Impact of Reagent Lot-to-Lot Variation on Internal Quality Control Results in Thyroid Hormone Analysis: A Retrospective Laboratory Study

[1]Kazi Reazuddin Ahmed, [1]Sudipto Das, [2]Moontaha Binte Rashid, [2]Ilteza Tabassum, [2]Jerin Sultana, [3]Afroza Naznin, [4]Azmal Kabir Sarker, [5]Zeenat Jabin

[1]Scientific Officer, [2]Medical Officer, [3]Senior Medical Officer, [4]Principal Medical Officer, [5]Professor & Chief Medical Officer, Institute of Nuclear Medicine and Allied Sciences (INMAS), Suhrawardy, Dhaka

Correspondence Address: Kazi Reazuddin Ahmed, Scientific Officer, INMAS, Suhrawardy, ShSMCH Campus, Sher-E-Bangla Nagar, Dhaka- 1207, E-mail: reazuddink614@gmail.com

ABSTRACT

Background: Reagent lot-to-lot variation is a recognized source of analytical variability in immunoassays. The magnitude and clinical relevance of this phenomenon are underexplored in routine laboratory practice.

Methods: This retrospective study analyzed internal quality control data for FT3, FT4, and TSH over a period of 12 months. Bias resulting from reagent lot shift was calculated by comparing mean QC values between shifting reagent lots. Both directional and absolute bias were calculated.

Results: The mean bias was 1.05% for FT3, 1.20% for FT4, and 0.83% for TSH. Despite low mean bias, substantial variability was observed, with mean absolute bias ranging from 5.08% to 6.01%. TSH demonstrated the highest maximum bias (15.06%). No consistent concentration-dependent pattern was observed across QC levels.

Conclusion: Reagent lot transitions introduced non-directional random variability in thyroid hormone analysis. The variability was often large enough to be clinically significant. These findings highlight the need for robust lot verification protocols to ensure analytical consistency and accurate clinical interpretation.
Keywords: Chemiluminescence Analyzer, Internal QC, reagent lot, hormone assay.

Bangladesh J. Nucl. Med. Vol. 28 No. 2 July 2025

DOI: https://doi.org/10.3329/bjnm.v28i2.89110

INTRODUCTION

Accurate and precise laboratory measurements are essential for reliable clinical decision-making, particularly in the context of serial monitoring of biochemical parameters (1). Thyroid function tests (TFTs), FT3, FT4, and TSH are three of the most requested analyses performed in clinical laboratories (2,3). Besides the evaluation of the health of the thyroid gland, they also assist in the diagnosis of various metabolic disorders. Thyroid hormones regulate basal metabolic rate, heart rate, and body temperature (4–6). Monitoring the thyroid hormone levels is essential in managing patients with cardiovascular issues or weight disorders. TFTs are used in newborn screening to prevent developmental delays. Monitoring of maternal thyroid health in pregnant women is vital for neurological development as well (7, 8). The accuracy of these tests relies on high-quality laboratory standards. Quality control (QC) metrics should be monitored extensively to ensure that the patient results are accurate and reliable for diagnosis and decision-making.

Analytical variation in laboratory testing arises from multiple sources, including instrument performance, calibration age, and reagent manufacturing-related factors (9). Studies indicate that immunoassays for thyroid hormones are susceptible to shifts due to reagent lot changes (10). TSH is often cited as the most stable among the three, but because it is used as the frontline screening tool, even a minor shift can reclassify patients (11–13). Some studies have noted that patient medians can show subtle shifts of up to 10% across multiple reagent lots (14). FT3 and FT4 often show more variation. Literature has documented that new reagent lots can introduce a bias of 20% or more compared to previous lots, which necessitates a complete re-evaluation of the laboratory's reference intervals (RIs) (15). Besides lot changes, "within-lot" instabilities require frequent recalibration to maintain long-term analytical stability. The CLSI EP26-A

Guideline is the primary industry standard for user evaluation of the acceptability of a reagent lot change (16, 17). It typically recommends using patient samples rather than just QC material, as QC samples may not always reflect the matrix effects seen in human serum. However, there remains a paucity of real-world longitudinal studies quantifying the magnitude and clinical relevance of reagent lot-to-lot bias in routine laboratory practice.

