The impact of rapid on-site evaluation (ROSE) on the quality and diagnostic value of thyroid nodule fine-needle aspirations

Short running title: The impact of ROSE

Raphaela Muri[1,2] MSc, Mafalda Trippel[3] MD, Urs Borner[4] MD, Sabine Weidner[5] MD, Roman Trepp[1] MD

[1] Department of Diabetes, Endocrinology, Nutritional Medicine and Metabolism, Inselspital, Bern University Hospital and University of Bern, Switzerland

raphaela.muri@insel.ch and roman.trepp@insel.ch

[2] Graduate School for Health Sciences, University of Bern, Switzerland

[3] Institute of Pathology, University of Bern, Switzerland

mafalda.trippel@pathology.unibe.ch

[4] Department of Otorhinolaryngology, Head and Neck Surgery, Inselspital, Bern University Hospital and University of Bern, Switzerland

urs.borner@insel.ch

[5] Department of Nuclear Medicine, Inselspital, Bern University Hospital and University of Bern, Switzerland

sabine.weidner@insel.ch

Corresponding author

PD Dr. med. Roman Trepp, Department of Diabetes, Endocrinology, Nutritional Medicine and Metabolism, Inselspital, Bern University Hospital and University of Bern, 3010 Bern, Switzerland, roman.trepp@insel.ch

2

Abstract

Background: Ultrasound-guided fine-needle aspiration (FNA) is the preferred method to evaluate the dignity of thyroid nodules. Nevertheless, the often-reported high non-diagnostic rate burdens affected patients and the healthcare system. Rapid on-site evaluation (ROSE) constitutes an addition to the thyroid FNA procedure, with various studies showing its beneficial effect on the Bethesda I non-diagnostic rate. We aimed to assess whether ROSE may reduce the rate of Bethesda categories III and V. Additionally, we examined the influence of ROSE on specimen quality.

Methods: We performed a retrospective cohort study, comparing Bethesda categorization and specimen quality in specimens subject to ROSE compared to those not subject to ROSE. We also evaluated aspects of specimen quality that differed according to the use of ROSE. We subcategorized Bethesda I into insufficient cellularity or artifacts, and Bethesda categories III and V into cellular without artifacts, sparsely cellular, or artifacts.

Results: We evaluated 5030 thyroid FNAs. ROSE was performed in 1304 (25.9%) cases, and ROSE was not utilized for 3726 (74.1%) specimens. The rate of Bethesda I non-diagnostic and Bethesda III categories was reduced in specimens subject to ROSE (4.3%, 56/1304), compared to non-ROSE (39.9%, 1487/3726, $p<.001$). The rate of both benign Bethesda II and malignant Bethesda VI diagnoses was 91.6% (1194/1270) in ROSE specimens, compared to 56.6% (1999/3530) in non-ROSE ($p<.001$). This was reflected by a significant improvement in diagnostic accuracy with ROSE ($AUC_{non-ROSE}=0.811$, $AUC_{ROSE}=0.895$, $p=.004$). The overall rate of specimens flawed by sparse cellularity in Bethesda categories III and V was 0.1% (1/1304) in ROSE specimens compared to 1.2% (45/3726) in non-ROSE ($p<.001$). The overall artifact rate was 0.3% (4/1304) for ROSE specimens and 2.5% (92/3726) for non-ROSE ($p<.001$).

Conclusions: ROSE significantly increased diagnostic accuracy by improving FNA specimens quantitatively and qualitatively. We suggest considering ROSE as standard of care for thyroid FNAs.

3

Introduction

Ultrasound-guided fine-needle aspiration (FNA) is the method of choice for evaluating the dignity of thyroid nodules. The Bethesda System for Reporting Thyroid Cytopathology (TBSRTC) categorizes aspirates of thyroid nodules on a six-level scale from Bethesda I to Bethesda VI (1). This standardized classification scheme guides physicians on the decision of the further treatment course. Bethesda II (benign) and Bethesda VI (malignant) results are considered a definitive cytological diagnosis, usually allowing a targeted recommendation on how to proceed. However, despite its status as the gold standard, thyroid FNA is frequently reported to produce high non-diagnostic rates – in some cases amounting to 40% (2). Aspirates classified as Bethesda I non-diagnostic usually require repetition of the FNA (1), increasing patient discomfort. Likewise, specimens classified as Bethesda III entail repeating the FNA or additional molecular testing – the latter increases the costs considerably. Surgery is usually advised if a thyroid FNA specimen is classified as Bethesda IV or V, but with a reduced probability of malignancy (3). Surgery is also occasionally performed in cases of two consecutive Bethesda III diagnoses. The malignancy rate in these three Bethesda categories vary between centers but might be as low as 6% for Bethesda III, 10% for Bethesda IV, and 45% for Bethesda V when noninvasive follicular thyroid neoplasm with papillary-like nuclear features (NIFTP) is not regarded as carcinoma (1, 4). In addition, determining the optimal primary surgical approach for Bethesda V is a matter of debate (3).

