Evidence evaluation report — Thyroid dysfunction

16 May 2017

Prepared by Ampersand Health Science Writing for the
Australian Government Department of Health

Contents

PROCESS OF THE REVIEW

Research questions

1. What is the prevalence and incidence of thyroid dysfunction in pregnancy, including population specific groups?
2 What is the diagnostic test accuracy of screening for thyroid dysfunction?
3 What are the benefits and harms of routine screening for thyroid dysfunction?
4 When should pregnant women be screened for thyroid dysfunction?
5 What interventions or treatments for thyroid dysfunction are effective and safe in pregnancy, and what advice should women receive?
6 What is the cost effectiveness of universal screening in pregnancy for hypothyroidism?
7 What are the additional considerations for Aboriginal and Torres Strait Islander women?

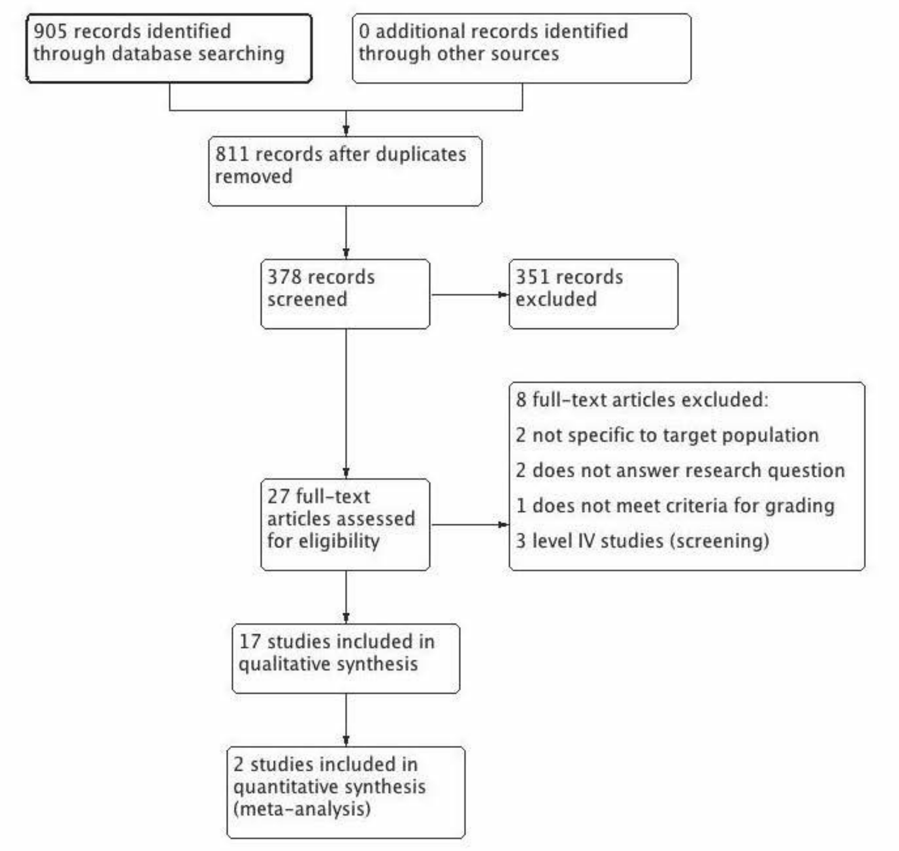
Search strategy

Databases searched:

- EMBASE (OVID) and MEDLINE (OVID) and PSYCHINFO (OVID) = 735
- COCHRANE LIBRARY = 47
- CINAHL = 123
- AUSTRALIAN INDIGENOUS HEALTHINFONET = 0

Date of searches: 15/04/2016

Dates searched: 2013 to present


905 records identified through database searching
0 additional records identified through other sources
811 records after duplicates removed
378 records screened
351 records excluded
8 full-text articles excluded:
2 not specific to target population
2 does not answer research question
1 does not meet criteria for grading
3 level IV studies (screening)
27 full-text articles assessed for eligibility
17 studies included in qualitative synthesis
2 studies included in quantitative synthesis (meta-analysis)

Prisma flow diagram

Full search strategies

MEDLINE AND EMBASE AND PSYCHINFO (OVID)

1. exp Thyroid Diseases/
2. (thyroid$ adj3 (defic$ or insuffic$ or diseas$)).tw.
3. (hypothyr$ or hypo-thyr$ or hyperthyr$ hyper-thyr$ or goitre$ or goiter$).tw.
4. (euthyr$ adj6 (autoimmun$ or autoanti$)).tw.
5. (grave$ adj6 (diseas$ or thyrotoxicos$ or hyperthyr$)).tw.
6. or/1-5
7. exp Pregnancy/
8. exp Pregnancy Complications/
9. exp Perinatal Care/
10. exp Prenatal Care/
11. (pregnan$ or antepart$ or prenatal$ or antenatal$ or perinatal$ or obstetric$ or maternal$).tw.
12. or/7-11
13. 6 and 12
14. exp Animals/ not Humans/
15. 13 not 14
16. 2013 to current

COCHRANE LIBRARY

1. MeSH descriptor: [Thyroid Diseases] explode all trees
2. (thyroid* near/3 (defic* or insuffic* or diseas*)):ti,ab,kw
3. (hypothyr* or hypo-thyr* or hyperthyr* hyper-thyr* or goitre* or goiter*):ti,ab,kw
4. (euthyr* near/6 (autoimmun* or autoanti*)):ti,ab,kw
5. (grave* near/6 (diseas* or thyrotoxicos* or hyperthyr*)):ti,ab,kw
6. #1 or #2 or #3 or #4 or #5
7. MeSH descriptor: [Pregnancy] explode all trees
8. MeSH descriptor: [Pregnancy Complications] explode all trees
9. MeSH descriptor: [Perinatal Care] explode all trees
10. MeSH descriptor: [Prenatal Care] explode all trees
11. (pregnan* or antepart* or prenatal* or antenatal* or perinatal* or obstetric* or maternal*):ti,ab,kw
12. #7 or #8 or #9 or #10 or #11
13. #6 and #12
14. 2013 to current

CINAHL

1. (MH "Thyroid Diseases+")
2. (thyroid* N3 (defic* or insuffic* or diseas*))
3. (hypothyr* or hypo-thyr* or hyperthyr* hyper-thyr* or goitre* or goiter*)
4. (euthyr* N6 (autoimmun* or autoanti*))
5. (grave* N6 (diseas* or thyrotoxicos* or hyperthyr*))
6. S1 OR S2 OR S3 OR S4 OR S5
7. (MH "Pregnancy+")
8. (MH "Pregnancy Complications+")
9. (MH "Perinatal Care")
10. (MH "Prenatal Care+)
11. (pregnan* or antepart* or prenatal* or antenatal* or perinatal* or obstetric* or maternal*)
12. S7 OR S8 OR S9 OR S10 OR S11
13. S6 AND S12
14. 2013 to current

AUSTRALIAN INDIGENOUS HEALTHINFONET

Title: thyroid OR hypothyroid OR hyperthyroid OR hypothyroidism OR hyperthyroidism OR goitre

2013-2016

Exclusion criteria

Full texts of papers within the review period and in English were reviewed. Exclusion criteria included:

- duplicate
- already included in high quality systematic reviews
- not specific to target population (eg specific to non-pregnant women or high-risk women only)

- does not answer research question
- does not meet criteria for grading (eg no outcomes reported, reporting too limited to establish risk of bias, conference abstract)
- narrative review or opinion paper (editorial, letter, comment).

As only low-level evidence was identified for research question 2, level IV studies were included. Level IV studies were excluded for research question 3 (3 papers) as higher-level evidence was identified. Only Level IV evidence was identified for research question 4.

Of 27 studies identified, 19 were included in the appraisal of the evidence.

Assigning level of evidence

Levels of evidence were assigned using the NHMRC levels (diagnostic accuracy for research question 2; screening intervention for research questions 3 and 4) and the definitions given below. No new evidence was identified for research questions 1, 5, 6 and 7.

Designations of levels of evidence according to type of research question

Level	Diagnostic accuracy	Screening intervention
I	A systematic review of level II studies	A systematic review of level II studies
II	A study of test accuracy with an independent, blinded comparison with a valid reference standard, among consecutive persons with a defined clinical presentation	A randomised controlled trial
III-1	A study of test accuracy with independent, blinded comparison with a valid reference standard, among non-consecutive persons with a defined clinical presentation	Pseudo-randomised controlled trial (ie alternate allocation or some other method)
III-2	A comparison with reference standard that does not meet the criteria required for Level II and III-1 evidence	A comparative study with concurrent controls: ▪ Non-randomised, experimental trial ▪ Cohort study ▪ Case-control study
III-3	Diagnostic case-control study	A comparative study without concurrent controls: ▪ Historical control study ▪ Two or more single arm study
IV	Study of diagnostic yield (no reference standard)	Case series

Source: NHMRC (2009) *NHMRC levels of evidence and grades of recommendations for developers of guidelines*.

