This article has been accepted for publication in JCLINPATH following peer review.

The definitive copyedited, typeset version is available online at

10.1136/jcp-2025-210174

# Adding interpretative comments to results of thyroid function tests from patients on thyroxine replacement does not improve management

Amy Mallorie[1], Tim James[1], Sureshni deFonseka[2], Gayani Weerasinghe[2], Dave Green[2], Brian Shine[1]

[1]Department of Clinical Biochemistry, Oxford University Hospitals NHS Foundation Trust

[2]Department of Clinical Biochemistry, Buckinghamshire Healthcare NHS Trust

**Corresponding author:** Brian Shine, Department of Clinical Biochemistry, John Radcliffe Hospital, Oxford University Hospitals NHS Foundation Trust, Oxford OX3 9DU

Brian.shine@ouh.nhs.uk

## Abstract

### Aims

To assess the impact of adding clinical comments to reports of thyroid function testing in patients treated for hypothyroidism.

### Methods

We compared thyroid function test results in primary care patients being treated for hypothyroidism from January 2016 to August 2023 at two NHS Trusts with similar demographics and using the same instruments, but with different interpretative comment policies. One laboratory, Buckinghamshire Health Trust (Bucks), adds interpretative comments whereas the other, Oxford University Hospitals (Oxford) does not. We used two outcome measures: the percentage of patients with thyroid stimulating hormone (TSH) within the reference interval on repeat testing; and the timing of repeat TSH testing samples, according to NICE guidance (NG145).

### Results

We identified 18,242 and 31,655 hypothyroid patients (9.0% and 7.7% of the population tested) in Bucks and Oxford, with a total of 121,961 and 247,639 tests over the evaluation period, respectively. The proportion of TSH results within the reference interval (83.4% in Bucks, 83.9% in Oxford) was similar in both Trusts, as was TSH concentration (median TSH concentration 1.60 [IQR 0.78–2.82] mU/L in Bucks, 1.68 [IQR 0.97–2.76] in Oxford). The interval between tests was shorter in Oxford, but differed significantly from NG145 in both Trusts. Differences were statistically significant for both outcome measures, but of questionable clinical significance.

### Conclusions

Adding interpretative comments to results of thyroid function tests does not appear to affect the distribution of TSH concentrations in primary care patients on thyroxine replacement or the intervals between tests in a clinically meaningful way.

## Key messages

**What is already known on this topic:**

The addition of interpretive comments to numeric results is a common practice in laboratories.

Thyroid function tests often have clinical comments added, particularly to results from patients treated for hypothyroidism to suggest changes of doses of levothyroxine and retest intervals.

Considerable time resource is undertaken on adding thyroid test comments as it is commonly requested test for a common clinical condition.

The impact of adding comments to thyroid function tests on patient outcomes is poorly understood.

**What this study adds**:

The proportion of results below, within, and above the reference interval for thyroid stimulating hormone are very similar in a service that adds comments to thyroid results compared to one that does not.

The retesting interval between monitoring samples differs slightly between services but show that there was significant testing both sooner and later than recommended within national guidance,

**How this study might affect research, practice or policy:**

This audit data would suggest that addition of comments to treated hypothyroidism patients test results may have limited impact and therefore may not be a cost-effective use of laboratory staff time

## Introduction

There is debate concerning whether laboratory personnel should add interpretative comments to results.[1,2] Those in favour feel that this is a valuable part of the laboratory service,[3] and that it has a positive effect on clinical management [4]. There is evidence that the practice has support from primary care and nursing staff.[5] Counterarguments include that clinicians are confident in interpreting results of common tests and that they understand the clinical situation of a patient better than laboratory-based personnel.[6]. Commenting on results without being fully aware of all a patient's circumstances may pose a risk to the patient. Alternative options to support clinicians' interpretation of laboratory results include providing advice through links to national or local guidelines. However, there is limited evidence to support the effectiveness of either individualised commentary or links to guidelines.

Thyroid function testing (TFT) is an area of testing where most laboratories in the UK add comments to the numeric results to aid interpretation. From a resource perspective, this can add significantly to the total cost of providing TFT as the activity is usually considerable. For instance, in a laboratory undertaking 100,000 thyroid test requests per year, adding comments to 10% of specimens, taking an extra 60 seconds is equivalent to nearly 170 hours per year, about 10% of a whole-time post.

A significant proportion of thyroid comments are added to reports where testing is being undertaken to monitor treated hypothyroidism, a condition thought to affect 2.2 million patients in the UK.[7] NICE Guideline NG145[8] recommends TSH assessment every 3 months until TSH stabilises (two readings 3 months apart), and then once a year following this.

