A Systematic Review and Meta-analysis of the Diagnostic Accuracy of Frozen Section for Parotid Gland Lesions

Robert L. Schmidt, MD, PhD, MMed, MBA,[1] Jason P. Hunt, MD,[2] Brian J. Hall, MD,[1] Andrew R. Wilson, MStat,[3] and Lester J. Layfield, MD[1]

Key Words: Parotid gland; Frozen section; Meta-analysis; Sensitivity; Specificity; Systematic review

DOI: 10.1309/AJCP2SD8RFQEUZJW

Downloaded from https://academic.oup.com/ajcp/article/136/5/729/1766053 by guest on 09 April 2024

Abstract

We conducted a systematic literature review using MEDLINE and Embase to identify articles on diagnostic accuracy of frozen section (FS) for salivary gland lesions published between January 1, 1985, and December 31, 2010. We also reviewed the reference lists of all identified articles and conducted a forward search using Scopus to identify all articles citing the reference set. Meta-analysis was used to produce a summary receiver operating characteristic (SROC) curve from which summary estimates of sensitivity and specificity were obtained. Study quality was assessed using the Quality of Diagnostic Accuracy Study (QUADAS) survey. The accuracy of FS was compared with that of fine-needle aspiration cytology using results from an earlier review.

A set of 13 studies (1,880 cases) with extractable data met our inclusion criteria. The summary estimates for the area under the SROC curve, FS sensitivity, and FS specificity are 0.99 (95% confidence interval [CI], 0.98-1.00), 0.90 (95% CI, 0.81-0.94), and 0.99 (95% CI, 0.98-1.00), respectively. FS has acceptable accuracy (90% sensitivity, 99% specificity) and is consistently accurate across study centers.

The diagnosis and treatment of salivary gland lesions proceeds in stages. In the initial stage, clinical, radiologic, and fine-needle aspiration cytology (FNAC) are used to categorize cases into surgical or nonsurgical groups **Figure 1** (decision 1). In general, patients with neoplastic disease are referred for surgical treatment, whereas patients with nonneoplastic disease receive other modes of treatment. In the second stage, neoplastic lesions must be categorized as benign or malignant, which influences the type and extent of surgery (Figure 1, decision 2). For example, the facial nerve can be spared in benign lesions. Thus, 2 important decisions rest on the initial diagnosis: (1) whether surgical treatment is required and (2) the type and extent of surgery.

FNAC has a well-established role in the initial, preoperative diagnosis of salivary gland lesions. It is safe, fast, well-tolerated, and minimally invasive; however, it is known to have several deficiencies. On average, FNAC has high specificity (97%), but the sensitivity is somewhat lower (80%).[1] Thus, a positive diagnosis by FNAC is quite reliable, but the false-negative rate associated with FNAC (20%) may be unacceptable. In addition, the accuracy of FNAC varies widely across practice sites.[1] Thus, there is a need to improve the reliability of FNAC.

Frozen section (FS) provides an opportunity to refine the presurgical diagnosis among patients who have been referred to surgery. FS is used for 3 general purposes[2] in the evaluation of salivary gland lesions: (1) clarify diagnosis, (2) check operative margins, and (3) determine whether nerve or neck involvement is present. FS would be routinely used in cases with a presurgical malignant diagnosis to check margins and assess nerve involvement, but the value of FS is less certain when the presurgical diagnosis is benign. In such cases, FS is

© American Society for Clinical Pathology

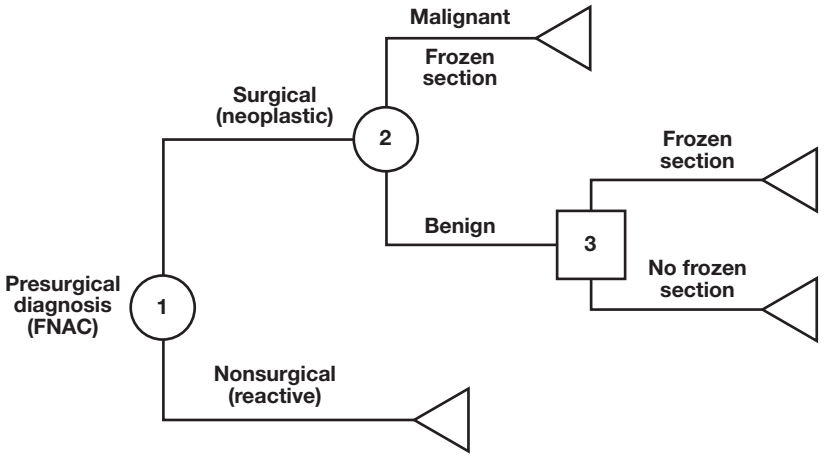

Figure 1 Decision diagram for the use of frozen section. The preoperative diagnosis divides patients into surgical and nonsurgical categories (decision 1). Patients with neoplastic disease are generally referred to surgery. For patients referred to surgery, frozen section will generally be used to assess margins and nerve involvement in patients diagnosed with malignant disease (decision 2). Frozen section can be used to refine the presurgical diagnosis in patients with benign neoplastic lesions (decision 3). The numbers in the diagram indicate the decision text. The square represents a decision node; circles represent chance nodes; and triangles indicate the end of a branch. FNAC, fine-needle aspiration cytology.

used to refine the diagnosis (Figure 1, decision 3), and, when used for this purpose, the usefulness of FS depends on the accuracy of FS and the accuracy of the presurgical diagnosis. Overall, FS can provide information that may supplement other diagnostic methods (eg, imaging and FNAC) or provide information that is complementary or unique to FS (eg, assessment of margins). Although FS provides multiple types of information that can be evaluated for accuracy, this review is concerned with the accuracy of FS for the classification of lesions with respect to malignancy.

