CIRUGÍA ESPAÑOLA

www.elsevier.es/cirugia

CIRUGÍA ESPAÑOLA

Original article

Efficiency of the Bethesda System for Thyroid Cytopathology☆

Ismael Mora-Guzmán,[a] José Luis Muñoz de Nova,[a,b,*] Cristina Marín-Campos,[a] José Antonio Jiménez-Heffernan,[c] Juan Julián Cuesta Pérez,[d] Marcos Lahera Vargas,[e] Emma Torres Mínguez,[a] Elena Martín-Pérez[a,b]

[a] Servicio de Cirugía General y del Aparato Digestivo, Hospital Universitario de La Princesa, Madrid, Spain
[b] Departamento de Cirugía, Universidad Autónoma de Madrid, Madrid, Spain
[c] Servicio de Anatomía Patológica, Hospital Universitario de La Princesa, Madrid, Spain
[d] Servicio de Radiodiagnóstico, Hospital Universitario de La Princesa, Madrid, Spain
[e] Servicio de Endocrinología, Hospital Universitario de La Princesa, Madrid, Spain

ARTICLE INFO

Article history:
Received 29 December 2017
Accepted 19 February 2018
Available online 29 June 2018

Keywords:
Thyroid cancer
Bethesda
Fine-needle aspiration
Cytology

ABSTRACT

Introduction: Fine-needle aspiration biopsies are a key tool for preoperative assessment of thyroid nodules, and the Bethesda system is the preferred method to report cytological analysis. The purpose of this study is to assess the efficiency of the Bethesda system to identify the malignancy risk of thyroid nodules.

Methods: Patients who underwent thyroid surgery between June 2010 and June 2017 were included. Samples were classified into six categories according to rates of malignancy associated with each diagnostic category. In order to investigate the correlation between categories, a statistical analysis compared the categories with pathology reports. Diagnostic indicators were calculated as a screening test (categories IV, V, VI as true-positive) and as a method to identify malignancy (V, VI as true-positive).

Results: In a series of 522 patients, we found 184 (35.2%) malignant tumors, papillary carcinoma being the most prevalent with 155 cases (84.2%). Malignant rates for diagnostic categories were: I, 0%; II, 1.5%; III, 6.4%; IV, 31%; V, 86.5%; VI, 100%. A robust correlation was identified between categories on statistical analysis. For the ‘‘screening test’’ analysis, sensitivity was 98.9%, specificity 84.4%, positive predictive value 69.6%, negative predictive value 99.5%, and diagnostic accuracy 88.2%. Analysing the accuracy to detect malignancy, values were: sensitivity 98.6%, specificity 97.6%, positive predictive value 93.5%, negative predictive value 99.5%, diagnostic accuracy 97.9%.

Conclusion: The Bethesda system is a clear and reliable approach to report thyroid cytology and therefore is an effective tool to identify malignancy risk and guide clinical management.

☆ Please cite this article as: Mora-Guzmán I, Muñoz de Nova JL, Marín-Campos C, Jiménez-Heffernan JA, Cuesta Pérez JJ, Lahera Vargas M, et al. Rendimiento del sistema Bethesda en el diagnóstico citopatológico del nódulo tiroideo. Cir Esp. 2018;96:363–368.

[*] *Corresponding author.*
E-mail address: jmunoz@salud.madrid.org (J.L. Muñoz de Nova).

2173-5077/

Rendimiento del sistema Bethesda en el diagnóstico citopatológico del nódulo tiroideo

Palabras clave:
Cáncer de tiroides
Bethesda
Punción-aspiración con aguja fina
Citología

RESUMEN

Introducción: La punción-aspiración con aguja fina es una pieza clave en la evaluación preoperatoria del nódulo tiroideo y el sistema Bethesda es el más aceptado para categorizar el análisis citológico. El objetivo del estudio es evaluar la validez del sistema Bethesda en la enfermedad nodular tiroidea para diagnosticar malignidad.

Métodos: Se incluye a los pacientes intervenidos de tiroides consecutivamente entre junio de 2010 y junio de 2017. Se realizó el análisis de la punción preoperatoria según el sistema Bethesda, correlacionando este dato con la histología definitiva para cada nódulo biopsiado. Los parámetros de prueba diagnóstica se calcularon como prueba de *screening* (verdadero positivo: categorías IV, V, VI) y como método para identificar malignidad (verdadero positivo: categorías V, VI).

