The Journal of Clinical Endocrinology & Metabolism, 2024, **109**, e2151–e2158
https://doi.org/10.1210/clinem/dgae528
Advance access publication 31 July 2024
Meta-Analysis

Defining Gestational Thyroid Dysfunction Through Modified Nonpregnancy Reference Intervals: An Individual Participant Meta-analysis

Joris A. J. Osinga,[1,2] Scott M. Nelson,[3] John P. Walsh,[4,5] Ghalia Ashoor,[6] Glenn E. Palomaki,[7] Abel López-Bermejo,[8,9] Judit Bassols,[10] Ashraf Aminorroaya,[11] Maarten A. C. Broeren,[12] Liangmiao Chen,[13] Xuemian Lu,[13] Suzanne J. Brown,[4] Flora Veltri,[14] Kun Huang,[15] Tuija Männistö,[16] Marina Vafeiadi,[17] Peter N. Taylor,[18] Fang-Biao Tao,[15] Lida Chatzi,[19] Maryam Kianpour,[11] Eila Suvanto,[20] Elena N. Grineva,[21] Kypros H. Nicolaides,[22] Mary E. D'Alton,[23] Kris G. Poppe,[14] Erik Alexander,[24] Ulla Feldt-Rasmussen,[25,26] Sofie Bliddal,[25,26] Polina V. Popova,[21] Layal Chaker,[1,2,27] W. Edward Visser,[1,2] Robin P. Peeters,[1,2] Arash Derakhshan,[1,2] Tanja G. M. Vrijkotte,[28] Victor J. M. Pop,[29] and Tim I. M. Korevaar[1,2]

[1]Department of Internal Medicine, Erasmus University Medical Center, 3000 CA Rotterdam, the Netherlands
[2]Academic Center for Thyroid Diseases, Erasmus University Medical Center, 3000 CA Rotterdam, the Netherlands
[3]School of Medicine, Dentistry and Nursing, University of Glasgow, G12 8QQ Glasgow, UK
[4]Department of Endocrinology and Diabetes, Sir Charles Gairdner Hospital, Nedlands, WA 6009, Australia
[5]Medical School, University of Western Australia, Crawley, WA 6009, Australia
[6]Harris Birthright Research Center for Fetal Medicine, King's College Hospital, SE5 9RS London, UK
[7]Department of Pathology and Laboratory Medicine, Women & Infants Hospital and Alpert Medical School at Brown University, RI 02903 Providence, USA
[8]Pediatric Endocrinology Research Group, Girona Biomedical Research Institute (IDIBGI), Dr. Josep Trueta Hospital, 17007 Girona, Spain
[9]Departament de Ciències Mèdiques, Universitat de Girona, 17003 Girona, Spain
[10]Maternal-Fetal Metabolic Research Group, Girona Biomedical Research Institute (IDIBGI), Dr. Josep Trueta Hospital, 17007 Girona, Spain
[11]Isfahan Endocrine and Metabolism Research Center, Isfahan University of Medical Sciences, 81745-33871 Isfahan, Iran
[12]Laboratory of Clinical Chemistry and Haematology, Máxima Medical Centre, 5504 DB Veldhoven, Netherlands
[13]Department of Endocrinology and Rui'an Center of the Chinese-American Research Institute for Diabetic Complications, Third Affiliated Hospital of Wenzhou Medical University, 325035 Wenzhou, China
[14]Endocrine Unit, Centre Hospitalier Universitaire Saint-Pierre, Université Libre de Bruxelles (ULB), 1000 Brussels, Belgium
[15]Department of Maternal, Child and Adolescent Health, Scientific Research Center in Preventive Medicine, School of Public Health, Anhui Medical University, 230032 Anhui, China
[16]NordLab, Oulu and Translational Medicine Research Unit, University of Oulu, 90570 Oulu, Finland
[17]Department of Social Medicine, School of Medicine, University of Crete, 710 03 Heraklion, Crete, Greece
[18]Thyroid Research Group, Systems Immunity Research Institute, Cardiff University School of Medicine, CF10 3EU Cardiff, UK
[19]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90089, USA
[20]Department of Obstetrics and Gynecology and Medical Research Center Oulu, University of Oulu, 90570 Oulu, Finland
[21]Institute of Endocrinology, Almazov National Medical Research Centre, 197341 Saint Petersburg, Russia
[22]Department of Women and Children's Health, Faculty of Life Sciences and Medicine King's College London, SE5 9RS London, UK
[23]Department of Obstetrics and Gynecology, Columbia University Irving Medical Center, New York, NY 10032, USA
[24]Division of Endocrinology, Hypertension and Diabetes, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA
[25]Department of Medical Endocrinology and Metabolism, Copenhagen University Hospital, Rigshospitalet, 2100 Copenhagen, Denmark
[26]Department of Clinical Medicine, Faculty of Health and clinical Sciences, Copenhagen University, 2100 Copenhagen, Denmark
[27]Department of Epidemiology, Erasmus University Medical Center, 3000 CA Rotterdam, the Netherlands
[28]Department of Public and Occupational Health, Amsterdam UMC, University of Amsterdam, Amsterdam Public Health Research Institute, 1081 HV Amsterdam, the Netherlands
[29]Department of Medical and Clinical Psychology, Tilburg University, 5000 LE Tilburg, the Netherlands

Correspondence: Joris Osinga, MD, Generation R, Wytemaweg 80, 3000 CA Rotterdam, the Netherlands. Email: j.osinga@erasmusmc.nl.

Received: 19 April 2024. **Editorial Decision:** 28 July 2024. **Corrected and Typeset:** 26 August 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of the Endocrine Society.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (https://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com. See the journal About page for additional terms.

Downloaded from https://academic.oup.com/jcem/article/109/11/e2151/7724966 by guest on 18 June 2025

Abstract

Background: Establishing local trimester-specific reference intervals for gestational TSH and free T4 (FT4) is often not feasible, necessitating alternative strategies. We aimed to systematically quantify the diagnostic performance of standardized modifications of center-specific nonpregnancy reference intervals as compared to trimester-specific reference intervals.