When used for diagnosis, FS can be used in several ways. First, FS can clarify diagnoses when other modes of preoperative diagnosis are inadequate or inconclusive. In such cases, FS is used as a stand-alone technique to provide a diagnosis. FS can also be used more routinely to confirm or refine a presurgical diagnosis. Although FS adds additional information, the incremental value of FS relative to other modes of preoperative diagnosis has not been determined, and it is unclear whether FS should be used routinely or limited to certain circumstances. In particular, the roles of FS and FNAC have been controversial. Some authors have concluded that FS is more reliable than FNAC and that FS should be used routinely to ensure the best patient management.[3] Others have concluded that FNAC and FS have similar accuracy and that both have a role in the evaluation of salivary gland lesions.[4] The answers to questions about the routine use of FS vs the use of FS only in specific circumstances require an accurate assessment of the relative diagnostic performance of presurgical diagnosis and FS.

The diagnostic accuracy of FS performance has significant implications for the surgical management of parotid tumors. Surgery is the preferred treatment of most neoplasms of the parotid. However, the extent of treatment differs based on the histopathologic type of neoplasm. Historically, the concerns of false-negatives with FNAC led to its minimal use in preoperative decision making. Thus, intraoperative decisions were based on FS analysis or on final histopathologic diagnoses with the potential of revision surgery.[5] Intraoperative FS may increase the extent of parotidectomy based on a malignant diagnosis and even determine the need for facial nerve sacrifice or the addition of a neck dissection. Thus, understanding the accuracy of FS has significant implications for its use in the intraoperative setting.

Systematic reviews are the foundation of evidence-based medicine and provide the basis for the development of guidelines for patient management. Numerous studies on the accuracy of FS for the diagnosis of parotid tumors have been published; however, this body of literature has never been adequately summarized. Although some studies have summarized selected results, the literature on FS has not been systematically reviewed. We recently completed a systematic review and meta-analysis on the diagnostic accuracy of FNAC for parotid gland lesions.[1] Our objective in this study was to conduct a similar review on the diagnostic accuracy of FS and to compare the accuracy of FS with FNAC for diagnosis of malignancy. To that end, we conducted a systematic review of the literature and used meta-analysis to develop a summary receiver operating characteristic (SROC) curve for the diagnostic performance of FS in the evaluation of parotid gland tumors. We also conducted a quality assessment of included articles to explore potential sources of bias and to provide recommendations to improve future studies.

Downloaded from https://academic.oup.com/ajcp/article/136/5/729/1766053 by guest on 09 April 2024

DOI: 10.1309/AJCP2SD8RFQEUZJW
© American Society for Clinical Pathology

Materials and Methods

We followed current guidelines for systematic review and meta-analysis of diagnostic studies.[6,7] A glossary of terms is provided in **Appendix 1**.[8,9]

Literature Search

We searched MEDLINE and Embase for studies evaluating the diagnostic accuracy of FS for parotid lesions published between January 1, 1985, and October 15, 2010, using a sensitive search strategy developed in consultation with an experienced medical reference librarian. Our search strategy was broad and included articles on head and neck lesions in addition to salivary glands. Language was not restricted. Scopus was used to perform a forward search to obtain articles citing the set of included articles. The references of all included articles were also searched to obtain additional studies.

Eligibility

Titles and abstracts were evaluated independently by 2 of us (R.L.S. and B.J.H.) for eligibility. Studies were eligible if they seemed to contain data on the diagnostic accuracy of FS for salivary gland tumors or head and neck tumors. Prospective and retrospective studies were eligible. Full reprints were obtained for all eligible studies.

Inclusion

Eligible studies were independently evaluated by 2 of us (R.L.S. and B.J.H.), and discrepancies were resolved by consensus. Studies were included if they contained accuracy data for the diagnosis of parotid gland lesions, contained histologically verified cases that could be extracted, and provided data that enabled lesions to be classified into broad categories (malignant vs benign).

We excluded case reports and studies with fewer than 10 cases. Eligible studies were included if accuracy data could be extracted in the form required for analysis (true-positives, false-positives, false-negatives, and true-negatives).

Data Extraction

Data extraction was completed independently by 2 of us (R.L.S. and B.J.H.), and discrepancies were resolved by consensus or by correspondence with authors of the study in question. Data from foreign-language articles (non-English) were extracted by pathologists with knowledge of the language, correspondence with study authors, or by a translator working in conjunction with one of us (R.L.S.). Inadequate or indeterminate biopsies were not counted in the calculation of accuracy. Diagnoses of "suspicious for malignancy" or "atypical" were counted as malignant. When results of a study were published more than once, we included only the most complete data. Non–salivary gland tumors (eg, lymphoma and metastases) were included in the analysis.