A sufficient number of cells within the aspirated specimen is the mandatory prerequisite for accurate categorization. In addition, preparation artifacts and heavy bloodstaining significantly impact the quality of the smears and hamper proper categorization. These decisive quantitative and qualitative characteristics are primarily influenced, on the one hand, by the expertise of the aspirator (5, 6) and, on the other hand, by the preparation technique of the acquired samples (7). Both aspects are particularly challenging in training settings and with a low procedural volume of thyroid FNAs (8).

Rapid on-site evaluation (ROSE) of thyroid FNAs offers immediate feedback on the representativeness and quality of the harvested sample. ROSE has repeatedly been shown to significantly reduce the non-diagnostic rate of thyroid specimens (e.g., 8, 9). For

4

example, in one study, the introduction of ROSE decreased the non-diagnostic rate by more than 30% (2). By contrast, few studies reported no significant influence of ROSE on specimen adequacy (10, 11).

However, most studies have focused on ROSE as a tool to reduce the non-diagnostic rate – which comes along with an increase in overall adequacy – but disregard that adequacy alone is not necessarily sufficient for a definitive diagnosis. Despite the clinical significance, studies on the influence of ROSE on each Bethesda category are lacking. Our aim was to evaluate the impact of ROSE on sample adequacy, sample quality, rate of Bethesda categorization, and diagnostic performance. We hypothesized that ROSE could reduce the number of Bethesda I, III, and V categorizations of thyroid FNAs by diminishing the rate of low cellularity specimens. We postulated that ROSE, in the presence of a cytopathologist, may improve the overall quality of the smears by reducing the rate of negatively affected specimens from preparation artifacts or heavy bloodstaining. We believed that quantitative and qualitative improvements in FNA specimens attributed to the use of ROSE could increase the frequency of Bethesda II and Bethesda VI diagnoses.

Materials and Methods

Study population

This study was reviewed and authorized by the local ethics committee Bern, Switzerland (2020-02231). Patient data were retrieved from the Institute of Pathology, University of Bern's electronic information system for all FNAs of thyroid nodules performed between January 2015 and December 2020. FNAs of lymph nodes, ectopic thyroid tissue, and thyroid bed after prior thyroidectomy were excluded from the analysis.

FNAs without ROSE were performed by residents and senior physicians of the Departments of Endocrinology, Nuclear Medicine, Otolaryngology, and Radiology of the Inselspital, Bern University Hospital, and specialists of the same disciplines and internal medicine in private practices. Per nodule, at least two needle passes are performed, and at least three smears with spray fixation are made. Subsequently, the aspiration needle is rinsed with CytoLyt® solution for later ThinPrep® or cell block. See supplemental methods for a detailed overview of the procedure with ROSE.

5

ROSE was performed only at the Department of Endocrinology by cytopathologists from the Institute of Pathology, University of Bern. Before April 2019, ROSE was restricted to selected cases, namely as part of repeat FNAs of nodules that had provided non-diagnostic (Bethesda I) or indeterminable (Bethesda III) results. However, in April 2019, the Department established a twice-weekly interdisciplinary thyroid FNA service in which ROSE is an integral part of every FNA (see supplemental table 4a+b for a comparison between ROSE and non-ROSE before and after it became standard of care at our hospital).

Final cytological analyses for all FNAs are conducted at the Institute of Pathology. Each cytological assessment is performed by a senior cytopathologist and routinely reviewed in an unblinded way by a second senior cytopathologist. The cytopathologists involved had 13–20 years of pathology experience.

Categorization of cytology results

Cytology specimens were categorized according to the Bethesda System for Reporting Thyroid Cytopathology (TBSRTC) (1). Bethesda I non-diagnostic was further subdivided into the two subcategories *acellularity / too low cellularity* or *artifacts*. Furthermore, we differentiated Bethesda III and Bethesda V results into the three subcategories *cellular without artifacts*, *sparsely cellular,* and *artifacts*. *Artifacts* were defined as heavy bloodstaining or preparation artifacts due to overly thick smears or air drying/spraying artifacts (spray fixation was only performed in the non-ROSE group).

We combined Bethesda I and III to estimate the frequency of nodules requiring another appointment to repeat the FNA. In addition, we analyzed sparse cellularity by combining the Bethesda III and V subcategory *sparsely cellular*. Lastly, we merged the subcategory *artifacts* of Bethesda I, III, and V to assess overall artifacts within ROSE and non-ROSE, respectively.

ROSE is not expected to significantly affect the Bethesda IV rate (10, 11). However, the rate of Bethesda IV may vary between different centers. Therefore, to facilitate comparability, we defined the sum of Bethesda II and VI as *definitive* cytology and the sum of Bethesda I non-diagnostic, III, and V as *non-definitive* cytology.

Histology

The patient data of FNAs with histological confirmation were retrieved from the electronic information system of the Institute of Pathology, University of Bern. Every

6

histological assessment is performed by a senior histopathologist and routinely reviewed in an unblinded way by a second senior histopathologist. Histological diagnoses were divided into benign, malignant, or NIFTP. Aspirates of thyroid nodules subsequently surgically removed and histologically evaluated formed the basis for further calculations on the risk of malignancy (ROM), sensitivity, specificity, and diagnostic accuracy. All analyses were conducted assuming NIFTP once as benign and once as malignant. We excluded incidental microcarcinomas (<10 mm) within larger nodules from these calculations.

