attention (N=45); Cognitive Behavioural Therapy (N=95)	VMS, Discontinuation,	Uterus, No uterus,
Duijts 2012	Treatments in study are not connected to network	Normal living/Usual care/Attention (N=103); Cognitive Behavioural Therapy (N=109)	VMS, Discontinuation,	Breast cancer/history
Mann 2012	Treatments in study are not connected to network	Normal living/Usual care/Attention (N=49); Cognitive Behavioural Therapy (N=47)	VMS,	Breast cancer/history
Saensak 2013	Study makes within-treatment comparison only	Relaxation (N=36); Relaxation (N=35)	Discontinuation,	Uterus, No uterus,
Notelovitz 2000	Treatments in study are not connected to network	Placebo (N=66); Oestrogen transdermal + progestogen transdermal Low (N=68); Oestrogen transdermal + progestogen transdermal Ave (N=67); Oestrogen transdermal + progestogen transdermal High (N=68)	VMS, Discontinuation,	Uterus, No uterus,
Nedeljkovic 2013	All trial arms are equal to zero	Placebo (N=10); Sham acupuncture (N=10); Chinese herbal medicine (N=10); Acupuncture (N=10)	Discontinuation,	Uterus, No uterus,
Bao 2014	Treatments in study are not connected to network	Sham acupuncture (N=24); Acupuncture (N=24)	Discontinuation,	Breast cancer/history
Nir 2007	Treatments in study are not connected to network	Sham acupuncture (N=17); Acupuncture (N=12)	Bleeding	Uterus, No uterus,

K.3.1.3 NMA Results for women with a uterus

Vasomotor symptoms

32 trials of 12 classes were included in the network of outcome of vasomotor symptoms with a total sample size of 4165 women with menopause (figure 64).

Table 6 presents the results of the conventional pair-wise meta-analyses (head to head comparisons) (upper-right section of table), together with the results computed by the NMA for every possible treatment comparison (lower-left section of table). Both results are presented as mean ratios (95% CrI). These results were derived from the random effects model with fixed dose effects (see Table 19). Figure 71 graphically presents the results computed by the NMA for each intervention versus placebo.

The combination of oestrogen plus progestogen via patches was found to be significantly better than placebo (MR 0.23 (0.09, 0.57) on relieving vasomotor symptoms for women in menopause. Although, the combination of oral oestrogen plus progesterone did not manage to achieve a statistically significant difference compared to placebo (MR 0.52 (0.25, 1.06), the point estimate suggests that it may have the same degree of efficacy for relieving vasomotor symptoms compared to placebo as the intervention of oestrogen plus progestogen via patches. In addition, the combination of oestrogen plus progestogen via patches was significantly more effective than raloxifene, SSRIs/SNRIs, isoflavones and Chinese herbal medicine in relieving vasomotor symptoms. Isoflavones and black cohosh were also found to be significantly better than placebo. In addition, black cohosh was found to be significantly better in achieving this outcome when compared to raloxifene. No other significant differences were found among other interventions in the network.

Due to the apparent differences in results between oral and non-oral oestrogen plus progestogen versus placebo, a sensitivity analysis was conducted to investigate if a study using a low dose of oral oestrogen plus progestogen may have lowered the pooled effect for this treatment. However, neither the point estimate nor the confidence interval appeared to be sensitive to this assumption.

Inconsistency was assessed in the closed loop between placebo, sham acupuncture and acupuncture, but no significant difference was found between results obtained through direct and indirect evidence.

In this analysis, non-oral oestrogen plus progestogen was found to have the highest probability (69.8%) of being the best treatment to relieve vasomotor symptoms among interventions with duration up to 26 weeks followed by Black cohosh (14.23%), tibolone (4.02%) and oral oestrogen plus progestogen therapy (3.73%) (Table 7). Median treatment rankings with their 95% CI are shown in Figure 77.

Table 6: Mean ratios (95% CrI) from conventional (white area) and network meta-analysis (grey area) for the frequency of vasomotor symptoms for women in menopause with uterus

	Placebo	Sham acupuncture	Oestrogen + progestogen non-oral	Oestrogen + progestogen oral	Tibolone	Raloxifine	SSRIs/SNRIs	Isoflavones	Chinese herbal medicine	Black cohosh	Multibotanicals	Acupuncture
Placebo		0.75 (0.27, 1.9)	**0.23 (0.09, 0.57)**	0.52 (0.25, 1.06)	0.55 (0.24, 1.29)	1.65 (0.61, 4.51)	0.84 (0.54, 1.31)	**0.62 (0.44, 0.87)**	0.95 (0.46, 1.9)	**0.4 (0.17, 0.9)**	0.71 (0.24, 2.07)	0.58 (0.23, 1.36)
Sham acupuncture	0.75 (0.27, 1.9)								1.28 (0.43, 3.98)			0.78 (0.35, 1.73)
Oestrogen + progestogen non-oral	**0.23 (0.09, 0.57)**	0.31 (0.08, 1.22)										
Oestrogen + progestogen oral	0.52 (0.25, 1.06)	0.69 (0.21, 2.43)	2.23 (0.7, 7.1)									
Tibolone	0.55 (0.24, 1.29)	0.74 (0.21, 2.78)	2.38 (0.69, 8.25)	1.07 (0.35, 3.25)								
Raloxifene	1.65 (0.61, 4.51)	2.22 (0.56, 9.26)	**7.12 (1.86, 27.63)**	3.19 (0.94, 11.04)	2.99 (0.81, 11.19)							
SSRIs/SNRIs	0.84 (0.54, 1.31)	1.13 (0.4, 3.44)	**3.63 (1.33, 9.93)**	1.63 (0.7, 3.81)	1.53 (0.59, 3.99)	0.51 (0.17, 1.52)						
Isoflavones	**0.62 (0.44, 0.87)**	0.83 (0.3, 2.45)	**2.67 (1.02, 7.05)**	1.2 (0.54, 2.69)	1.12 (0.45, 2.81)	0.38 (0.13, 1.08)	0.73 (0.42, 1.29)					
Chinese herbal medicine	0.95 (0.46, 1.9)	1.28 (0.43, 3.98)	**4.1 (1.29, 12.88)**	1.84 (0.66, 5.02)	1.73 (0.57, 5.18)	0.58 (0.17, 1.94)	1.13 (0.49, 2.56)	1.54 (0.69, 3.33)				0.61 (0.21, 1.74)
Black cohosh	**0.4 (0.17, 0.9)**	0.54 (0.15, 1.97)	1.72 (0.49, 5.79)	0.77 (0.25, 2.29)	0.72 (0.21, 2.35)	**0.24 (0.06, 0.87)**	0.47 (0.18, 1.19)	0.65 (0.25, 1.56)	0.42 (0.14, 1.25)			
Multibotanicals	0.71 (0.24, 2.07)	0.95 (0.23, 4.18)	3.05 (0.75, 12.45)	1.37 (0.37, 5)	1.28 (0.33, 5.05)	0.43 (0.1, 1.86)	0.84 (0.26, 2.7)	1.14 (0.37, 3.52)	0.74 (0.21, 2.73)	1.78 (0.46, 7.28)		
Acupuncture	0.58 (0.23, 1.36)	0.78 (0.35, 1.73)	2.51 (0.68, 8.7)	1.12 (0.34, 3.48)	1.05 (0.3, 3.53)	0.35 (0.09, 1.3)	0.69 (0.24, 1.81)	0.94 (0.34, 2.36)	0.61 (0.21, 1.74)	1.46 (0.42, 4.95)	0.82 (0.2, 3.24)	

Results in the top right diagonal of the table are the mean ratios and 95% credible intervals from the conventional meta-analyses of direct evidence between the column-defined treatments compared to the row-defined treatment. Mean ratio less than 1 favour the column-defined treatment.
Results in the bottom left are the mean ratios and 95% credible intervals from the random effect model with fixed dose effects of the NMA of direct and indirect evidence between the row-defined treatments compared to the column-defined treatment. Mean ratios less than 1 favour the row-defined treatment.
Numbers in bold denote statistically significant results (95% CI credible intervals do not include 1)

Figure 71: Forest plot showing mean ratios (with their 95% CrI) of NMA estimates for each intervention versus placebo

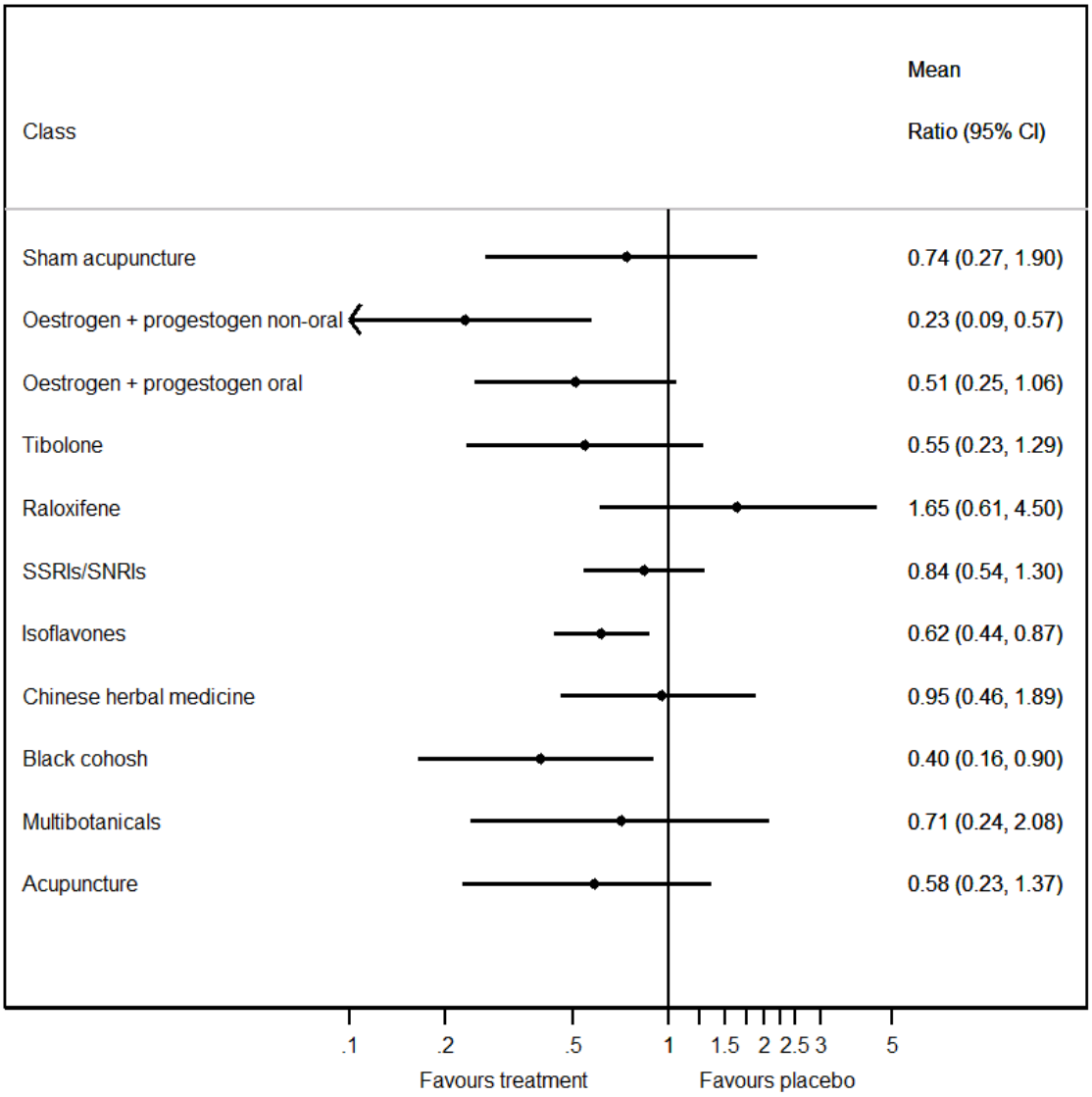


Table 7: Log mean ratios (with their 95% CI) of all interventions in the network and the probability of being the best treatment for achieving relief of vasomotor symptoms