We sought to evaluate if the results of TFTs in patients with hypothyroidism were different in two hospitals in adjacent counties of England where the populations were similar, but the policies on interpretative comments differed.

# Methods

## Hospital policies on interpretative comments

We examined thyroid function tests performed at Buckinghamshire Healthcare NHS Trust (Bucks) and compared this with tests performed by Oxford University Hospitals NHS Foundation Trust (Oxford). Both Trusts provide testing from two laboratory sites (in Aylesbury and High Wycombe for Bucks, Oxford and Banbury for Oxford). Both Trusts provide primary and secondary care services. Oxford also provides tertiary services in endocrinology and thyroid cancer to both regions. The approximate workloads for thyroid function testing are 80,000 specimens for a population of 553,100 (Bucks) and 160,000 specimens annually for a population of 650,000 (Oxford). Demographics were comparable with similar population density:[9] median age 42.3 years (49.0% males, 91.7% Caucasian) in Bucks, 40.2 years (49.9% males, 90.9% Caucasian).

In Bucks, thyroid function tests with any result outside the relevant reference interval are reviewed by a consultant chemical pathologist, who adds interpretative clinical comments. If all results are within the relevant reference interval, an automated comment, "Euthyroid or adequate replacement", is added. For patients who appear to be on thyroxine, common comments include "Thyroxine replacement may be inadequate; suggest increase thyroxine and retest after 6 weeks" for those above the reference interval, and "Thyroxine replacement may be too high; suggest reduce thyroxine and retest after 6 weeks" for those with values below the reference interval. An abnormal result is phoned out to the requester if the laboratory staff feels that this will improve care of the patient.

In Oxford, no comments are added to thyroid function tests. Instead, the duty biochemist (usually a junior doctor) may phone a GP with a result if it is very abnormal or appears unexpected.

## Assay characteristics, data collection and analysis

All sites used Abbott Architect i2000 analysers (Abbott Laboratories, Maidenhead, UK) for thyroid testing and used TSH as the initial test with follow up thyroid tests (Free T3 and Free T4) added depending upon the TSH result. Bucks used the manufacturer's suggested reference interval for TSH of 0.35–4.94 mU/L, while Oxford used a locally derived reference interval of 0.30–4.20 mU/L.

We obtained thyroid function data from the laboratory information systems of both BHT and Oxford laboratories for all adults tested in the audit period. Each record included the date of the specimen, the origin of the specimen, the clinical details given by the clinician, any comments added, and the results for TSH, free T4, and free T3. We included patients who had clinical details anywhere in the record that suggested they were on thyroxine replacement and included only tests that had been requested in primary care.

Data were imported into R, version 4.4.0 (R Core Team, 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/). We used the Tidyverse set of utilities for data

handling and graphical representation of data. We used Chi-squared tests to compare the proportion of patients with results within the reference interval and the proportion of patients with repeated tests within the recommended time-frame between Trusts.

## Outcome measures

*Proportion of results with a comment (Bucks)*

For Bucks, we calculated the proportions of specimens with a comment indicating adequate, under- or over-replacement and those with no comment.

*Proportion of results within the reference interval:*

For each patient, we calculated the median TSH concentration, and used this to construct a violin and box-plot plot for each laboratory service. We calculated the proportion of results below, within or above the reference interval.

*Interval between tests:*

NICE Guideline NG145[8] recommends measuring the TSH every 3 months until the TSH is stable (two similar measurements within the reference range 3 months apart) and then once a year. To allow for recall of patients, we defined an appropriate gap between tests as 10-14 months in "stable" patients, and 4 weeks to 4.5 months in "unstable" patients. We used violin plots to represent the distribution of values for both the TSH concentrations and the intervals between tests.

# Results

## Patient and sample characteristics

The dataset covered the period from January 2016 to August 2023. After elimination of 5506 patients in both hospitals' datasets and people under 18 years at first test, the Bucks dataset comprised 554,411 specimens from 202,397 patients, with 121,961 results from 18,242 patients (9.0% of the population tested) apparently on thyroxine replacement, with 1–48 (median 6) specimens per patient. The Oxford dataset contained 1,218,957 specimens from 410,721 patients, with 247,639 specimens from 31,655 hypothyroid patients (7.7%) from primary care, and 1–78 (median 7) specimens per patient. Patients were older in Bucks than in Oxford (median 61.0 years [19.5% male] versus 54.7 years [22.3% male]).