Statistical analyses

Continuous variables are presented as mean and standard deviation and categorical variables as proportions and percentages. Statistical analyses were performed with either a two-sample t-test, Pearson's chi-squared test, or Fisher's exact test. Sensitivity, specificity, and diagnostic accuracy were calculated as described in Bongiovanni et al. (12). Similarly to these authors, 37% of our cases with Bethesda III categorization had to undergo surgery. For this reason, we included this category in the sensitivity, specificity, and accuracy calculation. For comparison, these diagnostic metrics were also calculated without Bethesda III. In short, Bethesda II FNAs with benign histology were considered true negative samples. In contrast, true positives were defined as Bethesda (III,) IV, V, and VI FNAs with histologically confirmed malignancy. False negative samples included Bethesda II FNAs with malignant histology. Finally, false positive cases were defined as Bethesda (III,) IV, V, and VI FNAs, but histology confirmed benignancy. To compare sensitivity and specificity between non-ROSE and ROSE, we performed chi-squared tests on the true positives and false negatives (sensitivity) and the true negatives and false positives (specificity) of the two groups, respectively.

To evaluate diagnostic accuracy, receiver operating characteristic (ROC) curves were built, and the respective areas under the curve (AUC) were computed for each group (non-ROSE *vs*. ROSE). The difference in the AUC between non-ROSE and ROSE was evaluated using the DeLong method (13).

For percentages, Wilson's confidence intervals are calculated. For sensitivity calculations, exact binomial 95% confidence intervals are reported.

7

In order to determine independent predictors for a *definitive* (Bethesda II and VI) and a *non-definitive* (Bethesda I non-diagnostic, III, and V) cytology, we conducted post-hoc secondary multivariate logistic regression analysis with ROSE (yes/no), gender (male/female), and site (internal/external) as predictors and *definitive* cytology as the outcome variable. For each of these predictors, odds ratios and 95% confidence intervals were calculated.

P-values <.05 were considered statistically significant, and all reported *p*-values were corrected for multiple comparisons using the false-discovery-rate (FDR) method. The exact *p*-values for the chi-squared tests were calculated with the approach of Shan & Gerstenberger (14). All statistical analyses were performed on R (version 4.0.3).

Results

Patients

In this retrospective cohort study, the analysis included 5030 FNAs of 3140 patients. Of these, 3813 (75.8%) were obtained by physicians from the Inselspital, Bern University Hospital (i.e., internal), and 1217 (24.2%) by physicians in private practices (i.e., external). There were no significant differences between the non-ROSE cytology results obtained at the external and internal site, except for samples classified as Bethesda I cyst fluid only (internal 57/2517 (2.3%) *vs*. external 97/1209 (8.0%), (*p<.001*). Of the 5030 performed procedures, 3726 (74.1%) FNAs were conducted without ROSE (non-ROSE) and 1304 (25.9%) with ROSE. Between the two groups, there was no significant difference in age (non-ROSE 57.6±15.2 *vs*. ROSE 58.5±14.7 years, *p*=.074) or gender (male non-ROSE *vs*. male ROSE: 955/3726 (25.6%) *vs*. 366/1304 (28.1%) female non-ROSE *vs*. female ROSE: 2771/3726 (74.4%) *vs*. 938/1304 (71.9%) (*p*=.085). Inspecting the Bethesda categories separately, gender differences were found only within the non-ROSE group (see supplemental table 1a+b).

Cytology results without ROSE vs. with ROSE

An overview of ROSE for each Bethesda category is presented in table 1. Samples performed with ROSE showed significantly lower Bethesda I non-diagnostic (non-ROSE 1349/3726 (36.2%) *vs*. ROSE 29/1304 (2.2%), *p<.001*) and Bethesda III rates (non-ROSE 138/3726 (3.7%) *vs*. ROSE 27/1304 (2.1%), *p<.001*). There was no significant difference in the Bethesda IV rate between ROSE and non-ROSE (non-ROSE 42/3726 (1.1%) *vs*. ROSE

8

16/1304 (1.2%), *p*=.880). In examining the combined subcategories, the rate of need for a repeat FNA for Bethesda I non-diagnostic and Bethesda III categories was reduced in specimens subject to ROSE compared to non-ROSE (non-ROSE 1487/3726 (39.9%) *vs*. ROSE 56/1304 (4.3%), *p*<.001). Furthermore, the application of ROSE significantly decreased the rate of *sparse cellularity* (non-ROSE 45/3726 (1.2%) *vs*. ROSE 1/1304 (0.1%), *p*<.001) and *artifacts* (non-ROSE 92/3726 (2.5%) *vs*. ROSE 4/1304 (0.3%), *p*<.001) (see figure 1). Within Bethesda III and V, *cellularity without artifacts* was significantly higher with ROSE in both Bethesda III (non-ROSE 60/138 (43.5%) *vs*. ROSE 25/27 (92.6%), *p*<.001) and Bethesda V (non-ROSE 28/44 (63.6%) *vs*. ROSE 19/20 (95.0%), *p*=.038). Furthermore, ROSE demonstrated a significantly lower proportion of *sparse cellularity* (non-ROSE 41/138 (29.7%) *vs*. ROSE 1/27 (3.7%), *p*=.010) and *artifacts* (non-ROSE 37/138 (2.5%) *vs*. ROSE 1/27 (3.7%), *p*=.011) within Bethesda III (see figure 2).