	Median log mean ratios	95%CrI	Probability of being the best treatment	Median (95% CrI) treatment rank
Placebo	*Baseline treatment*		0.00%	10 (7-12)
Sham acupuncture	-0.30	(-1.32, 0.64)	1.44%	7 (2-12)
Oestrogen + progestogen non-oral	-1.46	(-2.37, -0.56)	69.82%	1 (1-5)
Oestrogen + progestogen oral	-0.67	(-1.4, 0.06)	3.73%	4 (1-10)
Tibolone	-0.60	(-1.45, 0.25)	4.02%	5 (1-11)

	Median log mean ratios	95%CrI	Probability of being the best treatment	Median (95% CrI) treatment rank
Raloxifine	0.50	(-0.49, 1.51)	0.04%	12 (6-12)
SSRIs/SNRIs	-0.17	(-0.61, 0.26)	0.01%	8 (4-11)
Isoflavones	-0.48	(-0.82, -0.13)	0.10%	6 (3-9)
Chinese herbal medicine	-0.05	(-0.78, 0.63)	0.09%	9 (4-12)
Black cohosh	-0.92	(-1.8, -0.11)	14.23%	3 (1-9)
Multibotanicals	-0.34	(-1.43, 0.73)	2.88%	7 (1-12)
Acupuncture	-0.54	(-1.49, 0.31)	3.64%	5 (1-11)

Discontinuation of treatment

21 trials of ten classes were included in the network of outcome of discontinuation of treatment with a total sample size of 4829 women with uterus in menopause (Figure 66).

Table 8 presents the results of the conventional pair-wise meta-analyses (head to head comparisons) (upper-right section of table), together with the results computed by the NMA for every possible treatment comparison (lower-left section of table). Both results are presented as odds ratios (95% CrI). These results were derived from the random effects model with fixed dose effects (Table 20). Figure 78 graphically presents the results computed by the NMA for each intervention versus placebo.

The combination of non-oral oestrogen plus progestogen was found to be significantly better than placebo on discontinuation of treatment for women in menopause. In addition, conjugated oestrogens plus bazedoxifene was only marginally significantly more effective than placebo in this outcome. SSRIs/SSNIs were found to be significantly worse than placebo on discontinuation of treatment in this population. Tibolone and SSRIs/SNRIs were both found to be significantly worse than non-oral oestrogen plus progestogen and conjugated oestrogens plus bazedoxifene for this outcome.

It was not possible to assess inconsistency in this network as no closed loops between treatments existed.

In this analysis, conjugated oestrogens plus bazedoxifene was found to have the highest probability (37.34%) of being the best treatment in relation to discontinuation of treatment among interventions with duration up to 26 weeks followed closely by valerian root (37.00%) (Table 9), though this is likely to be primarily due to the high uncertainty in estimates for valerian root. Median treatment rankings with their 95% CrI are shown in (Figure 78).

Table 8: Odds ratios (95% CrI) from conventional (white area) and network meta-analysis (grey area) for discontinuation of treatment for women in menopause with uterus

	Placebo	Oestrogen + progestogen oral	Conjugated oestrogens plus bazedoxifene	Tibolone	SSRIs/SNRIs	Gabapentin	Isoflavones	Chinese herbal medicine	Multibotanicals
Placebo		**0.61 (0.37, 0.99)**	**0.31 (0.1, 1)**	5.65 (0.94, 172.9)	**1.66 (1.07, 2.61)**	0.88 (0.63, 1.23)	0.95 (0.51, 1.76)	1.58 (0.42, 6.66)	0.5 (0.07, 4.53)
Oestrogen + progestogen oral	**0.61 (0.37, 0.99)**								
Conjugated oestrogens plus bazedoxifene	**0.31 (0.1, 1.00)**	0.52 (0.15, 1.83)							
Tibolone	5.65 (0.94, 172.9)	**9.36 (1.44, 294.6)**	**18.54 (2.07, 651.2)**						
SSRIs/SNRIs	**1.66 (1.07, 2.61)**	**2.73 (1.41, 5.33)**	**5.3 (1.53, 17.61)**	0.29 (0.01, 1.88)					
Gabapentin	0.88 (0.63, 1.23)	1.45 (0.8, 2.62)	2.81 (0.84, 8.99)	**0.16 (0.01, 0.97)**	0.53 (0.3, 0.92)				
Isoflavones	0.95 (0.51, 1.76)	1.56 (0.71, 3.45)	3.03 (0.81, 10.81)	0.17 (0.01, 1.14)	0.57 (0.27, 1.23)	1.08 (0.53, 2.18)			
Chinese herbal medicine	1.58 (0.42, 6.66)	2.61 (0.64, 11.89)	5.07 (0.88, 31.11)	0.27 (0.01, 2.87)	0.95 (0.24, 4.28)	1.8 (0.46, 7.9)	1.67 (0.39, 8.02)		
Multibotanicals	0.5 (0.07, 4.53)	0.82 (0.11, 7.87)	1.6 (0.17, 18.57)	0.08 (0.001, 1.55)	0.3 (0.04, 2.85)	0.57 (0.08, 5.28)	0.53 (0.07, 5.23)	0.32 (0.03, 4.11)	
Valerian root	0.4 (0.01, 5.4)	0.66 (0.02, 9.35)	1.26 (0.03, 21.86)	0.06 (0.001, 1.75)	0.24 (0.01, 3.41)	0.46 (0.01, 6.3)	0.42 (0.01, 6.16)	0.25 (0.01, 4.77)	0.76 (0.01, 20.45)

Results in the upper-right area are the odd ratios and 95% credible intervals from the conventional meta-analyses of direct evidence between the column-defined treatments compared to the row-defined treatment. Odd ratios less than 1 favour the column-defined treatment.
Results in lower-left are the odd ratios and 95% credible intervals from the random effect model with fixed dose effects of the NMA of direct and indirect evidence between the row-defined treatments compared to the column-defined treatment. Odd ratios less than 1 favour the row-defined treatment.
Numbers in bold denote statistically significant results (95% CI credible intervals do not include 1)

Figure 72: Forest plot showing odds ratios (with their 95% CrI) of NMA estimates for each

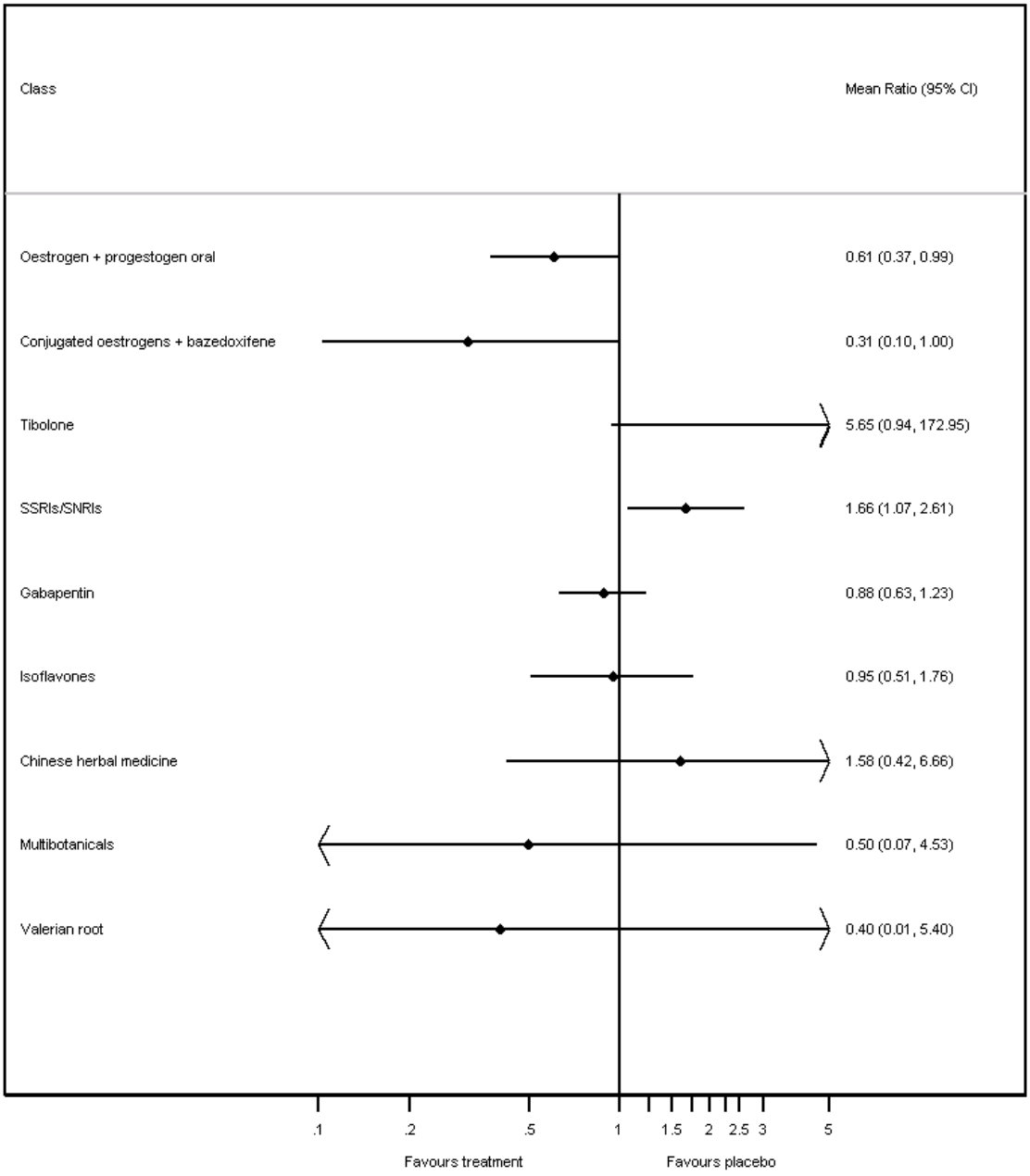


Table 9: Log odd ratios (with their 95% CrI) of all interventions in the network and the probability of being the best treatment for discontinuation of treatment

	Median log odds ratios	95%CrI	Probability of being the best treatment	Median (95% CrI) treatment rank
Placebo	Baseline treatment		0.00%	6 (4-8)
Oestrogen + progestogen oral	-0.50	(-0.99, -0.01)	2.83%	3 (1-6)
Conjugated oestrogens plus bazedoxifene	-1.16	(-2.28, 0.002)	37.34%	2 (1-6)

	Median log odds ratios	95%CrI	Probability of being the best treatment	Median (95% CrI) treatment rank
Tibolone	1.73	(-0.06, 5.15)	0.03%	10 (6-10)
SSRIs/SNRIs	0.50	(0.06, 0.96)	0.00%	8 (6-10)
Gabapentin	-0.13	(-0.46, 0.21)	0.08%	5 (3-8)
Isoflavones	-0.05	(-0.67, 0.57)	0.29%	6 (2-9)
Chinese herbal medicine	0.46	(-0.86, 1.9)	0.66%	8 (2-10)
Multibotanicals	-0.70	(-2.63, 1.51)	21.77%	3 (1-10)
Valerian root	-0.91	(-4.41, 1.69)	37.00%	2 (1-10)

Vaginal bleeding

5 trials of five classes were included in the network of outcome of bleeding with a total sample size of 1367 women with uterus in menopause (Figure 67).