Resultados: Se incluyó a 522 pacientes, de los que 184 (35,2%) presentaron un carcinoma en la histología definitiva; siendo el carcinoma papilar el más frecuente (84,2%). Los porcentajes de malignidad en el nódulo biopsiado para cada categoría Bethesda fueron: I, 0%; II, 1,5%; III, 6,4%; IV, 31%; V, 86,5% y VI, 100%. En el análisis como prueba de *screening*, se identificó una sensibilidad del 98,9%, especificidad del 84,4%, valor predictivo positivo del 69,6%, valor predictivo negativo del 99,5% y precisión diagnóstica global del 88,2%. En el análisis para detectar malignidad, los parámetros fueron: sensibilidad 98,6%, especificidad 97,6%, valor predictivo positivo 93,5%, valor predictivo negativo 99,5% y precisión diagnóstica global 97,9%.

Conclusiones: El sistema Bethesda es un método sencillo y reproducible en la categorización citológica del nódulo tiroideo, una herramienta útil en el manejo y eficaz para identificar el riesgo de malignidad.

Introduction

Nodular thyroid disease is very common in the population, especially in women and seniors, affecting up to 60%.[1,2] In recent decades, there has been a significant increase in the incidence of thyroid cancer, especially at the expense of microcarcinomas.[3,4] Thus, the annual incidence of thyroid cancer in the United States has tripled, with 40% for microcarcinomas.[5] One of the factors that has been associated with this fact is the constant increase in the use of cervical imaging tests, mainly thyroid ultrasound, which resulted in the identification of an increasing number of thyroid nodules (TN) susceptible to being biopsied using fine needle aspiration (FNA).[6] In order to standardize the terminology used for the description of thyroid cytology, in 2007 consensus recommendations were issued, known as the Bethesda system (BS).[7] This system was based on the creation of six categories associated with a certain risk of malignancy in each. Subsequent studies have described the results after adopting BS recommendations, with high agreement in the categorization of FNA samples.[8] However, there is a limitation inherent to the BS, which is the intra- and interobserver variability in the cytopathology study of TN.[9] Furthermore, the most frequent context in the general population is the presentation of multinodular goiter: a single patient may present several nodules that are able to be biopsied by FNA due to their ultrasound characteristics, which increases the complexity of the diagnostic-therapeutic process. The aim of the study is to evaluate the validity of BS in nodular thyroid disease to diagnose malignancy.

Methods

The study population consisted of consecutive patients who had undergone thyroid surgery between June 1, 2010 and June 30, 2017. Only those patients treated with a first thyroid intervention whose preoperative FNA had been performed at our hospital were included. Excluded from the study were patients with FNA performed at other centers, patients who underwent surgery without preoperative FNA or those with FNA not in accordance with the BS (Fig. 1). For all patients, the diagnostic-therapeutic protocol included anamnesis, physical examination, thyroid function work-up and a cervical ultrasound. In patients who met the criteria for FNA in accordance with the international guidelines of the American Thyroid Association,[10,11] this procedure was performed under ultrasound guidance by a radiologist assisted by an expert cytologist.

The samples were classified following the BS[7] recommendations, grouped into the six originally described categories: (I) unsatisfactory/non-diagnostic; (II) benign; (III) atypia of uncertain significance/follicular lesion of uncertain significance; (IV)

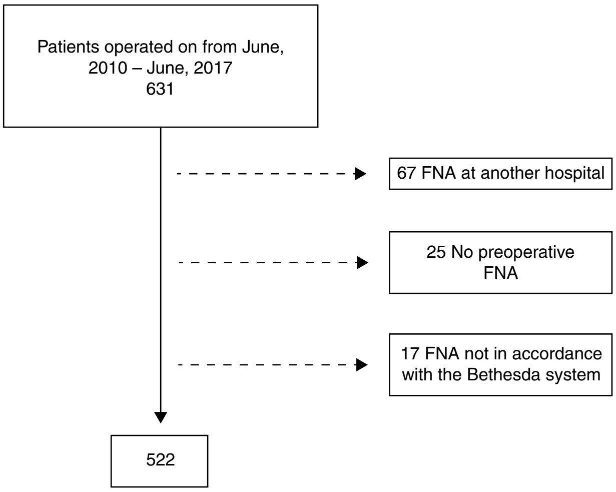
Patients operated on from June, 2010 – June, 2017
631
67 FNA at another hospital
25 No preoperative FNA
17 FNA not in accordance with the Bethesda system
522