The higher rate of diagnostic results and the lower rate of artifacts with ROSE resulted in a rise of benign Bethesda II results (non-ROSE 92/3726 (2.5%) *vs*. ROSE 4/1304 (0.3%), *p*<.001) and increased the rate of malignant Bethesda VI specimens (non-ROSE 92/3726 (2.5%) *vs*. ROSE 4/1304 (0.3%), *p*<.001) (table 1).

Accordingly, the ratio of *definitive* to *non-definitive* cytology improved significantly with ROSE (non-ROSE *definitive* 1999/3530 (56.6%) to *non-definitive* 1531/3530 (43.4%) *vs*. ROSE *definitive* 1194/1270 (94.0%) to *non-definitive* 76/1270 (6.0%); percentage difference (*definitive*): -37.4%, 95% CI [35.2%, 39.5%]; percentage difference (*non-definitive*): 37.4%, 95% CI [-39.5%, -35.2%], *p*<.001). ROSE (odds ratio: 10.5, 95% CI [8.4, 13.2], *p*<.001) and gender (odds ratio: 1.4, 95% CI [1.2, 1.6], *p*<.001) were significantly associated with *definitive* cytology.

Histology

The histological results for each Bethesda category can be found in table 2 (and supplemental table 2 for NIFTP=malignant). 893/5030 aspirates (17.8%) were followed by surgery. Histology was benign in 597/893 cases (66.9%), while 280/893 aspirates (31.4%) were histologically malignant (see supplemental table 5 for subtypes of malignant histological diagnoses). Only a small proportion (16/893, 1.8%) of specimens was classified as NIFTP.

9

Considering NIFTP as benign, the agreement on a benign diagnosis between the cytological and histological reports was significantly higher with ROSE than without ROSE (non-ROSE 225/463 (48.6%) *vs*. ROSE 134/150 (89.3%), percentage difference: -40.7%, 95% CI [-47.9%, -33.6%], *p*<.001). Similarly, histologically malignant nodules were more often cytologically malignant or suspicious for malignancy (Bethesda V and VI combined) with ROSE than without ROSE (non-ROSE 92/190 (48.4%) *vs*. ROSE 66/90 (73.3%), percentage difference: 24.9%, 95% CI [-37.3%, -12.5%], *p*<.001). The agreements on a benign and malignant diagnosis remained significantly higher with ROSE when NIFTP was considered as malignant (non-ROSE 93/199 (46.7%) *vs*. ROSE 67/97 (69.1%), percentage difference: 22.4%, 95% CI [-34.6%, -10.1%], *p*<.001; see also supplemental table 2).

Considering NIFTP as benign, specificity was significantly higher in the ROSE group than in the non-ROSE group (non-ROSE 80.1%, 95% CI [74.9%, 84.6%] *vs*. ROSE 90.5%, 95% CI [84.6%, 94.7%], *p*=.005). Additionally, diagnostic accuracy was significantly increased in the ROSE group compared to non-ROSE (AUC$_{non-ROSE}$=0.811, 95% CI [0.772, 0.850] *vs*. AUC$_{ROSE}$=0.895, 95% CI [0.854, 0.936]), *p*=.004) (table 3, upper row, and supplemental figure 1). Diagnostic accuracy was also significantly increased in the ROSE group when Bethesda III was not considered (AUC$_{non-ROSE}$=0.845, 95% CI [0.805, 0.885] *vs*. AUC$_{ROSE}$=0.907, 95% CI [0.866, 0.948], *p*=.036) (table 3, bottom row).

Considering NIFTP as malignant and including Bethesda III, specificity and diagnostic accuracy remained significantly higher in the ROSE group. Excluding Bethesda III, diagnostic accuracy was not significantly different anymore (supplemental table 3).

Discussion

Our study suggests that ROSE is helpful in the evaluation of thyroid nodules by FNA. The use of ROSE was associated with improvements in diagnostic conclusiveness, quality of cytological samples, and diagnostic accuracy.

