Accuracy of physical examination in the diagnosis of hypothyroidism: A cross-sectional, double-blind study

Indra R, Patil SS, Joshi R, Pai M,* Kalantri SP

Department of Medicine, Mahatma Gandhi Institute of Medical Sciences, Sevagram, Wardha - 442102, India and *Division of Epidemiology, School of Public Health, University of California, Berkeley, CA 94720, USA.

Correspondence:
SP Kalantri, MD
E-mail:
sp_kalantri@rediffmail.com

Received : 29-12-03
Review completed : 30-01-04
Accepted : 23-02-04
PubMed ID : 15047991
J Postgrad Med 2004;50:7-11

ABSTRACT

Background: Hypothyroidism is a common, potentially treatable endocrine disorder. Since hypothyroidism is not always associated with the signs and symptoms typically attributed to it, the diagnosis is often missed. Conversely, patients with typical signs and symptoms may not have the disease when laboratory tests are performed.

Aims: We aimed to determine the accuracy of physical examination in the diagnosis of hypothyroidism.

Setting and Design: Prospective, hospital-based, cross-sectional diagnostic study.

Material and Methods: Consecutive outpatients from the medicine department were screened and an independent comparison of physical signs (coarse skin, puffy face, slow movements, bradycardia, pretibial oedema and ankle reflex) against thyroid hormone assay (TSH and FT4) was performed.

Statistical Analysis: Diagnostic accuracy was measured as sensitivity, specificity, positive likelihood ratios, negative likelihood ratios and positive and negative predictive values.

Results: Of the 1450 patients screened, 130 patients (102 women and 28 men) underwent both clinical examination and thyroid function tests. Twenty-three patients (18%) were diagnosed to have hypothyroidism by thyroid hormone assays. No single sign could easily discriminate a euthyroid from a hypothyroid patient (range of positive likelihood ratio (LR+) 1.0 to 3.88; range of negative likelihood ratio (LR-): 0.42 to 1.0). No physical sign generated a likelihood ratio large enough to increase the post-test probability significantly. The combination of signs that had the highest likelihood ratios (coarse skin, bradycardia and delayed ankle reflex) was associated with modest accuracy (LR+ 3.75; LR- 0.48).

Conclusion: Clinicians cannot rely exclusively on physical examination to confirm or rule out hypothyroidism. Patients with suspected hypothyroidism require a diagnostic workup that includes thyroid hormone assays.

KEY WORDS: Hypothyroidism, physical examination, diagnosis, accuracy, sensitivity, specificity, likelihood ratio

The picture of a typical hypothyroid patient vividly painted in medical textbooks is seldom seen in clinical practice. What we often see is a presentation that is not always identified by the history and the physical examination. The diagnosis of hypothyroidism is sometimes missed because it is not always associated with the symptoms or signs attributed to it or because the clinical features manifest so slowly that clinicians may fail to notice them.[1,2] Also, the symptoms lack specificity and clinicians often attribute them to common non-thyroid diseases. Conversely, several individuals with non-specific symptoms are diagnosed to have hypothyroidism when evaluated with the help of thyroid function tests.[1] The U.S. Preventive Services Task Force recommends that clinicians remain alert to the subtle or non-specific nature of thyroid dysfunction and maintain a low threshold for the diagnostic evaluation of thyroid dysfunction.[3]

Can we rely on the clinical history and the physical examination alone to diagnose hypothyroidism? Several studies have evaluated this question.[4-11] Some studies retrospectively reviewed the medical records of patients and correlated clinical features with diagnoses.[5,6,8] Other studies were done as endocrine-clinic-based with limited generalisability.[4-6,8] Some studies included few men[5,9,10] or no men in the study population,[9] or included only elderly populations.[9,11] A few studies employed inadequate reference standards such as estimation of serum protein-bound iodine and cholesterol.[5,6,10] One study measured the thyroid hormones levels of only those patients who tested positive on a symptoms questionnaire.[6] We designed a cross-sectional, double-blind study to determine the diagnostic accuracy of physical examination in the diagnosis of hypothyroidism, in comparison to thyroid hormone assays, in a rural, tertiary hospital in India.

Material and Methods

Screening of the study population

Between April and September 2002, every Thursday and Saturday,

internal medicine residents (SSP and RJ) asked the following questions[6] to consecutive patients presenting to the Medicine outpatient department of a rural-based teaching hospital:

1. Do you feel less energetic than you felt a year ago?
2. Do you lack interest in your surroundings?
3. Has the skin of your arms or legs become more dry or rough during the past year?
4. Do you think you have put on weight in the last year?
5. Have you or any of your family or friends noticed that your voice has recently become huskier or weaker?