Table 10 presents the results of the conventional pair-wise meta-analyses (head to head comparisons) (upper-right section of table), together with the results computed by the NMA for every possible treatment comparison (lower-right section of table). Both results are presented as odds ratios (95% CrI). These results were derived from the fixed effects model (Table 21). Figure 73 graphically presents the results computed by the NMA for each intervention versus placebo.

No significant differences were found between any of the treatments in the network on outcomes of bleeding for women in menopause.

It was not possible to assess inconsistency in this network as no closed loops between treatments existed.

In this analysis, SSRIs/SNRIs were found to have the highest probability (66.34%) of being the best treatment in relation to vaginal bleeding among interventions with duration up to 26 weeks followed by gabapentin (25.96%) (Table 11). Median treatment rankings with their 95% CI are shown in (Figure 79).

Table 10: Odds ratios (95% CrI) from conventional (white area) and network meta-analysis (grey area) for bleeding for women in menopause with uterus

	Placebo	Oestrogen + progestogen oral	Tibolone	SSRIs/SNRIs	Gabapentin
Placebo		2.76 (0.68, 12.06)		0.2 (0.001, 4.6)	0.58 (0.06, 4.17)
Oestrogen + progestogen oral	2.76 (0.68, 12.06)		1.45 (0.35, 6.57)		
Tibolone	1.45 (0.35, 6.57)	0.53 (0.38, 0.73)			
SSRIs/SNRIs	0.2 (0.001, 4.6)	0.07 (0.001, 2.3)	0.13 (0.001, 4.42)		
Gabapentin	0.58 (0.06, 4.17)	0.21 (0.02, 2.36)	0.4 (0.03, 4.58)	3.01 (0.06, 1784)	

Results in the upper-right area are the odd ratios and 95% credible intervals from the conventional meta-analyses of direct evidence between the column-defined treatments compared to the row-defined treatment. Odd ratios less than 1 favour the column-defined treatment.
Results in lower-left are the odd ratios and 95% credible intervals from the random effect model with fixed dose effects of the NMA of direct and indirect evidence between the row-defined treatments compared to the column-defined treatment. Odd ratios less than 1 favour the row-defined treatment.
Numbers in bold denote statistically significant results (95% CI credible intervals do not include 1)

Figure 73: Forest plot showing odds ratios (with their 95% CrI) of NMA estimates for each intervention versus placebo.

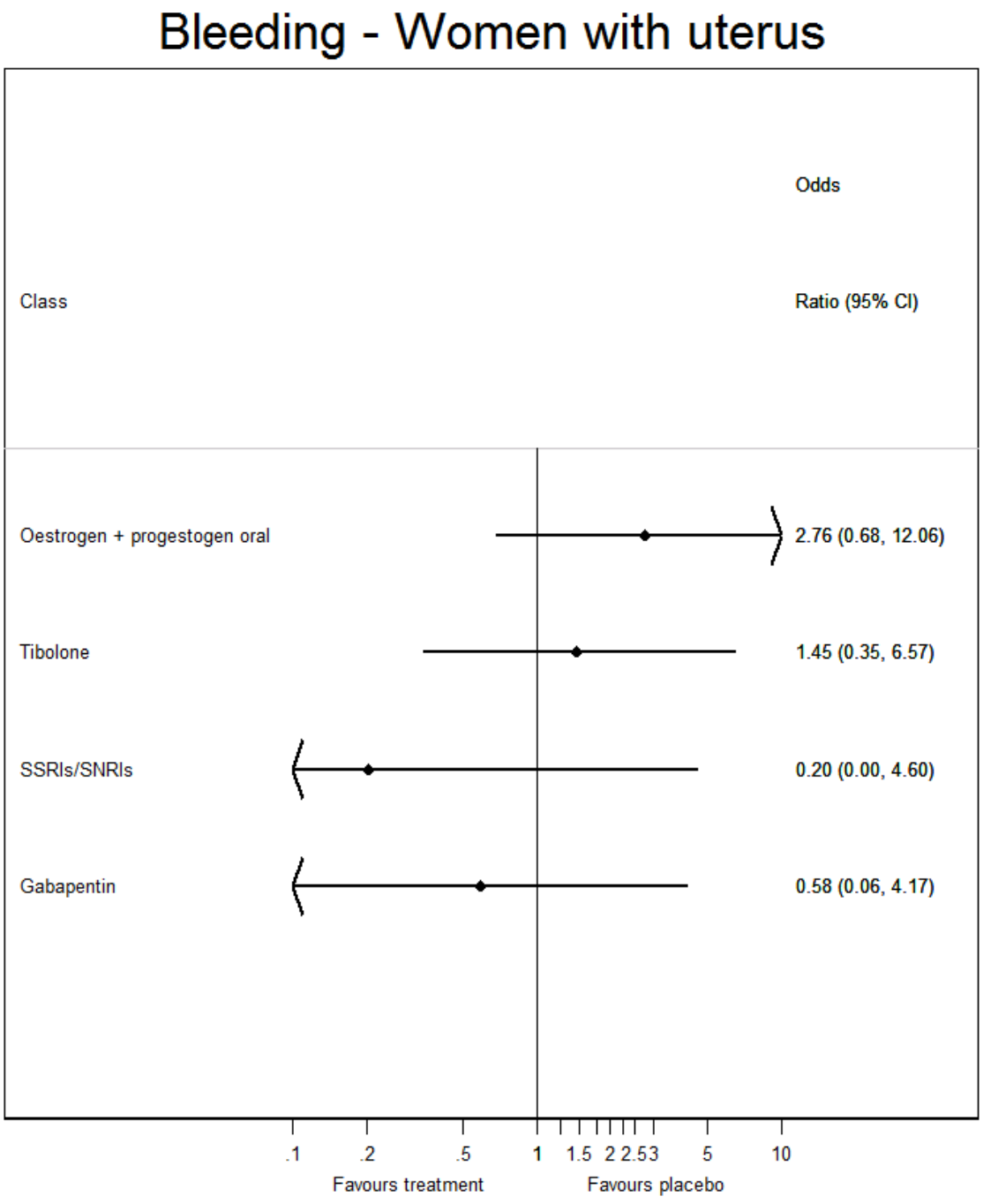


Table 11: Log odd ratios (with their 95% CrI) of all interventions in the network and the probability of being the best treatment for bleeding

	Median log odds ratios	95%CrI	Probability of being the best treatment	Median (95% CrI) treatment rank
Placebo	Baseline treatment		3.62%	
Oestrogen + progestogen oral	1.01	(-0.39, 2.49)	0.00%	3 (1-5)
Tibolone	1.01	(-0.39, 2.49)	4.08%	5 (3-5)
SSRIs/SNRIs	0.37	(-1.06, 1.88)	66.34%	4 (1-4)
Gabapentin	-1.59	(-7.78, 1.53)	25.96%	1 (1-5)

K.3.1.3.1 NMA Results for women without a uterus

Vasomotor symptoms

32 trials of nine classes were included in the network of outcome of vasomotor symptoms with a total sample size of 4165 women without uterus in menopause (Figure 65).

As mentioned previously (K.2.5.1) oestrogen alone was not included as a class intervention in the network given that the trials that have tested this intervention were either mixed population studies or did not give enough information on the estimation of relative effect (please refer to table of excluded studies for reasons for exclusion). All the trials contributed to this network included interventions -non hormonal pharmacological and non-pharmacological treatments- that also appeared in the network of vasomotor symptoms for women with uterus (please refer to NMA protocol for further details of inclusion of studies reported non HRT treatment).

Therefore the GDC decided not to consider the results of this network for decision making given the limitation of their generalibility in the clinical context. However, results of this NMA are reported as additional information (K.5.3). Further details are given in the LETR about the extrapolation of evidence on women with uterus to decision making for women without uterus (Chapter 7 in the full guideline).

Discontinuation of treatment

15 trials of eight classes were included in the network of outcome of discontinuation of treatment with a total sample size of 2672 women without a uterus (Figure 68).

Table 12 presents the results of the conventional pair-wise meta-analyses (head to head comparisons) (upper-right section of table), together with the results computed by the NMA for every possible treatment comparison (lower-right section of table). Both results are presented as odds ratios (95% CrI). These results were derived from the fixed effects model (Table 23). Figure 74 graphically presents the results computed by the NMA for each intervention versus placebo.

The only significant differences between treatments were for the comparisons of SSRIs/SNRIs versus placebo and SSRIs/SNRIs versus gabapentin, where SSRIs/SNRIs were found to be significantly worse in both instances for discontinuation of treatment for women in menopause.

It was not possible to assess inconsistency in this network as no closed loops between treatments existed.

In this analysis, non-oral oestrogen alone was found to have the highest probability (37.90%) of being the best treatment in relation to discontinuation of treatment among interventions with duration up to 26 weeks followed by valerian root (35.76%) (Table 13). Median treatment rankings with their 95% CrI are shown in Figure 81.