Quality Assessment

Quality assessment of articles written in English was conducted by using QUADAS.[10] Assessment was completed independently by 2 of us (R.L.S. and B.J.H.). We used a scoring form that we developed in our study of FNAC.[1] We reviewed items with discrepant scores. Discrepancies due to errors and misinterpretations were corrected. Discrepancies sometimes arose owing to differences in judgment. We discussed these items until consensus was reached. The consensus approach was infrequently required because the initial level of agreement was high.

Statistical Analysis

SROC curves were developed by using the hierarchical inverse-variance-weighted meta-analysis. We created new variables containing the diagnostic accuracy estimates and standard errors for each study.[8,11] Computations were done using Stata Statistical Software, Release 11 (2009; StataCorp, College Station, TX) and applying the metandi procedure.[12] We used the inconsistency (I^2) statistic to measure heterogeneity across studies. Heterogeneity is a measure of the between-study variation in studies and is used to assess whether the studies in a meta-analysis represent a single population (with similar accuracy) or several different populations.

Results

Literature Search

We screened 3,848 titles and abstracts to obtain a set of 551 eligible articles. The text of the eligible studies was screened to obtain 13 studies that met our inclusion criteria **Table 1**.[2-5,13-21] These studies included a total of 1,880 cases on the discrimination of benign vs malignant lesions.

Diagnostic Accuracy

The SROC curve for discrimination of benign and malignant lesions is shown in **Figure 2**, and the accuracy estimates are given in **Table 2**. FS has higher specificity than sensitivity. In addition, there is more variability in sensitivity than in specificity **Figure 3**. A test for study heterogeneity, I^2, was not significant ($P = .06$). The rate of inadequate or inconclusive results was 3.7% among the studies that reported such results.

Quality Assessment

A summary of the QUADAS quality assessment is given in **Table 3**. All studies were retrospective. The only QUADAS items that had variable results were items 1 (inclusion criteria), 2 (representativeness), and 13 (inconclusive results explained). All other items were identical among studies.

Downloaded from https://academic.oup.com/ajcp/article/136/5/729/1766053 by guest on 09 April 2024

© American Society for Clinical Pathology

DOI: 10.1309/AJCP2SD8RFQEUZJW

Table 1
Performance Data From 13 Included Studies

Study	No. of Cases	TP	FP	FN	TN	Inconclusive or Nondiagnostic	Sensitivity	Specificity	Study Location
Badoual et al,[13] 2006	694	91	6	24	573	12	0.79	0.99	France
Carvalho et al,[14] 1999	153	16	3	10	124	8	0.62	0.98	Brazil
Hwang and Brett,[15] 2003	36	2	0	0	34	0	1.00	1.00	Singapore
Ishida et al,[16] 2003	152	26	1	2	123	0	0.93	0.99	Japan
Iwai et al,[17] 1999	167	23	1	1	142	NR	0.96	0.99	Japan
Longuet et al,[18] 2001	94	9	0	3	82	NR	0.75	1.00	France
Arabi Mianroodi et al,[3] 2006	30	10	0	0	20	0	1.00	1.00	Australia
Seethala et al,[4] 2005	61	26	0	9	26	7	0.74	1.00	United States
Tew et al,[19] 1997	144	27	1	1	115	15	0.96	0.99	Australia
Upton et al,[5] 2007	155	22	1	1	131	0	0.96	0.99	United States
Wong,[2] 2002	19	9	1	0	9	12	1.00	0.90	Hong Kong
Zbaren et al,[20] 2008	110	63	2	5	40	0	0.93	0.95	United States
Zheng et al,[21] 1997	65	14	1	2	48	NR	0.88	0.98	China

FN, false-negatives; FP, false-positives; NR, not reported; TN, true-negatives; TP, true-positives.

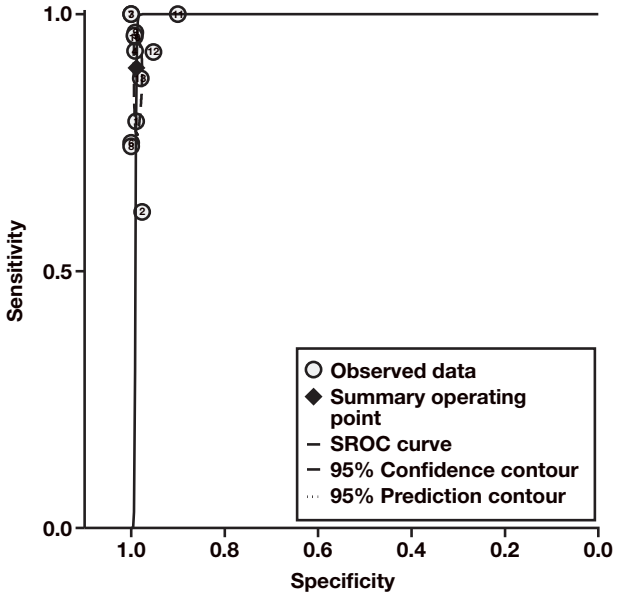

Figure 2 Summary operating receiver characteristic (SROC) curve for the diagnosis of malignancy by frozen section. The SROC curve shows that most studies have high specificity with variable sensitivity. The summary estimate is shown as the diamond. Individual studies are represented as circles. For the summary operating point, sensitivity, 0.90 (95% confidence interval [CI], 0.81-0.94); specificity, 0.99 (95% CI, 0.98-0.99). For the SROC curve, the area under the curve is 0.99 (95% CI, 0.98-1.00).