The present study aims to evaluate the impact of reagent lot-to-lot variation on internal QC performance in TFTs. By analyzing retrospective QC data for FT3, FT4, and TSH, we examined the analytical shifts across multiple reagent lot changes over 12 months. QC bias was calculated to quantify the magnitude of lot-to-lot result shift.

MATERIALS AND METHOD

Study Design and Materials

This retrospective analytical study was conducted on the internal QC performance of FT3, FT4, and TSH. The study utilized QC data generated on the Siemens Advia Centaur XPT Chemiluminescence Analyzer. The test reagents and calibrators were purchased from Siemens and stored in pharmacy refrigerators at 2–8°C until use. Lyophilized Lypochek Immunoassay Plus Control (lot # 40450) standards were purchased from Bio-Rad and stored in a pharmacy refrigerator at 2–8°C until reconstituted using distilled ampoule water. Reconstituted QC solutions were divided into 1.0 mL aliquot tubes and stored at ~-22.0°C in a medical freezer to avoid repeated freezing and thawing during routine QC checks. QC checks were performed at three concentration levels, low (level 1), normal (level 2), and high (level 3), for each analyte as part of routine QC checks. A total of 820 QC observations were included across all analytes and QC levels.

Data Collection and Structure

The dataset comprised QC results with associated variables necessary to identify reagent-related variation for each analyte. QC level, result, date of analysis, and primary reagent lot number were extracted from the XPT system database for the period of the 12 months of 2025. The data was then cleaned to remove diagnostic QC results.

Analysis

Lot-to-lot bias was assessed by comparing mean QC values between consecutive reagent lots. For each analyte and QC level, the mean control result was calculated within each reagent lot, and percentage bias between successive lots was determined using the following formula:

$$\text{Bias}\,\% = \frac{\text{Mean}_{\text{Lot n+1}} - \text{Mean}_{\text{Lot n}}}{\text{Mean}_{\text{Lot n}}} \times 100$$

Both directional bias and absolute bias were evaluated. Absolute bias was calculated as the absolute value of the percentage difference to reflect the magnitude of variation irrespective of direction. Bias calculations were performed separately for each QC level and analyte, and aggregated summaries were generated to describe overall assay performance.

Descriptive statistical methods were used to summarize the magnitude and distribution of lot-to-lot bias. For each analyte, mean, standard deviation (SD), minimum (Min.) and maximum (Max.) bias % were calculated. Bias values were further stratified by QC level to assess potential concentration dependence.

Modified Levey-Jennings plots were constructed for each analyte to visualize QC performance across reagent lots. Box plots of absolute bias were generated to compare variability across analytes.

Ethical Clearance

Ethical approval was waived as the study utilized anonymized internal quality control data without patient involvement.

Software

Data analysis was performed with python 3.13.2 on Jupyter notebook using Pandas and NumPy libraries. For visualization, Matplotlib library was used.

RESULTS

The Levey-Jennings plots provide a longitudinal visualization of internal QC performance across consecutive reagent lot changes for each analyte. FT3 has the fewest reagent lot changes (4 lots) during the study period. Both FT3 (Figure 1) and FT4 (Figure 2) show significant intra lot variability in routine QC. The shifts at transitions are not visibly as significant as the random variability.

