

Interobserver variation in the assessment of thyroid enlargement: A pitfall in surveys of the prevalence of endemic goitre

René Tonglet, Pierre Bourdoux, Michèle Dramaix, Philippe Hennart, and André-Marie Ermans

Abstract

An important concern in surveys of the prevalence of endemic goitre is the possibility of inaccuracy that can occur when estimating thyroid enlargement by clinical examination. We report here the results of a study from central Africa that demonstrated the importance of interobserver variation in the assessment of goitre size among experts and participants attending an international training course on iodine deficiency disorders (IDD). Agreement among raters estimated by calculation of the kappa statistic was poor on average when the five-stage goitre classification recommended by WHO was used. As expected, agreement on a single category was slightly better, especially when the category of interest was visible goitre. Our results suggest that palpation alone is not a reliable method for assessing the severity of IDD endemia. Other sources of inaccuracies are discussed and recommendations are made for better data collection on IDD in Africa and other regions.

Introduction

Iodine deficiency disorders (IDD) have recently been recognized as a potential target for eradication [1]. Current estimates by WHO indicate that about one billion people worldwide are at risk for iodine deficiency [1], but the severity of endemia differs strikingly from place to place [2]. Reliable data collection is therefore mandatory to monitor the planning and implementation of IDD control programmes at the regional level [3]. Regional health staffs should be able to organize their own surveys in a representative study population, providing data on the prevalence of goitre (clinical examination), the level of iodine intake (urinary iodine concentration), and, optimally, the pattern of thyroid function (serum thyroxin and thyrotropin levels) [4]. However, in many developing countries, especially in the WHO African region, because of the scarcity of biological data, assessment of the severity of IDD and of the urgency of their correction relies only on clinical observations [5]. Therefore, it is important in surveys of the prevalence of endemic goitre to assess the level of inaccuracy that can

occur when estimating thyroid enlargement in the field. Among the many sources of inaccuracies that can be detected, interobserver variation is extremely important.

Methods

In January 1992 the Centre Scientifique et Médical de l'Université Libre de Bruxelles pour ses Activités de Coopération convened the third International Training Course on the Prevention of Endemic Goitre in Bujumbura, Burundi, with the sponsorship of the International Council for Control of Iodine Deficiency Disorders (ICCIDD). This three-week training course gathered a panel of international experts and 30 participants from 15 French-speaking African countries. All the participants were involved in IDD control programmes at the national or regional level in their countries, and most of them were experienced.

During this training course, after lectures on the quantitative assessment of IDD, we organized a one-day workshop in the severely iodine-deficient province of Rutana, 120 km south-east from the capital, Bujumbura, where the national IDD survey in 1989 revealed a prevalence of goitre close to 60%, a median urinary iodine concentration below 20 µg/L, many cases of biological hypothyroidism, and occasional cases of overt cretinism [6]. About 400 out of the 1,600 inhabitants of Samahuge Hill met the experts and participants at the invitation of the local health authorities. The subjects and participants were randomly allocated to six groups, each headed by one of the experts.

Each group cautiously started with repeated examinations of a number of patients to standardize the clinical method of evaluation of goitre size. Goitre was defined as "a thyroid gland whose lateral lobes have a volume greater than the terminal phalanges of the thumbs of the person examined" [7]. Thyroid size was estimated according to the classification endorsed by WHO, PAHO, and ICCIDD: 0 = no goitre, 1a = palpable but not visible goitre, 1b = palpable and visible only when the neck is fully extended, 2 = goitre visible with the neck in normal position, 3 = very large goitre visible at

a distance of five meters [8]. When agreement had been achieved among the participants and their referee, 25 consecutive patients were independently examined by each member of the group and individual diagnoses were confidentially recorded on a special form.

We collected suitable data from 94 subjects, divided into four groups, with 25, 24, 21, and 24 subjects respectively. Each group was examined by a team including one referee (4 x 1) and a number of participants (6, 4, 5, and 3 respectively). Data from two additional groups had to be discarded because of confusion among the subjects.

The agreement between pairs of observers (each participant and the referee) was estimated by calculation of the kappa statistic as described by Fleiss [9]. The equation for this calculation is:

$$k = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the observed proportion of agreement and p_e is the proportion of agreement expected by chance. Kappa can vary between - 1 and + 1. As stated by Fleiss [9], "Values greater than 0.75 or so may be taken to represent excellent agreement beyond chance' values below 0.40 or so may be taken to represent poor agreement beyond chance, and values between 0.40 and 0.75 may be taken to represent fair to good agreement beyond chance." The standard error of kappa was also calculated according to Fleiss [9].