Fig. 1 – Diagram of patients included in the study.

follicular neoplasm/suspicion of follicular neoplasm; (V) suspected malignancy and (VI) malignant. The indication to repeat a needle-aspiration was limited to those cases with diagnostic categories I and III, and to benign punctures, but with a high degree of clinical-radiological suspicion. Once the needle-aspiration was done, surgery was indicated: in patients with categories IV, V and VI; in patients with persistent category I after repeated aspiration who presented a high degree of clinical-radiological suspicion; in patients with persistent category III after repeating needle-aspiration or after the initial aspiration if there was a high degree of suspicion; and in patients with category II, but who presented symptoms attributable to thyroid nodularity, hyperfunction, progressive growth of the TN or if any of the TN was >4 cm. The surgical technique used in each case was based on the individual characteristics of the patient, the BS categories and the location of the TN. In general, hemithyroidectomy was performed in the presence of unilateral nodules or millimetric contralateral nodules in categories I–IV. In the presence of symptomatic bilateral multinodular goiter, Graves' disease or categories V–VI, total thyroidectomy was selected.

The patient follow-up data and the final histological correlation were only available in patients with surgical management. If a patient had several FNA samples from different TN, the results of each aspiration and the corresponding histological results were analyzed separately. A thorough review of each evaluated TN was performed, carefully correlating the description of the ultrasound that guided the FNA (size and location) with the findings of the surgical piece to confirm the agreement between the biopsied TN with its respective definitive pathology diagnosis. Regarding the study design, a prospectively maintained database was analyzed that collected the diagnostic-therapeutic data of all the patients, particularly demographic data, size and ultrasound localization of the TN, BS diagnostic category (in cases of multiple aspirations in the same patient, only the highest BS risk category was included), operative data and pathology data. This study was approved by the Clinical Research Ethics Committee at our hospital.

Statistical Analysis

The statistical analysis was carried out using the SPSS® 23.0 program for Windows (SPSS Inc., Chicago, Illinois, USA). The results were expressed as percentages for categorical variables, and as mean and standard deviation for continuous variables, using the median and interquartile range for variables with asymmetrical distribution.

The correlation between the different diagnostic categories was assessed by comparing them with the final histological result, for which a linear logarithmic model (likelihood ratio) and a chi-squared model were applied, using symmetrical measures of association. The malignancy data used were calculated by assigning to each biopsied nodule its corresponding final histological diagnosis. Statistically significant differences were considered bilaterally with P values $<.05$. The phi correlation was used as a measure of the degree of association between categorical variables, whose values oscillate between +1 and −1. According to the strength of association: −1 indicates a strong negative association, +1 indicates a strong positive association and 0 indicates absence of association.

The diagnostic test parameters calculated were sensitivity, specificity, predictive values (positive predictive value [PPV], negative predictive value [NPV]) and diagnostic accuracy to detect malignancy by means of two analyses. In the analysis as a screening test (analysis I), the FNA results were considered an indication for surgery for suspected malignancy (BS II vs IV, V, VI categories). According to this analysis, the terms "positive" or "negative" constituted the existence or not of surgical indication for the statistical analysis. Categories I and III were excluded from this analysis because they may involve the repetition of FNA. A second analysis was performed, which measured the ability of the test to detect malignancy (analysis II) in the case of highly suspicious aspirations (categories V and VI) compared to patients with benign puncture (category II).