Table 12: Odds ratios (95% CrI) from conventional (white area) and network meta-analysis (grey area) for discontinuation of treatment for women in menopause without uterus

	Placebo	Oestrogen alone non-oral	SSRIs/SNRIs	Gabapentin	Isoflavones/Genistein/soy	Chinese herbal medicine	Multibotanicals	Valerian root
Placebo		0.37 (0.01, 3.1)	**1.66 (1.07, 2.61)**	0.88 (0.63, 1.23)	0.95 (0.51, 1.76)	1.59 (0.42, 6.66)	0.5 (0.07, 4.47)	0.4 (0.01, 5.36)
Oestrogen alone non-oral	0.37 (0.01, 3.1)							
SSRIs/SNRIs	**1.66 (1.07, 2.61)**	4.48 (0.51, 136)						
Gabapentin	0.88 (0.63, 1.23)	2.37 (0.28, 71.07)	**0.53 (0.3, 0.93)**					
Isoflavones	0.95 (0.51, 1.76)	2.57 (0.28, 79.99)	0.57 (0.27, 1.23)	1.08 (0.53, 2.18)				
Chinese herbal medicine	1.59 (0.42, 6.66)	4.46 (0.34, 166.1)	0.96 (0.24, 4.3)	1.81 (0.46, 7.85)	1.67 (0.38, 7.98)			
Multibotanicals	0.5 (0.07, 4.47)	1.44 (0.07, 73.26)	0.3 (0.04, 2.81)	0.57 (0.08, 5.2)	0.53 (0.07, 5.1)	0.32 (0.03, 4.03)		
Valerian root	0.4 (0.01, 5.36)	1.11 (0.02, 72.34)	0.24 (0.01, 3.36)	0.46 (0.01, 6.25)	0.42 (0.01, 6.08)	0.25 (0.01, 4.61)	0.76 (0.01, 20.17)	

Results in the upper-right area are the odd ratios and 95% credible intervals from the conventional meta-analyses of direct evidence between the column-defined treatments compared to the row-defined treatment. Odd ratios less than 1 favour the column-defined treatment.
Results in lower-left are the odd ratios and 95% credible intervals from the random effect model with fixed dose effects of the NMA of direct and indirect evidence between the row-defined treatments compared to the column-defined treatment. Odd ratios less than 1 favour the row-defined treatment.
Numbers in bold denote statistically significant results (95% CI credible intervals do not include 1)

Figure 74: Forest plot showing odds ratios (with their 95% CrI) of NMA estimates for each intervention versus placebo

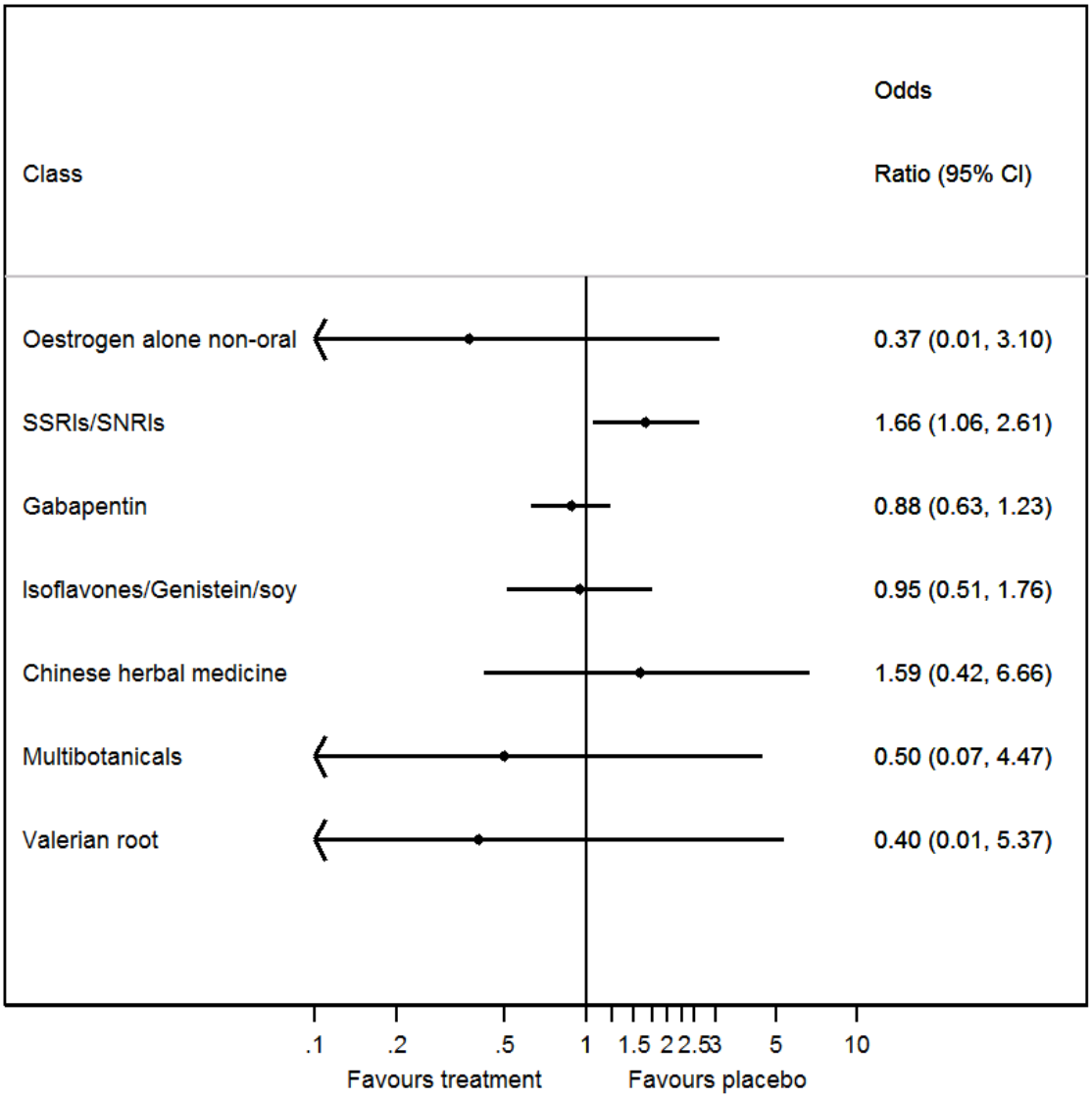


Table 13: Log odd ratios (with their 95% CrI) of all interventions in the network and the probability of being the best treatment for discontinuation of treatment

	Median log odds ratios	95%CrI	Probability of being the best treatment	Median (95% CrI) treatment rank
Placebo	Baseline treatment		0.08%	5 (3-7)
Oestrogen alone non-oral	-0.99	(-4.38, 1.13)	37.90%	2 (1-8)
SSRIs/SNRIs	0.50	(0.06, 0.96)	0.01%	7 (5-8)
Gabapentin	-0.13	(-0.46, 0.21)	1.18%	4 (2-7)
Isoflavones/Genistein/soy	-0.05	(-0.67, 0.57)	1.42%	5 (2-7)

	Median log odds ratios	95%CrI	Probability of being the best treatment	Median (95% CrI) treatment rank
Chinese herbal medicine	0.46	(-0.86, 1.9)	1.19%	7 (2-8)
Multibotanicals	-0.69	(-2.63, 1.5)	22.46%	2 (1-8)
Valerian root	-0.91	(-4.41, 1.68)	35.76%	2 (1-8)

The only trial that could match our agreed NMA protocol for the network of women without uterus that included oestrogen alone was Parsey 2000. The other trials that looked at oestrogen as intervention were excluded from the network of women without uterus due to their mixed population profile (included women with and without uterus in less than 2/3 of the population falling in each category) (please see table of excluded studies for more details). The RCT by Parsey 2000 compared the effectiveness of transdermal oestrogen versus oral conjugated oestrogens for the relief of vasomotor symptoms for a mixed population of 193 women with and without hysterectomy (70.3% with hysterectomy). This study was not included in the network of vasomotor symptoms for women without uterus as it didn't connect with other available treatments in the network. However, the Guideline Development Group, given the absence of information on the role of oestrogen alone for this population wished to see the evidence on the comparative effectiveness of these treatments to supplement the information on the NMA results and aid their decision making.

Very low quality evidence from this study (Parsey 2000) showed that there was no significant difference for the outcomes of vasomotor symptoms and treatment discontinuation between women who received transdermal oestrogen and oral conjugated oestrogens (Table 14).

Table 14: GRADE findings of comparison of transdermal oestrogen versus oral conjugated estrogens (pair-wise meta-analyis)

Quality assessment							No of patients		Effect			
No of studies	**Design**	**Risk of bias**	**Inconsistency**	**Indirectness**	**Imprecision**	**Other considerations**	**Transdermal oestrogen**	**Oral CEE**	**Relative (95% CI)**	**Absolute**	**Quality**	**Importance**
Hot flushes (follow-up mean 12 weeks; Better indicated by lower values)												
1	randomised trials	serious[1]	no serious inconsistency	no serious indirectness	very serious[2]	none	Final scores: 1.4 (1.1)/	Final scores: 1.4 (1.2)	Ratio of means: 0.92 (0.54 to 1.58)	-	Very low	CRITICAL
Discontinuation												
1	randomised trials	serious[1]	no serious inconsistency	no serious indirectness	very serious[2]	none	16/95 (16.8%)	15.3%	RR 1.1 (0.58 to 2.1)	15 more per 1000 (from 64 fewer to 168 more)	Very low	CRITICAL

1 Evidence was downgraded by 1 due to selection, performance, attrition, or detection bias
2 Evidence was downgraded by 2 due to very serious imprecision as 95% CI crossed two default MID (0.75 and 1.25)

K.3.1.4 NMA Results for women with breast cancer/history of breast cancer

Vasomotor symptoms

4 trials of five classes were included in the network of outcome of vasomotor symptoms with a total sample size of 686 women with breast cancer or history of breast cancer (Figure 69).

Table 15 presents the results of the conventional pair-wise meta-analyses (head to head comparisons) (upper-right section of table), together with the results computed by the NMA for every possible treatment comparison (lower-right section of table). Both results are presented as mean ratios (95% CrI). These results were derived from the random effects model (Table 24). Figure 75 graphically presents the results computed by the NMA for each intervention versus placebo.

No significant differences were found between any of the treatments in the network for relieving vasomotor symptoms for women in menopause.

It was not possible to assess inconsistency in this network as no closed loops between treatments existed. In this analysis, St John's Wort was found to have the highest probability (64.35%) of being the best treatment to relieve vasomotor symptoms among interventions, with the next highest probability being gabapentin (29.7%) (

Table 16). Median treatment rankings with their 95% CI are shown in **Figure 82**.

The GDG were concerned about the importance of pointing out to women the possible drug interactions of St John's Wort with prescribed medication. See:

http://www.cancerresearchuk.org/about-cancer/cancers-in-general/treatment/complementary-alternative/therapies/st-johns-wort.

Table 15: Mean ratios (95% CrI) from conventional (white area) and network meta-analysis (grey area) for the frequency of vasomotor symptoms for women in menopause with breast cancer or a history of breast cancer

	Placebo	SSRIs/SNRIs	Gabapentin	Isoflavones	St John's Wort
Placebo		1.35 (0.71, 3.71)	0.78 (0.5, 1.25)	1.08 (0.67, 1.74)	0.67 (0.34, 1.21)
SSRIs/SNRIs	1.35 (0.71, 3.71)				
Gabapentin	0.78 (0.5, 1.25)	0.57 (0.19, 1.28)			
Isoflavones	1.08 (0.67, 1.74)	0.8 (0.26, 1.79)	1.39 (0.71, 2.67)		
St John's Wort	0.67 (0.34, 1.21)	0.48 (0.14, 1.19)	0.85 (0.37, 1.79)	0.61 (0.27, 1.31)	

Results in the upper-right area are the mean ratios and 95% credible intervals from the conventional meta-analyses of direct evidence between the column-defined treatments compared to the row-defined treatment. Mean ratios less than 1 favour the column-defined treatment.
Results in lower-left are the mean ratios and 95% credible intervals from the random effect model with fixed dose effects of the NMA of direct and indirect evidence between the row-defined treatments compared to the column-defined treatment. Mean ratios less than 1 favour the row-defined treatment.
Numbers in bold denote statistically significant results (95% CI credible intervals do not include 1)

Figure 75: Forest plot showing mean ratios (with their 95% CrI) of NMA estimates for each intervention versus placebo