The categorical (yes or no) verbal responses to the questions were recorded. Patients with heart failure, anaemia, proteinuria, chronic renal failure, and laryngeal lesions were excluded by appropriate history and investigations. Those known to have hypothyroidism or those who were on thyroid replacement therapy and those who had had thyroidectomy were also excluded from the study.

Methods of physical examination

Patients who responded in the affirmative to any of the screening questions were referred to another internal medicine resident (RI) who was blind to the responses to the questions and findings of the physical examination. He elicited the following signs and recorded them as present or absent.

1. Coarse skin: the hands, forearms, and elbows were examined to judge if they felt rough and thick.
2. Sluggish movements: patients were asked to fold a 2-meter-long bed sheet. Those who took more than a minute to do so were considered to have sluggish movements.
3. Pulse rate: a resting pulse rate of less than 60/min was classified as bradycardia.
4. Pretibial oedema: the shin was pressed for thirty seconds to see if the pressure produced a pit.
5. Puffiness of the face: facial puffiness was detected by observing if the curve of the malar bone was obscured and the eyelids appeared boggy.
6. Ankle reflex: the contraction and the relaxation of the calf muscles were observed and the prolongation of the reflex was assessed by the naked eye.

A senior consultant (SPK) confirmed the physical signs; any disagreement in the interpretation of history or physical examination was sorted out by mutual discussion.

Measurement of the reference standard

All the screened patients had their blood drawn for free thyroxin (FT4) and thyroid stimulating hormone (TSH) levels on the day of the examination. Neither the nurse/technician who drew the blood samples nor the laboratory that analysed them had any access to the clinical data. The TSH levels were measured by a third generation, ultrasensitive radioimmunoassay (Thyrocare Technologies Limited, Mumbai, India). Free T4 levels were measured using a chemiluminescence assay. Patients with FT4 <0.7 ng/dL and TSH >7 IU/ml were judged to have hypothyroidism. We chose these standard cut-off points to exclude subclinical hypothyroidism, a condition characterised by raised TSH but normal FT4 values.[12]

The study design was cross-sectional: all patients, regardless of results of physical examination, underwent the reference standard test (thyroid hormone assays) at the same point in time. The investigator who performed the physical examination had no prior knowledge of the thyroid hormone assay results. The laboratory staff that performed the hormone assays had no knowledge of the patient's history and physical examination results. The study design, therefore, was double-blind. The institutional review board approved the study. The investigators explained the nature of the study to all the patients and obtained informed consent before enrolment.

Statistical analysis

Diagnostic accuracy was measured by the computation of the following test properties for each sign, and combination of signs, using standard methods: sensitivity, specificity, positive likelihood ratios (LR+), negative likelihood ratios (LR–), and positive and negative predictive values.[13] The precision of these estimates was evaluated by using 95% confidence intervals (95% CI).

The likelihood ratios were computed by means of sensitivity and specificity values. They indicate by how much a given test result will raise or lower the pre-test probability of the target disease.[13] An LR of 1 indicates that the post-test probability is the same as the pre-test probability (since pre-test odds x LR = post-test odds). Tests with LR values of close to 1 have limited clinical importance since they cannot help a clinician to rule in or rule out the target disease. Likelihood ratios of more than 1.0 increase the probability that the target disorder is present, and tests with large LR+ values may be useful for confirming the disease because they lead to large shifts in the post-test probabilities relative to pre-test probabilities. On the other hand, LRs which are <1.0 decrease the probability of the target disorder. Jaeschke *et al* provide the following rough guide for interpreting likelihood ratios:

1. Likelihood ratios of >10 or < 0.1 generate large and often conclusive changes from pre-test to post-test probability;
2. Likelihood ratios of 5-10 and 0.1-0.2 generate moderate shifts in pre-test to post-test probability;
3. Likelihood ratios of 2-5 and 0.5-0.2 generate small (but sometimes important) changes in probability; and
4. Likelihood ratios of 1-2 and 0.5-1 alter probability to a small (and rarely important) degree.