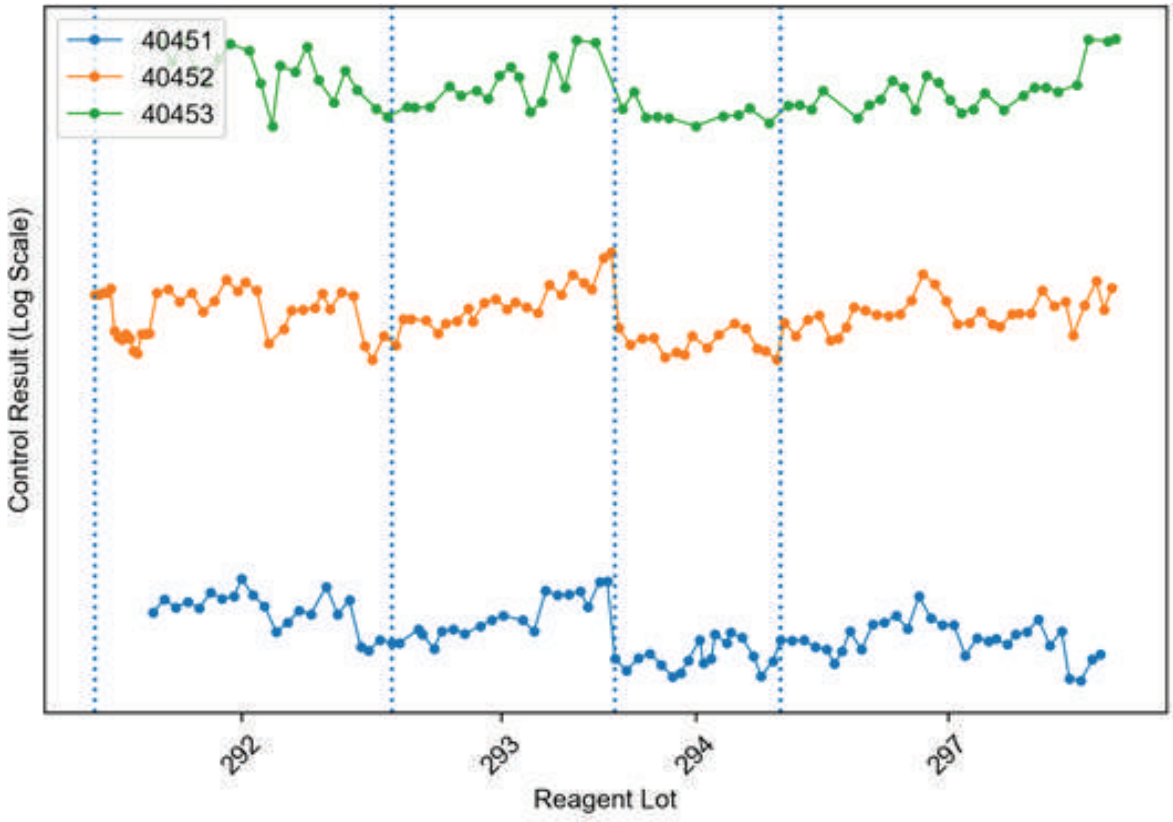

Figure 1: Levey-Jennings plot for FT3

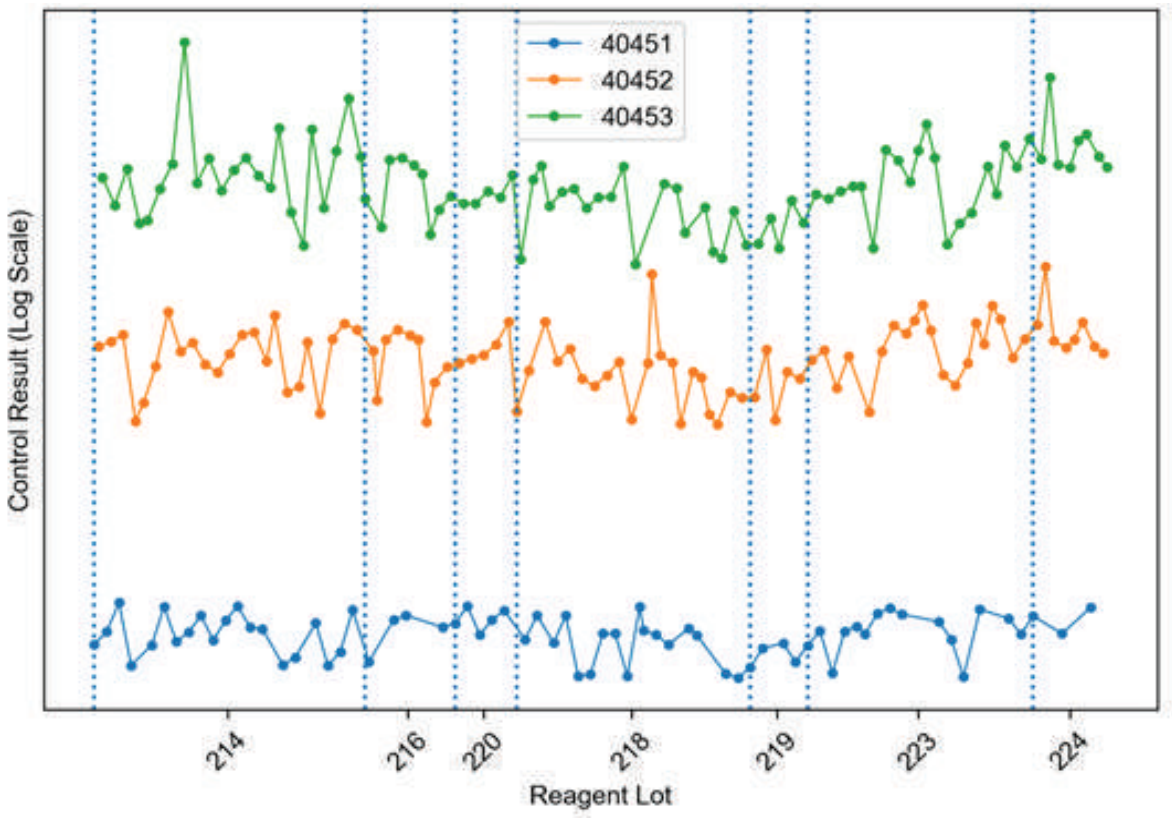

Figure 2: Levey-Jennings plot for FT4

The TSH plot (Figure 3) demonstrates the highest degree of analytical stability across multiple reagent lots (lot 3 to 11). The three control levels remain relatively tightly clustered. Compared to later lots, earlier lot (lot 4) appears more stable in the Levey-Jennings plot. After lot 4, there is significantly more random variability. There is no evidence of any significant shift with lot changes.

