Received: 7 May 2018 | Revised: 3 July 2018 | Accepted: 5 July 2018

DOI: 10.1002/dc.24037

WILEY

ORIGINAL ARTICLE

A comparison study of the reporting systems for salivary gland fine needle aspirations: Are they really different?

Diana Montezuma MD[1] | Sule Canberk MD, MIAC[2,3,4] | Ozlem Aydın MD, FIAC[4] | Mehmet Polat Dermirhas Med Stud[5] | André F. Vieira PhD[2,3] | Süha Goksel MD[4] | Ümit İnce MD[4] | Fernando Schmitt MD, PhD, FIAC[2,3,6]

[1]Department of Pathology, Portuguese Oncology Institute of Porto (IPO Porto), Porto, Portugal

[2]Unit of Molecular Pathology, IPATIMUP, Institute of Molecular Pathology and Immunology of University of Porto, Porto, Portugal

[3]Cancer signaling metabolism-Epithelial Interactions in Cancer, I3S, Instituto de Investigação e Inovação em Saúde, University of Porto, Porto, Portugal

[4]Department of Pathology subdivision of Cytopathology, Acibadem University, Turkey

[5]Medical Faculty of 4th year, Acibadem Mehmet Ali Aydınlar University, Istanbul, TR, Turkey

[6]Department of Pathology, Medical Faculty of Porto University, Porto, Portugal

Correspondence
Fernando Schmitt, Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), Rua Julio Amaral de Carvalho 45, 4200-135, Porto, Portugal.
Email: fernando.schmitt@ipatimup.pt

Funding information
A.F.V. is a FCT fellow (SFRH/BPD/90303/2012)

Background: Recently a new system for reporting salivary gland fine-needle aspiration (FNA) cytology was proposed, the *Milan System for Reporting Salivary Gland Cytopathology* (MSRSGC). Herein, we evaluated diagnostic accuracy of salivary gland FNA, comparing the system previously used in our hospital with the Milan system.

Methods: Salivary gland specimens obtained between 2011 and 2017 were reclassified according to MSRSGC. Risk of malignancy for each diagnostic category was determined. Diagnostic yield of both classifications was evaluated.

Results: The cases (n = 388) were classified according to the old system: nondiagnostic (n = 28), benign (n = 246), atypical (n = 36), neoplastic (n = 57), suspicious for malignancy (n = 7) and malignant (n = 14). The lesions were distributed according to the MSRSGC: nondiagnostic (n = 28), non-neoplastic (n = 89), atypia of undetermined significance (n = 39), benign neoplasm (n = 156), neoplasm of uncertain malignant potential (n = 55), suspicious for malignancy (n = 7) and malignant (n = 14). When considering only benign and malignant cases, both classifications showed the same sensitivity (62.5%), specificity (100%) and similar accuracy (95.8%). Comparison between the two systems showed no significant difference.

Conclusions: Salivary gland FNA has high diagnostic accuracy and assists clinical management independently of the reporting system used, however, in some cases, the use of Milan system could be beneficial, since it allows an enhanced category stratification.

KEYWORDS

cytology, fine needle aspiration, Milan system, reporting system, salivary gland

1 | INTRODUCTION

The frequency of malignant salivary tumors varies from 0.4 to 2.6 cases per 100 000 population.[1] In the United States, salivary gland malignancies accounted for 6% of head and neck cancers, and 0.3% of all malignancies.[1] Anyhow, most commonly, salivary gland tumors are benign neoplasms and many of the punctured lesions are non-neoplastic. Fine-needle aspiration (FNA) cytology is widely accepted as an easy, safe and cost-effective method for salivary gland diagnosis, enabling patient management in the pre-operatively setting.[2–4] It is commonly preferred to histological biopsy which may be associated with higher risk of infection and potential contamination of surgical planes.[5] Notwithstanding, FNA has some inherent limitations, only being able to establish a specific diagnosis in 60% to 75% of cases.[6] Difficulty in assessing salivary gland lesions FNA is related to the wide spectrum of tumor types and rarity of many of these entities.[7] Up until recently, there was no consensus concerning salivary gland FNA reporting. Beginning in 2015, and leading to the atlas edition in 2018, an international group of pathologists developed a classification scheme for reporting salivary gland FNA results: the "Milan System for Reporting Salivary

Diana Montezuma and Sule Canberk contributed equally to this study.