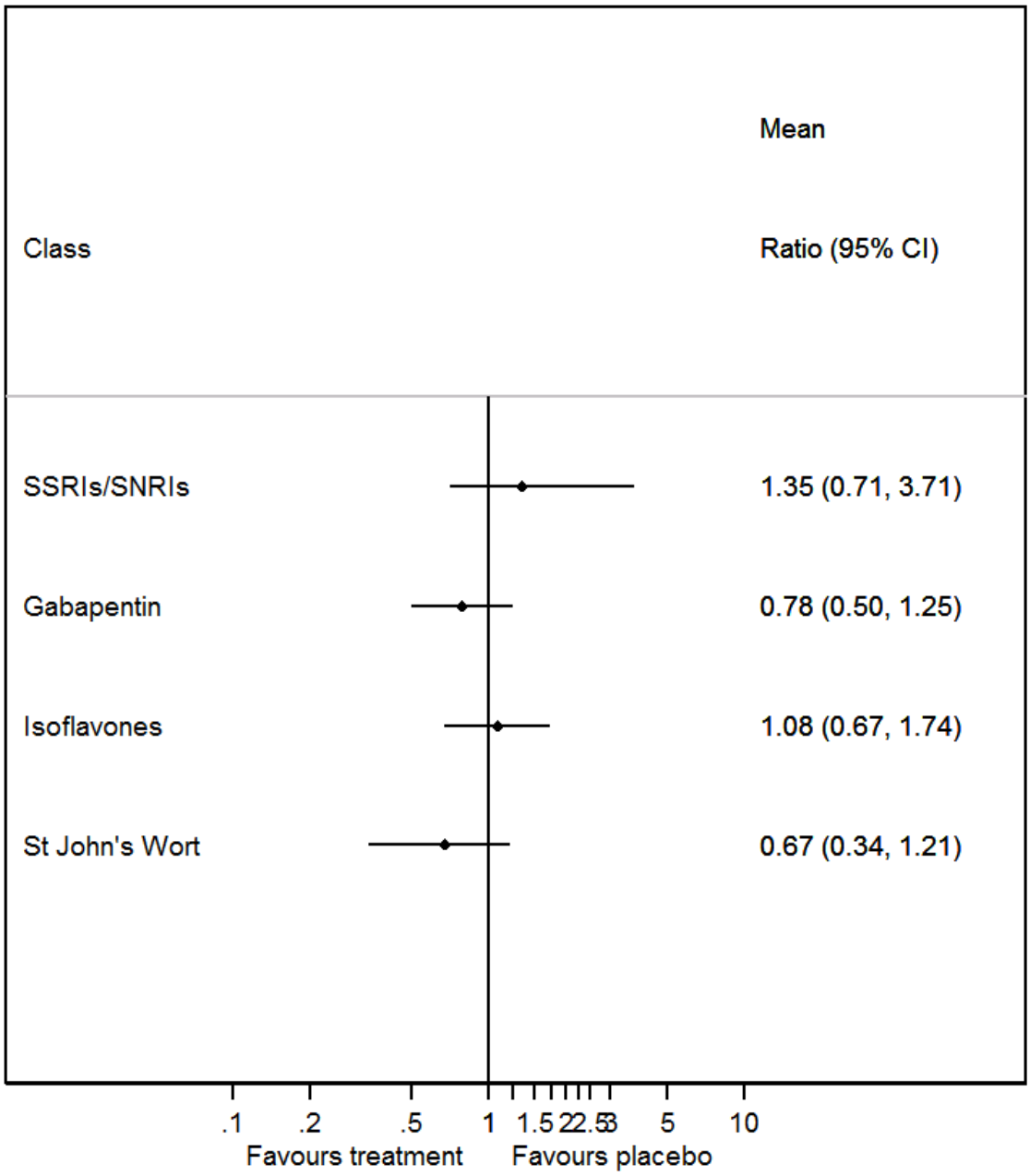


Table 16: Log mean ratios (with their 95% CrI) of all interventions in the network and the probability of being the best treatment for achieving relief of vasomotor symptoms

	Median log mean ratios	95%CrI	Probability of being the best treatment	Median (95% CrI) treatment rank
Placebo	*Baseline treatment*		0.61%	3 (2-5)
SSRIs/SNRIs	0.30	(-0.35, 1.31)	2.09%	5 (2-5)
Gabapentin	-0.25	(-0.7, 0.22)	29.70%	2 (1-4)
Isoflavones	0.08	(-0.4, 0.56)	3.25%	4 (1-5)
St John's Wort	-0.40	(-1.09, 0.19)	64.35%	1 (1-4)

Discontinuation of treatment

3 trials of four classes were included in the network of outcome of discontinuation of treatment with a total sample size of 639 women with uterus in menopause (Figure 70).

Table 17 presents the results of the conventional pair-wise meta-analyses (head to head comparisons) (upper-right section of table), together with the results computed by the NMA for every possible treatment comparison (lower-right section of table). Both results are presented as odds ratios (95% CrI). These results were derived from the fixed effects model (Table 25). Figure 76 graphically presents the results computed by the NMA for each intervention versus placebo.

No significant differences were found between any of the comparisons of treatments included in the network for discontinuation of treatment (Figure 70).

It was not possible to assess inconsistency in this network as no closed loops between treatments existed.

In this analysis, placebo was found to have the highest probability (43.04%) of being the best treatment in relation to discontinuation of treatment among interventions with duration up to 26 months (Table 18). Median treatment rankings with their 95% CrI are shown in **Figure 83**.

Table 17: Odds ratios (95% CrI) from conventional (white area) and network meta-analysis (grey area) for discontinuation of treatment for women in menopause with breast cancer or a history of breast cancer

	Placebo	SSRIs/SNRIs	Gabapentin	Isoflavones
Placebo		1.43 (0.2, 13.14)	1.35 (0.53, 3.93	2.67 (0.69, 13.64
SSRIs/SNRIs	1.43 (0.2, 13.14)			
Gabapentin	1.35 (0.53, 3.93)	0.95 (0.09, 8.76)		
Isoflavones	2.67 (0.69, 13.64)	1.89 (0.14, 23.48)	1.97 (0.35, 12.74)	

Results in the upper right area are the odd ratios and 95% credible intervals from the conventional meta-analyses of direct evidence between the column-defined treatments compared to the row-defined treatment. Odd ratios less than 1 favour the column-defined treatment.
Results in the down left are the odd ratios and 95% credible intervals from the fixed effect model of the NMA of direct and indirect evidence between the row-defined treatments compared to the column-defined treatment. Odd ratios less than 1 favour the row-defined treatment.
Numbers in bold denote statistically significant results (95% CI credible intervals do not include 1)

Figure 76: Forest plot showing odds ratios (with their 95% CrI) of NMA estimates for each intervention versus placebo

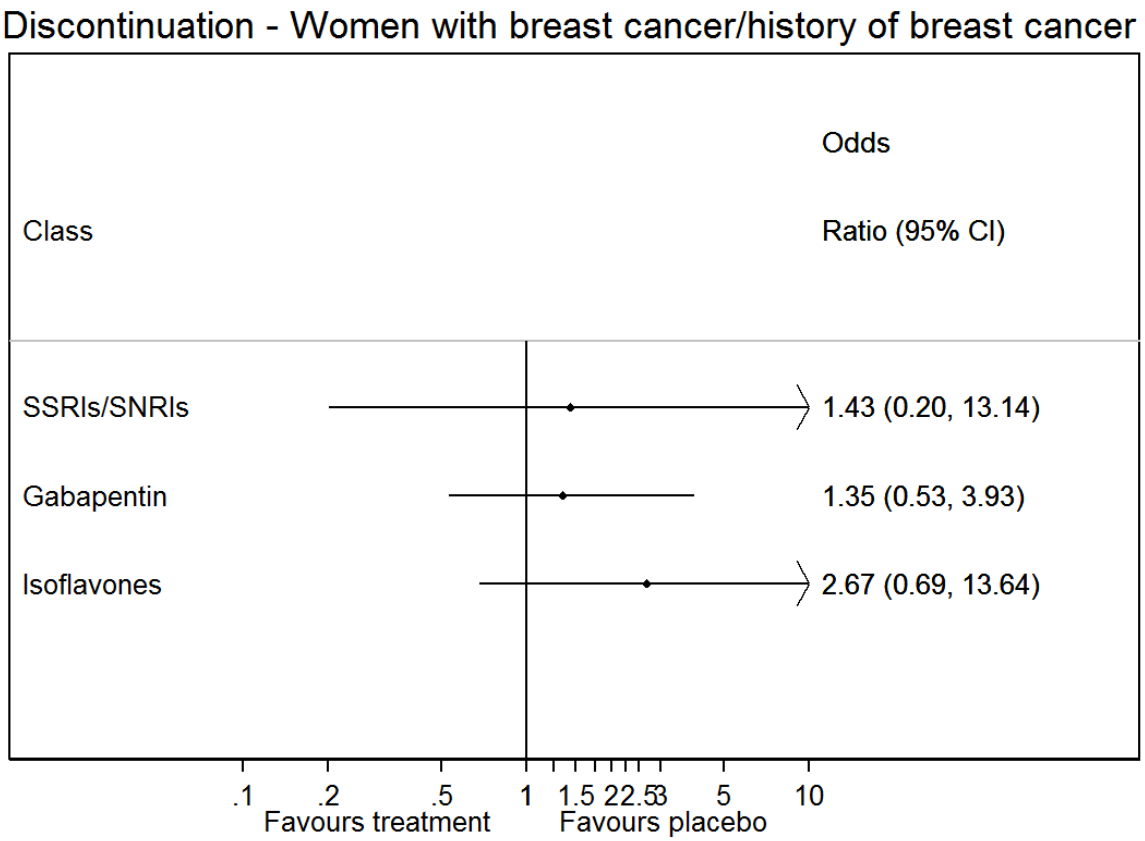


Table 18: Log odd ratios (with their 95% CrI) of all interventions in the network and the probability of being the best treatment for discontinuation of treatment

	Median log odds ratios	95%CrI	Probability of being the best treatment	Median (95% CrI) treatment rank
Placebo	*Baseline treatment*		43.04%	2 (1-3)
SSRIs/SNRIs	0.36	(-1.6, 2.58)	32.63%	3 (1-4)
Gabapentin	0.30	(-0.63, 1.37)	19.13%	2 (1-4)
Isoflavones	0.98	(-0.37, 2.61)	5.20%	4 (1-4)

K.4 Discussion

Ascertaining the most effective intervention for the treatment of women in menopause presents certain challenges. In order to overcome the difficulty of interpreting the conclusions from these numerous separate comparisons, NMAs were performed by including all the available evidence, given they met the inclusion criteria of the protocol.

The findings from the NMAs were used as the clinical base for the cost-effectiveness analysis in order to inform the GDG in decision making when developing recommendations for the most clinical and cost effective treatment for relieving short term symptoms for menopausal women.