Discussion

The overall accuracy of FS is quite good. It has high specificity (0.99), and, although the sensitivity is not as high (0.90), we believe it is clinically acceptable.

Sources of Bias and Heterogeneity

The degree of heterogeneity we observed in this collection of studies was borderline insignificant (P = .06). Tests for heterogeneity are known to be insensitive, so it is possible that there is some undetected heterogeneity in our collection of studies. In general, performance variation can be attributed to 4 factors: real differences between tests (population, test performance, reference test, and outcome measure), threshold effects, bias, and random variation. Understanding the causes of performance variation is important because it provides a basis to improve consistency and improve diagnostic performance.

Threshold Effects

Our statistical analysis suggests that most of the performance variation is due to threshold effects. In general, variation in a direction tangent to the SROC curve can be attributed to threshold effects, whereas variation in the direction perpendicular to the tangent of the SROC curve can be attributed to accuracy. In the evaluation of FS, a difference in accuracy would mean that pathologists differ in their ability to detect features and to interpret them correctly. A difference in threshold would mean that pathologists see the same features but use different criteria for diagnosing malignancy. As shown in Figure 2, almost all of the studies lie along the SROC curve, and the estimated percentage of variation due to threshold effects is close to 100%. Threshold effects have been recognized as an important source of variation in other studies.[22] Such variation might be minimized by more uniform application of diagnostic criteria.

Sources of Bias

The purpose of a quality assessment is to identify potential sources of bias and to estimate their impact. We identified 3 potential sources of bias: classification bias (item 3, Table 3), review bias (item 11, Table 3), and handling of inconclusive or inadequate results (item 13, Table 3).

Misclassification bias results from an imperfect reference test (ie, definitive histologic examination). There are

Downloaded from https://academic.oup.com/ajcp/article/136/5/729/1766053 by guest on 09 April 2024

DOI: 10.1309/AJCP2SD8RFQEUZJW
© American Society for Clinical Pathology

Table 2
Summary of Accuracy Estimates for Frozen Section and Comparison With FNAC

Parameter	Frozen Section		FNAC*	
	Point Estimate	95% CI	Point Estimate	95% CI
Sensitivity	0.90	0.81-0.94	0.80	0.76-0.83
Specificity	0.99	0.98-0.99	0.97	0.96-0.98
Positive likelihood ratio	80.6	47.5-137.0	28.6	20.7-42.0
Negative likelihood ratio	0.11	0.06-0.19	0.21	0.17-0.26
Area under summary receiver operating characteristic curve	0.99	0.98-1.00	0.96	0.94-0.97
Inconsistency, I^2	53	0-100	98	97-99

CI, confidence interval; FNAC, fine-needle aspiration cytology.
* Data from Schmidt et al.[1]

2 types of misclassification: differential and nondifferential. Differential misclassification occurs when the error rate of the "gold standard" is related to the result of the index test. For example, differential misclassification would occur if the error rate of definitive histologic examination was different for the samples called positive by FS compared with those called negative. Nondifferential misclassification occurs when the error rate of the gold standard is independent of the index test result (ie, FS). Although no data are available on the misclassification rate of parotid tumors, we believe that error rates are most likely nondifferential (ie, independent of the FS result). Renshaw et al[22] found a misclassification rate of about 3% in a wide range of surgical specimens. We investigated the effect of nondifferential misclassification on our summary estimates. To that end, we took the totals from all of the studies **Table 4** and conducted a sensitivity analysis **Figure 4**. Given the data in our study, nondifferential misclassification would cause

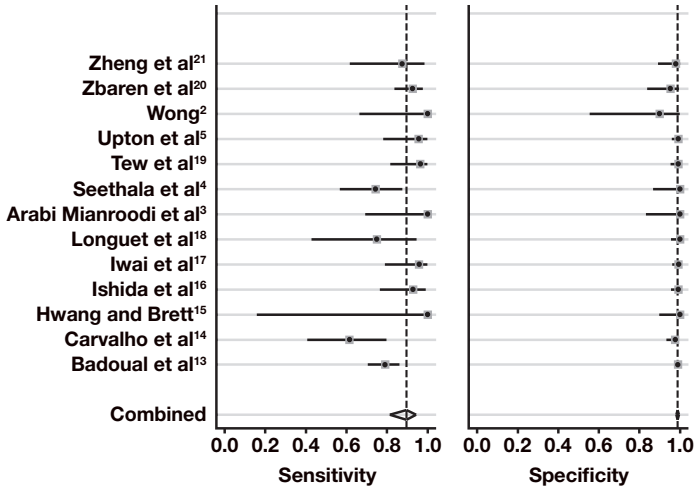

Figure 3 Sensitivity and specificity of included studies. Forest plot of diagnostic performance statistics. Boxes denote point estimates of sensitivity and specificity. Lines denote 95% confidence intervals. Diamonds represent 95% confidence intervals for the combined estimate from all studies.