Methods: We included prospective cohorts participating in the Consortium on Thyroid and Pregnancy. After relevant exclusions, reference intervals were calculated per cohort in thyroperoxidase antibody-negative women. Modifications to the nonpregnancy reference intervals included an absolute modification (per .1 mU/L TSH or 1 pmol/L free T4), relative modification (in steps of 5%) and fixed limits (upper TSH limit between 3.0 and 4.5 mU/L and lower FT4 limit 5-15 pmol/L). We compared (sub)clinical hypothyroidism prevalence, sensitivity, and positive predictive value (PPV) of these methodologies with population-based trimester-specific reference intervals.

Results: The final study population comprised 52 496 participants in 18 cohorts. Optimal modifications of standard reference intervals to diagnose gestational overt hypothyroidism were −5% for the upper limit of TSH and +5% for the lower limit of FT4 (sensitivity, .70, CI, 0.47-0.86; PPV, 0.64, CI, 0.54-0.74). For subclinical hypothyroidism, these were −20% for the upper limit of TSH and −15% for the lower limit of FT4 (sensitivity, 0.91; CI, 0.67-0.98; PPV, 0.71, CI, 0.58-0.80). Absolute and fixed modifications yielded similar results. CIs were wide, limiting generalizability.

Conclusion: We could not identify modifications of nonpregnancy TSH and FT4 reference intervals that would enable centers to adequately approximate trimester-specific reference intervals. Future efforts should be turned toward studying the meaningfulness of trimester-specific reference intervals and risk-based decision limits.

Key Words: thyroid gland, thyroid function tests, reference values, pregnancy, thyrotropin, thyroxine

Abbreviations: FT4, free T4; PI, prediction interval; PPV, positive predictive value.

Downloaded from https://academic.oup.com/jcem/article/109/11/e2151/7724966 by guest on 18 June 2025

Thyroid dysfunction during pregnancy is associated with a higher risk of miscarriage, preeclampsia, preterm birth, aberrant birthweight, and lower offspring IQ (1-6). Current international guidelines recommend defining gestational thyroid dysfunction according to population and pregnancy-specific TSH and free T4 (FT4) reference intervals, to take into account thyroid physiology during pregnancy, as well as differences in TSH and FT4 determinants between populations and the use of different laboratory assays (7-9). However, calculating such local reference intervals is generally not feasible for most centers (10, 11). In addition to the practical hurdles, most of the published reference intervals for TSH and FT4 are not in accordance with the current American Thyroid Association guidelines, as we recently exhibited by providing an overview of published TSH and FT4 reference intervals and methodologies, showing that most studies included used additional exclusion criteria based on health status, did not exclude TPOAb positive participants or used different percentile cutoffs (8). This is in part because of changing guidelines and in part because many centers use additional exclusion criteria or apply different reference limit cutoffs (8). These varying methodologies hamper the adoption of reference intervals from other centers, and as such, the vast majority of centers rely on nonpregnancy reference intervals for TSH with either a fixed limit approach (upper limit of 4.0 mU/L for TSH) or a subtraction approach (subtraction of 0.5 mU/L of the upper limit of TSH), whereas for FT4, varying local approaches are used including nonpregnancy reference intervals (12-14). These second-tier strategies are considered inferior compared to locally defined reference intervals (15-17). In a follow-up study, we showed that the use of a fixed upper TSH limit or the subtraction approach results in poor detection rates and high false-positive rates for (subclinical) hypothyroidism in early pregnancy with highly variable diagnostic performance between populations (sensitivity, 0.63-0.82; false discovery rate, 0.11-0.35) (18).

In search of a method that is both easy to implement in clinical practice and would better identify women with an abnormal thyroid function during pregnancy, we set out to investigate if it is possible to modify the center-specific nonpregnancy TSH and FT4 reference intervals so that these are useful in pregnancy. Such an approach could make the establishment of local pregnancy-specific reference intervals obsolete while it takes account of the local assay and preexisting laboratory harmonization efforts (19, 20). A useful diagnostic approach would need to fulfill certain conditions: (1) the diagnostic performance should at least perform better than currently recommended alternative methods (TSH upper limit of 4.0 mU/L or subtraction of 0.5 mU/L) (12, 13) and (2) the diagnostic performance should be reasonably consistent between populations.

In this individual participant meta-analysis, we aimed to modify the center-specific nonpregnancy reference intervals of TSH and FT4 in a standardized manner and study the sensitivity and the positive predictive value (PPV) compared to center-specific gestational reference intervals as calculated in accordance with the current international guidelines.

Methods

The study inclusion and eligibility procedures are described in detail previously (18). In short, eligible studies were those participating in the Consortium on Thyroid and Pregnancy (https://www.consortiumthyroidpregnancy.org). Exclusion criteria for participants were prepregnancy thyroid disease, pregnancy through in vitro fertilization/intracytoplasmic sperm injection, use of thyroid (interfering) medication, and multiple gestation. For this study, we followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines for Individual Patient Data and preregistered the study protocol (CRD42021270078), which can be found in the supplemental materials along with an outline of protocol deviations (21). Study quality and risk of bias were assessed using the Newcastle-Ottawa scale (Supplementary materials (21)). All cohorts were approved by a local review board and acquired participant informed consent or had been granted exemption from it by the local ethics committee.

Defining Gestational Thyroid Dysfunction

Nonpregnancy reference intervals were either published and/or provided by the principal investigator of the included cohorts and are assay-specific. We defined the trimesters as 0 to 13 weeks, > 13 to 27 weeks, and >27 weeks of gestation. For cohorts containing participants with repeated measurements, we used the first available sample for each trimester.