However, there were several challenges in the formulation of networks for this complex NMA and results should be interpreted with caution in view of the following limitations:

- The number of studies included for some comparisons was small and the majority of interventions were compared only to placebo. This had led to the very wide credible intervals in the MRs and ORs for specific interventions.
- Due to numerous outcomes selected for the review question of short term symptoms relief (vasomotor symptoms, low mood, anxiety, sexual activity, musculoskeletal symptoms, discontinuation of treatment, vaginal bleeding) the GDG had to prioritise either frequency or intensity as the selected measurement of VMS. Although it was recognised that frequency measures only one aspect of women's experience of VMS, due to the finite resources and time available to develop this guideline we had to be selective. In addition, during the planning of this review question, it was identified that intensity was measured using many different tools that could prevent any synthesis of evidence and metanalysis for decision making. Furthermore, in order to make the most use of the available data for the NMA, we had to prioritise at frequency rather than intensity of VMS (please refer to excluded list for this review question in Appendix G as only 38 out of 400 studies were excluded because they reported data in terms of intensity).
- The outcome of vasomotor symptoms was self-reported in the trials with a wide variation of baseline frequency of symptoms reported across the included trials.
- The majority of trials included in the networks were postmenopausal women
- Not all trials have a common comparator. It is important to note that comparisons with longer paths will have less precision. The 95% credible intervals of the log ORs were very wide in particular for most interventions comparing with placebo. This could be due to the lack of direct trial data to inform each comparison; and this led to a lot of uncertainty. Since indirect evidence is inherently less precise than direct evidence, the more links that are required to connect a treatment to the baseline comparator, the less precision there will be in the estimation of effect size for that treatment
- Other outcomes such as low mood, musculoskeletal and sexual function were only considered in pair wise meta-analyses and they were not prioritized in this NMA.
- This NMA examined each outcome independently; this analysis would benefit from multiple outcomes analysis especially because multiple outcomes are usually correlated. However, the methods on this type of analysis are still in development.
- This NMA did not address the sequence of interventions, i.e. first-, second- and third-line therapy; especially when there was treatment failure due to a number of reasons, e.g. sub-optimal/non response, intolerance.
- Many trials only reported outcomes at less than 26 weeks and follow up time was short. Therefore, this NMA will not be sufficient to determine the optimal choice of treatment in a lifetime perspective.

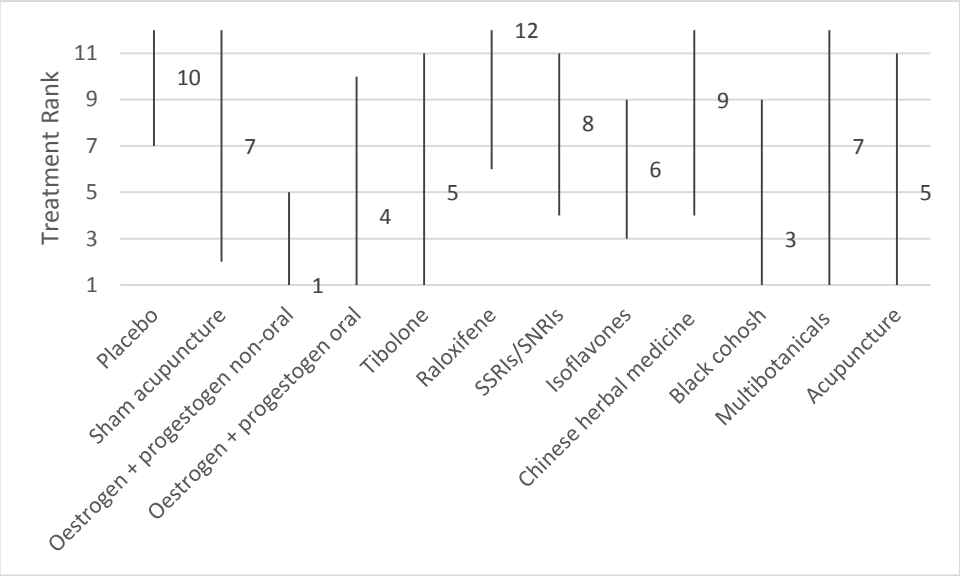
K.5 Additional information on networks

K.5.1 Treatment rankings

Women with a uterus

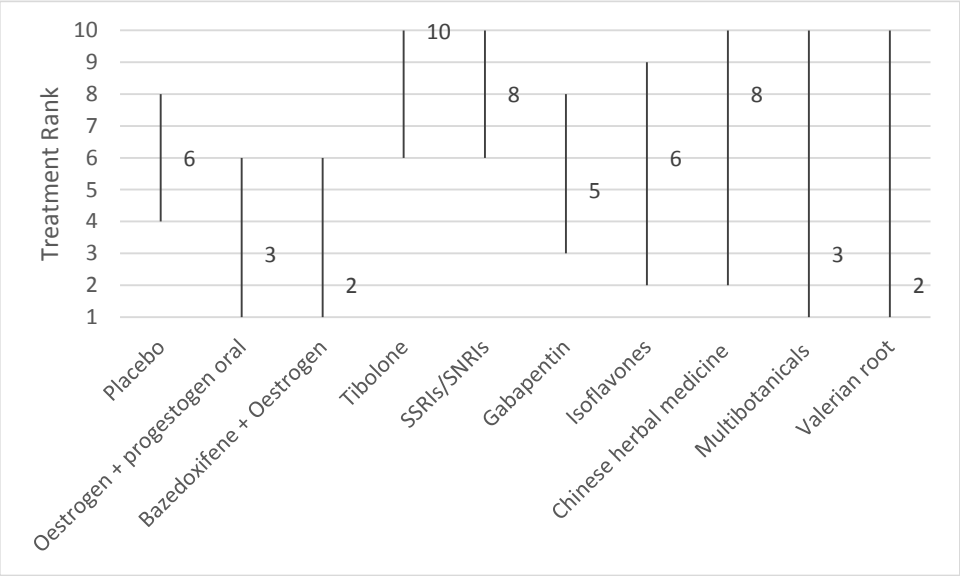
VMS

Figure 77: Median rankings (with their 95% CrI) for each intervention. Lower rank is associated with improvement in outcome



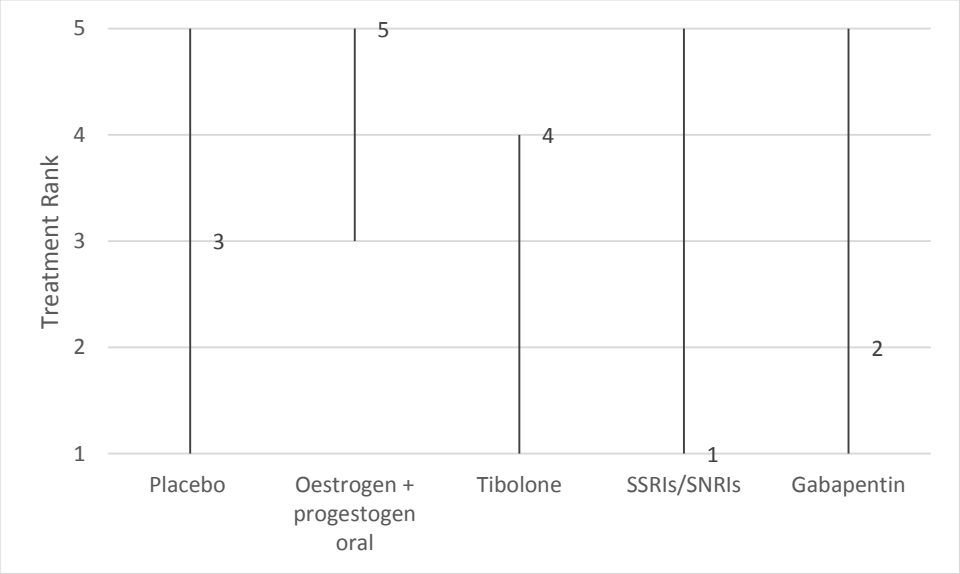
Discontinuation of treatment

Figure 78: Median rankings (with their 95% CrI) for each intervention. Lower rank is associated with improvement in outcome



Bleeding

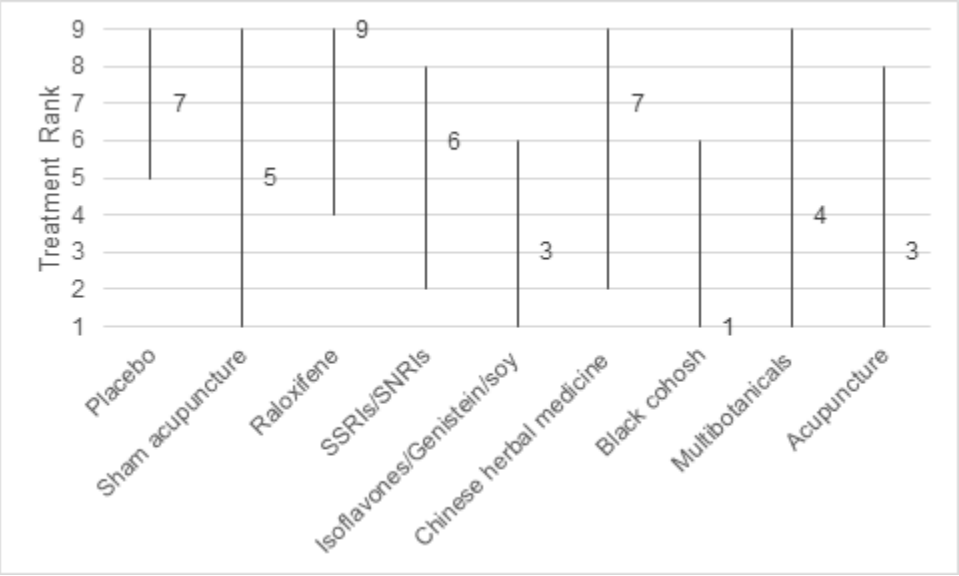
Figure 79: Median rankings (with their 95% CrI) for each intervention. Lower rank is associated with improvement in outcome



Women without a uterus

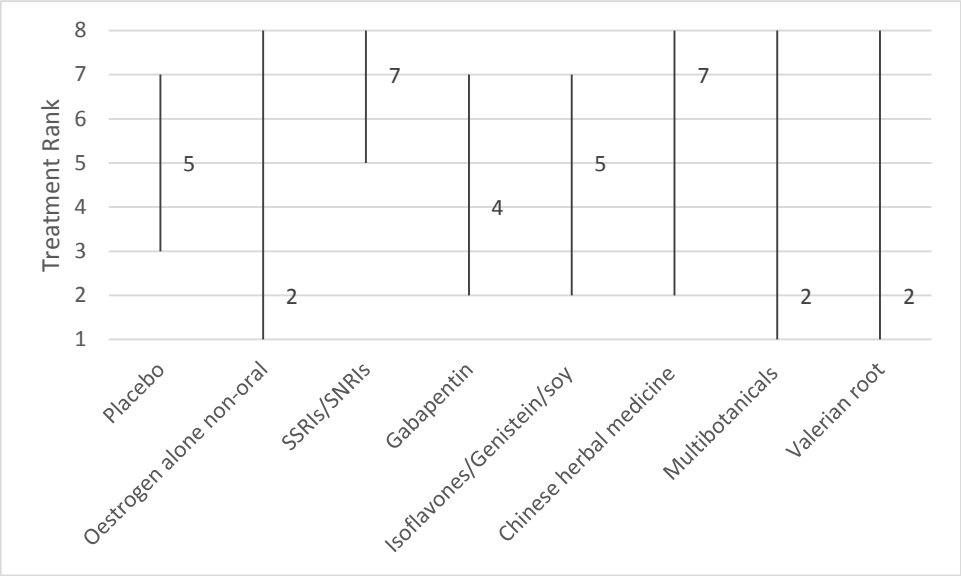
VMS

Figure 80: Median rankings (with their 95% CrI) for each intervention. Lower rank is associated with improvement in outcome