Study design definitions

- **A study of test accuracy with: an independent, blinded comparison with a valid reference standard, among consecutive patients with a defined clinical presentation** — a cross-sectional study where a consecutive group of people from an appropriate (relevant) population receive the test under study (index test) and the reference standard test. The index test result is not incorporated in (is independent of) the reference test result/final diagnosis. The assessor determining the results of the index test is blinded to the results of the reference standard test and vice versa.
- **A study of test accuracy with: an independent, blinded comparison with a valid reference standard, among non-consecutive patients with a defined clinical presentation** — a cross-sectional study where a non-consecutive group of people from an appropriate (relevant) population receive the test under study (index test) and the reference standard test. The index test result is not incorporated in (is independent of) the reference test result/final diagnosis. The assessor determining the results of the index test is blinded to the results of the reference standard test and vice versa.

- **Case series** — a single group of people exposed to the intervention (factor under study). **Post-test** – only outcomes after the intervention (factor under study) are recorded in the series of people, so no comparisons can be made. **Pre-test/post-test** – measures on an outcome are taken before and after the intervention is introduced to a series of people and are then compared (also known as a 'before- and-after study').
- **Case-control study** — people with the outcome or disease (cases) and an appropriate group of controls without the outcome or disease (controls) are selected and information obtained about their previous exposure/non-exposure to the intervention or factor under study.
- **Diagnostic (test) accuracy** – in diagnostic accuracy studies, the outcomes from one or more diagnostic tests under evaluation (the *index test*/s) are compared with outcomes from a *reference standard test*. These outcomes are measured in individuals who are suspected of having the condition of interest. The term *accuracy* refers to the amount of agreement between the index test and the reference standard test in terms of outcome measurement. Diagnostic accuracy can be expressed in many ways, including sensitivity and specificity, likelihood ratios, diagnostic odds ratio, and the area under a receiver operator characteristic (ROC) curve.
- **Diagnostic case-control study** – the index test results for a group of patients already known to have the disease (through the reference standard) are compared to the index test results with a separate group of normal/healthy people known to be free of the disease (through the use of the reference standard). In this situation patients with borderline or mild expressions of the disease, and conditions mimicking the disease are excluded, which can lead to exaggeration of both sensitivity and specificity. This is called spectrum bias because the spectrum of study participants will not be representative of patients seen in practice. *Note: this does not apply to well-designed population based case-control studies.*
- **Historical control study** – outcomes for a prospectively collected group of people exposed to the intervention (factor under study) are compared with either (1) the outcomes of people treated at the same institution prior to the introduction of the intervention (ie. control group/usual care), or (2) the outcomes of a previously published series of people undergoing the alternate or control intervention.
- **Non-randomised, experimental trial** - the unit of experimentation (eg. people, a cluster of people) is allocated to either an intervention group or a control group, using a non-random method (such as patient or clinician preference/availability) and the outcomes from each group are compared. This can include:
  - **a controlled before-and-after study**, where outcome measurements are taken before and after the intervention is introduced, and compared at the same time point to outcome measures in the (control) group.
  - **an adjusted indirect comparison**, where two randomised controlled trials compare different interventions to the same comparator ie. the placebo or control condition. The outcomes from the two interventions are then compared indirectly.
- **Prospective cohort study** — where groups of people (cohorts) are observed at a point in time to be *exposed or not exposed* to an intervention (or the factor under study) and then are followed prospectively with further outcomes recorded as they happen.
- **Pseudo-randomised controlled trial** - the unit of experimentation (eg. people, a cluster of people) is allocated to either an intervention (the factor under study) group or a control group, using a pseudo-random method (such as alternate allocation, allocation by days of the week or odd-even study numbers) and the outcomes from each group are compared.
- **Randomised controlled trial** — the unit of experimentation (eg. people, or a cluster of people4) is allocated to either an intervention (the factor under study) group or a control group, using a random mechanism (such as a coin toss, random number table, computer-generated random numbers) and the outcomes from each group are compared.

- **Retrospective cohort study** — where the cohorts (groups of people exposed and not exposed) are defined at a point of time in the past and information collected on subsequent outcomes, eg. the use of medical records to identify a group of women using oral contraceptives five years ago, and a group of women not using oral contraceptives, and then contacting these women or identifying in subsequent medical records the development of deep vein thrombosis.
- **Study of diagnostic yield** — these studies provide the yield of diagnosed patients, as determined by the index test, without confirmation of the accuracy of the diagnosis (ie. whether the patient is actually diseased) by a reference standard test.
- **Systematic literature review** — systematic location, appraisal and synthesis of evidence from scientific studies.
- **Two or more single arm study** – the outcomes of a single series of people receiving an intervention (case series) from two or more studies are compared.

Source: NHMRC (2009) *NHMRC levels of evidence and grades of recommendations for developers of guidelines.*

Selection of outcomes for GRADE analysis

Outcomes considered for inclusion comprised conditions known to be associated with thyroid dysfunction in pregnancy. Seven outcomes were selected on the basis of clinical impact.

Outcome	Importance	Inclusion
Neurosensory disability of the infant as child	9	☑
Diagnosis of hypothyroidism	7	☑
Diagnosis of hyperthyroidism	7	☑
Diagnosis of thyroid autoimmunity	5	☒
Pharmacological treatment for thyroid dysfunction	7	☒
Pre-eclampsia	7	☑
Preterm birth	7	☑
Fetal or neonatal death	9	☑
Miscarriage	7	☑
Gestational diabetes	5	☒
Gestational hypertension	5	☒
Macrosomia	5	☒
Placental abruption	5	☒
Apgar score <7 5 mins	5	☒
Mode of birth (Caesarean section)	5	☑

Key: 1 – 3 less important; 4 – 6 important but not critical for making a decision; 7 – 9 critical for making a decision

Evidence tables

1. *What is the prevalence and incidence of thyroid dysfunction in pregnancy, including population specific groups?*

Evidence summary

Results of previous review

Studies identified in the literature review conducted to inform the development of Module II of the Guidelines (Australian Health Ministers' Advisory Council 2014) were included in the narrative. It was noted that, with iodine fortification of bread in Australia most women would be entering pregnancy with adequate iodine intake. Moderate iodine deficiency was noted in some African countries, Afghanistan, Belarus and Vietnam and urinary iodine levels associated with a high risk of hyperthyroidism or autoimmune disease were identified in Brazil, Chile, Ecuador, Liberia and Uganda.

Results of current review

No prevalence studies of relevance to the Australian context were identified.

Additional information

The National Health Survey (ABS 2014) showed that in 2011–2012 (2 years after introduction of mandatory fortification) iodine levels were relatively low among women of childbearing age. Although women aged 16–44 years had sufficient iodine levels overall (a median UIC of 121.0 μg/L), around 18.3% had iodine levels <50 μg/L, compared with the national average of 12.8%. Likewise, nearly two thirds (62.2%) had a UIC <150 μg/L, which is the iodine level recommended by WHO for pregnant and breastfeeding women.

Advice to EWG

Incorporate results of the National Health Survey into the background section.

2 What is the diagnostic accuracy of testing for thyroid dysfunction?

Evidence summary

Results of the previous review

Studies identified in the literature review conducted to inform the development of Module II of the Guidelines (Australian Health Ministers' Advisory Council 2014) were included in the narrative. It was noted that pregnancy-specific ranges for thyroid-stimulating hormone (TSH) should be used.

Results of the current review

Test

One Level IV study (Foley et al 2013) with potential for bias reported on the validity of the dried blood spotted filter paper specimen to detect autoimmune thyroiditis and primary hypothyroidism in pregnant women but did not report on sensitivity or specificity compared to serum testing.

Reference ranges

Seven observational studies (Ekinci et al 2013; Moradi et al 2013; Bautista et al 2014; Bliddal et al 2014; Khalid et al 2014; Wilson et al 2014; Fan et al 2016) explored reference ranges for thyroid function tests in pregnancy. Of the six studies that determined reference ranges, one was conducted in Australia (Ekinci et al 2013), four were conducted in areas of iodine deficiency (Moradi et al 2013; Bautista et al 2014; Bliddal et al 2014; Khalid et al 2014) and two were conducted among a largely Hispanic population (Wilson et al 2014; Bryant et al 2015).