Cytological and histological results without ROSE vs. with ROSE

In our study, the percentage of non-diagnostic specimens was 36.2% without ROSE and 2.2% with ROSE. Medina Chamorro and colleagues (2) reported a similar decrease in the non-diagnostic rate from 40.0% to 9.5% by adding ROSE. The initial inadequacy rate determines how much can be gained from the use of ROSE (15), and institutions with high inadequacy rates without ROSE have significantly more room for improvement (16). We

10

recognize that the initial non-diagnostic rate at our hospital of 36.2% was comparatively high. Still, such high rates are not uncommon (2, 17–19) and must be anticipated in the context of a training facility. Nevertheless, despite a high initial non-diagnostic rate, a Bethesda I rate of 2.2% with ROSE is in no way inferior to studies conducted with experienced operators with a high procedural volume. For instance, Houdek and colleagues (8) reported a 4.2% non-diagnostic rate with ROSE where pathologists and radiologists with high procedural volume performed the FNAs. Apart from a significantly improved non-diagnostic rate with ROSE, the authors could also demonstrate that, contrary to the assertion of Witt and Schmidt (15), experienced clinicians can very well benefit from ROSE. They reported a significant decrease of the non-diagnostic rate from 12.5% to 5.1% from a group of specialists who routinely perform thyroid FNAs. Therefore, the assumption that ROSE is only beneficial for inexperienced clinicians should not be extrapolated to any institution and expert group. ROSE can likewise be a valuable tool in reducing FNA inadequacy rates in training hospitals with high procedural volume such as ours.

In our study, the Bethesda III rate was 3.7% without ROSE and 2.1% with ROSE. When combining Bethesda I non-diagnostic and Bethesda III, ROSE reduced the need to repeat the FNA by nearly tenfold. A reduction in the need for repeat FNAs can significantly diminish the burden on patients and lower health care costs. By contrast, the rate of *definitive* benign and malignant diagnoses was significantly higher in the ROSE group (86.5% and 5.1%, respectively) than in the non-ROSE group (51.1% and 2.6%, respectively). As previously reported (10, 11), the addition of ROSE had no impact on the Bethesda IV rate. However, the Bethesda IV rate was relatively low in our study compared to other reports (10, 19).

In a meta-analysis, Lan et al. (21) reported a pooled sensitivity of 72% (ranging from 50% to 94%) and a pooled specificity of 99% (ranging from 32% to 100%) for FNAs. In another meta-analysis, Bongiovanni et al. (12) reported an overall sensitivity rate of 97.2% (with the inclusion of Bethesda III as true positive) and an overall specificity rate of 50.7% for FNA of thyroid nodules. Compared to the latter, the sensitivity of ROSE in our study was lower but still above the sensitivity reported by Lan et al. (21). However, the specificity rate in the ROSE group in our study appeared to be higher than reported by Bongiovanni et

al. (12) but slightly lower than calculated by Lan et al. (21). Lastly, while the diagnostic accuracy of both our groups corresponded to previous studies (21–23), it was significantly higher in the ROSE group.

Bethesda subcategories

We observed lower rates of inferior FNAs in specimens subject to ROSE compared to those not subject to ROSE. This reduction is likely related to the presence of a cytopathologist performing ROSE. A good smearing and staining technique with the omission of spray fixation prevents artifactual changes to the aspirate. Likewise, the lower rate of sparse cellularity can be attributed to the cytopathologists' assistance during the FNA.

Risk of malignancy

Overall, risk of malignancy in our sample was comparable to other studies. For example, Inabnet et al. (20) conducted a large multicenter study with over 20,000 thyroid patients with cytological and histological assessments. Interestingly, the overall ROM rate in Bethesda category I of 22% in the current study was similar to Inabnet et al. (20) and Pastorello et al. (18), who reported 19% and 24%, respectively. However, when comparing the ROM rate between the ROSE and non-ROSE groups, all Bethesda categories except for Bethesda II were higher in the ROSE group. This could be due to the higher diagnostic accuracy of ROSE.

Cost considerations of ROSE

A handful of studies have conducted analyses on the cost-effectiveness of ROSE with varying results (see Schmidt et al. (24) for an overview). Cost fluctuations depending on the country and institution need to be considered. At our hospital, a regular FNA amounts to approximately US$300 per patient. Extending an FNA consultation with ROSE adds US$100. In comparison, the costs of molecular testing for Bethesda III can amount to up to US$5000 per nodule. We, therefore, believe that ROSE may be more cost-effective than molecular testing.

Limitations

There are some limitations to report for this study. Needle passes per nodule were not documented and could not be collected due to the study's retrospective nature. The FNAs were performed by different physicians, and there were also different

12

cytopathologists involved. Additionally, the cytopathologists and histopathologists were not blinded when reviewing the specimens. We further acknowledge that the retrospective pre-post analysis of the implementation of ROSE utilized in this study has its flaws, as the allocation of ROSE could be biased. Indeed, a prospective randomized assignment of ROSE would be the most accurate method to evaluate its influence on specimen adequacy.

Conclusion

This retrospective study on over 5000 thyroid specimens demonstrated that the use of ROSE with FNA was associated with improvements in specimen quality, definitive diagnosis rate of benign and malignant nodules, and diagnostic accuracy compared to FNA without ROSE. We, therefore, suggest considering implementing ROSE as standard of care, especially when the rate of *definitive* cytology is below 90%. The need for additional human resources should not overshadow the benefits of ROSE concerning patient comfort, prevention of unnecessary surgery, and the financial burden on the health care system.