Reference intervals, thyroid dysfunction (overt and subclinical hypothyroidism), and diagnostic test properties were calculated separately for each cohort to account for inter-population differences. All reference intervals were calculated as the 2.5th to 97.5th percentiles in TPOAb-negative participants. Our primary aim was to optimize the diagnosis of thyroid dysfunction states for which treatment is indicated or should be considered based on current guidelines, and thus we limited analyses to overt and subclinical hypothyroidism (13). A treatment indication was defined as either (1) overt hypothyroidism, (2) subclinical hypothyroidism with TSH > 10 mU/L, or (3) subclinical hypothyroidism with TPOAb positivity. A treatment consideration was defined as (1) TSH between 2.5 mU/L and the upper reference limit with concomitant TPOAb positivity or (2) subclinical hypothyroidism without TPOAb positivity (13). Treatment of hyperthyroidism was outside the scope of this study because gestational hyperthyroidism is often considered physiological and we do not have data available to differentiate between gestational transient thyrotoxicosis and Graves' hyperthyroidism (13). The prevalence of thyroid dysfunction and diagnostic performance measures were calculated according to several methods: (1) a relative modification of the nonpregnancy upper limit of TSH varying from −5% to −40% in steps of 5%, with modifications to the lower limit of FT4 varying from −20% to +20% in steps of 5% (relative modification approach); (2) a subtraction from the nonpregnancy upper limit of TSH varying from −0.1 to −1.0 mU/L, with modification of the nonpregnancy lower limit of FT4 varying from −5 to +5 pmol/L (−0.39 to +0.39 ng/dL; absolute modification approach); and (3) using fixed upper limits for TSH, varying from 3.0 to 4.5 mU/L, and fixed lower limits for FT4, varying from 5 to 15 pmol/L (0.39-1.17 ng/dL; fixed limit approach). The choice for the range of modifications was based on previous recommendations (eg, the fixed upper limit of 4.0 mU/L for TSH and 0.5 subtraction from this limit) and the optimal diagnostic performance in this study to keep the results organized. The results for each method were compared to the reference standard (trimester-specific reference intervals), as is currently advised in international guidelines (12, 13).

Diagnostic Performance Measures

The diagnostic performance of each assessed combination is described using the sensitivity (equivalent to true-positive rate, true-positive rate among all with the disease according to the trimester-specific method) and the PPV (equivalent to 1-false discovery rate, true positives among all with a positive test result). Presenting the PPV, rather than the specificity, was preferred because the PPV is more informative with regard to false positives for outcomes with a low prevalence (22). The aim was to maximize both diagnostic performance markers, which poses a challenge because maximizing sensitivity and the PPV is often a tradeoff.

The primary outcome was a single diagnostic performance measure, the F-score (also referred to as F1-score), which is a combined measure of PPV (also referred to as "precision") and sensitivity (also referred to as "recall") (23). A higher F-score denotes a better overall diagnostic performance.

Prediction intervals and the I^2 statistic are presented to illustrate the expected inter-population variation in diagnostic performance and between-study heterogeneity (21, 24). Prediction intervals are an attempt to predict future individual values whereas CIs give an indication of where the mean value lies. To facilitate comparison of diagnostic performance markers between methods, interactive heatmaps were constructed and can be found online (25).

Statistical Analyses

Diagnostic performance measures were calculated using 2 × 2 contingency tables (confusion matrices) per cohort and pooled using random intercept logistic regression models using maximum likelihood for modeling between-study heterogeneity. This approach was chosen because it outperforms conventional 2-step inverse-variance approaches for sparse event datasets (26, 27). For each alternative approach, the sensitivity, PPV, and F-scores were calculated and compared with the trimester-specific approach. All analyses were performed using R statistical software version 4.2.2 (28), specifically using the packages "meta" (29), "ggplot2," (30) and "heatmaply" (31).

Results

After exclusions, the final study population comprised 52 496 participants included in 18 cohorts (Fig. 1), of whom 8.6% were TPOAb positive (range across cohorts 5.7-17.1%; Supplementary Table 1 (21)). The prevalence of thyroid function test abnormalities (in the first and second trimester, respectively) according to the trimester-specific approach was 0.5% and 0.3% for overt hypothyroidism and 3.4% and 3.2% for subclinical hypothyroidism. The inclusion process and maternal demographics are described in detail previously (18). Cohort-specific prevalence of thyroid disease, reference limits, iodine status, and assay information can be found in Supplementary Tables 2-6 (21). All figures are accompanied

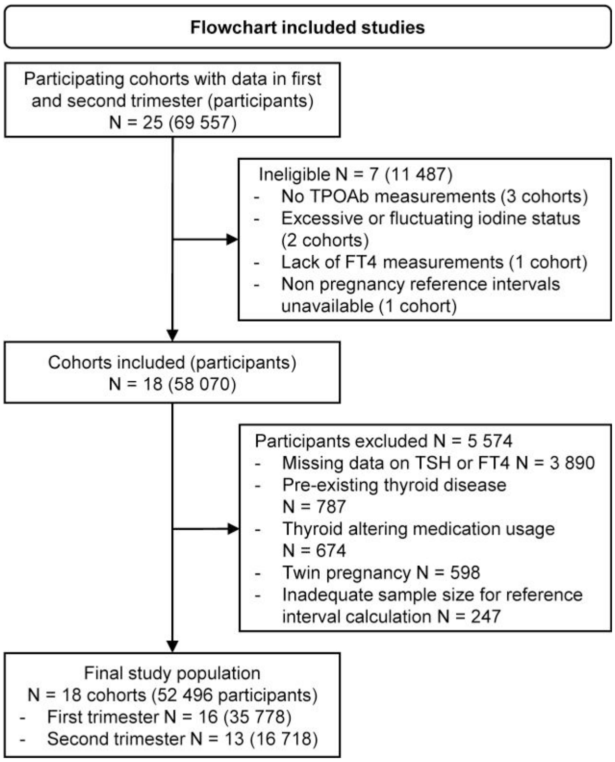
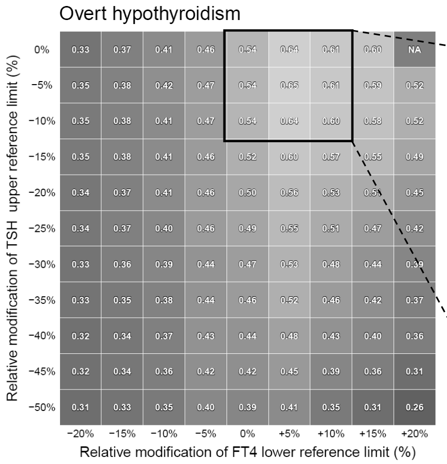

Flowchart included studies
Participating cohorts with data in first and second trimester (participants)
N = 25 (69 557)
Ineligible N = 7 (11 487)
- No TPOAb measurements (3 cohorts)
- Excessive or fluctuating iodine status (2 cohorts)
- Lack of FT4 measurements (1 cohort)
- Non pregnancy reference intervals unavailable (1 cohort)
Cohorts included (participants)
N = 18 (58 070)
Participants excluded N = 5 574
- Missing data on TSH or FT4 N = 3 890
- Pre-existing thyroid disease N = 787
- Thyroid altering medication usage N = 674
- Twin pregnancy N = 598
- Inadequate sample size for reference interval calculation N = 247
Final study population
N = 18 cohorts (52 496 participants)
- First trimester N = 16 (35 778)
- Second trimester N = 13 (16 718)

Figure 1. Flowchart of included cohorts and participants. Reprinted by Osinga et al (18).