Table 3
Summary of QUADAS Survey Results

QUADAS Item	Description	Yes	Unclear	No
1	Were the selection criteria clearly described?	10	2	1
2	Was the spectrum of patients representative of the patients who will receive the test in practice?	6	7	0
3	Is the reference standard likely to correctly classify the target condition?	0	13	0
4	Is the time period between the reference standard and the index test short enough to be reasonably sure that the target condition did not change between tests?	13		
5	Did the whole sample or a random selection of the sample receive verification using the reference standard of diagnosis?	13	0	0
6	Did patients receive the same reference standard regardless of the index test result?	13	0	0
7	Was the reference standard independent of the index test (ie, the index test did not form part of the reference standard)?	13	0	0
8	Was the execution of the index test described in sufficient detail to permit replication of the test?	13	0	0
9	Was the execution of the reference standard described in sufficient detail to permit replication?	13	0	0
10	Were the index results interpreted without knowledge of the results of the reference standard?	13	0	0
11	Were the reference standard results interpreted without knowledge of the results of the index test?	0	0	13
12	Were the same clinical data available when test results were interpreted as would be available when the test was used in practice?	13	0	0
13	Were uninterpretable or intermediate results reported?	8	5	0
14	Were withdrawals from the study explained?	—	—	—

Downloaded from https://academic.oup.com/ajcp/article/136/5/729/1766053 by guest on 09 April 2024

© American Society for Clinical Pathology

DOI: 10.1309/AJCP2SD8RFQEUZJW

an underestimation of sensitivity and would have relatively little impact on specificity.

Review bias occurs when the results of one test can influence the interpretation of another. Review bias is called diagnostic review bias when the interpretation of the reference test is made with knowledge of the results of the index test. Test review bias occurs when the interpretation of the index test is made with knowledge of the results of the reference test. All of the studies we included were retrospective studies using data obtained under actual clinical conditions. Pathologists were not blinded to the results of FS when making the final diagnosis. Thus, there is some potential for FS results to influence the interpretation of the reference test. This diagnostic review bias would tend to increase sensitivity and specificity; however, we believe that this effect is relatively minor because final histologic study is generally weighed much more heavily than FS when making a final diagnosis.

Table 4
Summary of Totals for 13 Included Studies

	Histologic Diagnosis	
Frozen Section	**Malignant**	**Benign**
Malignant	338	17
Benign	58	1,467

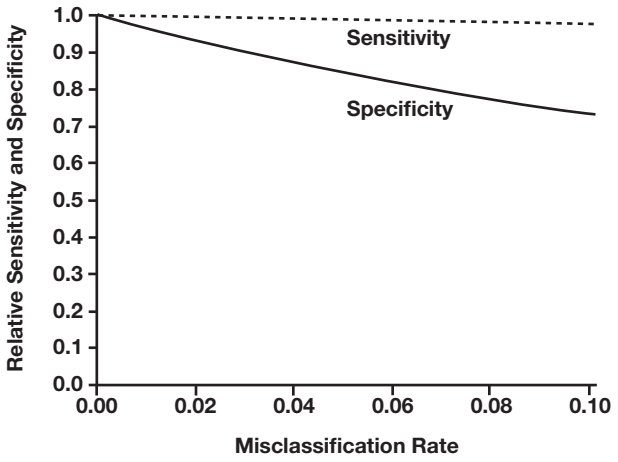

Figure 4 The effect of misclassification on the observed sensitivity and specificity. Summary totals for all included studies (Table 4) were used as a reference. The values shown in the graph are the calculated sensitivity and specificity relative to the initial sensitivity and specificity in Table 4. For example, the sensitivity based on the summary totals is 0.85. At a misclassification rate of 0.05, the observed sensitivity would be 0.72. The sensitivity shown in the graph is the ratio of the observed sensitivity to the actual sensitivity, 0.72/0.85 = 0.85.

Inconclusive results can have a significant impact on accuracy statistics. Thus, it is important that such results are reported in a consistent manner. The percentage of inconclusive or inadequate results was 3.7% for FS in our set of studies. This finding is in contrast with an inconclusive-inadequacy rate between 8.1% and 9.5% that we found in our review of FNAC.[1] It is important that studies report inconclusive and inadequate results and report them in a standard format to facilitate comparisons between studies. This is an area requiring improvement.

There are a few additional areas in which improved reporting or study design would reduce potential bias. In some cases, authors could improve descriptions of the criteria used to select patients. In particular, it is important for authors to indicate whether the study included only consecutive patients from a given period. Diagnostic performance is sensitive to subtle effects related to referral patterns, so it is also important for authors to provide a brief description of the patient population. Although the patient populations were not described in some studies, we saw no studies in which referral patterns were likely to be problematic.