Acknowledgment

Contributions

RM gathered and analyzed the data, wrote the manuscript; RT designed the study, checked and analyzed the data, and wrote the manuscript; MT designed the study, gathered and checked the data and contributed to the manuscript; UB & SW reviewed the study protocol and contributed to the manuscript

References

1. Cibas ES, Ali SZ 2017 The 2017 Bethesda System for Reporting Thyroid Cytopathology. Thyroid 27:1341–1346.

2. Medina Chamorro FM, Calle JA, Stein JE, Merchancano L, Mendoza Briñez AM, Pulido Wilches AA 2018 Experience of the Implementation of Rapid On-Site Evaluation in Ultrasound-Guided Fine-Needle Aspiration Biopsy of Thyroid Nodules. Curr Probl Diagn Radiol 47:220–224.

3. Haugen BR 2017 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: What is new and what has changed? Cancer 123:372–381.

4. Yaprak Bayrak B, Eruyar AT 2020 Malignancy rates for Bethesda III and IV thyroid nodules: A retrospective study of the correlation between fine-needle aspiration cytology and histopathology. BMC Endocr Disord 20:1–9.

5. De Fiori E, Rampinelli C, Turco F, Bonello L, Bellomi M 2010 Role of operator experience in ultrasound-guided fine-needle aspiration biopsy of the thyroid. Radiol Medica 115:612–618.

6. Ghofrani M, Beckman D, Rimm DL 2006 The value of onsite adequacy assessment of thyroid fine-needle aspirations is a function of operator experience. Cancer 108:110–113.

7. Bellevicine C, Vigliar E, Malapelle U, Pisapia P, Conzo G, Biondi B, Vetrani A, Troncone G 2016 Cytopathologists can reliably perform ultrasound-guided thyroid fine needle aspiration: A 1-year audit on 3715 consecutive cases. Cytopathology 27:115–121.

8. Houdek D, Cooke-Hubley S, Puttagunta L, Morrish D 2021 Factors affecting thyroid nodule fine needle aspiration non-diagnostic rates: a retrospective association study of 1975 thyroid biopsies. Thyroid Res 14.

9. De Koster EJ, Kist JW, Vriens MR, Rinkes IHMB, Valk GD, De Keizer B 2016 Thyroid ultrasound-guided fine-needle aspiration: The positive influence of on-site adequacy assessment and number of needle passes on diagnostic cytology rate. Acta Cytol 60:39–45.

14

10. Jiang D, Zang Y, Jiang D, Zhang X, Zhao C 2019 Value of rapid on-site evaluation for ultrasound-guided thyroid fine needle aspiration. J Int Med Res 47:626–634.

11. Aly AK, Ali MA, Sharma A, Gubbels MA, Zhao X, Ahmed A, Aurit S, Stavas J 2021 Rapid On-site Evaluation (ROSE) for Fine Needle Aspiration of Thyroid: Is It Helpful? SciMedicine J 3:1–7.

12. Bongiovanni M, Spitale A, Faquin WC, Mazzucchelli L, Baloch ZW 2012 The Bethesda system for reporting thyroid cytopathology: A meta-analysis. Acta Cytol. Karger Publishers.

13. DeLong ER, DeLong DM, Clarke-Pearson DL 1988 Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics 44:837.

14. Shan G, Gerstenberger S 2017 Fisher's exact approach for post hoc analysis of a chi-squared test. PLoS One 12:e0188709.

15. Witt BL, Schmidt RL 2013 Rapid Onsite Evaluation Improves the Adequacy of Fine-Needle Aspiration for Thyroid Lesions: A Systematic Review and Meta-Analysis. Thyroid 23:428–435.

16. Schmidt RL, Witt BL, Lopez-Calderon LE, Layfield LJ 2013 The Influence of Rapid Onsite Evaluation on the Adequacy Rate of Fine-Needle Aspiration Cytology. Am J Clin Pathol 139:300–308.

17. Jing X, Michael CW, Pu RT 2008 The clinical and diagnostic impact of using standard criteria of adequacy assessment and diagnostic terminology on thyroid nodule fine needle aspiration. Diagn Cytopathol 36:161–166.

18. Pastorello RG, Destefani C, Pinto PH, Credidio CH, Reis RX, Rodrigues T de A, Toledo MC de, De Brot L, Costa F de A, do Nascimento AG, Pinto CAL, Saieg MA 2018 The impact of rapid on-site evaluation on thyroid fine-needle aspiration biopsy: A 2-year cancer center institutional experience. Cancer Cytopathol 126:846–852.

19. Zhu W, Michael CW 2007 How important is on-site adequacy assessment for thyroid FNA? An evaluation of 883 cases. Diagn Cytopathol 35:183–186.

20. Inabnet WB, Palazzo F, Sosa JA, Kriger J, Aspinall S, Barczynski M, Doherty G, Iacobone M, Nordenstrom E, Scott-Coombes D, Wallin G, Williams L, Bray R, Bergenfelz A 2020 Correlating the Bethesda System for Reporting Thyroid

15

Cytopathology with Histology and Extent of Surgery: A Review of 21,746 Patients from Four Endocrine Surgery Registries Across Two Continents. World J Surg 44:426–435.

21. Lan L, Luo Y, Zhou M, Huo L, Chen H, Zuo Q, Deng W 2020 Comparison of Diagnostic Accuracy of Thyroid Cancer With Ultrasound-Guided Fine-Needle Aspiration and Core-Needle Biopsy: A Systematic Review and Meta-Analysis. Front Endocrinol (Lausanne). Frontiers.