Downloaded from https://academic.oup.com/jcem/article/109/11/e2151/7724966 by guest on 18 June 2025



	−20%	−15%	−10%	−5%	0%	+5%	+10%	+15%	+20%
0%	0.33	0.37	0.41	0.46	0.54	0.64	0.61	0.60	NA
−5%	0.35	0.38	0.42	0.47	0.54	0.65	0.61	0.59	0.52
−10%	0.35	0.38	0.41	0.47	0.54	0.64	0.60	0.58	0.52
−15%	0.35	0.38	0.41	0.46	0.52	0.60	0.57	0.55	0.49
−20%	0.34	0.37	0.41	0.46	0.50	0.56	0.53	0.51	0.45
−25%	0.34	0.37	0.40	0.46	0.49	0.55	0.51	0.47	0.42
−30%	0.33	0.36	0.39	0.44	0.47	0.53	0.48	0.44	0.39
−35%	0.33	0.35	0.38	0.44	0.46	0.52	0.46	0.42	0.37
−40%	0.32	0.34	0.37	0.43	0.44	0.48	0.43	0.40	0.36
−45%	0.32	0.34	0.36	0.42	0.42	0.45	0.39	0.36	0.31
−50%	0.31	0.33	0.35	0.40	0.39	0.41	0.35	0.31	0.26

F-Score	Modification		Sensitivity				Positive Predictive Value			
	ULTSH	*LLFT4*		*Confidence interval*	*Prediction interval*	*I^2 Statistic*		*Confidence interval*	*Prediction interval*	*I^2 Statistic*
0.54	None	None	0.49	0.33-0.65	0.08-0.91	59%	0.65	0.45-0.81	0.13-0.96	26%
0.54	-5%	None	0.53	0.32-0.73	0.05-0.96	48%	0.65	0.46-0.81	0.15-0.95	17%
0.54	-10%	None	0.54	0.34-0.73	0.05-0.96	44%	0.63	0.44-0.79	0.14-0.95	30%
0.64	None	+5%	0.65	0.46-0.80	0.10-0.97	67%	0.65	0.54-0.74	0.17-0.94	48%
0.65	-5%	+5%	0.70	0.47-0.86	0.06-0.99	64%	0.64	0.54-0.74	0.18-0.94	45%
0.64	-10%	+5%	0.71	0.49-0.86	0.07-0.99	64%	0.62	0.46-0.76	0.17-0.93	55%
0.61	None	+10%	0.75	0.50-0.91	0.05-0.99	51%	0.56	0.45-0.67	0.25-0.83	35%
0.61	-5%	+10%	0.83	0.53-0.95	0.03-1.00	44%	0.55	0.44-0.66	0.25-0.82	33%
0.60	-10%	+10%	0.84	0.55-0.95	0.03-1.00	46%	0.53	0.42-0.65	0.22-0.82	39%

Figure 2. Diagnostic performance of modified nonpregnancy reference intervals for overt hypothyroidism using relative modification. Diagnostic performance for relative modifications of nonpregnancy reference intervals for the diagnosis of overt hypothyroidism, presented as F-scores. The zoomed-in section presents additional diagnostic performance markers for selected modifications, of which an interactive version can be found online (https://www.consortiumthyroidpregnancy.org/heatmaps).

by supplemental tables (21) containing the diagnostic performance markers for each specific combination (Fig. 2 is an explanatory example of the diagnostic markers presented). To facilitate comparison of diagnostic performance measures, an interactive version of the heatmaps including other diagnostic performance measures can be found online and is also referred to throughout, as an alternative to the supplemental tables (21) (https://www.consortiumthyroidpregnancy.org/heatmaps (25)).

Diagnostic Performance of Alternative Approaches

Using the relative modification approach in the first trimester, the highest F-scores for overt hypothyroidism were achieved with a relative subtraction of 5% for the upper reference limit of TSH and a relative addition of 5% for the lower reference limit of FT4 (F-score 0.65; Fig. 3A). The associated sensitivity was 0.70 (95% CI, 0.47-0.86; 95% prediction interval [PI], 0.06-0.99; I^2 64%), and the PPV was 0.64 (CI, 0.54-0.74; PI, 0.18-0.94; I^2 45%; Fig. 3A, Supplementary Table 7 (21), Interactive figures (25)). For subclinical hypothyroidism, the highest F-scores were achieved with a relative subtraction of 20% for the upper reference limit of TSH and a relative subtraction of 15% for the lower reference limit of FT4 (F-score, 0.69; Fig. 3B). Associated sensitivity was 0.91 (CI, 0.67-0.98; PI, 0.02-1.00; I^2 95%) and PPV was 0.71 (CI, 0.58-0.80; PI, 0.20-0.96; I^2 95%; Supplementary Table 8 (21), Interactive figures (25)).

Using the absolute modification approach in the first trimester, the highest F-scores for overt hypothyroidism were achieved with a subtraction of either −0.1, −0.2, or −0.3 mU/L for the upper limit of TSH and an addition of +1 pmol/L to the lower limit of FT4 and (F-score, 0.62; Fig. 3C). Associated sensitivity (for upper limit TSH, −0.2 mU/L) was 0.74 (CI, 0.52-0.89; PI, 0.08-0.99; I^2 66%) and PPV was 0.57 (CI, 0.45-0.68; PI, 0.24-0.84; I^2 39%; Supplementary table 9 (21), Interactive figures (25)). For subclinical hypothyroidism, the highest F-scores were achieved with a subtraction of −0.8 mU/L from the upper limit of TSH and a subtraction of either −1, −2, −3, −4, or −5 pmol/L from the lower limit of FT4 (F-score, 0.64; Fig. 3D). Associated sensitivity (for lower limit FT4, −4 pmol/L) was 0.91 (CI, 0.61-0.98; PI, 0.01-1.00; I^2 95%) and PPV was 0.68 (CI, 0.55-0.78; PI, 0.20-0.95; I^2 95%; Supplementary table 10 (21), Interactive figures (25)).