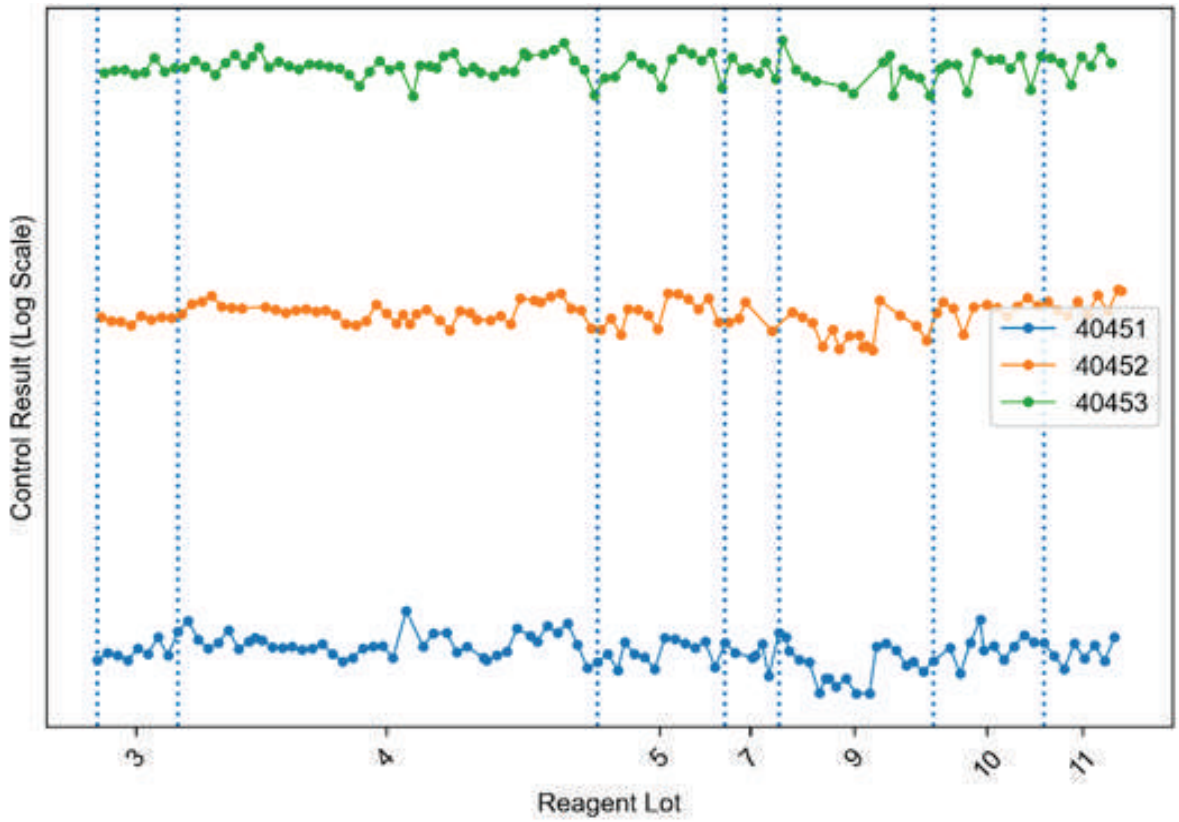

Figure 3: Levey-Jennings plot for TSH

The quantitative analysis of reagent lot transitions revealed varying degrees of shift across the three thyroid tests (Table 1). While TSH showed the lowest overall mean bias (0.83%), it demonstrated the highest maximum bias (15.06%). FT4 showed the highest mean bias (1.20%) and a significant maximum absolute bias (11.12%). The relatively high SD (6.62) aligns with the noisy Levey-Jennings plot (Figure 2)