TABLE 1 Milan system for reporting salivary gland cytopathology: Diagnostic categories definitions and ROM[a]

Diagnostic category	Definitions	ROM, % (range)
I. **Nondiagnostic**	Insufficient cellular material.	25 (0-67)
II. **Non-neoplastic**	Benign entities such as sialadenitis, reactive lymph node, granulomas and infection.	10 (0-20)
III. **Atypia of undetermined significance**	Containing limited atypia; indefinite for a neoplasm.	20 (10-35)
IV. **Neoplastic** A. **Benign Neoplasm** B. **SUMP**	Benign neoplasms diagnosed based on established cytologic criteria. Neoplasm; however, diagnosis of a specific entity cannot be made.	<5 (0-13) 35 (0-100)
V. **Suspicious for malignancy**	Highly suggestive of, but not unequivocal for malignancy.	60 (0-100)
VI. **Malignant**	Specimens diagnostic of malignancy.	90 (57-100)

SUMP salivary gland neoplasm of uncertain malignant potential.
[a] Categories and ROM range described in the literature.[9]

Gland Cytopathology" (MSRSGC).[8,9] This system defines a six-tiered classification scheme and the categories include nondiagnostic (ND), non-neoplastic (NN), atypia of undetermined significance (AUS), neoplasm (benign neoplasm, BN, and salivary gland neoplasm of uncertain malignant potential, SUMP), suspicious for malignancy (SM) and malignant (M) (Table 1). Each diagnostic category is associated with a known malignancy risk (Table 1). Herein, we applied the MSRSGC to salivary gland specimens diagnosed at our institution and compared it with the previously used classification system. In addition, risks of malignancy (ROM) for each category were assessed. Our previously used classification system corresponded to an in-house classification and categories were defined as follows: nondiagnostic, benign, atypical, neoplastic, suspicious for malignancy and malignant.

2 | MATERIAL AND METHODS

2.1 | Case collection and evaluation

The study was planned as exempt by the ethics committee and conforms to the standards of the Helsinki Accord. Since the study is retrospectively designed, no protected health information was used and patient consent was not obtained.

All salivary gland cytology specimens from January 2011 to December 2017 of the Acıbadem University pathology archive were retrieved (388 cases). Additional information were collected from the clinical files and are showed in Table 2. Histologic samples were considered the gold standard and were available in 104 cases (26.8%). All cytological cases were retrospectively reviewed by two experienced cytopathologists (S.C. and O.A.) and categorized by the recent MSRSGC. The conventional classification system was recorded considering the previous case reports.

TABLE 2 Patient's demographics and sample characteristics

Case cohort	
Samples, *n*	388
Patient age, mean (min-max)	46.01 (7-77)
Males, *n* (%)	55 (51.4)
Females, *n* (%)	52 (48.6)
Histological follow up, *n* (%)	104 (26.8)
Nodule size, when present (mm), mean (min-max)	22.37 (4.5-50)
Punctured lesion location, *n* (%)	
Parotid	93 (84.5)
Submandibular gland	17 (15.5)
Presence of mass lesion, *n* (%)	99 (92.5)
Rapid onsite evaluation	76 (71.0)
NA nonavailable.	

Non availabe information about gender in the request order form in 281 cases and location of the lesion in 278 cases.

2.2 | Statistical analysis

Statistical analysis was executed using NCSS (Number Cruncher Statistical System) 2007 (Kaysville, Utah). Standard descriptive analysis was performed. Risk of malignancy was defined for each category as number of confirmed malignant cases/total number of cases in diagnostic category. Sensitivity, specificity, positive and negative predictive values (PPV, NPV) and accuracy ratios were calculated using histologic diagnosis as the gold standard. The cytologic diagnostic categories were successively set as diagnostic threshold, that is, "cut-point." Categories with risks of malignancy equal to or greater than the "cut-point" category were combined as a group, which was defined as positive test whereas categories with risks of malignancy lower than the "cut-point" category were combined as a group, which was defined as negative test and for each of these combinations, the sensitivity, specificity, accuracy rate and Youden index were calculated.