Using the fixed-limit approach in the first trimester, the highest F-scores for overt hypothyroidism were achieved with an upper limit of TSH of either 3.8, 3.9, 4.0, 4.1, and 4.4 mU/L and a lower limit of FT4 of 12 pmol/L (F-score, 0.65; Fig. 3E). Associated sensitivity (for upper limit TSH, 4.0 mU/L) was 0.83 (CI, 0.70-0.91; PI, 0.41-0.97; I^2 0%) and PPV was 0.50 (CI, 0.32-0.68; PI, 0.05-0.95; I^2 70%; Supplementary Table 11 (21), Interactive figures (25)). For subclinical hypothyroidism the highest F-scores were achieved with an upper limit of TSH of 3.2 mU/L and a lower limit of FT4 of either 5, 6, 7, or 8 pmol/L (F-score, 0.70; Fig. 3F). Associated sensitivity (for lower limit FT4, 8 pmol/L) was 0.99 (CI, 0.88-1.00; PI, 0.03-1.00; I^2 91%) and PPV was 0.66 (CI, 0.51-0.79; PI, 0.11-0.97; I^2 96%; Supplementary Table 12 (21), Interactive figures (25)).

Additional Analyses

In the second trimester, maximum F-scores were similar for the relative modification method, the absolute modification approach and the fixed-limit approach (Supplementary Fig. S1A-F (21)). However, comparing the diagnostic performance measures of individual studies, the variability between studies was very high, as reflected by overlapping CIs for all methods, based on the wide prediction intervals and based on high I^2 statistics for higher F-scores (Supplementary Tables 13-18 (21)). The diagnostic performance of alternative methods to detect women for whom levothyroxine treatment is indicated and those for whom treatment should be considered, according to American Thyroid Association guidelines, in the first trimester and second trimester were similar based on overlapping CIs (Supplementary Figs. S2 and 3; Supplementary Tables 19-30 (21)).

Discussion

In this study, we systematically evaluated multiple standardized procedures to modify nonpregnancy TSH and FT4

Downloaded from https://academic.oup.com/jcem/article/109/11/e2151/7724966 by guest on 18 June 2025

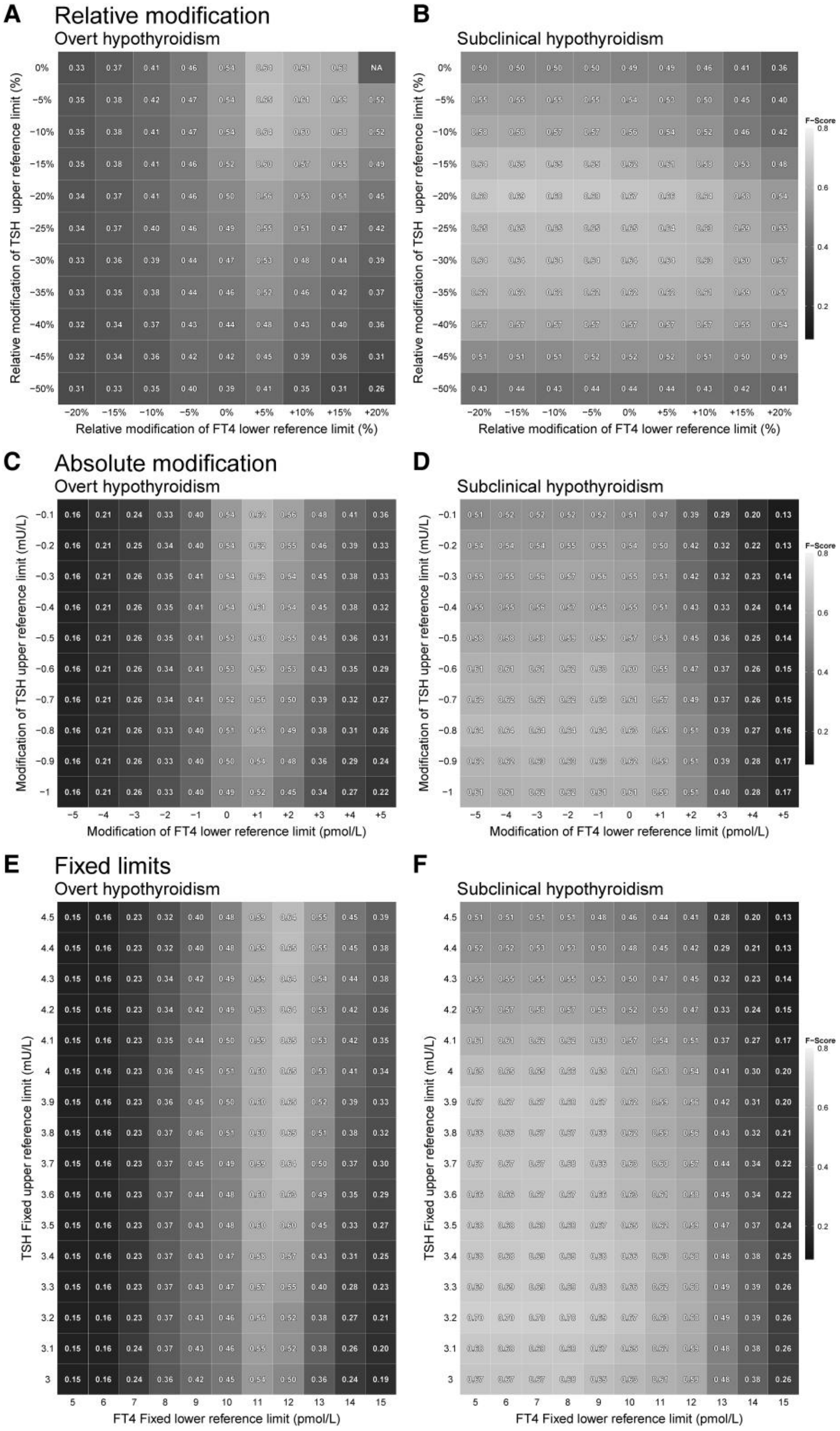


Figure 3. Diagnostic performance of modified nonpregnancy reference intervals for overt and subclinical hypothyroidism. Diagnostic performance of modified nonpregnancy reference intervals are presented using a relative modification (A, B), absolute modifications (C, D), and fixed limits (E, F) for overt and subclinical hypothyroidism, respectively, of which an interactive version can be found online (https://www.consortiumthyroidpregnancy.org/heatmaps).

reference intervals with the aim of diagnosing the same individuals as having an abnormal gestational thyroid function in line with the "gold-standard" approach of center-specific and trimester-specific reference intervals. Despite our efforts, we were unable to identify a standardized procedure that achieved a satisfactory balance between sensitivity and PPV for gestational thyroid dysfunction without considerable variability across different populations. These results underscore the inherent challenge in balancing precise identification of gestational thyroid dysfunction with the practical limitations

Downloaded from https://academic.oup.com/jcem/article/109/11/e2151/7724966 by guest on 18 June 2025

of applying these diagnostic strategies universally in clinical settings and indicate that calculating local center and pregnancy-specific reference intervals for TSH and FT4 should still be considered as current best practice.