Results

In the study period, 631 patients were treated. Excluded from the study were 67 patients with FNA performed at another hospital, 25 patients without preoperative FNA, and 17 patients in whom the FNA report was not done in accordance with the BS (Fig. 1). Thus, out of the 522 patients included, 433 (83%) were women, with a mean age of 51.8±16 years. The median TN size evaluated preoperatively was 2.5 cm (1.6–4). The most frequent cytology among the operated patients was category II (49%), with very similar percentages of patients operated on with categories III, IV, V and VI (14.9, 13.6, 7.1 and 11.5%, respectively). In 316 cases (60.5%), total thyroidectomy was performed; the remainder had hemithyroidectomies with isthmectomies (39.5%). Dissection of the central compartment was associated in 66 cases (12.6%). Regarding histological results, 184 malignancies (35.2%) were identified; papillary carcinoma was the most frequent tumor with 155 cases (84.2%), 42 of which were incidental microcarcinomas (27.1% of the total papillary carcinomas and 8% of the total number of patients treated surgically). The remaining

Table 1 – Percentage of Malignancy in the Diagnostic Categories.

Diagnostic category	No. of cases (%)	Risk of malignancy, No. of cases (%)	Risk of malignancy in TN excluding miCPin, No. of cases (%)	Risk of malignancy in TN evaluated, No. of cases (%)
I	17 (3.3)	6 (35.3)	6 (35.3)	0 (0)
II	259 (49.6)	37 (14.3)	12 (4.6)	4 (1.5)
III	78 (14.9)	18 (23.1)	9 (11.5)	5 (6.4)
IV	71 (13.6)	30 (42.3)	24 (33.8)	22 (31)
V	37 (7.1)	33 (89.2)	32 (86.5)	32 (86.5)
VI	60 (11.5)	60 (100)	60 (100)	60 (100)

miCPin: incidental papillary microcarcinoma (*n*=42).

neoplasms identified were 19 follicular carcinomas (10.3%), right medullary carcinomas (4.3%), an anaplastic carcinoma (0.5%) and a thyroid lymphoma (0.5%).

As for the percentages of malignancy in the different BS categories, after excluding incidental microcarcinomas, malignancy rates for category II, III, IV, V and VI were 4.6%, 11.5%, 33.8%, 86.5% and 100%, respectively (Table 1). In category I, the rate of malignancy was 35.3%, but in no case was this due to the preoperatively biopsied nodule, while in the overall series 86.4% of patients presented the tumor on the nodule that had been biopsied preoperatively. Thus, the rates of malignancy attributable to the biopsied nodule for categories II, III, IV, V and VI were 1.5%, 6.4%, 31%, 86.5% and 100%, respectively. By analyzing the differences between the percentages of malignancy in each of the different categories, we found a strong correlation in practically all of the comparisons (Table 2). Only statistically significant differences were not detected between categories I and II (*P*=1.000) and between categories I and III (*P*=.581).

Regarding the performance of BS, when we analyzed its utility as a screening test (analysis I: category II vs IV+V+VI), we found a sensitivity for detecting malignancy of 98.9%, with a specificity of 84.4%, a PPV of 69.6%, a NPV of 99.5% and an overall diagnostic accuracy of 88.2% (Table 3). In highly suspicious aspirations (analysis II: category II vs V+VI), this analysis increased the overall accuracy of the test up to 97.9% (sensitivity 98.6%, specificity 97.6%, 93.5 PPV% and NPV 99.5%).

Discussion

The most commonly used method for the description and categorization of thyroid FNA samples is the BS.[8,10] It is based on six categories, for each of which there is an estimated risk of thyroid cancer.[12] Our study seeks to review this risk, comparing cytology findings with the only possible gold standard: the definitive histological study of patients who have undergone thyroid surgery. The malignancy rates on which the statistical study has been based do not take incidental microcarcinomas into account, since most of them will have an indolent clinical course. In addition, we have considered only the existence of malignancy on the biopsied nodule, since we sought to define the capacity of the cytology study to identify the malignancy, not that of the ultrasound selection of the nodule to be biopsied. Despite this, we would like to point out that only 20 patients (3.8%) had a tumor >1 cm in one of the non-biopsied nodules. For identifying malignancy, our analysis shows the existence of a strong correlation

Table 2 – Statistical Analysis of the Categories of the Bethesda System.