Unlike FNAC studies, verification bias (QUADAS item 5) is not a source of bias in FS studies because all samples (negative and positive) are referred for definitive histologic examination. Thus, we do not believe that verification bias is an important source of bias in the studies we included in our analysis.

Random Variation

The I^2 statistic indicates that 53% of the variation is due to between-study variation (real differences between studies) and that the remainder is attributable to within-study, or random variation. The within- vs between-study variation is shown in Figure 3. This result is in contrast with the high level of performance variation seen in FNAC, in which 98% of the variation was estimated to be due to between-study variation.[1]

Overall, the variability in sensitivity seen between FS studies is most likely due to a combination of threshold differences and random variation. The studies were homogeneous with respect to real differences (population, index test performance, reference test performance, and outcome measures), so this is an unlikely source of variation. We identified only minor potential sources of bias (review bias and misclassification bias). These sources of bias would tend to be fairly uniform across studies and would most likely lead to an underestimate of sensitivity.

Comparison of the Accuracy of FS and FNAC

We use the findings from our previous study of the diagnostic accuracy of FNAC[1] for the diagnosis of parotid gland lesions for comparison with FS diagnosis of lesions in the same anatomic location.

Downloaded from https://academic.oup.com/ajcp/article/136/5/729/1766053 by guest on 09 April 2024

DOI: 10.1309/AJCP2SD8RFQEUZJW
© American Society for Clinical Pathology

The area under the SROC (AUSROC) curve is a measure of accuracy that can be used to compare 2 different methods. The diagnostic accuracy of FS (AUSROC, 0.99; 95% confidence interval [CI], 0.98-1.00) is significantly higher than the diagnostic accuracy of FNAC (AUSROC, 0.96; 95% CI, 0.94-0.97). The specificity of FS is quite high (0.99). The specificity values of FS and FNAC are both clinically acceptable (0.99 for FS vs 0.97 for FNAC). Thus, a positive result obtained by either technique is quite reliable. As with FNAC, the sensitivity of FS is lower than the specificity. While the sensitivities of FS and FNAC are not statistically significantly different, the sensitivity of FS was slightly higher (0.90; 95% CI, 0.81-0.94) than the sensitivity of FNAC (0.80; 95% CI, 0.75-0.83). Thus, FS might be used to refine the diagnosis of cases referred to surgery with a nonmalignant diagnosis by FNAC (Figure 1, decision 3).

A possible policy would be to accept an FNAC diagnosis of malignant as correct but to check an FNAC diagnosis of benignancy and to use FS only to check the cases referred to surgery. Assuming that FNAC and FS are independent, the overall sensitivity and specificity of such a policy would be 0.98 and 0.95, respectively, and would cut the false-negative rate from 20% (FNAC only) to 5% (combined FNAC and FS). Although such a policy has the potential to reduce false-negatives, it would also increase costs by increasing the rate of use of FS; however, clinical and radiologic analysis will often support or call into question the fine-needle aspiration and guide the decision on the need for a post-FNAC FS. The trade-offs should be investigated through a formal cost-effectiveness analysis.

The performance variability (heterogeneity) seen in our study of FS is much less than the heterogeneity seen in our previous study of FNAC.[1] The I^2 statistic[23] is a measure of the percentage of total performance variability that can be attributed to between-study variability vs within-study variability that results from the natural underlying error rate. For FS, the inconsistency statistic was 53% (CI, 0%-100%) and was not statistically significant. In contrast, the inconsistency statistic for FNAC was highly significant at 98% (CI, 97%-99%). Thus, there is much greater between-study variability for FNAC than for FS. We were unable to determine the reason for the high level of heterogeneity seen in FNAC studies; however, our results suggest that the diagnostic accuracy of FNAC may be more operator-dependent than FS.

Salivary gland neoplasms are diagnostically challenging at the time of FNAC interpretation and FS analysis. Many salivary gland neoplasms have cytologically low-grade nuclear morphologic features. Thus, interpretation of FNAC material relies on the evaluation of tumor fragments with retained architectural features and extracellular and intracellular materials (mucin and myxoid matrix) for identification of several tumor types. Certain neoplasms demonstrate considerable morphologic overlap, making cytologic distinction difficult. These neoplastic types include cellular pleomorphic adenomas, monomorphic adenomas (especially basal adenomas), low-grade mucoepidermoid carcinomas, adenoid cystic carcinomas, and polymorphous low-grade adenocarcinomas. Nuclear features of these neoplasms may overlap, making cytologic separation difficult. In addition, in some aspiration specimens, there are insufficient architectural clues for definitive diagnosis. Properly performed FS has the advantage over FNAC in that architectural features such as capsular and perineural invasion are retained and when present favor carcinoma or adenoma. The retention of architectural clues in FS specimens undoubtedly improves the sensitivity and specificity of FS in comparison with FNAC.

FS provides valuable information for surgeons and intraoperative decision making because it often determines the extent of treatment of parotid neoplasms. On the most basic level, it is important to distinguish malignant from benign disease.

Benign disease often requires only partial parotidectomy, frequently superficial parotidectomy. Even limited tumor resection with preservation of normal parotid tissue has been proposed as adequate treatment for benign disease.[24,25] With a sensitivity of fine-needle aspiration biopsy near 80%, FS provides additional confirmation of benign disease during resection that allows for limited surgery.