One study (Bryant et al 2015) compared pregnancy outcomes of women with initial TSH >4.5 mU/L with those with initial TSH 3–4.5 mU/L and found that adverse outcomes were associated with the former, suggesting that TSH >4.5 mU/L may be a more clinically relevant threshold.

Two studies found significant differences in reference ranges resulting from different blood sampling methods (sequential vs non-sequential)(Fan et al 2016) or different immunoassays (Bliddal et al 2014; Fan et al 2016).

Advice to EWG

The identified evidence for Research Question 2 is insufficient to support a recommendation on diagnostic test accuracy but may inform the narrative.

2.1 Test

Study ref	Design	LoE	N	Aim, setting, population, measurements	Results	Comments
(Foley et al 2013)	Diagnostic yield	IV	494	**Aim**: to confirm the validity of the dried blood spotted (DBS) filter paper specimen as a very practical and accurate method to detect autoimmune thyroiditis and primary hypothyroidism in pregnant women **Setting:** Pittsburgh, Pennsylvania **Population**: first-trimester pregnant women with no exclusion criteria. **Measurements**: Finger stick blood was applied to filter paper, dried in room air, eluted, and promptly tested for TSH and TAb. A total of 178 of the pregnant women (36%) were tested in the early postpartum. Women with abnormal results had confirmatory serum tests.	Abnormal TSH values (>4.0 mU/L) and/or positive Tab — pregnant women: 18.4% (n=91); postpartum women: 24.2% (n=43) TSH values >2.5 mU/L — pregnant women: 28.3%. All subjects with TSH values >4.0 mU/L tested positive for TAb. 18 women (3.6%) who tested normal during pregnancy tested abnormal in the postpartum. Results are comparable to serum data in this population published in the literature.	Sensitivity and specificity not reported. Authors affiliated with Neo Gen

2.2 Reference ranges

Study ref	Design	LoE	N	Aim, setting, population, measurements	Results	Comments
(Bryant et al 2015)	Diagnostic yield	IV	26,518	**Aim**: To evaluate pregnancy outcomes of hypothyroidism identified in a population-based prenatal testing program **Setting**: Parkland Hospital, Dallas, Texas **Population**: All women attending the hospital for antenatal care. **Observations**: This is a secondary analysis of a prospective prenatal population-based study in which serum thyroid analytes were obtained. Initial testing thresholds were intentionally inclusive (TSH >3.0 mU/L; free thyroxine <0.9 ng/dL); those who tested positive were referred for confirmatory testing. Hypothyroidism was identified and treated if TSH level was >4.5 mU/L and if fT4 level was <0.76 ng/dL. Perinatal outcomes in these women and those who tested positive but were not confirmed as having hypothyroidism (and were untreated) were compared with women with euthyroidism. Outcomes were then analyzed according to initial TSH levels.	At initial testing 24,584 women (93%) were euthyroid, and 284 women (1%) had abnormal initial values that suggested hypothyroidism. Of those referred, 232 (82%) underwent repeat testing, and 47 (0.2% initially tested) were confirmed to have hypothyroidism. Perinatal outcomes of women with treated overt hypothyroidism were similar to women with euthyroidism. Higher rates of pregnancy-related hypertension were identified in the 182 women with unconfirmed hypothyroidism when compared with women with euthyroidism (P<0.001); however, this association was seen only in women with initial TSH >4.5 mU/L (aOR 2.53; 95%CI 1.4–4.5).	

Study ref	Design	LoE	N	Aim, setting, population, measurements	Results	Comments
(Bliddal et al 2014)	Case-control	III-3	356	**Aim:** To establish two independent sets of longitudinal gestational age-specific reference ranges, and subsequently evaluate the use of these when applied to another comparable cohort. **Setting:** Copenhagen University Hospital **Population**: Comparison of two longitudinal prospective cohort studies including 255 (cohort 1) and 101 (cohort 2) healthy antibody-negative pregnant women attending prenatal care in a mildly iodine-deficient area. **Observations**: Different immunoassays were used to measure thyroid hormone levels in the two cohorts. Thyroid hormone reference ranges were established for every 5 weeks of gestation. Differences between cohorts were explored through mixed-model repeated measures regression analyses. By applying reference ranges from one cohort to the other, the proportion of women who would be misclassified by doing so was investigated.	TSH increased and free thyroxine (FT4) decreased as pregnancy progressed. Results indicated highly significant differences between cohorts in free triiodothyronine (F=21.3, P<0.001) and FT4 (F=941, P<0.001). TSH levels were comparable (P=0.09). Up to 90.3% of the women had FT4 levels outside their laboratory's nonpregnant reference range, and up to 100% outside the other cohort's gestational-age-specific reference ranges. Z-score-based reference ranges markedly improved comparison between cohorts. Even in the same region, the use of gestational-age-specific reference ranges from different laboratories led to misclassification. Up to 100% of maternal FT4 levels fell outside the other cohort's reference range despite similar TSH levels.	May not be applicable to the Australian context.

Study ref	Design	LoE	N	Aim, setting, population, measurements	Results	Comments
(Fan et al 2016)	Case-control	III-3	647	**Aim**: To compare nonsequential and sequential methods for the evaluation of maternal thyroid function. **Setting**: International Peace Maternity and Child Health Hospital, Shanghai **Population**: Han Chinese women with no history of thyropathy or autoimmune disease, no goitre, TPOAb negative and no use of medicine affecting the thyroid hormone. The study area is an iodine- stable and adequate area. **Measurements**: Serum thyroid stimulating hormone (TSH) and free thyroxine (FT4) were measured using the Abbott and Roche kits. There were 447 and 200 patients enrolled in the nonsequential and sequential groups, respectively. The central 95% range between the 2.5th and the 97.5th percentiles was used as the reference interval for the thyroid function parameter	The nonsequential group exhibited a significantly larger degree of dispersion in the TSH reference interval during the 2nd and 3rd trimesters as measured using both the Abbott and Roche kits (all $P < 0.05$). The TSH reference intervals were significantly larger in the nonsequential group than in the sequential group during the 3rd trimester as measured with both the Abbott (4.95 vs. 3.77 mU/L, $P < 0.001$) and Roche kits (6.62 vs. 5.01 mU/L, $P = 0.004$). The nonsequential group had a significantly larger FT4 reference interval as measured with the Abbott kit during all trimesters (12.64 vs. 5.82 pmol/L; 7.96 vs. 4.77 pmol/L; 8.10 vs. 4.77 pmol/L, respectively, all $P < 0.05$), whereas a significantly larger FT4 reference interval was only observed during the 2nd trimester with the Roche kit (7.76 vs. 5.52 pmol/L, $P = 0.002$).	

Study ref	Design	LoE	N	Aim, setting, population, measurements	Results	Comments
(Bautista et al 2014)	Diagnostic yield	IV	200	**Aim:** To establish trimester-specific, laboratory assay specific normative thyroid function test values **Setting**: Philippine General Hospital (PGH) out-patient services **Population**: Healthy third-trimester pregnant Filipino women (Philippines is classified as having an inadequate iodine intake with a mildly deficient iodine status). **Measurements**: Serum TSH, FT4, FT3, and TPOAb were measured. Reference intervals are based on 2.5th and 97.5th percentiles for TSH, FT4, and FT3 among TPOAb-negative third-trimester pregnant Filipino patients.	Reference ranges for TSH, FT4 and FT3 in TPOAb-negative third-trimester pregnant population are as follows: • TSH= 0.2-3.0 uIU/mL; • FT4 = 9.16-18.64 pmol/L • FT3= 2.09-3.7 pmol/L.	May not be applicable to the Australian context.