Comparison of diagnostic categories	Chi-squared[a]	phi[b]	LR[c]	DF	*P*
DC I vs DC II vs DC III vs DC IV vs DC V vs DC VI	365.84	0.84	374.37	5	<.001
DC II vs DC III vs DC IV vs DC V vs DC VI	352.35	0.83	365.05	4	<.001
DC II vs DC VI	294.44	−0.96	278.51	1	<.001
DC II vs DC V	218.66	−0.86	148.52	1	<.001
DC II vs DC IV	66.55	−0.45	52.83	1	<.001
DC II vs DC III	5.46	−0.13	4.52	1	.034
DC II vs DC I	0.27	−0.03	0.51	1	1.000
DC III vs DC VI	119.22	−0.93	153.70	1	<.001
DC III vs DC V	74.74	−0.80	78.03	1	<.001
DC III vs DC IV	15.13	−0.32	15.98	1	<.001
DC III vs DC I	1.15	−0.11	2.03	1	.581
DC IV vs DC VI	66.15	−0.71	85.30	1	<.001
DC IV vs DC V	29.97	−0.53	32.52	1	<.001
DC IV vs DC I	7.02	−0.28	11.07	1	.005
DC V vs DC VI	8.55	−0.30	10.08	1	.007
DC V vs DC I	36.09	−0.82	43.69	1	<.001
DC VI vs DC I	77.00	−1.00	81.29	1	<.001

DC: diagnostic category; DF: degrees of freedom; LR: likelihood ratio.

[a] Chi-squared model to evaluate the association between categorical variables.

[b] Phi correlation coefficient used to evaluate the strength of association in categorical variables.

[c] Likelihood ratio (log-linear model) used to evaluate association in categorical variables.

Table 3 – Diagnostic Test Parameters of the Bethesda System.

Parameter	Analysis I (%)[a]	Analysis II (%)[b]
Sensitivity	98.9 (87/88)	98.6 (72/73)
Specificity	84.4 (205/243)	97.6 (205/210)
PPV in DC VI	100 (43/43)	100 (43/43)
PPV in DC V	85.3 (29/34)	85.3 (29/34)
PPV in DC IV	31.3 (15/48)	–
PPV	69.6 (87/125)	93.5 (72/77)
NPV	99.5 (205/206)	99.5 (205/206)
Rate of false negatives	1.1 (1/88)	1.4 (1/73)
Rate of false positives	15.6 (38/243)	2.4 (5/210)
Diagnostic accuracy	88.2 (292/331)	97.9 (277/283)

[a] Considers cases in DC IV+V+VI true positives and cases in DC II true negatives.
[b] Considers cases in DC V+VI true positives and cases in DC II true negatives.

for each of the BS categories, with differences between almost all of them (Table 2). This could justify maintaining the six categories, as occurred in the recent revision of the BS.[13] The percentages of malignancy observed for each category were, in general, within the limits described.[9,10]

Only 3.3% of the patients who underwent surgery had category I, a figure well below reports of other authors (7–26%).[8,14] We believe that this figure is based on the fact that all FNA were guided by ultrasound, avoiding other aspirations, while a cytologist evaluated the quality of the material obtained in situ. The percentage of malignancy for the nodule biopsied within this category was 0%, which represents an ideal percentage if we take into account that the results of this category affect the importance of obtaining satisfactory material for cytological analysis.

The malignancy rate associated with category III was 6.4%, similar to the originally proposed rate.[7] Although subsequent studies presented rates of up to 48% for this category, this can be attributed to the selection of surgical patients and the inclusion of incidental neoplasms in the analysis.[15] To try to assess this wide variation of malignancy within category III, it is recommended to assess the quotient between category III and category VI patients,[16] whose ideal value should be between 1 and 3. Values above 3 would indicate an overuse of category III, while values lower than 1 would be due to a low use of this category, with the consequent risk of loss of sensitivity for detecting malignancy. In our series, this quotient was 1.3, which brings us closer to the most efficient part of the recommended range.