Discontinuation of treatment

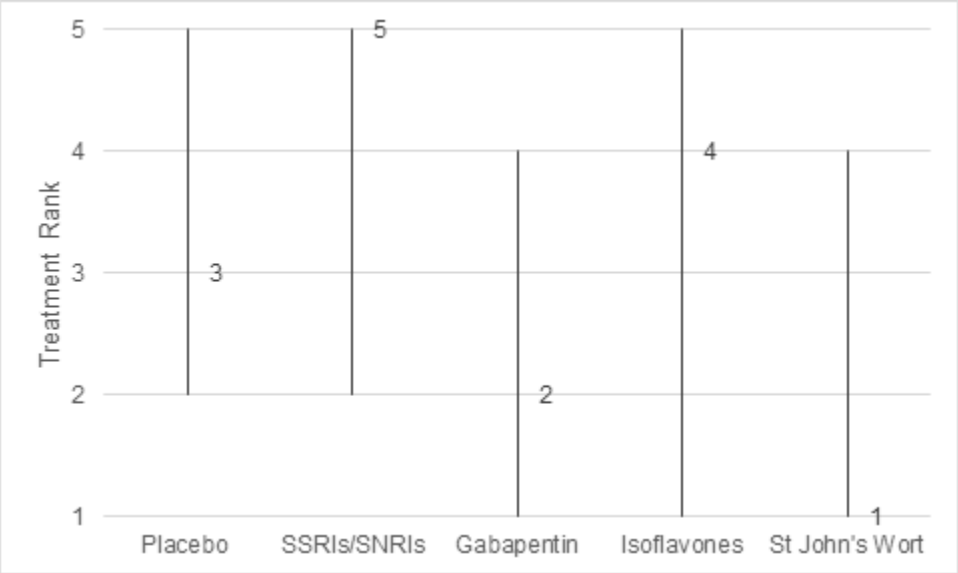
Figure 81: Median rankings (with their 95% CrI) for each intervention. Lower rank is associated with improvement in outcome



Women with breast cancer/history of breast cancer

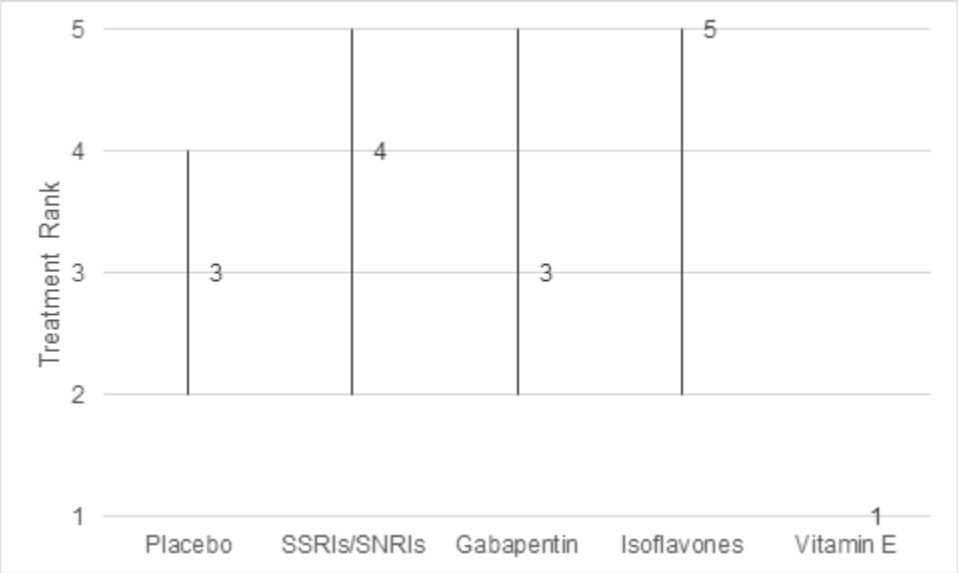
VMS

Figure 82: Median rankings (with their 95% CrI) for each intervention. Lower rank is associated with improvement in outcome



Discontinuation of treatment

Figure 83: Median rankings (with their 95% CrI) for each intervention. Lower rank is associated with improvement in outcome



K.5.2 Model fit

Women with a uterus

VMS

Both fixed and random effects models were fitted. Table 19 presents results of between-study heterogeneity for the random effect model with fixed dose effects, the random effect model with exchangeable dose effects, and goodness of fitness of all three models. DIC suggested that there was not more than a 5 point difference between the random effects model with fixed dose effects and the random effects model with exchangeable dose effects. In addition, the residual deviance showed that the random effects model with fixed dose effects fitted the data similarly to the random effects model with exchangeable dose effects, as the residual deviance (76.12 vs. 76.2) were both similar to the number of unconstrained data points, 74. Therefore, the results of the random effects model with fixed dose effects are presented for this network, as the simpler of the two models.

Table 19: Measures of fitness of fixed (FE) and random (RE) effects models

	FE model	RE model (Fixed dose effects)	RE model (Exchangeable dose effects)
Measure of between study heterogeneity			
Standard deviation on the log MRs scale (SD) (95% CrI)		0.50 (0.37, 0.70)	0.46 (0.32, 0.65)
Measure of common within-class variance			
Standard deviation on the log MRs scale (SD) (95% CrI)			0.26 (0.02, 0.70)
Measure of goodness-of-fit			
Residual Deviance (r)*	293.1	76.12	76.2
Deviance information criteria (DIC)	668.635	469.4	470.06

FE model: fixed effect model, RE model: random effect model,^ Values of SD from 0.1 to 0.5 are reasonable, from 0.5 to 1.0 are considered fairly high and greater than 1.0 represent extreme heterogeneity.
** Compared to 74 data points*

Discontinuation of treatment

Both fixed and random effects models were fitted. **Table 20** presents results of between-study heterogeneity for the random effect model with fixed dose effects, the random effect model with exchangeable dose effects, and goodness of fitness of all three models. DIC suggested that there was not more than a 5 point difference between any of the models. However, the residual deviance for the random effects model with fixed dose effects (45.41) was slightly closer to the number of unconstrained data points (45) than either of the other models (47.71 and 45.95 for the fixed effects and random effects with exchangeable dose effects respectively).Therefore, the results of the random effects model with fixed dose effects are presented for this network.

Table 20: Measures of fitness of fixed (FE) and random (RE) effects models

	FE model	RE model (Fixed dose effects)	RE model (Exchangeable dose effects)
Measure of between study heterogeneity			
Standard deviation on the log MRs scale (SD) (95% CrI)		0.25 (0.01, 0.70)	0.25 (0.01, 0.72)

	FE model	RE model (Fixed dose effects)	RE model (Exchangeable dose effects)
Measure of common within-class variance			
Standard deviation on the log MRs scale (SD) (95% CrI)			0.25 (0.01, 1.07)
Measure of goodness-of-fit			
Residual Deviance (r)*	47.71	45.41	45.95
Deviance information criteria (DIC)	243.77	245.10	246.74

FE model: fixed effect model, RE model: random effect model,^ Values of SD from 0.1 to 0.5 are reasonable, from 0.5 to 1.0 are considered fairly high and greater than 1.0 represent extreme heterogeneity.
** Compared to 45 data points*

Bleeding

Both fixed and random effects models with an empirical prior for heterogeneity were fitted. Table 21 presents results of between-study heterogeneity for the random effect model with fixed dose effects and goodness of fitness for both models. A random effect model with exchangeable treatment effects was not fitted as the data were too sparse. DIC suggested that there was not more than a 5 point difference between either of the models. Although the residual deviance for the random effects model (11.45) was slightly closer to the number of unconstrained data points (10) the fixed effects model (12.7), the estimate of heterogeneity for the random effects model was reasonably unstable, and was strongly influenced by the prior. Therefore, the results of the fixed effects model are presented for this network.

Table 21: Measures of fitness of fixed (FE) and random (RE) effects models

	FE model	RE model (Fixed dose effects)
Measure of between study heterogeneity (with empirical prior)		
Standard deviation on the log ORs scale (SD) (95% CrI)	-	0.15 (0.03, 0.54)
Measure of goodness-of-fit		
Residual Deviance (r)*	12.7	11.45
Deviance information criteria (DIC)	59.2	58.45

FE model: fixed effect model, RE model: random effect model,^ Values of SD from 0.1 to 0.5 are reasonable, from 0.5 to 1.0 are considered fairly high and greater than 1.0 represent extreme heterogeneity.
** Compared to 10 data points*

Women without a uterus

VMS

Both fixed and random effects models were fitted. Table 22 presents results of between-study heterogeneity for the random effect model with fixed dose effects, the random effect model with exchangeable dose effects, and goodness of fitness of all three models. DIC suggested that there was not more than a 5 point difference between the random effects model with fixed dose effects and the random effects model with exchangeable dose effects. In addition, the residual deviance showed that the random effects model with fixed dose effects fitted the data similarly to the random effects model with exchangeable dose effects, as the residual deviance (63.06 vs. 63.31) were both similar to the number of unconstrained data points, 61. Therefore, the results of the random effects model with fixed dose effects are presented for this network, as the simpler of the two models.

Table 22: Measures of fitness of fixed (FE) and random (RE) effects models

	FE model	RE model (Fixed dose effects)	RE model (Exchangeable dose effects)
Measure of between study heterogeneity			
Standard deviation on the log MRs scale (SD) (95% CrI)		0.47 (0.33, 0.70)	0.47 (0.32, 0.71)
Measure of common within-class variance			
Standard deviation on the log MRs scale (SD) (95% CrI)			0.27 (0.01, 1.22)
Measure of goodness-of-fit			
Residual Deviance (r)*	194.2	63.06	63.31
Deviance information criteria (DIC)	503.38	388.83	387.12

FE model: fixed effect model, RE model: random effect model,^ Values of SD from 0.1 to 0.5 are reasonable, from 0.5 to 1.0 are considered fairly high and greater than 1.0 represent extreme heterogeneity.
** Compared to 61 data points*

Discontinuation of treatment

Both fixed and random effects models with an empirical prior for heterogeneity were fitted. Table 23 presents results of between-study heterogeneity for the random effect model with fixed dose effects and goodness of fitness for both models. A random effect model with exchangeable treatment effects was not fitted as the data were too sparse. DIC suggested that there was not more than a 5 point difference between either of the models. The residual deviances for both fixed effect and random effect models (32.38 vs 31.44) were close to the number of unconstrained data points, 32. Therefore, as the simpler model, the results of the fixed effects model are presented for this network.

Table 23: Measures of fitness of fixed (FE) and random (RE) effects models

	FE model	RE model (Fixed dose effects)
Measure of between study heterogeneity (with empirical prior)		
Standard deviation on the log ORs scale (SD) (95% CrI)	-	0.138 (0.033, 0.486)
Measure of goodness-of-fit		
Residual Deviance (r)*	32.38	31.44
Deviance information criteria (DIC)	157.247	157.185

FE model: fixed effect model, RE model: random effect model,^ Values of SD from 0.1 to 0.5 are reasonable, from 0.5 to 1.0 are considered fairly high and greater than 1.0 represent extreme heterogeneity.
** Compared to 32 data points*

Women with breast cancer/history of breast cancer

VMS

Both fixed and random effects models with an empirical prior for heterogeneity were fitted. Table 24 presents results of between-study heterogeneity for the random effect model with fixed dose effects and goodness of fitness for both models. A random effect model with exchangeable treatment effects was not fitted as the data were too sparse. Though DIC for the random effects model was slightly lower than for the fixed effects model, there was not more than a 5 point difference between them. However, the residual deviance for the random effects model (9.78 was slightly closer to the number of unconstrained data points (9) than

the residual deviance for the fixed effects model (10.79). Therefore, the results from the random effects model with an empirical prior on heterogeneity are presented for this network.