## Comments added in Bucks

In Bucks, 73.3% of specimens had a comment "Euthyroid or adequate replacement", and 5.8% and 6.0% a comment indicating over- or under-replacement, respectively, 12.1% another comment, and 2.7% no comment.

## TSH values

TSH values from Bucks and Oxford were very similar (Figure 1) with median TSH concentrations of 1.60 (interquartile range 0.78–2.82) mU/L in Bucks, and 1.68 (0.97–2.76) mU/L in Oxford. In Bucks, 83.4% of patients had a median TSH concentration within the reference interval (0.3–4.9mU/L), with 7.3% higher and 9.3% lower. In Oxford, 83.9% had a median TSH concentration within the reference interval (0.3–4.2 mU/L), with 9.7% higher and 6.4% lower. The proportion within the reference interval was significantly lower in Bucks ($p < 0.0001$, Chi-squared test).

## Retest interval

The median interval between specimens was significantly lower in Oxford (134 and 280 days for "stable" and "not stable" patients, respectively) than in Bucks (243 and 370 days) (Figure 2). For "stable" patients, a higher proportion of patients in Bucks had repeat tests in the appropriate interval (64.3% in Bucks versus 46.4% in Oxford, $p < 0.0001$, Chi-squared test), whereas, for patients who were "not stable", the proportion was higher in Oxford (44.5% in Oxford versus 3.3% in BHT, $p < 0.0001$, Chi-squared test).

## Discussion

There is a longstanding debate on the benefits of adding interpretive comments and there has been a call for both evidence of effectiveness and improved quality assurance of the process.[6] To develop this further, an IFCC working group has been established to address harmonisation and quality.[10] However, studies on effectiveness of this activity on patient outcome are limited. Kilpatrick[4] observed mixed results from adding comments, but felt that there was a reduction in the proportion of people who were under-replaced.

In the current study we have shown that the median concentration of TSH is similar in hypothyroid patients in primary care on thyroxine replacement in two NHS Trusts with similar populations using the same analytical method but different strategies for adding interpretative comments to thyroid function tests (Figure 1). Although the proportion within range was greater in Oxford based on statistical testing, this may not be clinically significant. The two Trusts used slightly different reference intervals over the period of this survey, and using the same reference intervals may have increased the disparity between the proportion of results within the reference interval. Since this dataset was collected, the two Trusts have harmonised their reference intervals as part of a change to a common laboratory information system and this will provide an opportunity to reassess the impact on the outcomes assessed.

The interval between specimens was shorter in Oxford for both “stable” and “not stable” patients, with a lower proportion within the appropriate interval for the “stable” group, but a higher proportion in the “not stable” group. Although there were statistically significant differences for both subgroups, they may not represent clinically significant differences. In any case, it is apparent that many patients at both sites appear to be tested sooner or later than the guidelines recommend in both ‘stable’ and ‘not stable’ groups. This is consistent with a previous study,[11] in which laboratory data of repeat testing of thyroid monitoring was reviewed at two UK sites, and which concluded that most thyroid testing requests were outside the recommended intervals.

Because it is difficult to measure outcomes of the effectiveness of interpretative comments, using repeat TSH concentration and retesting intervals was a pragmatic choice. It is likely that patient preferences, levels of engagement with health care, patients’ changing health status and access to general practice appointments and phlebotomy also play important roles in achieving control of hypothyroidism.

The aim of adding comments to of laboratory results is to improve interpretation. While we have found no evidence of benefit in the context of monitoring patients with hypothyroidism, we do not claim that this can be extrapolated to other clinical areas. Hypothyroidism is a common condition and one with which clinicians in primary care are likely to have reasonable familiarity. The Office for National Statistics reported that, in 2022, there was an average of 2,300 patients per full time GP[12]. In the UK population 3.5% are reported to have hypothyroidism[7] which would mean each GP is likely to be monitoring around 80 patients. It is possible that comments appended to less commonly requested tests, where there is less familiarity in general practice, may have a greater impact on patient care.

We hope that this study may challenge expectations for laboratories to provide comments for patients on thyroxine replacement. As the population of treated hypothyroid patients continues to grow,[7] it is important to consider whether the addition of comments to thyroid results in the current manner is cost effective and sustainable against the resource constraints within the NHS.