Study ref	Design	LoE	N	Aim, setting, population, measurements	Results	Comments
(Ekinci et al 2013)	Diagnostic yield	IV	130	**Aim**: To establish reference intervals for thyroid stimulating hormone (TSH), free thyroxine (fT4) and to track intraindividual changes in thyroid function throughout pregnancy. **Setting**: Mercy Hospital for Women, Melbourne **Population**: Women without antithyroid peroxidase antibodies. **Measurements**: Thyroid function was determined at T1 (9–13 wks); T2 (22–26 wks); T3 (35–39 wks) and postpartum (8–12 wks). A subgroup (n=47) was used to track intraindividual changes using postpartum as non-pregnant state (baseline).	For trimesters 1–3, TSH (median (2.5th, 5th, 95th and 97.5th percentile)) was 0.77 (0.03, 0.05, 2.33, 3.05), 1.17 (0.42, 0.47, 2.71, 3.36) and 1.35 (0.34, 0.42, 2.65, 2.83) mIU/L, respectively. Free T4 (mean (95%CI)) was 10.7 (5.9–15.5), 8.1 (4.9–11.3), 7.8 (4.5–11.0) pmol/L, respectively. In T2 and T3, 36% and 41% of the fT4 values, respectively, fell below the non-pregnancy lower normal limit. In the subgroup assessed for longitudinal changes, of the women with baseline TSH4median, 71–75% remained at or below the corresponding median for trimesters 1–3. Of the women with baseline fT44median, 69–81% also remained at or below the corresponding median for trimesters 1–3. High correlation was observed at different trimesters and baseline for TSH (Spearman's r: 0.593–0.846, P<0.001) and for fT4 (r: 0.480–0.739, P<0.001). Use of trimester-specific RIs would prevent misclassification of thyroid function during pregnancy. In the majority of women, TSH and fT4 tracked on the same side of the median distribution, from a non-pregnant baseline, throughout pregnancy.	Australian study

Study ref	Design	LoE	N	Aim, setting, population, measurements	Results	Comments
(Khalid et al 2014)	Diagnostic yield	IV	351	**Aim**: To establish trimester-specific thyroid function reference intervals throughout pregnancy, and to examine the prevalence of thyroid autoimmunity in otherwise euthyroid women. **Setting**: Cork University Hospital, Ireland **Population**: Pregnant women (median age 30) attending a large, tertiary referral maternity hospital. Patients with known thyroid disorders, autoimmune disease, recurrent miscarriage, hyperemesis gravidarum and pre-eclampsia were excluded. **Measurements**: TFTs were analysed in the CUH biochemistry laboratory using Roche Modular E170 electrochemiluminescent immunoassay. Trimester-specific reference ranges (2.5th, 50th and 97.5th centiles) were calculated.	TSH concentrations showed slightly increasing median centile throughout gestation (. Free thyroxine (T4) and T3 decreased throughout gestation. Derived reference ranges (2.5th to 97.5th percentiles) for TSH exclusive of TPO positive women were: Week 12: 0.2–3.0 Week 14–16: 0.2–3.1 Week 18–22: 0.3–3.1 Week 24: 0.4–3.1 Week 26–28: 0.4–3.2 Week 30–32: 0.5–3.2 Week 34: 0.5–3.3 Week 36–38: 0.6–3.3 Week 40: 0.7–3.3	

Study ref	Design	LoE	N	Aim, setting, population, measurements	Results	Comments
(Moradi et al 2013)	Diagnostic yield	IV	584	**Aim:** To determine the reference ranges for free triiodothyronine (FT3), free thyroxin (FT4) and thyroid stimulating hormone (TSH) in Iranian pregnant women **Setting**: Akbarabadi University Hospital, Iran **Population**: 162 in first trimester and 422 in the third trimester from an area with low iodine intake. Women with a known history of diabetes mellitus, thyroid disease, thyroid or anti-thyroid drugs consumption and family history of thyroid disease were excluded. **Measurements**: A single blood sample from 584 pregnant women was analysed for thyroid function. Serum levels of TSH, FT4, FT3, total T4 (TT4), T3 resin uptake (T3RU) and anti-thyroid peroxidase antibody (TPO Ab) were measured. Urinary iodine was determined in some cases.	The 2.5th and 97.5th percentiles values were used to determine the reference ranges for FT3, FT4, TT4, T3RU and TSH. These values for TPOAb negative women were: • FT3=1.4–2.9 pmol/L, • FT4=7.1–18 pmol/L; • TT4=7.2–13.5 micro g/dL • T3RU=20.0–30.0 % • TSH=0.5–3.9 micro g/L. The level of urinary iodine in 80.5% of the subjects was less than normal.	May not be applicable to the Australian context. Reference ranges not specific to stage of gestation.

Study ref	Design	LoE	N	Aim, setting, population, measurements	Results	Comments
(Wilson et al 2014)	Diagnostic yield	IV	17,298	**Aim**: To establish a gestational-age specific curve for serum total thyroxine (T4) levels and to compare pregnancy outcomes of euthyroid women with those identified to have subclinical hypothyroidism (SCH) defined by an elevated TSH level in conjunction with either total T4 or free T4 determinations. **Setting**: Parkland Hospital, Dallas, Texas **Population**: All women presenting for prenatal care. Women with overt thyroid disorders were excluded. **Measurements**: The normal distribution of serum total T4 levels were determined by quantile curves for those tested in the first 20 weeks and who were delivered of a singleton infant weighing at least 500 g. Pregnancy outcomes for women with an elevated TSH and normal *total T4* concentrations were analysed and compared with those of women identified to have SCH defined by normal *free T4* levels.	Of women tested, serum total T4 increased into the second trimester and plateaued around 16 weeks. The upper threshold for total T4 ranged from 12.6 to 16.4 µg/dL, and the lower threshold ranged from 5.3 to 8.0 µg/dL. When combined with elevated TSH levels, free or total T4 determinations are equally sensitive to identify women with SCH who are at increased risk for preterm birth (P = 0.007) and placental abruption (P = 0.013) when compared with euthyroid women.	Study population largely Hispanic women; findings may not be generalisable to other populations. Testing only conducted in the first 20 weeks of gestation.

2.3 Excluded studies for research question 2

Study	Reason for exclusion
Ashoor, G., O. Muto, et al. (2013). "Maternal thyroid function at gestational weeks 11-13 in twin pregnancies." Thyroid 23(9): 1165-1171.	Not specific to target population
Amouzegar, A., M. Khazan, et al. (2014). "An assessment of the iodine status and the correlation between iodine nutrition and thyroid function during pregnancy in an iodine sufficient area." European Journal of Clinical Nutrition 68(3): 397-400.	Does not answer research question
Yoshihara, A., J. Y. Noh, et al. (2015). "Serum human chorionic gonadotropin levels and thyroid hormone levels in gestational transient thyrotoxicosis: Is the serum hCG level useful for differentiating between active Graves' disease and GTT?" Endocrine Journal 62(6): 557-560.	Does not answer research question

3 What are the benefits and harms of routine testing for thyroid dysfunction?

Evidence summary

Results of the previous review

The literature review conducted to inform Module II of the Guidelines (Australian Health Ministers' Advisory Council 2014) found insufficient evidence that identifying and treating thyroid dysfunction in pregnancy improves maternal or fetal outcomes. The guidelines recommended against routine testing (Grade B; based on a systematic review and the two RCTs that are included in the Cochrane review identified in this review) and recommended testing of women with symptoms of or high risk for thyroid dysfunction (Grade B; based on a systematic review and one of the RCTs included in the Cochrane review).

Results of the current review

The review identified a Cochrane review (Spencer et al 2015), one randomised controlled trial (RCT) (Ma et al 2016) that was not included in the Cochrane review and had a high risk of bias, and five observational studies (Qian et al 2013; Ahmed et al 2014; Granfors et al 2014; Yang et al 2014; Nazarpour et al 2016).

Universal testing versus case finding

While universal testing versus case finding for thyroid dysfunction increased diagnosis and subsequent treatment, the Cochrane review (Spencer et al 2015) found no clear differences for the primary outcomes: pre-eclampsia or preterm birth. No clear differences were seen for secondary outcomes, including miscarriage and fetal or neonatal death; data were lacking for the primary outcome, neurosensory disability for the infant as a child, and for many secondary outcomes.

The five observational studies compared universal testing with case finding (Qian et al 2013; Ahmed et al 2014; Granfors et al 2014; Yang et al 2014; Nazarpour et al 2016) but four did not report on maternal or infant outcomes (other than thyroid dysfunction). These studies all found that universal testing increased the number of women diagnosed with thyroid dysfunction. One study that was underpowered to analyse outcomes found no increase in prevalence of adverse outcomes among untested women who had elevated TSH levels and equal prevalence of elevated TSH and overt hypothyroidism in targeted tested and untested women (Granfors et al 2014).

Universal testing versus no testing

The Cochrane review (Spencer et al 2015) found that, though universal testing versus no testing for hypothyroidism similarly increased diagnosis and subsequent treatment, no clear difference was seen for the primary outcome: neurosensory disability for the infant as a child (IQ < 85 at three years); data were lacking for the other primary outcomes: pre-eclampsia and preterm birth, and for the majority of secondary outcomes.

The RCT (Ma et al 2016), compared universal testing with no testing, found a lower risk of miscarriage ($p<0.001$) and macrosomia ($p<0.001$) and a higher risk of Caesarean section ($p=0.002$) among women who were tested and treated than among the control group. There was no significant difference in the risk of preterm birth ($p=0.72$).