Kappas were computed on 5 x 5 matrices to determine agreement across all categories of the WHO goitre classification for each pair of observers. Paired kappas were also computed to determine agreement on a single category after collapsing each original matrix in 2 x 2 tables including either stage 0 versus stages 1 a -3 (no goitre versus goitre), or stages 0-1b versus stages 2 and 3 (no visible goitre versus visible goitre). In addition, averaged group kappas were computed for each group to determine agreement on each of the three categorical scales of interest. Unfortunately, the present study did not allow us to determine agreement among the referees.

Results

As an example, table 1 displays the overall matrix of the individual categorical assessments made by the participants and their referee in one of the four groups. This illustrates the wide range of variation we observed between raters using the WHO goitre classification. In this example, complete consensus was obtained for only 1 patient in 25.

Table 2 shows that the paired kappas computed to determine agreement on the WHO classification varied from 0.11 to 0.54 (mean 0.38, SD 0.11). This may be interpreted to represent poor to fair agreement beyond chance.

Agreement on a single category was slightly better on average when the category of interest was the absence/presence of goitre (mean 0.48, SD 0.20, range 0.17-0.89) (table 3) or the absence/presence of a visible goitre (mean 0.50, SD 0.19, range 0.37-1.00) (table 4).

Table 5 allows comparison between the averaged group kappas computed for each group to determine agreement on the different goitre categorizations. Agreement among all observers on the original WHO classification was poor, whereas agreement on a single category was fair to good.

Discussion

According to Kleinbaum et al. [10], inaccuracy in estimating population characteristics may result from either random error, a precision problem essentially attributable to sampling variation, or systematic error, a validity problem attributable to methodological aspects of the study, including the quality of the information obtained.

The validity of the clinical method of estimating goitre size has already been challenged by the development of thyroid ultrasonography, which is now an accepted reference test or gold standard for the diagnosis of thyroid enlargement [11, 12]. When comparing diagnoses made either by sonographic volumetry or by clinical examination, ranges of thyroid volume determined sonographically overlap widely in all age groups with volume estimated by palpation [13]. Accordingly, several studies have demonstrated that average errors rose up to 30% when palpation was compared with ultrasonography [14,16]. From observations made in a large sample of Swedes and Germans, Gutekunst et al. calculated that the sensitivity and specificity of palpation in adults were 91% and 63.5% respectively [13]. Consequently, if the true but unknown goitre prevalence in the study population were, say, 10%, the prevalence computed after clinical examination would rise to 42%. This figure would be 64% if the true but unknown prevalence were 50%. Palpation could be even less reliable in younger subjects [13].

Regardless of these characteristics of the clinical method, we must avoid overlooking the possibility of another undetected flaw in assessing goitre size. Our results underline the poor reliability of the clinical method attributable to wide interobserver variation. Agreement among raters is poor, on average, when the

WHO classification is used. The apparent precision gained from the use of this five-stage scale could be spurious even when examinations are made by high-level managers of IDD programmes such as the participants in the 1992 training course. As was to be expected, the agreement among the raters was better when goitre size was limited to a single category, especially when the single category was visible goitre. Visible goitre—a nosologic entity depicted without ambiguity—is more easily identified than a normal or slightly enlarged thyroid, and accuracy in detecting it could probably be easily reinforced by sustained training.

It should be borne in mind that goitre prevalence is critically influenced by age and sex, with a pattern fairly typical of the local endemia [2]. Therefore, additional information would be necessary to evaluate the prevalence dependency of the kappa statistic in epidemiological surveys of endemic goitre. The kappa coefficient can be shown to depend not only on the sensitivity and the specificity of the diagnosis test but also on the true prevalence of the characteristic in the population studied [16]. The kappa value is lowered when the true prevalence approaches either 0 or 100%, and rises to a higher level when the prevalence is in the medium range [17]. A low kappa value combined with a high prevalence of one of the categories could indicate that the classification system cannot be used properly in that particular situation [18, 19]. Consequently, when the prevalence of small goitres (stages 1a-1b in schoolchildren is as high as it is in most surveys from the African region [5], one may question whether the current commitment made to organize prevalence surveys among schoolchildren is sound [8].

Since there are so many sources of inaccuracies in goitre prevalence surveys, we recommend that more efforts should be made to collect unequivocal biological data in a representative sample of the population. One hundred or so urine specimens are probably more valuable to demonstrate the severity of IDD endemia than the accumulation of questionable data on clinical goitre [20]. On top of this, determination of the prevalence of visible goitre in young adults, including women of childbearing age, and case notification followed by biological confirmation of overt cretinism in young subjects could yield positive conclusions which would be invaluable for the assessment of endemia. However, if such a rapid assessment demonstrates the existence of IDD, additional data should be collected in the different age and sex groups of the whole population. These would be particularly necessary for the objective follow-up of interventions.