Table 1: Summary of Lot-to-Lot Bias (n = number of lot transitions)

Test Parameter	Mean Bias (%)	SD	Mean Absolute Bias (%)	Max. Absolute Bias (%)	Min. Bias (%)	Max. Bias (%)
FT3 (n=3)	1.05	5.89	6.01	9.83	4.64	9.86
FT4 (n=6)	1.20	6.62	5.08	11.12	3.57	10.63
TSH (n=6)	0.83	6.72	5.51	15.06	10.13	15.06

The impact of the lot changes was not uniform across all three standard levels (Table 2). Level 2, which represents the physiological range, FT3 (7.00%) and TSH (6.72%) showed the highest mean absolute bias, while FT4 (4.04%) showed the lowest absolute bias. Relatively small values of mean bias for all three analytes suggest that the bias is random and non-directional.

Table 2: Bias by QC Level

Test Parameter	QC Level	Mean Bias (%)	Mean Absolute Bias (%)
FT3	1	2.34	5.43
	2	-0.43	7.00
	3	1.24	5.60
FT4	1	0.85	4.37
	2	1.34	4.04
	3	1.41	6.84
TSH	1	-0.09	6.60
	2	1.70	6.72
	3	0.87	3.20

Figure 4 provides a visual synthesis of the analytical stability for FT3, FT4, and TSH, highlighting the spread and central tendency of the shifts observed. FT3 displays the most compact interquartile range (IQR), with the median absolute bias near 5.6%. FT4 has the lowest median absolute bias (~4.4%), but a significant upper whisker reaching near 7.0%. TSH exhibits the largest overall spread, with an IQR that spans from 4.9% to 6.7, and a median value near 6.6%. It also shows the longest downward whisker (~3.3%).