Results

Of the total of 1450 patients screened, 130 (102 women and 28 men) were found eligible for the study (Figure 1). The mean age of the study population was 44 years (standard deviation (SD) 13; range 14-75). Two patients (1.53%) were aged <20 years, 49 (37.6%) were aged 20–39 years, 59 (45.38%) were aged 40–59 years, and 20 (15.3%) were aged 60 years or above. The mean TSH in the entire study population was 15.9 (SD 27.8; range 0.06-110.5)]). Twenty-three patients (18%) were detected to have hypothyroidism by the thyroid hormone assays. Of the 23 patients (mean age 46, SD 15, range 14-70), 20 (87%) were women. On an average, the hypothyroid subjects were no older than those who were euthyroid. The mean TSH among the hypothyroid patients was 61.4 (SD 33.0; range 7.7-110.5). This prevalence of 18% was our best estimate of the pre-test probability of hypothyroidism in our patient population. Table 1 summarises the diagnostic accuracy of physical signs associated with hypothyroidism. None of the signs, when considered in isolation, had likelihood ratios that would result in conclusive shifts in post-test probabilities. No single finding, when absent, provided sufficient evidence against the diagnosis of hypothyroidism (negative likelihood ratios ranging from 0.42 to 1.0).

In patients with suspected hypothyroidism, the findings most likely to detect hypothyroidism were bradycardia (LR+ 3.88), abnormal ankle reflex (LR+ 3.41), and coarse skin (LR+ 2.3).

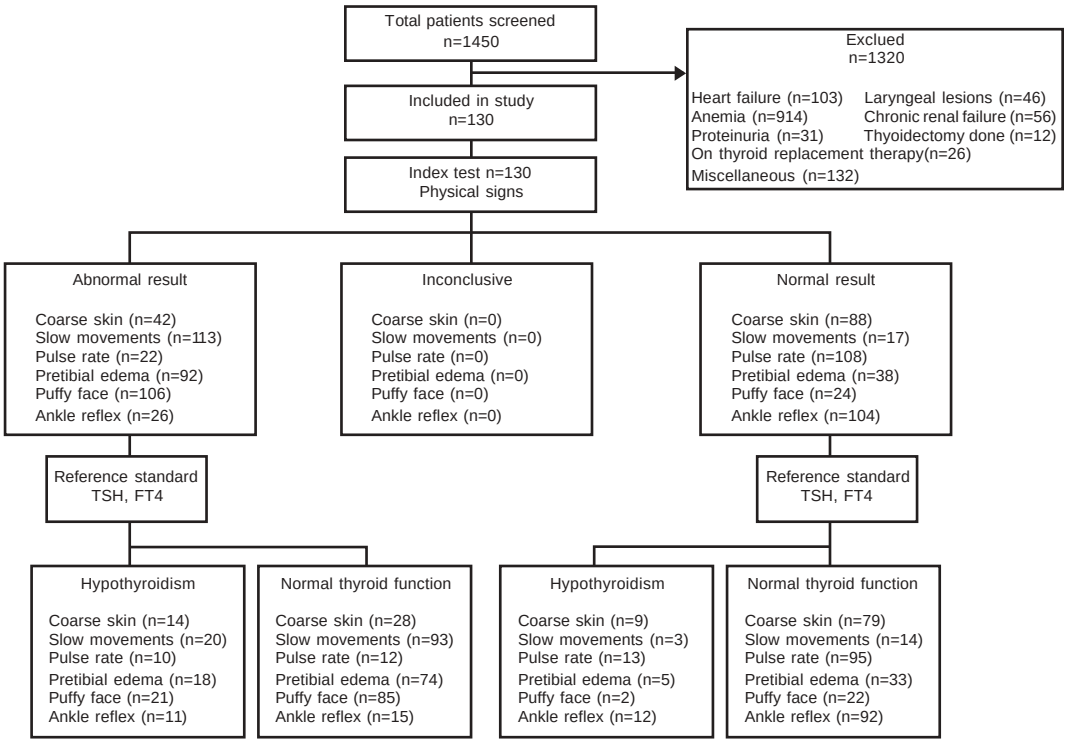

Figure 1: The study design

In a *post hoc* analysis, we evaluated the accuracy of the combination of these three signs. The LR+ was 3.75 and LR– 0.48 (Table 1). These results indicate modest accuracy for this combination of signs.

Discussion

Although several studies[4-11] have assessed the accuracy of clinical variables for the diagnosis of hypothyroidism, most studies have had methodological limitations. A study of the diagnostic properties of the clinical examination for thyroid disease should prospectively recruit consecutive subjects presenting with clinical features suggestive of hypothyroidism, and it should evaluate the clinical features blindly and independently with the reference standard of diagnosis. The lack of blinding may cause a clinician to over-interpret physical signs that he or she expects to see, and would also induce a bias in the interpretation of clinical features.[14] Some studies examining the diagnostic accuracy of clinical features for diagnosing hypothyroidism retrospectively reviewed the medical records and depended on the primary care physicians' records of the history and physical examination.[4-7] By enrolling mostly elderly individuals[9-11] very few men[5,9,10] or exclusively women,[8] and patients attending endocrine clinics,[4-6,8] the studies introduced a spectrum of bias in their designs. The results of these studies may not be applicable to a general population.