Based on the sensitivity and specificity associated with various combinations aforementioned, a receiver operating characteristic (ROC) curve was plotted and the area under the receiver operating characteristic curve (AUROC) was subsequently calculated. The value of AUROC was interpreted as 0.5 (no discriminatory power) to 1.0 (perfect discrimination). The AUROC for the MSRSGC and for the old classification were compared by the DeLong method.[10]

3 | RESULTS

The patients' demographic data and lesions characteristics are summarized in Table 2. The aspirates were assigned six-categories using the old classification: nondiagnostic (ND), benign (B), atypical (A), neoplastic (N), suspicious for malignancy (SM) and malignant (M). The new Milan system also comprises six categories, described in Table 1. A total of 388 salivary gland specimens were classified according to the old system as: ND 7.2% (*n* = 28); B 63.4% (*n* = 246); A 9.3% (*n* = 36);

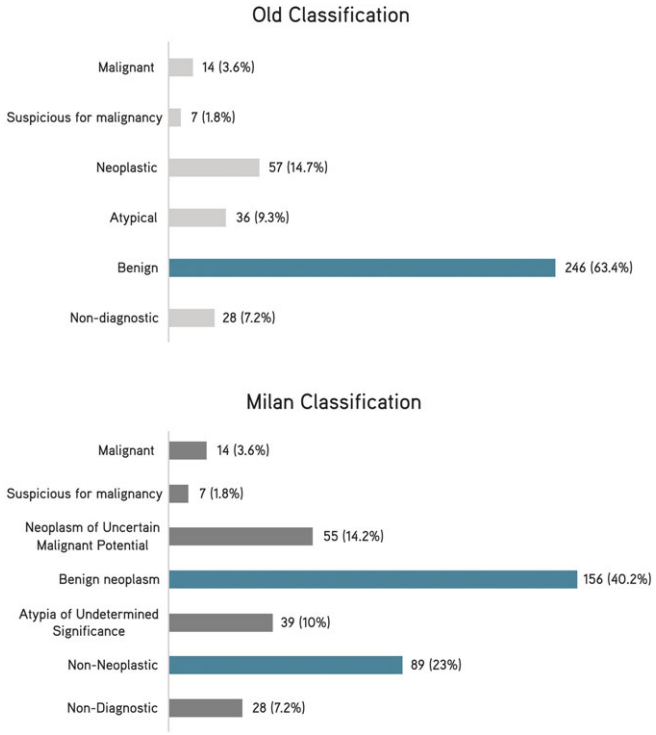

FIGURE 1 Diagnostic categories using the old and the new (Milan) classification systems. The old classification category B corresponds fairly to the NN and BN categories of the Milan classification [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Sensitivity, specificity, predictive value and accuracy (considering only benign and malignant cases) of both classifications

	Old classification % (95% CI)	Milan classification % (95% CI)
Sensitivity	62.5 (24.5-91.5)	62.5 (24.5-91.5)
Specificity	100 (94.6-100)	100 (94.5-100)
PPV	100 (47.8-100)	100 (47.8-100)
NPV	95.7 (88.0-99.1)	95.6 (87.6-99.1)
Accuracy	96.0 (88.8-99.2)	95.9 (88.5-99.1)

N 14.7% (*n* = 57); SM 1.8% (*n* = 7) and M 3.6% (*n* = 14) (Figure 1). The samples were reevaluated and reassigned according to the new Milan system as: ND 7.2% (*n* = 28); NN 23% (*n* = 89); AUS 10% (*n* = 39); BN 40.2% (*n* = 156); SUMP 14.2% (*n* = 55); SM 1.8% (*N* = 7) and M 3.6% (*n* = 14) (Figure 1). When applying the binary system to calculate diagnostic yield (considering only benign and malignant cases), both classifications showed the same sensitivity (62.5%) and specificity (100%). PPV, NPV and diagnostic accuracy were also very similar (Table 3). When grouping AUS, NUMP, SM and M categories altogether as a positive test result, using the Milan system, the sensitivity, specificity and accuracy rate associated with this combination were 70.59%, 88.89% and 86%, respectively. This combination yielded the maximum value of the Youden index (0.5948), when comparing with other diagnostic groupings, meaning the best combination of the sensitivity and specificity. Similar calculations for the old classification were performed. Likewise, the best diagnostic combination was achieved when grouping A, N, SM and M categories as a positive test result, with