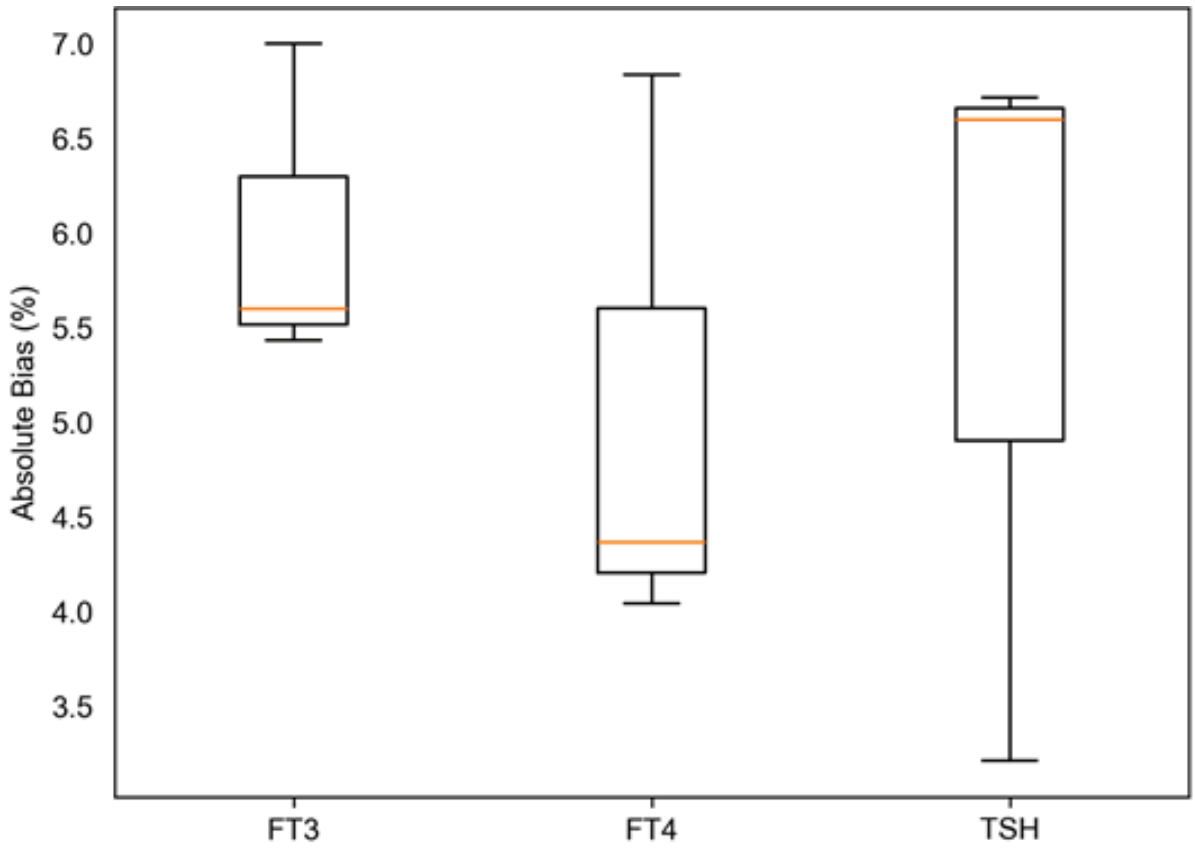

Figure 4: Lot-to-Lot absolute bias distribution

DISCUSSION

The study shows that TSH may experience significant bias (15.06%) when shifting reagent lots. The relatively small values for overall mean bias across all three analytes suggest that these shifts are largely random and non-directional rather than representing a consistent upward or downward drift over time. As shown in Table 2, concentration had negligible influence on the lot-to-lot shift result bias. The 15.06% maximum bias observed for TSH is concerning. Such large shifts may mimic clinically significant changes in patient status (18,19). The widespread high median absolute bias for TSH (~6.6%) shown in Figure 4 represents diagnostic risks for subclinical patients.

Table 3: Comparison of mean absolute bias with regulatory guidelines.

Test	Mean Absolute Bias (%)	Clinical Laboratory Improvement Amendments (CLIA) recommended total allowable error (%) (20)	Royal College of Pathologists of Australasia (RCPA) recommended total allowable error (%) (21)
FT3	6.01	± 25%	± 20%
FT4	5.08	± 15%	± 12%
TSH	5.51	± 20%	± 20%

The magnitudes of the mean absolute bias are all within recommended values (Table 3). However, the possibility of combined effect of lot bias and biological variation violating those values advocates for routine re-evaluation of reference intervals following major reagent changes to maintain analytical stability (22,23).

CONCLUSION

This retrospective study quantified significant lot-to-lot variation in TFTs over a 12-month period. While TSH appeared most stable visually, it demonstrated the highest susceptibility to extreme bias events. The findings reveal that reagent lot transitions introduce non-directional, concentration-independent bias that may impact results. Clinical laboratories should move beyond basic internal QC to include robust lot-to-lot verification protocols of new reagent lots and establish internal reference intervals to ensure uncompromised clinical decision-making.