Our study design had some methodological advantages. We used the cross-sectional design in our study and made an independent, blind comparison between physical examination findings and the hormone assay. We also avoided verification (workup) bias in our study by ensuring that all eligible patients, irrespective of their physical examination findings, were tested for hormone levels. We chose FT4 and TSH levels, the most appropriate reference standard for the study.

Attia *et al* argue that when researchers examine a large number of signs and symptoms in a relatively small population, chance

Table 1: Accuracy of physical examination findings in the diagnosis of hypothyroidism

Sign	Sensitivity (95% CI)*	Specificity (95% CI)	LR+† (95% CI)	LR–‡ (95% CI)	PPV	NPV
Coarse skin	60.9 (38.5, 80.3)	73.8 (64.4, 81.9)	2.33 (1.47, 3.67)	0.53 (0.34, 0.84)	33.79	89.58
Slow movements	87 (66.4, 97.2)	13.1 (7.3, 21)	1 (0.84, 1.19)	1 (0.84, 1.19)9*	18.02	82.11
Bradycardia	43.5 (23.2, 65.5)	88.8 (81.2, 94.1)	3.88 (1.91, 7.87)	0.64 (0.31, 1.29)	46.02	87.74
Pretibial oedema	78.3 (56.3, 92.5)	30.8 (22.3, 40.5)	1.13 (0.88, 1.45)	0.7 (0.55, 0.9)	19.90	86.61
Puffiness of the face	91.3 (72, 98.9)	20.6 (13.4, 29.5)	1.15 (0.98, 1.35)	0.42 (0.36, 0.5)	20.15	91.52
Delayed ankle reflex	47.8 (26.8, 69.4)	86 (77.9, 91.9)	3.41 (1.81, 6.43)	0.61 (0.32, 1.14)	42.84	88.24
Coarse skin, bradycardia, and ankle reflex	60.0 (18.24, 92.65)	84 (76.78, 89.66)	3.75 (1.65, 8.52)	0.48 (0.16, 1.40)	45.15	90.54

*95% CI: 95% Confidence Interval, †LR+: likelihood ratio of a positive test, ‡LR –: likelihood ratio of a negative test, PPV: positive predictive value, NPV: negative predictive value

alone may influence the study results.[14] Studies that depend on physical signs and symptoms elicited before diagnostic test results generate lower likelihood ratios—in the range of 2 to 3. Our results agree with these observations. Only coarse skin (LR+ 2.3), bradycardia (LR+ 3.88) and abnormal ankle reflex (LR+ 3.41) were predictive of hypothyroidism in our study, and even these three features had small likelihood ratios. No symptom or sign definitively ruled out the disease (LR- range from 0.42 to 1.0). A study that evaluated 16 symptoms for the diagnosis of hypothyroidism found that only three current symptoms [hoarse voice (LR 4.2), dry skin (LR 1.3), and muscle cramps (LR 2.2)] differed between case and control subjects.[15] Another study has shown that in patients with suspected thyroid disease, the findings arguing the most for hypothyroidism were coarse skin (LR+ 5.6), hypothyroid speech (LR+ 5.4), cool and dry skin (LR+ 4.7), bradycardia (LR+ 4.1), and pretibial oedema (LR+ 2.8).[16] In a retrospective review of 982 patient charts, Schectman *et al* found a poor correlation between clinical features and thyroid disease.[7] The authors collected data from the primary physicians' records, and whether or not the physicians specifically sought the clinical features in their patients is unclear.

Rather than evaluating individual signs and the symptoms, investigators have evaluated the accuracy of combinations of signs and symptoms of thyroid disease.[4,17] In a retrospective chart review of 500 patients seen in a thyroid clinic, the presence of more than five symptoms and signs significantly predicted thyroid disease (LR+ 18.6), while the lack of signs and symptoms (<2 signs or symptoms) argued against it (LR = 0.11).[4] The prevalence of thyroid disease was 4% in the study but the reference standard to diagnose thyroid disease has not been clearly defined. Drake *et al* in a review of 135 family practice charts found that when patients lacked symptoms and signs, they were unlikely to have thyroid disease (LR = 0.11).[17] However, it is not clear what proportion of the patients with thyroid disease had hypothyroidism in this study.

Our *post hoc* analysis of the combination of signs indicated only modest accuracy for the combination of coarse skin, bradycardia and delayed ankle reflex. It is unlikely that even this combination can make a meaningful difference in the post-test probabilities. However, these signs could be useful in identifying those patients who might benefit from thyroid function tests.