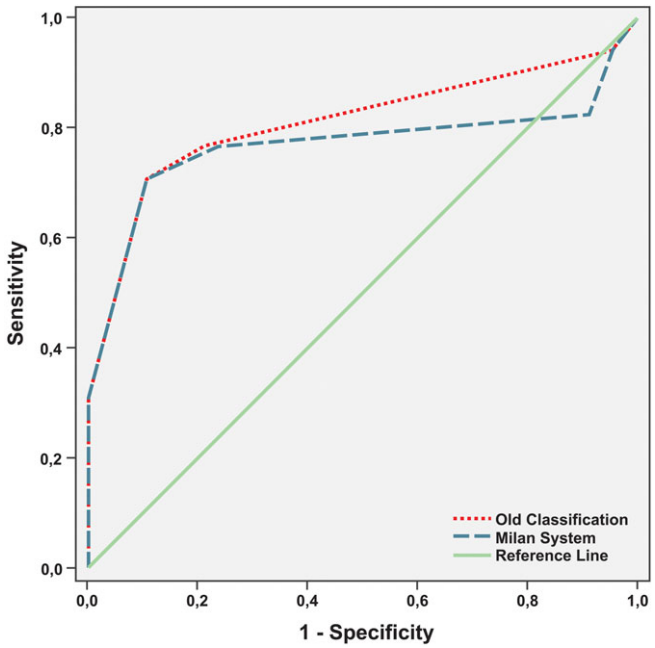

FIGURE 2 ROC curve for the Milan classification compared to ROC curve for the old classification [Colour figure can be viewed at wileyonlinelibrary.com]

sensitivity, specificity and accuracy rates of 70.59%, 88.89% and 86.00%, respectively. The corresponding Youden index was 0.5948. Figure 2 demonstrates the ROC curve of the Milan system comparing with the ROC curve for the old classification. The AUROC for the old and the new classifications were 0.808 (95% Confidence Interval [CI], 0.721-0.878) and 0.767 (95% CI, 0.675-0.843), respectively; and there was no statistically significant difference between them (*P* = 0.148). The ROM were calculated for both the old system and the Milan system and are detailed in Tables 4 and 5.

4 | DISCUSSION

The management of salivary gland lesions is usually based on clinical and imaging findings in association with cytological evaluation. FNA is commonly reported to have fewer complications when compared to core biopsy and with less potential for tumor seeding.[5] Nonetheless, some recent studies question this assumption and defend the use of core biopsy as safe and with potentially superior diagnostic yield.[11,12] Notwithstanding, FNA is certainly a less costly option which can be performed on a one-stop clinic setting and most studies find it an excellent pre-surgical triage tool, distinguishing accurately neoplastic from non-neoplastic lesions.[13]

In our series, we included mainly parotid and submandibular lesions. There were not sublingual or minor salivary gland lesions represented in our sample. The majority of cases corresponded to benign / non-neoplastic lesions, as is the case for most studies.[14–16] The B category in the old classification encompassed the most number of cases (*n* = 246), since it included both non-neoplastic lesions and benign neoplasms, corresponding fairly to the NN and the BN categories of the Milan classification (245 cases in total). The one case that was considered benign in the old classification and was not

TABLE 4 Correlation between FNA results and histological gold standard using the old classification

	Nondiagnostic (28)	Benign (246)	Atypical (36)	Neoplastic (57)	Suspicious for malignancy (7)	Malignant (14)
Resection	4	69	9	15	2	5
Benign	3	66	8	9	1	0
Malignant	1	3	1	6	1	5
ROM	25%	4.3%	11.1%	40%	50%	100%

TABLE 5 Correlation between FNA results and histological gold standard using the old classification