REFERENCES

1. Plebani M. Charting the course of medical laboratories in a changing environment. Clinica Chimica Acta 2002;319:87–100. https://doi.org/10.1016/S0009-8981(02)00028-1.
2. Javaid U, Kennedy D, Addison C, Tsatlidis V, Razvi S. Frequency, determinants and costs of thyroid function testing in a laboratory serving a large population. Eur J Endocrinol 2022;186:553–60. https://doi.org/10.1530/EJE-21-1172.
3. Kiel S, Ittermann T, Völzke H, Chenot J-F, Et. A. Frequency of thyroid function tests and examinations in participants of a population-based study. BMC Health Services Research 2020. https://doi.org/10.1186/s12913-020-4910-7.
4. Kim B. Thyroid Hormone as a Determinant of Energy Expenditure and the Basal Metabolic Rate. Thyroid® 2008;18:141–4. https://doi.org/10.1089/thy.2007.0266.
5. Vargas-Uricoechea H, Bonelo-Perdomo A, Sierra-Torres CH. Effects of thyroid hormones on the heart. Clínica e Investigación En Arteriosclerosis 2014;26:296–309. https://doi.org/10.1016/j.arteri.2014.07.003.
6. Silva JE. Thyroid Hormone and the Energetic Cost of Keeping Body Temperature. Biosci Rep 2005;25:129–48. https://doi.org/10.1007/s10540-005-2882-9.
7. Porterfield SP, Hendrich CE. The Role of Thyroid Hormones in Prenatal and Neonatal Neurological Development—Current Perspectives. Endocr Rev 1993;14:94–106. https://doi.org/10.1210/edrv-14-1-94.
8. Escobar GM de, Obregón MJ, Rey FE del. Maternal thyroid hormones early in pregnancy and fetal brain development. Best Practice & Research Clinical Endocrinology & Metabolism 2004;18:225–48. https://doi.org/10.1016/j.beem.2004.03.012.
9. McCormack JP, Holmes DT. Your results may vary: the imprecision of medical measurements. BMJ 2020;368:m149. https://doi.org/10.1136/bmj.m149.
10. Algeciras-Schimnich A, Bruns DE, Boyd JC, Bryant SC, La Fortune KA, Grebe SK. Failure of Current Laboratory Protocols to Detect Lot-to-Lot Reagent Differences: Findings and Possible Solutions. Clin Chem 2013;59:1187–94. https://doi.org/10.1373/clinchem.2013.205070.
11. Helfand M, Crapo LM. Screening for Thyroid Disease. Ann Intern Med 1990;112:840–9. https://doi.org/10.7326/0003-4819-112-11-840.
12. Jefferies WMcK, Levy RP, Storaasli JP. Use of the TSH Test in the Diagnosis of Thyroid Disorders. Radiology 1959;73:341–4. https://doi.org/10.1148/73.3.341.
13. Matyjaszek-Matuszek B, Pyzik A, Nowakowski A, Jarosz MJ. Diagnostic methods of TSH in thyroid screening tests. Ann Agric Environ Med 2013;20:731–5.
14. Park JM, Kim JT, Park JY. Consideration of the Application of Measurement Uncertainty as an Acceptable Criterion For Reagent Lot-to-lot Comparability Test in Nuclear Medicine. Korean J Clin Lab Sci 2025;57:56–65. https://doi.org/10.15324/kjcls.2025.57.1.56.

15. Zweig MH, Robertson EA. Lot-to-lot variation in three thyroid-related radioimmunoassay kits. Clin Chem 1980;26:646–8. https://doi.org/10.1093/clinchem/26.5.646.
16. Clinical and Laboratory Standards Institute. User Evaluation of Acceptability of a Reagent Lot Change; Approved Guideline 2013.
17. Katzman BM, Ness KM, Algeciras-Schimnich A. Evaluation of the CLSI EP26-A protocol for detection of reagent lot-to-lot differences. Clinical Biochemistry 2017;50:768–71. https://doi.org/10.1016/j.clinbiochem.2017.03.012.
18. Abildgaard A, Knudsen CS, Bjerg LN, Lund S, Støy J. Lot variation and inter-device differences contribute to poor analytical performance of the DCA Vantage™ HbA1c POCT instrument in a true clinical setting. Clinical Chemistry and Laboratory Medicine (CCLM) 2022;60:127–34. https://doi.org/10.1515/cclm-2021-0720.
19. Houcke SKV, Stepman HCM, Thienpont LM, Fiers T, Stove V, Couck P, Anckaert E, Gorus F. Long-term stability of laboratory tests and practical implications for quality management. Clinical Chemistry and Laboratory Medicine (CCLM) 2013;51:1227–31. https://doi.org/10.1515/cclm-2012-0820.
20. Centers for Medicare & Medicaid Services (CMS). Clinical Laboratory Improvement Amendments of 1988 (CLIA) Proficiency Testing Regulations Related to Analytes and Acceptable Performance. Federal Register 2022;87:41194–242.
21. Royal College of Pathologists of Australasia Quality Assurance Programs (RCPAQAP). Allowable Limits of Performance. RCPAQAP Resource Centre 2023. https://rcpaqap.com.au/resource-centre/.
22. Ozarda Y, Higgins V, Adeli K. Verification of reference intervals in routine clinical laboratories: practical challenges and recommendations. Clinical Chemistry and Laboratory Medicine (CCLM) 2019;57:30–7. https://doi.org/10.1515/cclm-2018-0059.
23. Ozarda Y, Sikaris K, Streichert T, Macri J. Distinguishing reference intervals and clinical decision limits – A review by the IFCC Committee on Reference Intervals and Decision Limits. Critical Reviews in Clinical Laboratory Sciences 2018;55:420–31. https://doi.org/10.1080/10408363.2018.1482256.