Advice to EWG

More evidence is needed to assess the benefits or harms of different testing methods for thyroid dysfunction in pregnancy on maternal, infant and child health outcomes.

The identified evidence for Research Question 3 does not alter the existing recommendation against routine testing for thyroid dysfunction but the evidence base is strengthened by the inclusion of the Cochrane review.

Update narrative to incorporate Cochrane review (Spencer et al 2015).

Evidence statements

Universal testing vs case finding

- Universal testing for thyroid dysfunction identifies more women with hypothyroidism than case finding (high quality evidence) and more women with hyperthyroidism are identified (moderate quality evidence).
- The rate of preterm birth does not differ substantially between women who undergo case finding for thyroid dysfunction and those who are universally tested (high quality evidence).

- Rates of miscarriage, pre-eclampsia and neonatal death are not clearly different between women who undergo case finding for thyroid dysfunction and those who are universally tested (moderate quality evidence).

Universal testing vs no testing

- Universal testing for thyroid dysfunction identifies more women with hypothyroidism than no testing (moderate quality evidence)
- Prevalence of neurosensory disability of the infant is not clearly different between the two groups (moderate quality evidence).
- Rates of miscarriage are lower and Caesarean section are higher among women universally tested for thyroid dysfunction compared to those not tested (low quality evidence).
- Rates of preterm birth are not clearly different between women universally tested for thyroid dysfunction and those not tested (very low quality evidence).

Recommendation

Do not routinely test pregnant women for thyroid dysfunction.

Consensus-based recommendation

Recommend thyroid testing to pregnant women who are at increased risk of thyroid dysfunction.

Summary of findings

Universal testing compared to case finding for thyroid dysfunction in pregnancy

Patient or population: Pregnant women
Setting: Italy
Intervention: Universal testing
Comparison: Case finding

Outcomes	Anticipated absolute effects* (95% CI)		Relative effect (95% CI)	№ of participants (studies)	Quality of the evidence (GRADE)	Comments
	Identification with case finding	Identification with Universal testing				
Hypothyroidism	9 per 1,000	**28 per 1,000** (17 to 46)	**RR 3.15** (1.91 to 5.20)	4562 (1 RCT)	⊕⊕⊕⊕ HIGH	
Hyperthyroidism	1 per 1,000	**4 per 1,000** (1 to 17)	**RR 4.50** (0.97 to 20.82)	4562 (1 RCT)	⊕⊕⊕◯ MODERATE [1]	
Miscarriage	45 per 1,000	**41 per 1,000** (31 to 54)	**RR 0.90** (0.68 to 1.19)	4516 (1 RCT)	⊕⊕⊕◯ MODERATE [1]	
Pre-eclampsia	37 per 1,000	**32 per 1,000** (24 to 44)	**RR 0.87** (0.64 to 1.18)	4516 (1 RCT)	⊕⊕⊕◯ MODERATE [1]	
Preterm birth	66 per 1,000	**65 per 1,000** (52 to 81)	**RR 0.99** (0.80 to 1.24)	4516 (1 RCT)	⊕⊕⊕⊕ HIGH	
Fetal and neonatal death	6 per 1,000	**5 per 1,000** (2 to 12)	**RR 0.92** (0.42 to 2.02)	4516 (1 RCT)	⊕⊕⊕◯ MODERATE [1]	
Neurosensory disability of the infant	No data available					

***The risk in the intervention group** (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI).
CI: Confidence interval; **RR:** Risk ratio

GRADE Working Group grades of evidence
High quality: We are very confident that the true effect lies close to that of the estimate of the effect
Moderate quality: We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different
Low quality: Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect
Very low quality: We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect

[1] Wide CI crossing line of no effect

Source: Adapted from (Spencer et al 2015) (outcome miscarriage included).

Universal testing compared to no testing for thyroid dysfunction in pregnancy

Patient or population: Pregnant women
Setting: UK and Italy, China
Intervention: Universal testing
Comparison: No testing

Outcomes	Anticipated absolute effects* (95% CI)		Relative effect (95% CI)	№ of participants (studies)	Quality of the evidence (GRADE)	Comments
	Risk with no testing	Risk with universal testing				
Diagnosis of hypothyroidism in pregnancy	0 per 1,000	**0 per 1,000** (0 to 0)	**RR 772.92** (104.06 to 5741.25)	23,510 (2 RCTs)	⊕⊕⊕◯ MODERATE [1]	
Diagnosis of hyperthyroidism in pregnancy	**No data available**					
Miscarriage	85 per 1,000	**31 per 1,000** (20 to 49)	**RR 0.36** (0.23 to 0.58)	1,671 (1 RCT)	⊕⊕◯◯ LOW [1,2]	
Caesarean section	335 per 1,000	**410 per 1,000** (363 to 460)	**RR 1.22** (1.08 to 1.39)	1,671 (1 RCT)	⊕⊕◯◯ LOW [1,2]	
Neurosensory disability of the infant	0 per 1,000	**0 per 1,000** (0 to 0)	**RR 0.85** (0.6 to 1.22)	794 (1 RCT)	⊕⊕⊕◯ MODERATE [3]	
Pre-eclampsia	**No data available**					
Preterm birth	48 per 1,000	**44 per 1,000** (28 to 69)	**RR 0.92** (0.59 to 1.44)	1,671 (1 RCT)	⊕◯◯◯ VERY LOW [1,2,3]	
Fetal and neonatal death	**No data available**					

***The risk in the intervention group** (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI).

CI: Confidence interval; **RR:** Risk ratio; **OR:** Odds ratio

GRADE Working Group grades of evidence
High quality: We are very confident that the true effect lies close to that of the estimate of the effect
Moderate quality: We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different
Low quality: Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect
Very low quality: We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect

1. Clustering not taken into account in analysis of one RCT (n=1,671)
2. Significant difference in baseline characteristics
3. Wide CI crossing the line of no effect

Source: Adapted from (Spencer et al 2015) (outcomes miscarriage, Casearean section and preterm birth included).

3.1 Universal testing versus case finding

Thyroid dysfunction

Study ref	Design	LoE	N	Aim, setting, population, measurements/observations	Results	Limitations/ Comments
(Spencer et al 2015)	SLR	I	26,408 (2 RCTs)	In one trial (n=4562), before 11 weeks' gestation, women in the universal testing group, and 'high-risk' women in the case finding group had their sera tested for TSH, fT4 and TPO-Ab; women with hypothyroidism (TSH > 2.5 mIU/litre) received levothyroxine; women with hyperthyroidism (undetectable TSH and elevated fT4) received antithyroid medication.	Compared with the case finding group, more women in the universal testing group were diagnosed with hypothyroidism (RR 3.15, 95%CI 1.91 to 5.20; n=4562; GRADE: high), with a trend towards more women being diagnosed with hyperthyroidism (RR 4.50, 95% CI 0.97 to 20.82; 4562 women; P = 0.05; GRADE: moderate).	Cochrane review
Observational studies						
(Ahmed et al 2014)	Cohort	III-2	168	**Aim:** To compare universal vs targeted testing for thyroid dysfunction and to estimate the prevalence of hypothyroidism. **Setting**: Ain Shams University Hospital, Cairo **Population**: Healthy pregnant women aged >18 with a singleton pregnancy. Women who had used medications that could interfere with the results of thyroid function tests for at least 6 months prior were excluded. **Observations**: Based on data collection and laboratory testing, women were divided into high- (n=64) and low-risk (n=104) groups according to the most recent Endocrine Society clinical practice guidelines, as well as into groups by trimester for application of American Thyroid Association guidelines.	No statistically significant differences were found between the high- and low-risk groups regarding prevalence of either clinical (12.5 vs 13.5%) or subclinical (43.8 vs 42.3%) hypothyroidism (p=0.97), and no significant differences were found regarding the prevalence of hypothyroidism in the first (50.6%), second (60.4%), or third trimester (63.3%)(p=35). Use of the most recent Endocrine Society clinical practice guidelines led to missed detection of clinical or subclinical hypothyroidism in 34.5% of pregnant women.	No studies of iodine status in Egypt have been conducted since 1992, when prevalence of goitre in mothers and preschool-aged children was 6–7%.