Current recommendations on gestational reference interval definitions for TSH and FT4 are time and resource consuming and are not feasible for most centers worldwide. The modification of nonpregnancy reference intervals for the use in pregnancy could overcome feasibility problems. However, in the current study, we show that the variability in TSH and FT4 distributions leads to unacceptable variation in diagnostic performance between cohorts. A possible explanation for this variation is that even the nonpregnancy TSH and FT4 reference intervals are not an adequate reflection of the distribution of thyroid function tests for a population if they are based on the manufacturer's recommendation rather than local laboratory-specific establishment of the intervals. Methods for determining reference intervals in pregnancy and outside pregnancy often differ because current recommendations on establishing reference limits in pregnancy include the local population and are by definition a reflection of local TSH and FT4 distributions (12-14), whereas reference limits outside pregnancy are often supplied by the assay manufacturer, who mostly established reference intervals in selected, nonpregnant populations (32, 33). Global harmonization efforts for TSH and FT4 assays by the International Federation of Clinical Chemistry and Laboratory Medicine Committee for Standardization of Thyroid Function Tests are ongoing to address this issue outside of pregnancy, which could lead to an attenuation of this mismatch (19, 20).

We also show that for overt hypothyroidism and subclinical hypothyroidism, different and sometimes opposing modifications of the reference limits of TSH and FT4 were needed to achieve maximum diagnostic performance. For instance, when reviewing the relative modifications needed to achieve the best diagnostic performance for overt hypothyroidism in the first trimester, we find that the best F-score of 0.65 is achieved with the upper limit of TSH −5% and the lower limit of FT4 + 5% (Fig. 2A), whereas the best F-score for subclinical hypothyroidism of 0.69 is achieved with the upper limit of TSH −20% and the lower limit of FT4 −15% (Fig. 2B). We previously showed that the use of trimester-specific reference intervals for FT4 are most important for the correct diagnosis of overt hypothyroidism, whereas for subclinical hypothyroidism, the use of trimester-specific reference intervals for TSH are more important (18), which could explain the current results. This finding suggests that a uniform rule established to diagnose both overt and subclinical disease would be good at diagnosing one at the cost of incorrectly diagnosing the other. We also observe that the trends in diagnostic performance for a treatment indication (Supplementary Fig. S2A and C, 2E (21)) mostly overlap with the trend in diagnostic performance for subclinical hypothyroidism (Fig. 2B and D, 2F). This is because most women with a treatment indication present with subclinical hypothyroidism with TPOAb positivity (73.6%) rather than overt hypothyroidism (25.4%) or subclinical hypothyroidism with TSH > 10 (1.1%; data not shown). Because the prevalence of subclinical hypothyroidism is much higher than of overt hypothyroidism, it can be expected that the best diagnostic performance of a test to detect a treatment indication is reached with the same modifications as for subclinical hypothyroidism. This concept is important for future recommendations on universal reference limits because diagnosing overt hypothyroidism, an entity with an evident treatment indication, is generally prioritized in diagnostic strategies for gestational thyroid dysfunction. However, failing to identify the more prevalent subclinical disease could also lead to decreased benefits of (selective) screening. Although we found no method with an agreeable tradeoff in terms of diagnostic performance, it is important to realize that the interpretation of diagnostic performance of a test depends on the prior probability of disease (34). This is a highly relevant concept when thinking about differences between generalized population screening (with a low prior probability) vs high-risk case-based screening (with higher prior probabilities). For example, for a hypothetical diagnostic test with a sensitivity of 0.75 and a specificity of 0.99 (roughly equal to the tests assessed in our study), a pretest probability of 3% would result in a postpositive test probability (or PPV) of 70% and a false discovery rate of 30%. Using the same sensitivity and specificity, a pretest probability of 10% would result in a postpositive test probability of 89% with a false discovery rate of 11%. The current study population consists of population-based cohort studies as a reflection of the general population, which have a low prior probability of disease equal to the population prevalence and similar to a universal screening approach. One option to improve how alternative reference interval strategies could identify those with an abnormal thyroid function would be to increase the prior probability of disease (34). This can be achieved by optimizing the identification of high-risk subgroups and a risk-based screening approach, which could improve the accuracy of diagnostic strategies (35). Thus, the implementation of universal screening will be inherently associated with the lowest prior probability of disease and the highest rates of both over and underdiagnosis, especially if alternative strategies are used to define thyroid function test abnormalities.

The heterogeneity between populations (as denoted by wide prediction intervals and high I^2 statistics) underline that calculating local center and pregnancy-specific reference intervals for TSH and FT4 should still be considered as current best practice. However, other strategies for the improvement of the diagnosis of gestational thyroid dysfunction might prove more effective. The trimester-specific approach is currently accepted as the best diagnostic method for diagnosing thyroid dysfunction in pregnancy, but the pragmatic division of the gestational period in trimesters does not necessarily reflect the physiological changes of thyroid function tests during pregnancy (36-38). Further studies are needed to assess which gestational period reference intervals should be based on to optimally identify the women at increased risk of adverse events because of thyroid dysfunction, or if any form of standardization to gestational age should be abandoned altogether. Current reference interval definitions are based on outlying percentiles of TSH and FT4 distributions (2.5th and 97.5th percentiles), values above or below those cutoffs were later shown to be associated with adverse pregnancy outcomes (39). With increasing data availability in the literature, the ideal way to establish reference values would be to turn this methodology around and base the cutoffs on the risk of adverse outcomes, similar to other fields (40, 41). Obvious adverse pregnancy events would be those associated with thyroid function tests in previous studies such as preterm birth and offspring IQ scores (3, 4, 6). Because we did not identify an adequate or easily implementable methodology to approach trimester-specific reference intervals in the current

Downloaded from https://academic.oup.com/jcem/article/109/11/e2151/7724966 by guest on 18 June 2025

study, our group will aim to establish risk-based decision limits.