	Nondiagnostic (28)	Non-neoplastic (89)	Benign neoplasm (156)	Atypia of undetermined significance (39)	Neoplasm of uncertain malignant potential (55)	Suspicious for malignancy (7)	Malignant (14)
Resection	4	6	61	11	15	2	5
Benign	3	4	60	10	9	1	0
Malignant	1	2	1	1	6	1	5
ROM	25%	33.3%	1.6%	9.1%	40%	50%	17.3%

included in the NN or BN categories of the Milan classification corresponded to a cystic lesion with some atypical cells that was assigned as atypical by the new system. One important difference between the Milan classification and our previous classification is the subdivision of the neoplastic category in BN and SUMP and the addition of a NN category, which allows a more refined diagnosis and correlation with ROM. The cases corresponding to ND, SM and M categories remained unaltered between both classifications, which is understandable as they probably correspond to the most unambiguous categories. The A and AUS categories also maintained a similar number of cases (A: 36 cases and AUS: 39 cases). The SUMP category cases (using the Milan system) almost overlapped with the cases assigned to N category of the old classification.

We evaluated ROM using both the old and the new classification systems (Tables 4 and 5). When comparing our results using the Milan system and the reported ROM values for this classification we find that we have a good concordance overall. NN category showed a superior ROM value (33.3%) than expected (10%),[9] which can be explained since only six cases on this category had corresponding histological specimen, and four turned out to be malignant (1 adenoid cystic carcinoma and 3 metastases to salivary gland). All these cases had a cystic component, which might have contributed to sampling error. A more representative sample will probably avert this issue. Our AUS ROM value (9.1%) was on the lower limit of the reported ROM for this category (10%-35%).[9] The BN category had a ROM of 1.63%, in accordance to the expected value of less than 5%. ND, SUMP, SM and M categories ROM values were identical to the literature reported values.[9] A recent article, by Thyryayi et al., has reported lower ROM values for the ND, NN, AUS, BN and SUMP categories with 100% ROM for SM and M categories.[17]

When evaluating sensitivity, specificity, PPV, NPV and accuracy rates (for detecting malignancy) we found that both classifications presented similar values, almost perfectly overlapping (refer to Table 3). The reported values in most series for FNA sensitivity and specificity vary between 86%-100% and 90%-100%, respectively.[7,9,18] In our study, specificity was 100% (using both classifications), but the sensitivity was lower than expected (62.5%). Albeit some other studies also reported lower sensitivity values (such as 57%)[19] it can be difficult to compare these values between studies due to methodological differences (namely different approaches to the statistical analysis). Probably the considerable number of nonmalignant lesions in our sample might also have contributed to this outcome. In accordance to the relatively low sensitivity and high specificity values, NPV was lower (95.5%) than PPV (100%). Of note, one limitation of our study is the relatively small number of cases (n = 104) with corresponding histology. Another important remark is that our study is based only on specimens with corresponding histological specimens. This can lead to verification bias and cause overestimation of sensitivity and underestimation of specificity. This can affect the absolute estimates but the comparison between systems will still be valid because the bias would affect both systems in the same way.

When grouping AUS, NUMP, SM and M categories altogether (using the Milan classification) as a positive test result, the sensitivity, specificity and accuracy rate were 70.59%, 88.89% and 86%, respectively. This combination yielded the maximum value of the Youden index, meaning the best combination of the sensitivity and specificity. Similarly, for the old classification, the best diagnostic combination was achieved when grouping A, N, SM and M categories as a positive test result, with identical values comparing to the new system.

When comparing the old and new classification systems, the AUROC were 0.808 and 0.767, respectively, and there was no statistically significant difference between them (P = 0.148), meaning that both systems performed similarly in our sample, with a good discriminatory value.

5 | CONCLUSION

Salivary gland FNA has a high diagnostic accuracy and is valuable in the pre-surgical setting, independently of the classification system used. Nonetheless, in difficult or ambiguous cases the use of the Milan system could be beneficial since it allows for a better stratification of diagnostic categories and corresponding malignancy risk values, namely with the distinction of benign neoplasms from other benign entities and the introduction of a SUMP category.