22. Crowe A, Linder A, Hameed O, Salih C, Roberson J, Gidley J, Eltoum IA 2011 The impact of implementation of the Bethesda System for Reporting Thyroid Cytopathology on the quality of reporting, "risk" of malignancy, surgical rate, and rate of frozen sections requested for thyroid lesions. Cancer Cytopathol 119:315–321.

23. Theoharis C, Adeniran AJ, Roman S, Ann Sosa J, Chhieng D 2013 The impact of implementing the bethesda system for reporting of thyroid FNA at an academic center. Diagn Cytopathol 41:858–863.

24. Schmidt RL, Walker BS, Cohen MB 2015 When is rapid on-site evaluation cost-effective for fine-needle aspiration biopsy? PLoS One 10:1–20.

16

Table 1: **Cytology results without ROSE and with ROSE**

	non-ROSE (n=3726)	ROSE (n=1304)	Percentage Difference (95% CI)	*p*-value
Bethesda I cyst fluid only ***	154 (4.1%)	18 (1.4%)	2.7% (1.8%, 3.7%)	<.001
Bethesda I non-diagnostic ***	1349 (36.2%)	29 (2.2%)	34.0% (32.2%, 35.8%)	<.001
Bethesda II ***	1903 (51.1%)	1128 (86.5%)	-32.4% (-37.9%, -32.9%)	<.001
Bethesda III **	138 (3.7%)	27 (2.1%)	1.6% (0.6%, 2.7%)	.007
Bethesda IV	42 (1.1%)	16 (1.2%)	-0.1% (-0.8%, 0.6%)	.880
Bethesda V	44 (1.2%)	20 (1.5%)	-0.3% (-1.2%, 0.5%)	.454
Bethesda VI ***	96 (2.6%)	66 (5.1%)	-2.5% (-3.8%, -1.1%)	<.001

Percentages in relation to the whole sample of ROSE and non-ROSE, respectively. CI: 95% confidence interval for the difference between the percentages. Significance levels after FDR-correction: *p<.05, **p<0.01, ***p<.001.

17

Table 2: **Histology of nodules according to Bethesda category and malignancy rate (NIFTP ≠ malignant) of cytologies with histological confirmation**

	Histological confirmation	Benign	NIFTP	Malignant	**Risk of Malignancy**
Bethesda I cyst fluid only (n=172)	26 (24 / 2)	25 (23 / 2)	0 (0 / 0)	1 (1 / 0)	**4%** **(4% / 0%)**
Bethesda I non-diagnostic (n=1378)	205 (202 / 3)	157 (157 / 0)	2 (2 / 0)	46 (43 / 3)	**22%** **(21% / 100%)**
Bethesda II (n=3031)	395 (251 / 144)	354 (224 / 130)	5 (1 / 4)	36 (26 / 10)	**9%** **(10% / 7%)**
Bethesda III (n=165)	61 (50 / 11)	29 (26 / 3)	6 (4 / 2)	26 (20 / 6)	**42%** **(40% / 55%)**
Bethesda IV (n=58)	39 (27 / 12)	25 (18 / 7)	1 (1 / 0)	13 (8 / 5)	**33%** **(30% / 42%)**
Bethesda V (n=64)	49 (33 / 16)	6 (5 / 1)	1 (0 / 1)	42 (28 / 14)	**86%** **(85% / 88%)**

18

Bethesda VI (n=162)	118 (66 / 52)	1 (1 / 0)	1 (1 / 0)	116 (64 / 52)	**98% (97% / 100%)**
Total (n=5030)	893 (653 / 240)	597 (454 / 143)	16 (9 / 7)	280 (190 / 90)	**31% (29% / 38%)**

Numbers in brackets show numbers or percentages of non-ROSE and ROSE, respectively. NIFTP = noninvasive follicular thyroid neoplasm with papillary-like nuclear features.

19

Table 3: **Sensitivity, specificity, and diagnostic accuracy for the ROSE and non-ROSE group (NIFTP = benign)**

	Sensitivity				Specificity				Diagnostic accuracy (AUC)			
	Non-ROSE	ROSE	X^2	p	Non-ROSE	ROSE	X^2	p	Non-ROSE	ROSE	D	p
Bethesda III, IV, V, and VI viewed as true positives	82.2% (75.0%, 88.0%) (120/146)	88.5% (79.9%, 94.4%) (77/87)	1.7	.197	80.1% (74.9%, 84.6%) (225/281)	90.5% (84.6%, 94.7%) (134/148)	7.8	.005	0.811 (0.772, 0.850)	0.895 (0.854, 0.936)	-2.90	.004
Bethesda IV, V, and VI viewed as true positives	79.4% (71.3%, 86.1%) (100/126)	87.7% (78.5%, 93.9%) (71/81)	2.4	.125	89.6% (85.1%, 93.1%) (225/251)	93.7% (88.4%, 97.1%) (134/143)	1.9	.169	0.845 (0.805, 0.885)	0.907 (0.866, 0.948)	-2.10	.036

Sensitivity, specificity, and diagnostic accuracy calculated once with Bethesda III, IV, V, and VI (upper row) and once without the addition of Bethesda III (bottom row). Noninvasive follicular thyroid neoplasm with papillary-like nuclear features (NIFTP) included as benign.