In this study, we were able to leverage a large international dataset of multiple population-based prospective cohort studies to assess novel strategies for diagnosing thyroid dysfunction in pregnancy. The interpretation of the results of this study are limited to populations with sufficient or mild-to-moderate iodine deficiency because studies with excessive status were excluded and no studies were performed in an area of severe iodine deficiency. Additionally, multiple differences between the included study populations, including differences in iodine supplementation, assays, and determinants of thyroid function tests, could have contributed to the variability in diagnostic performance of the nonpregnancy reference interval adaptations assessed in this study. Adaptations of nonpregnancy reference limits could be more accurate in specific populations, which we were not able to assess with sufficient power. Nonetheless, this study reflects common practice because these factors naturally vary between populations. The results of the current study may not be optimally generalizable to present-day populations because the inclusion periods for the majority of included cohorts were between the years 2000 and 2015. It is likely that determinants of thyroid function and assay calibrations standards have changed over time (42). It can, however, be expected that large inter-population differences, as demonstrated in this study, are still present to this day. Ongoing harmonization efforts by the International Federation of Clinical Chemistry and Laboratory Medicine could improve the diagnostic performance of alternative strategies and future studies could assess if a generalizable rule is more effective in cohorts established after the start of the harmonization efforts.

In conclusion, this is the first study to systematically quantify the diagnostic performance of standardized modifications of nonpregnancy TSH and FT4 reference intervals in pregnancy. We show that standardized modifications have poor overlap in diagnostic accuracy compared with cohort and trimester-specific reference intervals, resulting in considerable variation in diagnostic performance between populations. Future efforts should be turned toward studying the meaningfulness of trimester-specific, pregnancy-specific reference intervals and the establishment of risk-based decision limits.

Acknowledgments

The authors gratefully acknowledge all participants, general practitioners, hospitals, and midwives for their important contribution to the establishment of the cohorts and the resulting works. Acknowledgments for individual cohorts are listed in the supplemental materials (21).

Funding

Netherlands Organization for Scientific Research (grant 401.16.020 and 016.176.331) to R.P.P.

Disclosures

P.T. reports a travel grant from Society for Endocrinology (leadership development award). E.N.G. received speaker's fees and payment for expert testimony from Merck and consulting fees from Brunel Rus. T.G.M.V. reports grants from the Netherlands Organization for Health Research and Development. L.C. received travel support by Pfizer. S.M.N. has received consultancy, speakers' fees, or travel support from Access Fertility, Beckman Coulter, Ferring Pharmaceuticals, Merck, Modern Fertility, Roche Diagnostics, and The Fertility Partnership. S.M.N. also reports payments for medical–legal work and investment in The Fertility Partnership. T.I.M.K. reports lectureship fees from Berlin-Chemie, Goodlife Healthcare, Institut Biochimique SA, Merck, and Quidel. U.F.R.'s research salary was sponsored by an unrestricted grant from Kirsten and Freddy Johansen's Fund and U.F.R. reports lecture fees from Merck, Darmstadt. S.B.'s research salary was sponsored by the Capital Region of Denmark's Research Foundation and the Novo Nordisk Foundation (ID 0077221). S.B. received a lecture fee from Merck and Novo Nordisk. All other authors declare no competing interests.

Data Availability

The data that support the findings of this study are not publicly available due to local, national, and international restrictions aimed to protect the privacy of research participants.

References

1. Derakhshan A, Peeters RP, Taylor PN, *et al.* Association of maternal thyroid function with birthweight: a systematic review and individual-participant data meta-analysis. *Lancet Diabetes Endocrinol*. 2020;8(6):501-510.
2. Toloza FJK, Derakhshan A, Mannisto T, *et al.* Association between maternal thyroid function and risk of gestational hypertension and pre-eclampsia: a systematic review and individual-participant data meta-analysis. *Lancet Diabetes Endocrinol*. 2022;10(4):243-252.
3. Levie D, Korevaar TIM, Bath SC, *et al.* Thyroid function in early pregnancy, child IQ, and autistic traits: a meta-analysis of individual participant data. *J Clin Endocrinol Metab*. 2018;103(8):2967-2979.
4. Thompson W, Russell G, Baragwanath G, Matthews J, Vaidya B, Thompson-Coon J. Maternal thyroid hormone insufficiency during pregnancy and risk of neurodevelopmental disorders in offspring: a systematic review and meta-analysis. *Clin Endocrinol (Oxf)*. 2018;88(4):575-584.
5. Han Y, Gao X, Wang X, *et al.* A systematic review and meta-analysis examining the risk of adverse pregnancy and neonatal outcomes in women with isolated hypothyroxinemia in pregnancy. *Thyroid*. 2023;33(5):603-614.
6. Korevaar TIM, Derakhshan A, Taylor PN, *et al.* Association of thyroid function test abnormalities and thyroid autoimmunity with preterm birth: a systematic review and meta-analysis. *JAMA*. 2019;322(7):632-641.
7. Krassas GE, Poppe K, Glinoer D. Thyroid function and human reproductive health. *Endocr Rev*. 2010;31(5):702-755.
8. Osinga JAJ, Derakhshan A, Palomaki GE, *et al.* TSH and FT4 reference intervals in pregnancy: a systematic review and individual participant data meta-analysis. *J Clin Endocrinol Metab*. 2022;107(10):2925-2933.
9. Springer D, Bartos V, Zima T. Reference intervals for thyroid markers in early pregnancy determined by 7 different analytical systems. *Scand J Clin Lab Invest*. 2014;74(2):95-101.
10. Negro R, Attanasio R, Papini E, *et al.* A 2018 Italian and Romanian survey on subclinical hypothyroidism in pregnancy. *Eur Thyroid J*. 2018;7(6):294-301.
11. Toloza FJK, Ospina NMS, Rodriguez-Gutierrez R, *et al.* Practice variation in the care of subclinical hypothyroidism during pregnancy: a national survey of physicians in the United States. *J Endocr Soc*. 2019;3(10):1892-1906.