Study ref	Design	LoE	N	Aim, setting, population, measurements/observations	Results	Limitations/ Comments
(Granfors et al 2014)	Cohort	III-2	5,254	**Aim**: To evaluate the efficacy of a targeted thyroid testing approach during pregnancy. **Setting**: Uppsala County, Sweden. **Population**: Data were derived from the population-based Uppsala Biobank of Pregnant Women, in which blood samples are collected in conjunction with the routine ultrasound testing in gestational week 17–19. **Observations**: On review of medical records, women who were tested for thyroid dysfunction during pregnancy in clinical practice were identified (n=891). From the remaining untested women, 1,006 women were randomly selected for analyses of thyrotropin (TSH), free thyroxine levels, and thyroid peroxidase antibodies. Thyroid-stimulating hormone levels in both groups were analysed with regard to trimester-specific upper reference levels as recommended by the International Endocrine Society Guidelines.	The proportion of trimester-specific TSH elevation was 12.6% in the targeted thyroid testing group and 12.1% in the untested group (P=0.8; OR 1.04, 95%CI 0.79-1.37). The proportion of overt hypothyroidism was 1.1% and 0.7% in the groups, respectively (P=0.4; OR 1.57, 95% CI 0.55-4.45). The prevalence of trimester-specific elevated TSH and overt hypothyroidism was equal in targeted thyroid tested and untested women.	

Study ref	Design	LoE	N	Aim, setting, population, measurements/observations	Results	Limitations/ Comments
(Nazarpour et al 2016)	Cohort	III-2	1,600	**Aim:** To compare universal testing with targeted high-risk case findings for early diagnosis of thyroid disorders in Iranian pregnant women. **Setting:** Iran **Population**: Pregnant women in their first trimester, of whom 656 (44.3%) had at least one risk factor for thyroid diseases and were eligible for the targeted high-risk case finding (high-risk group) approach, while 55.7% had no risk factors (low-risk group; n=944). **Observations**: A checklist, including all related risk factors recommended by the American Thyroid Association, was completed for all participants. Serum concentrations of thyroxine (T4), T-uptake, TSH and thyroid peroxidase antibody (TPOAb) were measured and thyroid status was documented, based on hormonal measurements and clinical examinations.	Using the universal testing approach, there were 974 women (65.8%) with normal thyroid status and 506 participants (34.2%) with thyroid disturbances, including overt hyperthyroidism (0.7%), overt hypothyroidism (1.1%), subclinical hypothyroidism (30.1%; positive TPOAb (5.5%) and negative TPOAb (24.6%); and euthyroid and positive TPOAb (2.3%). Of women with thyroid dysfunction, 64.4% were in the high-risk group and 35.6% were in the low-risk group (P<0.0001). The targeted high-risk case finding approach overlooks about one-third of pregnant women with thyroid dysfunction. If ongoing prospective trials provide evidence on the efficacy of treating subclinical hypothyroidism in pregnancy, in populations with a low prevalence of presumed risk factors, the targeted high-risk case finding approach will be proven inefficient.	

Study ref	Design	LoE	N	Aim, setting, population, measurements/observations	Results	Limitations/ Comments
(Qian et al 2013)	Cohort	III-2	1,889	**Aim**: To explore the influence of different testing strategies on the prevalence of thyroid dysfunction and the missed diagnosis during pregnancy. **Setting**: International Peace Maternity and Child Health Hospital, Shanghai **Population**: Pregnant women (13-27 weeks) divided into high-risk and low-risk groups according to backgrounds collected by questionnaire. High-risk women accounted for 10.69%. **Observations**: We detected the prevalence of thyroid dysfunction in high-risk groups and low-risk pregnant women by normal reference range during the second trimester.	Using targeted high-risk case testing strategy, misdiagnosis rate of pregnancy with hyperthyroidism, subclinical hyperthyroidism, pregnancy with hypothyroidism, subclinical hypothyroidism, low T4 syndrome and positive TPOAb were 87.5% (14 cases), 87.08% (155 cases), 87.08% (155 cases), 83.93% (47 cases), 89.47% (17 cases) and 88.35% (91 cases), respectively. There was no statistically significant difference between the high-risk and low-risk groups in the prevalence of thyroid dysfunction. Therefore, we believe that universal testing of pregnant women can effectively reduce misdiagnosis rate of thyroid dysfunction. Further, we recommend universal testing for thyroid function in second trimester of pregnancy.	Another study conducted at the same hospital (Fan et al 2016) notes that the study area is an iodine- stable and adequate area

Study ref	Design	LoE	N	Aim, setting, population, measurements/observations	Results	Limitations/ Comments
(Yang et al 2014)	Cohort	III-2	3,882	**Aim**: To evaluate the effectiveness of the targeted high risk case-finding approach for identifying women with thyroid dysfunction during the first and second trimesters of pregnancy. **Setting**: Third Hospital Affiliated to Wenzhou Medical University, China **Population**: Exclusion criteria included women > 28 weeks gestation and women who lived locally <5 years (the local region is iodine-adequate). **Observations**: Levels of thyroid stimulating hormone (TSH), free thyroxine (FT4), and thyroid peroxidase antibodies (TPOAb) were measured during the first and second trimester. All tested women were divided into high risk or non-high risk groups, based on their history, findings from physical examination, or other clinical features suggestive of a thyroid disorder. Diagnosis of thyroid disorders was made according to the standard trimester-specific reference intervals. The prevalence of thyroid disorders in each group was determined, and the feasibility of a testing approach focusing exclusively on high risk women was evaluated to estimate the ability of finding women with thyroid dysfunction.	The prevalence of overt hypothyroidism or hyperthyroidism in the high-risk group was higher than in the non-high risk group during the first trimester (0.8% vs 0, chi2 = 7.10, p = 0.008; 1.6% vs 0.2%, chi2 = 7.02, p=0.008, respectively). The prevalence of hypothyroxinemia or TPOAb positivity was significantly higher in the high-risk group than in the non-high risk group during the second trimester (1.3% vs 0.5%, chi2 = 4.49, p = 0.034; 11.6% vs 8.4%, chi2 = 6.396, p = 0.011, respectively). The total prevalence of hypothyroidism or hyperthyroidism and of subclinical hypothyroidism or hyperthyroidism were not statistically different between the high risk and non-high risk groups, for either the first or second trimester. The high risk testing strategy failed to detect the majority of pregnant women with thyroid disorders. Therefore, we recommend universal testing of sTSH, FT4, and TPOAb during the first and second trimester.	

Maternal and infant outcomes

Study ref	Design	LoE	N	Aim, setting, population, measurements/observations	Results	Limitations/ Comments
(Spencer et al 2015)	SLR	I	26,408 (2 RCTs)	In one trial (universal testing vs case finding) (n=4562), before 11 weeks' gestation, women in the universal testing group, and 'high-risk' women in the case finding group had their sera tested for TSH, fT4 and TPO-Ab; women with hypothyroidism (TSH > 2.5 mIU/litre) received levothyroxine; women with hyperthyroidism (undetectable TSH and elevated fT4) received antithyroid medication.	In the first trial, no clear differences were seen in the risks of pre-eclampsia (RR 0.87, 95% CI 0.64–1.18; n=4516; GRADE: moderate) or preterm birth (RR 0.99, 95% CI 0.80–1.24; n=4516; GRADE: high). More women in the universal group received pharmacological treatment for thyroid dysfunction (RR 3.15, 95% CI 1.91–5.20; n=4562). No clear differences between groups were observed for miscarriage (RR 0.90, 95% CI 0.68–1.19; n=4516; GRADE: moderate), fetal and neonatal death (RR 0.92, 95% CI 0.42–2.02; n=4516; GRADE: moderate) or other secondary outcomes.	Cochrane review

Study ref	Design	LoE	N	Aim, setting, population, measurements/observations	Results	Limitations/ Comments
(Granfors et al 2014)	Cohort	III-2	5,254	**Aim**: To evaluate the efficacy of a targeted thyroid testing approach during pregnancy. **Setting**: Uppsala County, Sweden. **Population**: Data were derived from the population-based Uppsala Biobank of Pregnant Women, in which blood samples are collected in conjunction with the routine ultrasound screening at 17–19 wks. **Observations**: On review of medical records, women who were tested for thyroid dysfunction during pregnancy were identified (n=891). From the remaining untested women, 1,006 women were randomly selected for analyses of thyrotropin (TSH), free thyroxine levels, and thyroid peroxidase antibodies. Thyroid-stimulating hormone levels in both groups were analysed with regard to trimester-specific upper reference levels as recommended by the International Endocrine Society Guidelines.	While the study was underpowered to analyse outcomes, no increase in prevalence of adverse outcomes among untested women who had elevated TSH levels was found: preterm birth (p=0.4); post-term delivery (p=0.4); gestational hypertension or pre-eclampsia (p=0.8); birth weight (p=0.1), length of gestation (p=0.6); pH<7.05 at delivery (p=0.2); Apgar score <7 at 5 min (p=1.0).	