For malignant parotid disease, difficult decisions are made intraoperatively, including management of the facial nerve and regional lymph nodes. While sacrificing the facial nerve is not necessary in benign disease, involvement of the facial nerve by malignancy will require sacrifice of 1 or more branches of the nerve. Preoperative dysfunction of the facial mimetic muscles is indicative of tumor involvement of the facial nerve, but there may be nerve involvement in people with normal preoperative facial movement. A malignant diagnosis increases a surgeon's comfort level in the indication for facial nerve sacrifice, and this may require an FS diagnosis.

After a portion of the facial nerve is resected, FS is used to ensure proximal and distal control of the perineural tumor spread before a primary neural reconstruction. When FS identifies tumor at the stylomastoid foramen, a mastoidectomy may be required to control the proximal facial nerve margin. In regard to management of regional lymph nodes, an FS diagnosis of malignant disease, other than low-grade disease, requires treatment of the regional lymphatics by neck dissection or irradiation.[26,27] Several authors favor a neck dissection in the treatment of regional lymphatics when the parotid mass is known to be malignant.[28,29] For people with a history of squamous cell carcinoma of the skin, an FS diagnosis of squamous cell carcinoma is diagnostic of regional lymphatic disease necessitating a neck dissection in addition to the parotidectomy for the treatment of regional

Downloaded from https://academic.oup.com/ajcp/article/136/5/729/1766053 by guest on 09 April 2024

© American Society for Clinical Pathology

DOI: 10.1309/AJCP2SD8RFQEUZJW

metastatic spread. Also, a more aggressive parotidectomy with removal of the entire superficial parotid is prudent with malignancy because the majority of parotid lymph nodes are contained within the superficial lobe. The results of this study show that FS is a reliable technique and should provide assurance to surgeons who depend on FS to make important intraoperative decisions.

Conclusions

The overall accuracy of FS is clinically acceptable (90% sensitivity, 99% specificity) and shows consistent results between study centers. The accuracy of FS is somewhat higher than that of FNAC. It could be used to confirm a diagnosis for patients who are referred to surgery following a negative FNAC diagnosis, but there is concern about the low sensitivity of FNAC.

From the [1]Department of Pathology, University of Utah, Salt Lake City; [2]Division of Otolaryngology–Head and Neck Surgery, University of Utah Health Sciences Center, Salt Lake City; and [3]ARUP Laboratories Institute of Experimental and Clinical Pathology, Salt Lake City.

Address reprint requests to Dr Schmidt: Dept of Pathology, University of Utah, 15 North Medical Dr East, Salt Lake City, UT 84112.

Acknowledgments: We thank Chie Minoda for translation of Japanese articles and Mary Youngkin for assistance with the literature search.

Appendix 1
Glossary of Terms

Term	Definition
Area under the curve (AUC)	A measure of diagnostic accuracy that is independent of a particular threshold or cutoff point. In general, it is difficult to compare diagnostic accuracy between studies because differences could represent real differences in accuracy (ability to detect features) or differences in threshold. The AUC represents the average sensitivity for all levels of specificity and thereby removes threshold effects. The AUC is a valid method to compare the diagnostic accuracy of tests because it removes the effect of the choice of threshold. The AUC is the region that underlies the ROC curve or the SROC curve and varies between 0.5 (no discrimination) to 1.0 (perfect discrimination).
Bias	A systematic difference between an observed measurement and the true value. For a diagnostic accuracy study, bias occurs when there is a methodological factor that leads to a systematic difference between the observed sensitivity and specificity vs the true sensitivity and specificity. Common sources of bias in diagnostic studies include verification bias, spectrum bias, misclassification bias, and review bias.
Embase	An electronic database similar to PubMed but with broader coverage of foreign language journals, particularly in Europe.
Forward search	A search to identify more recent studies that have cited a set of known publications. A forward search is also known as a citation search. Citation databases include Scopus, Google Scholar, Web of Knowledge, and the Ovid interface for MEDLINE.
Heterogeneity	A measure of the between-study variation in studies. Measures of heterogeneity assess whether the included studies represent a single population (with similar diagnostic accuracy) or several different populations with different diagnostic performance. For example, a homogeneous set of studies might have sensitivities and sensitivities clustered around 80 and 80 whereas a heterogeneous set of studies might have 2 clusters of studies centered around sensitivity and specificity values of 60 and 65 and 95 and 90. The assessment of heterogeneity is an important component of meta-analysis and can lead to the identification of factors that cause variation in diagnostic accuracy between study centers.
Hierarchical analysis	Hierarchical models are a generalization of multiple linear regression. In multiple regression, an outcome variable is modeled at a single level. For example, y = a + bx, where a and b are parameters that are determined by best fit techniques. In hierarchical regression, models are also constructed for the coefficients (eg, a = c + dz). In this study, we used a hierarchical model[8] to construct a summary ROC curve. At one level, we model the variation within studies and, at another level, we model the variation between studies. The STATA metandi function applies the method of Macaskill[9] to formulate a summary ROC curve.
Inconsistency statistic (I^2)	A statistical measure of heterogeneity that varies from 0 to 100. The value describes the proportion of total variation in study results that is due to heterogeneity rather than chance. A value of 0 means that the set of studies is homogeneous and the observed variation is due to chance (random variation), whereas a value of 100 means that the observed variation between studies is due to heterogeneity.
Index test	Refers to the test under study. In the present study, FNAC is the index test, and histologic examination is the reference test.
Meta-analysis	A method for combining study results to obtain better estimates (by increasing sample size) and to investigate factors that cause variability in study results.
Metandi	A procedure for meta-analysis of diagnostic accuracy studies that is contained in Stata. Stata is a general purpose statistical package.
Misclassification bias	Results from errors in the reference test (the "gold standard"). In FNAC studies, the reference test is histologic examination or clinical follow-up. Neither of these reference standards is error-free. Errors in the reference standard can lead to systematic differences (bias) in the observed sensitivity and specificity of the index test (FNAC).