Acknowledgements

The third International Course on the Prevention of Endemic Goitre was supported by the government of Burundi; the Administration Générale de la Coopération au Développement (AGCD), Belgium; the Centre Scientifique et Médical de l'Université de Bruxelles pour ses Activités de Coopération (CEMUBAC), Belgium; the International Council for Control of Iodine Deficiency Disorders (ICCIDD); the Fonds International de Coopération Universitaire (FICU); several national offices of the United Nations Children's Fund (UNICEF) in Africa; and the Regional Office for Africa of the World Health Organization (WHO/AFRO), Congo.

This study was supported by a grant from the Van Buren Foundation to René Tonglet. Special thanks are expressed to Dr. Christine Goset, former national nutrition adviser in Bujumbura, Burundi.

References

1. Dunn JT. Iodine deficiency—the next target for elimination? *N Engl J Med* 1992;326:267-68.
2. Ermans AM. Disorders of iodine deficiency. In: Ingbar SH, Braverman LE, eds. *Werner's The thyroid*. Philadelphia, Pa, USA: Lippincott, 1986:705-19.
3. Dunn IT, van der Haar F. A practical guide to the correction of iodine deficiency. Wageningen, Netherlands: International Council for Control of Iodine Deficiency Disorders, 1990.
4. Bourdoux P, Ermans AM. Quantitative assessment of iodine deficiency: a proposed classification. In: Asian and Oceanian Thyroid Association Symposium on Iodine Deficiency Disorders, April 24-25, 1989, Tianjin. Tianjin, China: Tianjin Medical College and Tianjin Institute of Endocrinology, 1989:45 (abstract).
5. Lemaire B. Databank on endemic goiter for the WHO African region. Brazzaville, Congo: World Health Organization, Regional Office for Africa, 1991.
6. Goset C. Enquête sur les troubles dus à la carence en iode au Burundi. Projet de lutte contre les maladies transmissibles et carencielles. Bujumbura, Burundi: Ministère de la Santé Publique, 1991.
7. Perez C, Scrimshaw NS, Muñoz JA. Technique to endemic goiter surveys. In: *Endemic goiter*. WHO Monograph Series, no. 44. Geneva: WHO, 1960:369-83.
8. Delange F, Bastani S, BenMiloud M. Definitions of endemic goiter and cretinism, classification of goiter size and severity of endemias, and survey techniques. In: Dunn JT, Pretell EA, Daza CH, Viteri FE, eds. *Towards the eradication of endemic goiter, cretinism, and iodine deficiency*. Washington, DC: Pan American Health Organization, 1986:373-76.
9. Fleiss JL. Statistical methods for rates and proportions. New York: John Wiley & Sons, 1981:212-36.

10. Kleinbaum DO, Kupper LL, Morgenstern H. Epidemiologic research: principles and quantitative methods. New York: Van Nostrand Reinhold, 1982:185-89.
11. Hegedüs L, Perrild H, Poulsen LR et al. The determination of thyroid volume by ultrasound and its relationship to body weight, age, and sex in normal subjects. *J Clin Endocrinol Metab* 1983;56(2):260-63.
12. Hegedüs L. Thyroid size determined by ultrasound: influence of physiological factors and non-thyroidal disease. *Dan Med Bull* 1990;37(3):249-63.
13. Gutekunst R, Smolarek H, Hasenpusch U et al. Goiter epidemiology: thyroid volume, iodine excretion, thyroglobulin and thyrotropin in Germany and Sweden. *Acta Endocrinologica (Copenhagen)* 1986;112: 494-501.
14. Tannahill AJ, Hooper MJ, England M, Ferriss JB, Wilson GM. Measurement of thyroid size by ultrasound, palpation and scintiscan. *Clin Endocrinol* 1978;8:483-86.
15. Igl W, Lukas P, Leisner F et al. Sonographische Volumenbestimmung der Schilddrüse: Vergleich mit anderen Methoden. *Nuklearmedizin* 1981;20:64-71.
16. Gutekunst R. The value and application of ultrasonography in goiter survey. *IDD Newsletter* 1990;6(4):3-5.
17. Thompson WD, Walter SD. A reappraisal of the kappa coefficient. *J Clin Epidemiol* 1988;41(10):949-58.
18. Gjørup T. The kappa coefficient and the prevalence of a diagnosis. *Methods Inform Med* 1988;27:184-86.
19. Donker DK. Interobserver variation in the assessment of fetal heart recordings. Amsterdam: VU University Press, 1991.
20. Bourdoux P. Measurement of iodine in the assessment of iodine deficiency. *IDD Newsletter*. Feb. 1988:8-12