3.2 Universal testing versus no testing

Thyroid dysfunction

Study ref	Design	LoE	N	Aim, setting, population, measurements/observations	Results	Limitations/ Comments
(Spencer et al 2015)	SLR	I	26,408 (2 RCTs)	In the other trial (n=21,846), before 15 + 6 weeks' gestation, women in the universal testing group had their sera tested; women who tested 'positive' (TSH > 97.5th percentile, fT4 < 2.5th percentile, or both) received levothyroxine.	Compared with the no testing group, more women in the universal testing group tested 'positive' for hypothyroidism (RR 998.18, 95% CI 62.36 to 15,978.48; 21,839 women; GRADE: high). More women in the universal testing group received pharmacological treatment for thyroid dysfunction (RR 1102.90, 95% CI 69.07 to 17,610.46; n=1,050); 10% had their dose lowered because of low TSH, high fT4 or minor side effects.	Cochrane review
(Ma et al 2016)	RCT	II	1,671	**Aim**: To evaluate the effect of subclinical hypothyroidism (SCH) testing and intervention on pregnancy outcomes. **Setting**: Peking Union Medical College Hospital (testing group; n=675) and Haidian Maternal & Child Health Hospital (control group whose serum was stored and measured shortly after delivery; n=996). **Population**: Exclusion criteria included age <18 years, ectopic pregnancy, twin pregnancy and known thyroid disease. **Observations**: Thyrotropin levels 2.5–10 mIU/L and free T4 levels in normal range were considered SCH. Some of the testing group were treated with levothyroxine. The others had regular follow-up visits.	2.5<TSH≤10 was found in 252/996 from the control group and 167/675 from the testing group. In the testing group, 105 SCH and 4 hypothyroid women received thyroid hormone replacement therapy.	High risk of bias; see Section 3.3

Maternal and infant outcomes

Study ref	Design	LoE	N	Aim, setting, population, measurements/observations	Results	Limitations/ Comments
(Spencer et al 2015)	SLR	I	26,408 (2 RCTs)	In the other trial (universal testing vs no testing) (n=21,846), before 15 + 6 weeks' gestation, women in the universal testing group had their sera tested; women who tested 'positive' (TSH > 97.5th percentile, fT4 < 2.5th percentile, or both) received levothyroxine.	No data were provided for pre-eclampsia or preterm birth, the trial reported rates of 5.6% and 7.9% for the testing and no testing groups respectively. No clear difference was seen for neurosensory disability of the infant as a child (RR 0.85, 95% CI 0.60–1.22; n=794; GRADE: moderate). No clear differences were observed for other secondary outcomes, including developmental delay/intellectual impairment at 3 years.	Cochrane review
(Ma et al 2016)	RCT	II	1,671	**Aim**: To evaluate the effect of subclinical hypothyroidism (SCH) testing and intervention on pregnancy outcomes. **Setting**: Peking Union Medical College Hospital (testing group; n=675) and Haidian Maternal & Child Health Hospital (control group whose serum was stored and measured shortly after delivery; n=996). **Population**: Exclusion criteria included age <18 years, ectopic pregnancy, twin pregnancy and known thyroid disease. **Observations**: Thyrotropin levels 2.5–10 mIU/L and free T4 levels in normal range were considered SCH. Some of the testing group were treated with levothyroxine. The others had regular follow-up visits.	The risk of miscarriage (OR=0.343, 95% CI 0.21–0.56; p<0.001) and fetal macrosomia (OR=0.46, 95% CI 0.28–0.74; p<0.001) was lower, and of caesarean (OR=1.38, 95% CI 1.13–1.69; p=0.002) was higher in the testing group than the control group. Differences in other pregnancy complications — gestational diabetes (p=0.046; gestational hypertension (p=0.72); premature rupture of the membranes (p=0.433); placental abruption (p=0.631); protracted active phase (p=0.991); postpartum haemorrhage (p=0.507) — or outcomes — preterm birth (p=0.722); fetal growth restriction (p=0.235); breech position (p=0.139) — were not significant.	High risk of bias; see Section 3.3

3.3 Evaluation of limitations of randomised controlled trials for research question 3

Study limitation	Judgement	Support for judgement
(Ma et al 2016)		
Random sequence generation	High risk	'Cluster randomization was utilized. Pregnant women from PUMCH were assigned to a testing group, in which measurements of thyroid function (thyrotropin, FT3, FT4) and thyroid antibody (TPOAb, TGAb) were obtained in early pregnancy. Pregnant women from HMCHH were assigned to a control group, in which serum was stored at –20°C. Levels of free T4, TPOAb and thyrotropin were assessed from the serum shortly after delivery.' Clustering not taken into account in analysis.
Allocation concealment	Low risk	'Participants were blinded as to which group they were a part of, control or testing, while investigators were aware. This method helps physicians to follow-up with patients in both the groups.'
Blinding	Low risk	'A single-blind method of this clinical trial was utilized. Due to the objective nature of the data collected including pregnancy outcomes, single-blinding would not lead to severe bias.'
Incomplete outcome data	Unclear risk	Attrition not discussed.
Selective reporting	Low risk	Pre-specified outcomes reported.
Other limitations	High risk	Baseline characteristics differed significantly between the groups.

3.4 Level IV studies for research question 3

Study	Study design
Jouyandeh, Z., S. Hasani-Ranjbar, et al. (2015). "Universal screening versus selective case-based screening for thyroid disorders in pregnancy." Endocrine 48(1): 116-123.	Review of heterogeneous observational studies
Ohashi, M., S. Furukawa, et al. (2013). "Risk-based screening for thyroid dysfunction during pregnancy." Journal of Pregnancy 2013: 619718.	Case series
Sarapatkova, H., J. Sarapatka, et al. (2013). "What is the benefit of screening for thyroid function in pregnant women in the detection of newly diagnosed thyropathies?" Biomedical Papers of the Medical Faculty of Palacky University in Olomouc, Czech Republic 157(4): 358-362.	Case series

3.5 Excluded studies for research question 3

Study	Reason for exclusion
Khan, I., J. K. Witczak, et al. (2013). "Preconception thyroid-stimulating hormone and pregnancy outcomes in women with hypothyroidism." Endocrine Practice 19(4): 656-662.	Not specific to target population
Pombar-Perez, M., M. Penin-Alvarez, et al. (2013). "[Impact of the application of the American Thyroid Association criteria on the diagnosis of hypothyroidism in pregnant women in Vigo, Spain]." Revista Peruana de Medicina Experimental y Salud Publica 30(3): 428-431.	Abstract only (full article in Spanish)

4 When should pregnant women be tested for thyroid dysfunction?

Evidence summary

Results of the previous review

In the literature review conducted to inform Module II of the Guidelines (Australian Health Ministers' Advisory Council 2014), one high-level evidence paper was found to support testing in early pregnancy and up to the end of the second trimester. No recommendation on the timing of testing was made and the paper informed the narrative.

Results of the current review

Only Level IV evidence was identified and findings were inconsistent.

One study (Ekinci et al 2015) found that gestation-dependent loss of TPOAb/TGAb positivity and reduction in diagnostic accuracy for predicting post-partum thyroid dysfunction limits the value of testing in the second and third trimesters.

Another study (Ong et al 2014) found that adding TSH to first trimester screening tests identifies mainly minor elevations in TSH, which do not predict adverse pregnancy outcomes.

Advice to EWG

The identified evidence does not alter the previous guidance and may inform the narrative.