20

Numbers in brackets indicate 95% confidence interval. Proportions are described below the 95% confidence intervals (for sensitivity: true positives/(true positives +false negatives); for specificity: true negatives/(true negatives + false positives)). X^2 = Chi-squared, $D=D$-statistic.
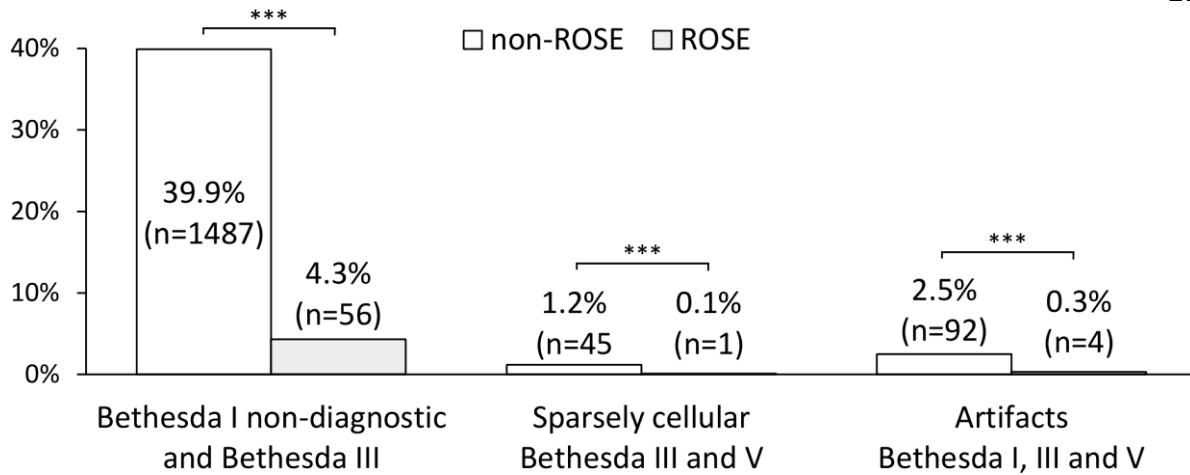
21

Figure 1:   **Effect of ROSE on the estimated need for a repetition of the FNA and the subcategories *sparsely cellular* and *artifacts***

Percentages in relation to the whole sample of ROSE and non-ROSE, respectively. The corresponding 95% confidence intervals and FDR-corrected p-values for each group are as follows: Bethesda I non-diagnostic and Bethesda III: CI [0.336, 0.376], p=.000; Sparsely cellular Bethesda III and V: CI [0.399, 0.043], p=.001; Artifacts Bethesda I, III, and V: CI [0.015, 0.028], p=.000. Significance levels after FDR-correction: *p<.05, **p<.01, ***p<.001).
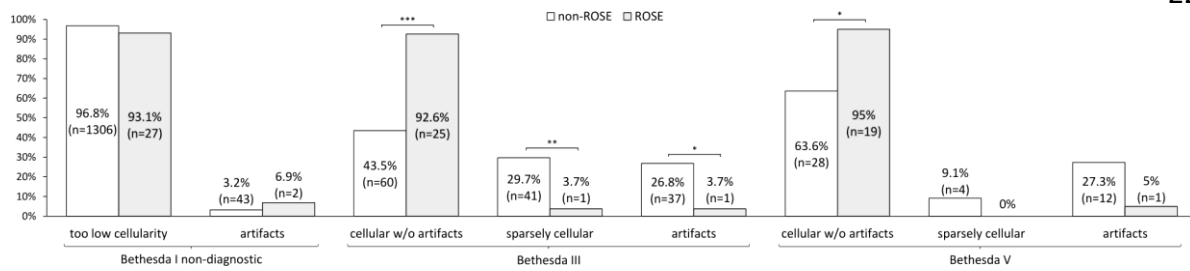
Figure 2: **Effect of ROSE on cellularity and artifacts in Bethesda I non-diagnostic, III, and V, respectively**

Percentages in relation to Bethesda I non-diagnostic, III and V, respectively. The corresponding 95% confidence intervals and FDR-corrected p-values for each group are as follows: Bethesda I non-diagnostic too low cellularity: CI [-0.073, 0.147], p=.488; Bethesda I non-diagnostic artifacts: CI [-0.147, 0.073], p=.488; Bethesda III cellular without artifacts: CI [-0.642, -0.340], p=.000; Bethesda III sparsely cellular: CI [0.134, 0.387], p=.010; Bethesda III artifacts: CI [0.106, 0.356], p=.011; Bethesda V cellular without artifacts: CI [-0.521, -0.106], p=.038; Bethesda V sparsely cellular: CI [-0.030, 0.212], p=.300; Bethesda V artifacts: CI [0.024, 0.422], p=.073. Significance levels after FDR-correction: *p<.05, **p<.01, ***p<.001.