Downloaded from https://academic.oup.com/jcem/article/109/11/e2151/7724966 by guest on 18 June 2025

12. Lazarus J, Brown RS, Daumerie C, Hubalewska-Dydejczyk A, Negro R, Vaidya B. 2014 European thyroid association guidelines for the management of subclinical hypothyroidism in pregnancy and in children. *Eur Thyroid J*. 2014;3(2):76-94.
13. Alexander EK, Pearce EN, Brent GA, *et al.* 2017 guidelines of the American thyroid association for the diagnosis and management of thyroid disease during pregnancy and the postpartum. *Thyroid*. 2017;27(3):315-389.
14. Thyroid disease in pregnancy: ACOG practice bulletin, number 223. *Obstet Gynecol*. 2020;135(6):e261-e274.
15. Bliddal S, Feldt-Rasmussen U, Boas M, *et al.* Gestational age-specific reference ranges from different laboratories misclassify pregnant women's thyroid status: comparison of two longitudinal prospective cohort studies. *Eur J Endocrinol*. 2014;170(2): 329-339.
16. Liu J, Yu X, Xia M, *et al.* Development of gestation-specific reference intervals for thyroid hormones in normal pregnant northeast Chinese women: what is the rational division of gestation stages for establishing reference intervals for pregnancy women? *Clin Biochem*. 2017;50(6):309-317.
17. Mehran L, Amouzegar A, Delshad H, *et al.* Trimester-specific reference ranges for thyroid hormones in Iranian pregnant women article. *J Thyroid Res*. 2013;2013:651517.
18. Osinga JAJ, Derakhshan A, Feldt-Rasmussen U, *et al.* TSH and FT4 reference interval recommendations and prevalence of gestational thyroid dysfunction: quantification of current diagnostic approaches. *J Clin Endocrinol Metab*. 2024;109(3):868-878.
19. Thienpont LM, Van Uytfanghe K, De Grande LAC, *et al.* Harmonization of serum thyroid-stimulating hormone measurements paves the way for the adoption of a more uniform reference interval. *Clin Chem*. 2017;63(7):1248-1260.
20. Thienpont LM, Van Uytfanghe K, Van Houcke S, *et al.* A progress report of the IFCC committee for standardization of thyroid function tests. *Eur Thyroid J*. 2014;3(2):109-116.
21. Osinga JAJ, Derakhshan A, Korevaar TIM. Data from: reference intervals. Consortium on thyroid and pregnancy. Updated July 18, 2023. https://www.consortiumthyroidpregnancy.org/reference intervals
22. Lutgendorf MA, Stoll KA. Why 99% may not be as good as you think it is: limitations of screening for rare diseases. *J Matern Fetal Neonatal Med*. 2016;29(7):1187-1189.
23. Goutte C, Gaussier E. *A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation*. Springer; 2005:345-359.
24. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ*. 2011;342:d549.
25. Osinga JAJ, Derakhshan A, Peeters RP, Korevaar TIM. Data from: Heatmaps. Consortium on Thyroid and Pregnancy. Updated January 31, 2024. Accessed January 31, 2024. https://www.consortiumthyroidpregnancy.org/heatmaps
26. Lin L, Chu H. Meta-analysis of proportions using generalized linear mixed models. *Epidemiology*. 2020;31(5):713-717.
27. Stijnen T, Hamza TH, Ozdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Stat Med*. 2010;29(29): 3046-3067.
28. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2022.
29. Balduzzi S, Rucker G, Schwarzer G. How to perform a meta-analysis with R: a practical tutorial. *Evid Based Ment Health*. 2019;22(4):153-160.
30. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York; 2009.
31. Galili T, O'Callaghan A, Sidi J, Sievert C. Heatmaply: an R package for creating interactive cluster heatmaps for online publishing. *Bioinformatics*. 2018;34(9):1600-1602.
32. Brochure. Roche Diagnostics GmbH. Reference Intervals for Children and Adults Elecsys Thyroid Tests 2009.
33. Brochure. Abbott Diagnostic Division. Architect system TSH ref 7K62. 2010.
34. Bours MJ. Bayes' rule in diagnosis. *J Clin Epidemiol*. 2021;131: 158-160.
35. Osinga JAJ, Liu Y, Männistö T, *et al.* Risk factors for thyroid dysfunction in pregnancy: an individual participant data meta-analysis. *Thyroid*. 2024;34(5):646-658 .
36. Korevaar TIM, Medici M, Visser TJ, Peeters RP. Thyroid disease in pregnancy: new insights in diagnosis and clinical management. *Nat Rev Endocrinol*. 2017;13(10):610-622.
37. Glinoer D, de Nayer P, Bourdoux P, *et al.* Regulation of maternal thyroid during pregnancy. *J Clin Endocrinol Metab*. 1990;71(2): 276-287.
38. Andersen SL, Andersen S, Carle A, *et al.* Pregnancy week-specific reference ranges for thyrotropin and free thyroxine in the north Denmark region pregnancy cohort. *Thyroid*. 2019;29(3): 430-438.
39. van den Boogaard E, Vissenberg R, Land JA, *et al.* Significance of (sub) clinical thyroid dysfunction and thyroid autoimmunity before conception and in early pregnancy: a systematic review. *Hum Reprod Update*. 2011;17(5):605-619.
40. HAPO Study Cooperative Research Group; Metzger BE, Lowe LP, *et al.* Hyperglycemia and adverse pregnancy outcomes. *N Engl J Med*. 2008;358(19):1991-2002.
41. Xu Y, Derakhshan A, Hysaj O, *et al.* The optimal healthy ranges of thyroid function defined by the risk of cardiovascular disease and mortality: systematic review and individual participant data meta-analysis. *Lancet Diabetes Endocrinol*. 2023;11(10):743-754.
42. Van Uytfanghe K, Ehrenkranz J, Halsall D, *et al.* Thyroid stimulating hormone and thyroid hormones (triiodothyronine and thyroxine): an American thyroid association-commissioned review of current clinical and laboratory status. *Thyroid*. 2023;33(9): 1013-1028.

Downloaded from https://academic.oup.com/jcem/article/109/11/e2151/7724966 by guest on 18 June 2025