Study ref	Design	LoE	N	Intervention/observation/comparison population	Results	Limitations
(Ekinci et al 2015)	Diagnostic yield	IV	154	**Aim:** To assess optimal test timing of TPOAb/TGAb for the detection of Hashimoto's thyroiditis and post-partum thyroid dysfunction (PPTD) **Setting:** Mercy Hospital for Women, a tertiary obstetric hospital in Melbourne **Population**: Healthy women <13 weeks' gestation. Exclusion criteria included past history of thyroid disease, thyroid hormone replacement therapy, type 1 diabetes mellitus, past history of intravenous drug abuse or presence of major systemic illness. Women with twin pregnancy and miscarriage prior to 20 weeks in the current pregnancy were also excluded from the analysis. **Measurements**: Serum TPOAb, TGAb, TSH and fT4 were measured at T1, T2, T3 and postpartum. Post-partum thyroid dysfunction (PPTD) was defined if TSH deviated from the assay's nonpregnant reference interval. Longitudinal random-effect logistic regression was used to investigate the association between time and positive/ negative thyroid autoantibody status. Samples from 140 women at T1 (12.0: 10.3-13.0) (median: IQR weeks' gestation); 95 at T2 (24.3: 23.0-25.9), 79 at T3 (35.9: 34.8-36.7) and 83 at PP (12.4: 10.8-14.6 weeks post-partum) were attained.	At T1, 13 (9%) and 15 (11%) women had positive TPOAb and TGAb, respectively. At T2, the odds of having a positive TPOAb were 96% lower [OR = 0.04 (95% CI: 0.02-0.8; P = 0.03)] and 97% lower at T3 [OR = 0.03 (95% CI: 0.001-0.6; P = 0.02)] than at T1. Similarly, at T3 the odds of having a positive TGAb were 99.4% lower [OR = 0.006 (95% CI: 0-0.3; P = 0.01)] at T2, and 99.5% lower [OR = 0.005 (95% CI: 0-0.4; P = 0.02)] than at T1. The ROC analysis diagnostic ORs for a positive TPOAb and/or TGAb to predict PPTD were 7.8 (95% CI: 2.2-27.6) at T1, 1.2 (95% CI: 0-8.9) at T2, 2.0 (95% CI: 0-16.8) at T3, and 12.2 (95% CI: 3.3-44.9) post-partum. A significant proportion of pregnant women lose their thyroid autoantibody positivity after T1. The gestation-dependent loss of TPOAb/TGAb positivity and reduction in diagnostic accuracy for predicting PPTD limits the value of testing at T2 and T3.	

Study ref	Design	LoE	N	Intervention/observation/comparison population	Results	Limitations
(Ong et al 2014)	Diagnostic yield	IV	2,411	**Aim**: To determine if thyroid function tests performed with first trimester screening predicts adverse pregnancy outcomes. **Setting**: Western Australia **Population**: Women in with singleton pregnancies attending first trimester screening between 9 and 14 weeks gestation. **Measurements**: We evaluated the association between TSH, free T4, free T3, thyroid antibodies, free beta human chorionic gonadotrophin (beta-hCG) and pregnancy associated plasma protein A (PAPP-A) with a composite of adverse pregnancy events as the primary outcome. Secondary outcomes included placenta praevia, placental abruption, pre-eclampsia, pregnancy loss after 20 weeks gestation, threatened preterm labour, preterm birth, small size for gestational age, neonatal death, and birth defects.	TSH exceeded the 97.5th percentile for the first trimester (2.15 mU/L) in 133 (5.5%) women, including 22 (1%) with TSH above the nonpregnant reference range (4 mU/L) and 5 (0.2%) above 10 mU/L. Adverse outcomes occurred in 327 women (15%). TSH and free T4 did not differ significantly between women with or without adverse pregnancy events. On the multivariate analysis, neither maternal TSH >2.15 mU/L nor TSH as a continuous variable predicted primary or secondary outcomes. Adding TSH to first trimester screening tests identifies mainly minor elevations in TSH, which do not predict adverse pregnancy outcomes. A small number of women with more significant hypothyroidism are also detected who should be routinely treated with thyroid replacement therapy. However, it remains uncertain whether testing is justified in order to detect these cases.	

5 What interventions or treatments for thyroid dysfunction are effective and safe in pregnancy, and what advice should women receive?

Evidence summary

Results of the previous review

Some evidence on treatments for overt hypothyroidism, subclinical hypothyroidism and overt hyperthyroidism was identified in the literature review conducted to inform Module II of the Guidelines (Australian Health Ministers' Advisory Council 2014). However, it was agreed by the expert working group that treatment was beyond the scope of the guidelines. A statement on current practice in Australia was included.

Results of the current review

No new evidence identified.

6 What is the cost effectiveness of universal testing in pregnancy for hypothyroidism?

Evidence summary

Results of the previous review

One study conducted in North America was identified in the literature review conducted to inform Module II of the Guidelines (Australian Health Ministers' Advisory Council 2014). A separate economic analysis conducted to inform the development of the guidelines found insufficient evidence that treatment of thyroid dysfunction reduces adverse maternal or fetal outcomes and no economic evaluations relevant to Australia.

Results of the current review

No new evidence identified.

7 What are the additional considerations for Aboriginal and Torres Strait Islander women?

Evidence summary

Results of the previous review

This question was not asked in the literature review conducted to inform Module II of the Guidelines (Australian Health Ministers' Advisory Council 2014).

Results of the current review

No evidence identified.

References

ABS (2014) 4364.0.55.006 - Australian Health Survey: Biomedical Results for Nutrients, 2011-12. Canberra: Australian Bureau of Statistics. Available at: www.abs.gov.au

Ahmed IZ, Eid YM, El Orabi H et al (2014) Comparison of universal and targeted screening for thyroid dysfunction in pregnant Egyptian women. *Eur J Endocrinol* 171(2): 285-91.

Australian Health Ministers' Advisory Council (2014) Clinical Practice Guidelines: Antenatal care — Module II. Canberra: Australian Government Department of Health. Available at: http://www.health.gov.au/antenatal

Bautista AA, Antonio MQ, Jimeno C et al (2014) Reference Intervals in Thyroid Function Tests in the Third Trimester in Pregnant Filipino Women. *Phillipino J Int Med* 52(3): 1–5.

Bliddal S, Feldt-Rasmussen U, Boas M et al (2014) Gestational age-specific reference ranges from different laboratories misclassify pregnant women's thyroid status: comparison of two longitudinal prospective cohort studies. *Eur J Endocrinol* 170(2): 329-39.

Bryant SN, Nelson DB, McIntire DD et al (2015) An analysis of population-based prenatal screening for overt hypothyroidism. *Am J Obstet Gynecol* 213(4): 565 e1-6.

Ekinci EI, Lu ZX, Sikaris K et al (2013) Longitudinal assessment of thyroid function in pregnancy. *Ann Clin Biochem* 50(Pt 6): 595-602.

Ekinci EI, Chiu WL, Lu ZX et al (2015) A longitudinal study of thyroid autoantibodies in pregnancy: the importance of test timing. *Clin Endocrinol (Oxf)* 82(4): 604-10.

Fan J-X, Yang S, Qian W et al (2016) Comparison of the Reference Intervals Used for the Evaluation of Maternal Thyroid Function During Pregnancy Using Sequential and Nonsequential Methods. *Chinese Med J* 129(7): 785–81.

Foley TP, Jr., Henry JJ, Hofman LF et al (2013) Maternal screening for hypothyroidism and thyroiditis using filter paper specimens. *J Womens Health (Larchmt)* 22(11): 991-6.

Granfors M, Akerud H, Skogo J et al (2014) Targeted thyroid testing during pregnancy in clinical practice. *Obstet Gynecol* 124(1): 10-5.

Khalid AS, Marchocki Z, Hayes K et al (2014) Establishing trimester-specific maternal thyroid function reference intervals. *Ann Clin Biochem* 51(Pt 2): 277-83.

Ma L, Qi H, Chai X et al (2016) The effects of screening and intervention of subclinical hypothyroidism on pregnancy outcomes: a prospective multicenter single-blind, randomized, controlled study of thyroid function screening test during pregnancy. *J Matern Fetal Neonatal Med* 29(9): 1391-4.

Moradi S, Gohari MR, Aghili R et al (2013) Thyroid function in pregnant women: iodine deficiency after iodine enrichment program. *Gynecol Endocrinol* 29(6): 596-9.

Nazarpour S, Tehrani FR, Simbar M et al (2016) Comparison of universal screening with targeted high-risk case finding for diagnosis of thyroid disorders. *Eur J Endocrinol* 174(1): 77-83.

Ong GS, Hadlow NC, Brown SJ et al (2014) Does the thyroid-stimulating hormone measured concurrently with first trimester biochemical screening tests predict adverse pregnancy outcomes occurring after 20 weeks gestation? *J Clin Endocrinol Metab* 99(12): E2668-72.

Qian W, Zhang L, Han M et al (2013) Screening for thyroid dysfunction during the second trimester of pregnancy. *Gynecol Endocrinol* 29(12): 1059-62.

Spencer L, Bubner T, Bain E et al (2015) Screening and subsequent management for thyroid dysfunction pre-pregnancy and during pregnancy for improving maternal and infant health. *Cochrane Database Syst Rev*(9): CD011263.

Wilson KL, Casey BM, McIntire DD et al (2014) Is total thyroxine better than free thyroxine during pregnancy? *Am J Obstet Gynecol* 211(2): 132 e1-6.

Yang H, Shao M, Chen L et al (2014) Screening strategies for thyroid disorders in the first and second trimester of pregnancy in China. *PLoS One* 9(6): e99611.