ORCID

Fernando Schmitt http://orcid.org/0000-0002-3711-8681

REFERENCES

1. Barnes L, Evenson JW, Reichart P, Sidransky D. *Tumours of the Salivary Glands. WHO Classification of Head and Neck Tumours. IARC WHO Classification of Tumours*. 1st ed. ; Lyon: IARC Press; 2005:209-281.
2. Layfield LJ, Glasgow BJ. Diagnosis of salivary gland tumors by fine-needle aspiration cytology: a review of clinical utility and pitfalls. *Diagn Cytopathol*. 1991;7:267-272.
3. Frable MA, Frable WJ. Fine-needle aspiration biopsy of salivary glands. *Laryngoscope*. 1991;101:245-249.
4. Shaha AR, Webber C, DiMaio T, Jaffe BM. Needle aspiration biopsy in salivary gland lesions. *Am J Surg*. 1990;160:373-376.
5. Cibas ES, Ducatman BS, eds. *Cytology: Diagnostic Principles and Clinical Correlates* . Philadelphia: Elsevier Health Sciences; 2014.
6. Ellis GL, Auclair PL, eds. *Tumors of the Salivary Glands. AFIP Atlas of Tumor Pathology, 4th Series, Fascicle 9*. Silver Spring, MD: ARP Press; 2008.
7. Wei S, Layfield LJ, LiVosi VA, et al. Reporting of fine needle aspiration (FNA) specimens of salivary gland lesions: a comprehensive review. *Diagn Cytopathol*. 2017;45:820-827.
8. Baloch ZW, Faquin WC, Layfield LJ. Is it time to develop a tiered classification scheme for salivary gland fine-needle aspiration specimens? *Diagn Cytopathol*. 2017;45:285-286.
9. Faquin WC, Rossi ED, eds. *The Milan System for Reporting Salivary Gland Cytopathology*. Cham: Springer International Publishing AG; 2018.
10. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing areas under two or more correlated reciever operating characteristics curves: a nonparamentric approach. *Biometrics*. 1988;44:837-845.
11. Haldar S, Sinnott JD, Tekeli KM, Turner SS, Howlett DC. Biopsy of parotid masses: review of current techniques. *World J Radiol*. 2016;8:501-505.
12. Song IH, Song JS, Sung CO, et al. Accuracy of core needle biopsy versus fine needle aspiration cytology for diagnosing salivary gland tumors. *J Pathol Transl Med*. 2015;49:136-143.
13. Mairembam P, Jay A, Beale T, et al. Salivary gland FNA cytology: role as a triage tool and an approach to pitfalls in cytomorphology. *Cytopathology*. 2016;27:91-96.
14. Naz S, Hashmi AA, Khurshid A, et al. Diagnostic role of fine needle aspiration cytology (FNAC) in the evaluation of salivary gland swelling: an institutional experience. *BMC Res Notes*. 2015;27(8):101.
15. Singh Nanda KD, Mehta A, Nanda J. Fine-needle aspiration cytology: a reliable tool in the diagnosis of salivary gland lesions. *J Oral Pathol Med*. 2012;41:106-112.
16. Tryggvason G, Gailey MP, Hulstein SL, et al. Accuracy of fine-needle aspiration and imaging in the preoperative workup of salivary gland mass lesions treated surgically. *Laryngoscope*. 2013;123:158-163.
17. Thiryayi SA, Low YX, Shelton D, et al. A retrospective three-year study of salivary gland fine needle aspiration cytology with categorization using the Milan reporting system. *Cancer Cytopathol*. 2017;125:767-775.
18. Jain R, Gupta R, Kudesia M, et al. Fine needle aspiration cytology in diagnosis of salivary gland lesions: a study with histologic comparison. *Cytojournal*. 2013;10:5.
19. Zerpa V, Gonzáles MT, Porras G, et al. Diagnostic accuracy of fine needle aspiration cytology in parotid tumours. *Acta Otorrinolaringol Esp*. 2014;65:157-161.

How to cite this article: Montezuma D, Canberk S, Aydın O, et al. A comparison study of the reporting systems for salivary gland fine needle aspirations: Are they really different? *Diagn Cytopathol*. 2018;46:859–863. https://doi.org/10.1002/dc.24037