Our study had limitations. Firstly, the precision of some of our estimates indicates that our sample size was not large. Secondly, most signs and symptoms are subjective and open to measurement error (intraobserver variability). Unfortunately, we did not systematically collect data on the reproducibility of the signs evaluated. A clinical examination done by an experienced resident may be even less reliable than an evaluation by a more skilled and experienced attending physician. The physical signs were however confirmed by a senior consultant in our study. Similarly, slowness of movements and delayed relaxation of ankle reflex posed problems for consistent interpretation. However, since the resident and the consultant who evaluated the patients had no access to the laboratory data at the time of the history and physical examination, it is likely that the measurement error was not correlated with the disease status (random misclassification), which is known to affect the accuracy of a diagnostic test.[18] Another limitation pertains to external validity. Since our participants were predominantly rural Indian women, our results may have limited generalisability. Lastly, since the study patients were pre-screened, our method of patient recruitment might have led to a higher prevalence of hypothyroidism.

Conclusion

In conclusion, our study suggests that physical signs when considered in isolation have poor diagnostic accuracy for hypothyroidism. Even combinations of signs do not appear to have high accuracy. Important treatment decisions, therefore, cannot be made purely on the basis of physical findings. However, since selected signs (such as coarse skin, bradycardia and delayed ankle reflex) are associated with modest accuracy, clinicians could use physical examination to generate and revise their estimates of pre-test probabilities and use the information to select those patients who will benefit most from thyroid hormone assays. This strategy is likely to maximize the number of patients in whom clear diagnostic decisions can be made.

Acknowledgments

MP receives training support from the Fogarty AIDS International Training Program, University of California, Berkeley, USA.

References

1. Cooper DS. Clinical practice. Subclinical hypothyroidism. N Engl J Med 2001; 345:260-5.
2. Larsen PR, Ingbar SH. The thyroid gland. In: William's Textbook of Endocrinology. *In:* Wilson JD, Foster DW, ed. Philadelphia: WB Saunders Company; 1992. 357-487.
3. US Preventive services task force. Screening for thyroid disease. *In:* Guide to Clinical Preventive Services. Baltimore: Williams and Wilkins. 1996. p. 209-18.
4. White GH, Walmsley RN. Can the initial clinical assessment of thyroid function be improved? Lancet 1978;ii:933-5.
5. Watanakunakorn C, Hodges RE, Evans TC. Myxedema. Arch Intern Med 1965; 116:183-90.
6. Gardner MJ, Barker DJ. Diagnosis of hypothyroidism: a comparison of statistical techniques. BMJ 1975;2:260-2.
7. Schectman JM, Kallenberg GA, Shumacher RJ, Hirsch RP. Yield of hypothyroidism in symptomatic primary care patients. Arch Intern Med 1989;149:861-4.
8. Zulewski H, Muller B, Exer P, Miserez AR, Staub JJ. Estimation of tissue hypothyroidism by a new clinical score: evaluation of patients with various grades of hypothyroidism and controls. J Clin Endocrinol Metab 1997;82:771-6.
9. Bagchi N, Brown TR, Parish RF. Thyroid dysfunction in adults over age 55 years. A study in an urban US community. Arch Intern Med 1990;150:785-7.
10. Bahemuka M, Hodkinson HM. Screening for hypothyroidism in elderly inpatients. BMJ 1975;2:601-3.
11. Sawin CT, Castelli WP, Hershman JM, McNamara P, Bacharach P. The aging thyroid: Thyroid deficiency in the Framingham Study. Arch Intern Med 1985;145:1386-8.
12. Dayan CM. Interpretation of thyroid function tests. Lancet 2001;357:619-24.
13. Jaeschke R, Guyatt G, Lijmer J. Diagnostic tests. *In:* Users' guides to the medical literature. A manual for evidence-based clinical practice Edited by Guyatt G, Rennie D. Chicago: AMA Press; 2002. p. 121-40.
14. Attia J, Margetts P, Guyatt G. Diagnosis of thyroid disease in hospitalized patients: a systematic review. Arch Intern Med 1999;159:658-65.
15. Canaris GJ, Steiner JF, Ridgway EC. Do traditional symptoms of hypothyroidism correlate with biochemical disease? J Gen Intern Med 1997;12:544-50.
16. Barker DJ, Bishop JM. Computer-based screening system for patients at risk of hypothyroidism. Lancet 1969;ii:835-8.
17. Drake JR, Miller DK, Evans RG. Cost-effectiveness of thyroid function tests. Arch Intern Med 1982;142:1810-2.
18. Walter SD, Irwig L, Glasziou PP. Meta-analysis of diagnostic tests with imperfect reference standards. J Clin Epidemiol 1999;52:943-51.