Interpretative comments inserted by a consultant chemical pathologist or clinical scientist could be replaced by reflex comments depending upon the result of the TSH and FT4 measurements. We feel that this may mislead the requestor, and that an automated comment with a link to advice may be more useful and in keeping with the aims of personalised medicine. There is also the danger that clinicians may ignore routinely added interpretative comments because of the volume of test results that they are asked to look at and acknowledge. While Barlow found that GPs and nurse practitioners liked comments,[5] when we have asked GPs who use the Oxford laboratories, they said they were confident of interpreting results of thyroid function tests.

There are several limitations to this study: we think that we identified most people on levothyroxine using the clinical details provided but accept that this was an imperfect method; we did not perform a cost-benefit analysis, but this might be a next step, since modelling the benefits and costs of comments might reveal the critical factors for which further research might answer the question as to whether adding comments is a good use of a limited resource.

In summary, our data show no clinical benefit from adding interpretative comments to the results of requests for thyroid function tests in patients taking replacement thyroid hormones for hypothyroidism. We cannot recommend that laboratories spend resource on adding interpretative comments to thyroid function test results. Adding links to advice may be an alternative, but is untested.

## Contributorship

BS and TJ conceived the study. The data were collected by BS and DG. BS analysed the data. TJ and BS were the lead writers, with contributions from all other authors. All authors had access to the data. BS is the guarantor.


## Funding Statement

The authors received no funding for this work.


## Competing interests

The authors have no competing interests

## Ethics approval

The assessment was undertaken as an internal audit against requirements of laboratory service effectiveness ISO15189:2012, clause 4.14.5.

## Data sharing

The authors will be happy to share anonymised data for any reasonable purpose. Contact Dr Brian Shine (brian.shine@ouh.nhs.uk).


## Acknowledgements

None


### Patient and public involvement

No patient involvement.

## References


1. Marshall WJ., Challand GS. Provision of interpretative comments on biochemical report forms. Ann Clin Biochem. 2000;37:758–763
2. Challand G., Osypiw J. Interpretation in clinical biochemistry: an external quality assurance scheme. *EJIFCC*. 2004 Jun 17;15(2):35–38.
3. Vasikaran S, Loh TP. Interpretative commenting in clinical chemistry with worked examples for thyroid function test reports. *Pract Lab Med.* 2021;26:e00243.
4. Kilpatrick ES. Can the addition of interpretative comments to laboratory reports influence outcome? An example involving patients taking thyroxine. *Ann Clin Biochem.* 2004;41:227–229.
5. Barlow IM. Are biochemistry interpretative comments helpful? Results of a general practitioner and nurse practitioner survey. *Ann Clin Biochem*. 2008;45:88–90.
6. Young IS. Interpretive comments on clinical biochemistry reports *J Clin Path*.2005;58:575.
7. Razvi S, Korevaar T, Taylor P. Trends, Determinants, and Associations of Treated Hypothyroidism in the United Kingdom, 2005-2014. *Thyroid*. 2019;29:174-182
8. Thyroid disease: assessment and management, NICE guideline [NG145] Published 20 November 2019
9. https://www.ons.gov.uk/aboutus/transparencyandgovernance/freedomofinformationfoi/averagestatisticsbycounty (accessed 25/7/2023)
10. Vasikaran S, Sikaris K, Kilpatrick E, French J, Badrick T, Osypiw J, Plebani M; IFCC WG Harmonization of Quality Assessment of Interpretative Comments. EQA for interpretive commentsAssuring the quality of interpretative comments in clinical chemistry*. Clin Chem Lab Med*. 2016;54:1901-1911.
11. Scargill JJ, Livingston M, Holland D, Duff CJ, Fryer AA, Heald AH. Monitoring Thyroid Function in Patients on Levothyroxine. Assessment of Conformity to National Guidance and Variability in Practice. *Exp Clin Endocrinol Diabetes*. 2017;125:625-633.
12. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthcaresystem/articles/trendsinpatienttostaffnumbersatgppracticesinengland/2022#:~:text=Across%20England%2C%20the%20number%20of,overall%20(1%2C800%20to%201%2C700) (accessed 16/5/2025).

## Figure legends

***Figure 1*** *Combined violin and box plot of distribution of median TSH concentrations for Bucks and Oxford (logarithmic scale), with upper and lower reference values for Bucks (dotted lines) and Oxford(dashed lines), and proportion of patients with values within, above (High) and below (Low) the reference interval*

***Figure 2*** *Combined violin and box plot of distributions of interval between tests (log scale) stratified by hospital and whether "Stable" (two previous values within the reference interval more than 75 days apart) or "Not stable", with proportion of tests within expected interval (OK), early or late*