Regarding the diagnostic test parameters of the BS (when assessed as a screening test, that is), needle-aspirations indicating a surgical intervention (Table 3) had an observed sensitivity of 98.9% and an NPV of 99.5%. These data are similar to the Bongiovanni et al.[9] study in terms of sensitivity, although our NPV greatly improves the average indicated in this study (99.5 vs 47%). This last datum is of special importance, since the main objective of preoperative FNAP is to rule out the existence of malignancy in order to reduce the number of unnecessary surgeries for this reason. In the second part of the analysis of the diagnostic test parameters of the BS, we have considered their capacity to ensure the existence of malignancy (categories V and VI), obtaining in this case a specificity of 97.6% and a PPV of 93.5%, with an overall accuracy of 97.9%. These data allow the BS to be defined as a very reliable tool when it comes to confirming the existence of malignancy. One of the limitations of the present study is those inherent to cytological analyses. Pathologists should be alert to the possibility of more errors in the analysis of cystic lesions, multinodular goiter or overlapping lesions with similar cytomorphological characteristics, such as the presence of reactive follicular cells or lesions with Hürthle cells. These findings mainly appear in the context of categories I and III, precisely those involved in the only comparisons between BS categories that did not show significant differences. However, we believe that the small number of patients that made up category I limits the ability to detect differences.

Conflict of Interest

The authors have no conflict of interests to declare.

REFERENCES

1. Burman KD, Wartofsky L. Clinical practice. Thyroid nodules. N Engl J Med. 2015;373:2347–56.
2. Guth S, Theune U, Aberle J, Galach A, Bamberger CM. Very high prevalence of thyroid nodules detected by high frequency (13 MHz) ultrasound examination. Eur J Clin Invest. 2009;39:699–706.
3. Davies L, Morris LG, Haymart M, Chen AY, Goldenberg D, Morris J, et al. American Association of Clinical Endocrinologists and American College of Endocrinology disease state clinical review: the increasing incidence of thyroid cancer. Endocr Pract. 2015;21:686–96.
4. Vaccarella S, Franceschi S, Bray F, Wild CP, Plummer M, Dal Maso L. Worldwide thyroid-cancer epidemic? The increasing impact of overdiagnosis. N Engl J Med. 2016;375:614–7.
5. Siegel R, Ma J, Zou Z, Jemal A. Cancer statistics, 2014. CA Cancer J Clin. 2014;64:9–29.
6. Ross DS. Predicting thyroid malignancy. J Clin Endocrinol Metabol. 2006;91:4253–5.
7. Ali SZ, Cibas ES. The Bethesda system for reporting thyroid cytopathology. Definitions criteria and explanatory notes. Nueva York: Springer; 2010.
8. Bongiovanni M, Spitale A, Faquin WC, Mazzucchelli L, Baloch ZW. The Bethesda system for reporting thyroid cytopathology: a meta-analysis. Acta Cytol. 2012;56:333–9.
9. Cibas ES, Baloch ZW, Fellegara G, Livolsi VA, Raab SS, Rosai J, et al. A prospective assessment defining the limitations of thyroid nodule pathologic evaluation. Ann Intern Med. 2013;159:325–32.
10. Haugen BR, Alexander EK, Bible KC, Doherty GM, Mandel SJ, Nikiforov YE, et al. 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer. Thyroid. 2016;26:1–133.
11. Cooper DS, Doherty GM, Haugen BR, Kloos RT, Lee SL, Mandel SJ, et al. Revised American Thyroid Association management guidelines for patients with thyroid nodules and differentiated thyroid cancer. Thyroid. 2009;19:1167–214.

12. Ali SZ. Thyroid cytopathology: Bethesda and beyond. Acta Cytol. 2011;55:4–12.

13. Cibas ES, Ali SZ. The 2017 Bethesda system for reporting thyroid cytopathology. Thyroid. 2017;27:1341–6.

14. Theoharis CG, Schofield KM, Hammers L, Udelsman R, Chhieng DC. The Bethesda thyroid fine-needle aspiration classification system: year 1 at an academic institution. Thyroid. 2009;19:1215–23.

15. Iskandar ME, Bonomo G, Avadhani V, Persky M, Lucido D, Wang B, et al. Evidence for overestimation of the prevalence of malignancy in indeterminate thyroid nodules classified as Bethesda category III. Surgery. 2015;157:510–7.

16. Krane JF, Vanderlaan PA, Faquin WC, Renshaw AA. The atypia of undetermined significance/follicular lesion of undetermined significance malignant ratio: a proposed performance measure for reporting in the Bethesda system for thyroid cytopathology. Cancer Cytopathol. 2012;120:111–6.