Table 24: Measures of fitness of fixed (FE) and random (RE) effects models

	FE model	RE model (Fixed dose effects)
Measure of between study heterogeneity (with empirical prior)		
Standard deviation on the log ORs scale (SD) (95% CrI)	-	0.14 (0.03, 0.48)
Measure of goodness-of-fit		
Residual Deviance (r)*	10.79	9.78
Deviance information criteria (DIC)	54.763	51.366

FE model: fixed effect model, RE model: random effect model,^ Values of SD from 0.1 to 0.5 are reasonable, from 0.5 to 1.0 are considered fairly high and greater than 1.0 represent extreme heterogeneity.
** Compared to 9 data points*

Discontinuation of treatment

Both fixed and random effects models with an empirical prior for heterogeneity were fitted. Table 25 presents results of between-study heterogeneity for the random effect model with fixed dose effects and goodness of fitness for both models. A random effect model with exchangeable treatment effects was not fitted as the data were too sparse. DIC suggested that there was not more than a 5 point difference between either of the models. The residual deviances for fixed effect and random effect models (9.42 vs 9.34) were both close to the number of unconstrained data points, 9. Therefore, as the simpler model, the results of the fixed effects model are presented for this network.

Table 25: Measures of fitness of fixed (FE) and random (RE) effects models

	FE model	RE model (Fixed dose effects)
Measure of between study heterogeneity (with empirical prior)		
Standard deviation on the log ORs scale (SD) (95% CrI)	-	0.13 (0.03, 0.51)
Measure of goodness-of-fit		
Residual Deviance (r)*	9.42	9.34
Deviance information criteria (DIC)	46.03	46.10

FE model: fixed effect model, RE model: random effect model,^ Values of SD from 0.1 to 0.5 are reasonable, from 0.5 to 1.0 are considered fairly high and greater than 1.0 represent extreme heterogeneity.
** Compared to 9 data points*

K.5.3 Full NMA results for vasomotor symptoms in women without uterus

32 trials of nine classes were included in the network of outcome of vasomotor symptoms with a total sample size of 4165 women without uterus in menopause (Figure 65).

Table 26 presents the results of the conventional pair-wise meta-analyses (head to head comparisons) (upper-right section of table), together with the results computed by the NMA for every possible treatment comparison (lower-left section of table). Both results are presented as mean ratios (95% CrI). These results were derived from the random effects model with fixed dose effects (Table 22). Figure 84 graphically presents the results computed by the NMA for each intervention versus placebo.

Isoflavones and black cohosh were both found to be significantly better than placebo on relieving vasomotor symptoms for women in menopause. Black cohosh was also found to be significantly better than raloxifine. No other significant differences were found among other interventions in the network.

In this analysis, black cohosh was found to have the highest probability (57.32%) of being the best treatment to relieve vasomotor symptoms among interventions (Table 27).

Table 26: Mean ratios (95% CrI) from conventional (white area) and network meta-analysis (grey area) for the frequency of vasomotor symptoms for women in menopause without uterus

	Placebo	Sham acupuncture	Raloxifine	SSRIs/SNRIs	Isoflavones/Genistein/soy	Chinese herbal medicine	Black cohosh	Multibotanicals	Acupuncture
Placebo		0.75 (0.27, 1.85)	1.65 (0.63, 4.35)	0.84 (0.55, 1.29)	**0.62 (0.44, 0.86)**	0.95 (0.47, 1.86)	**0.4 (0.17, 0.88)**	0.7 (0.25, 1.99)	0.58 (0.23, 1.33)
Sham acupuncture	0.75 (0.27, 1.85)					1.28 (0.44, 3.88)			0.58 (0.23, 1.33)
Raloxifine	1.65 (0.63, 4.35)	2.21 (0.59, 8.96)							
SSRIs/SNRIs	0.84 (0.55, 1.29)	1.13 (0.41, 3.34)	0.51 (0.18, 1.45)						
Isoflavones/Genistein/soy	**0.62 (0.44, 0.86)**	0.82 (0.32, 2.38)	0.37 (0.13, 1.04)	0.73 (0.43, 1.26)					
Chinese herbal medicine	0.95 (0.47, 1.86)	1.28 (0.44, 3.88)	0.58 (0.17, 1.85)	1.13 (0.5, 2.51)	1.55 (0.71, 3.26)				0.61 (0.21, 1.68)
Black cohosh	**0.4 (0.17, 0.88)**	0.54 (0.16, 1.91)	**0.24 (0.07, 0.84)**	0.48 (0.18, 1.16)	0.66 (0.26, 1.51)	0.42 (0.14, 1.2)			
Multibotanicals	0.7 (0.25, 1.99)	0.95 (0.24, 4.05)	0.43 (0.1, 1.78)	0.84 (0.27, 2.58)	1.15 (0.38, 3.4)	0.74 (0.22, 2.6)	1.75 (0.49, 6.72)		
Acupuncture	0.58 (0.23, 1.33)	0.78 (0.36, 1.69)	0.35 (0.09, 1.26)	0.69 (0.25, 1.75)	0.95 (0.35, 2.29)	0.61 (0.21, 1.68)	1.45 (0.44, 4.72)	0.82 (0.21, 3.1)	

(a) Results in the upper-right area are the mean ratios and 95% credible intervals from the conventional meta-analyses of direct evidence between the column-defined treatments compared to the row-defined treatment. Mean ratio less than 1 favour the column-defined treatment.

(b) Results in the lower-left area are the mean ratios and 95% credible intervals from the random effect model with fixed dose effects of the NMA of direct and indirect evidence between the row-defined treatments compared to the column-defined treatment. Mean ratios less than 1 favour the row-defined treatment.

(c) Numbers in bold denote statistically significant results (95% CI credible intervals do not include 1)

Figure 84: Forest plot showing mean ratios (with their 95% CrI) of NMA estimates for each intervention versus placebo

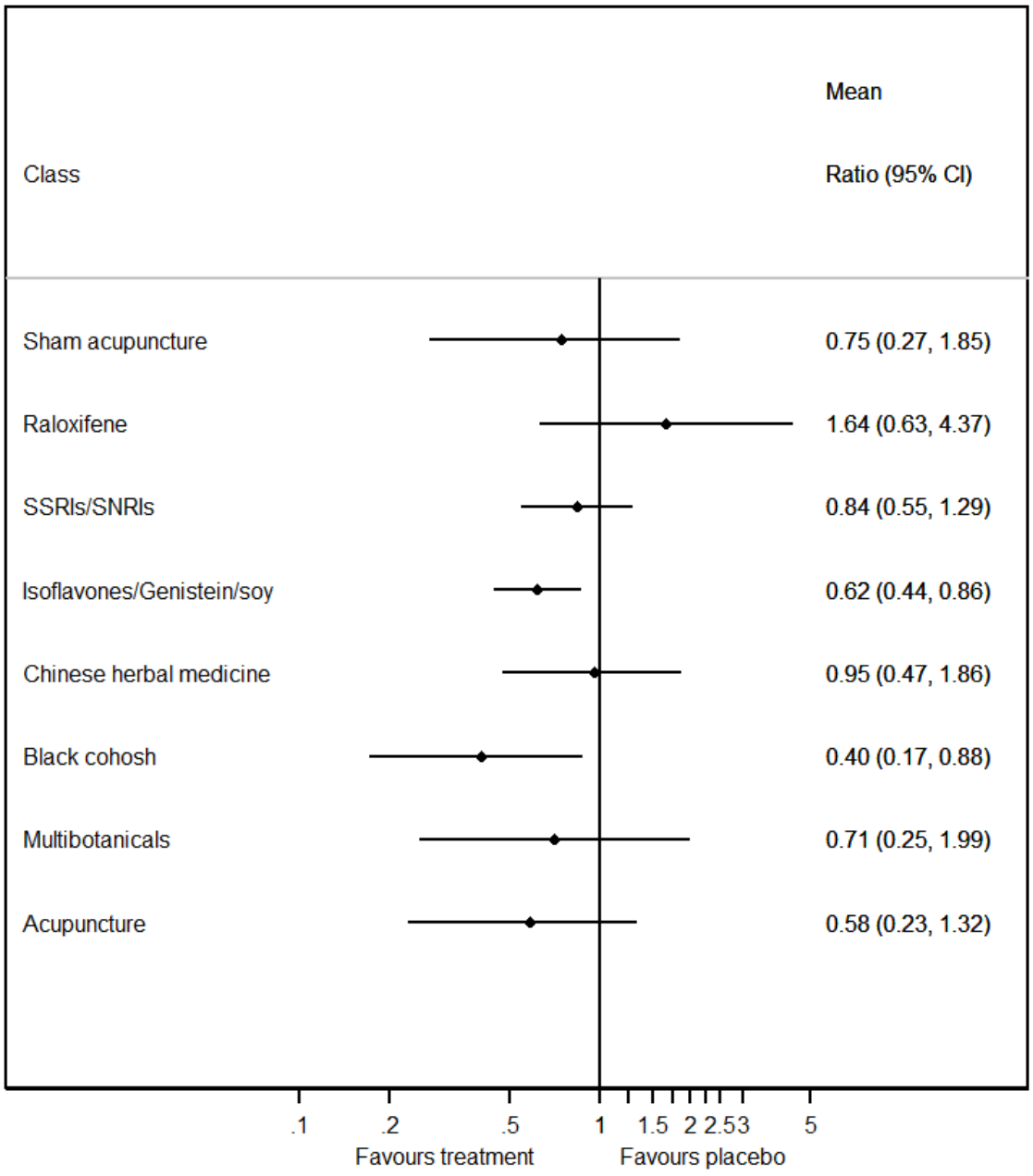


Table 27: Log mean ratios (with their 95% CrI) of all interventions in the network and the probability of being the best treatment for achieving relief of vasomotor symptoms

	Median log mean ratios	95%CrI	Probability of being the best treatment	Median (95% CrI) treatment rank
Placebo	*Baseline treatment*		0.00%	7 (5-9)

	Median log mean ratios	95%CrI	Probability of being the best treatment	Median (95% CrI) treatment rank
Sham acupuncture	-0.29	(-1.31, 0.62)	6.13%	5 (1-9)
Raloxifine	0.50	(-0.45, 1.48)	0.30%	9 (4-9)
SSRIs/SNRIs	-0.17	(-0.6, 0.25)	0.50%	6 (2-8)
Isoflavones/Genistein/soy	-0.48	(-0.81, -0.15)	4.92%	3 (1-6)
Chinese herbal medicine	-0.05	(-0.75, 0.62)	1.07%	7 (2-9)
Black cohosh	-0.90	(-1.76, -0.13)	57.32%	1 (1-6)
Multibotanicals	-0.34	(-1.38, 0.69)	12.50%	4 (1-9)
Acupuncture	-0.54	(-1.49, 0.28)	17.26%	3 (1-8)

Appendix L: Health economics

The health economics is presented in a separate document.

Appendix M: Absolute risk references

The references for the calculations of absolute risk are presented in a separate document.