FNAC, fine-needle aspiration cytology; ROC, receiver operating characteristic.

Downloaded from https://academic.oup.com/ajcp/article/136/5/729/1766053 by guest on 09 April 2024

DOI: 10.1309/AJCP2SD8RFQEUZJW
© American Society for Clinical Pathology

References

1. Schmidt RL, Hall BJ, Wilson AR, et al. A systematic review and meta-analysis of the diagnostic accuracy of fine needle aspiration cytology for parotid gland lesions. *Am J Clin Pathol*. 2011;136:45-59.
2. Wong DS. Frozen section during parotid surgery revisited: efficacy of its applications and changing trend of indications. *Head Neck*. 2002;24:191-197.
3. Arabi Mianroodi AA, Sigston EA, Vallance NA. Frozen section for parotid surgery: should it become routine? *ANZ J Surg*. 2006;76:736-739.
4. Seethala RR, LiVolsi VA, Baloch ZW. Relative accuracy of fine-needle aspiration and frozen section in the diagnosis of lesions of the parotid gland. *Head Neck*. 2005;27:217-223.
5. Upton DC, McNamar JP, Connor NP, et al. Parotidectomy: ten-year review of 237 cases at a single institution. *Otolaryngol Head Neck Surg*. 2007;136:788-792.
6. Leeflang MMG, Deeks JJ, Gatsonis C, et al. Systematic reviews of diagnostic test accuracy. *Ann Intern Med*. 2008;149:889-897.
7. Deeks JJ, Bossuyt PM, Gatsonis C, eds. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*. Version 1.0.0. The Cochrane Collaboration; 2009. Available at http://srdta.cochrane.org/handbook-dta-reviews. Accessed January 15, 2011.
8. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med*. 2001;20:2865-2884.
9. Macaskill P. Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis. *J Clin Epidemiol*. 2004;57:925-932.
10. Whiting P, Rutjes AWS, Reitsma JB, et al. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol*. 2003;3:25. doi:10.1186/1471-2288-3-25.

Term	Definition
QUADAS (Quality of Diagnostic Accuracy Study)	A survey instrument designed to assess the methodological quality and reporting quality of diagnostic accuracy studies. The survey is a questionnaire that covers common sources of bias in diagnostic accuracy studies.
Reference test	Refers to the gold standard test. For FNAC studies, the reference test is usually histology or, sometimes, clinical follow-up.
Review bias	Occurs when the results of the index test (FNAC) are known when the reference test (histologic examination) is interpreted. Knowledge of the index test can influence the interpretation of the reference test. Blinding can remove this source of bias; however, most FNAC studies are retrospective and not blinded.
Receiver operating characteristic (ROC) curve	A plot that shows the ability of a test to discriminate between 2 diagnostic categories. Sensitivity is plotted against 1 – specificity to show the tradeoff that occurs by varying the test criteria (threshold). For a chemical test, the threshold is varied by changing the cutoff value. For surgical pathology specimens, thresholds are related to multiple features such as number of mitoses and cell sizes. Each point on the ROC curve represents a combination of sensitivity and specificity. The ROC curve shows the tradeoff between sensitivity and specificity for a particular test. Each point on the ROC curve is obtained by varying the cutoff point for positive and negative cases. The points on an ROC curve are based on the same features (accuracy) but represent different interpretations of their meaning. Two different points on the ROC curve will have different sensitivity and specificity owing to differences in threshold, but they will have the same accuracy.
Scopus	A large electronic abstract and citation database.
Summary ROC (SROC) curve	The SROC is similar to the ROC except that an SROC is obtained by combining the results of multiple studies.
Systematic review	A transparent, comprehensive, and reproducible approach to obtain evidence to answer a research question. The search strategy for a systematic review should survey multiple data sources and should not be limited with respect to the language in which the study report is published. Criteria for selecting studies should be stated in advance. Evaluation for the inclusion of studies should be completed independently by 2 persons. The process generally begins by screening abstracts and titles to determine eligibility, which is then followed by a more detailed review of full articles. Data should be extracted independently in duplicate. Critical appraisal of the included studies is an important part of a systematic review. A quality assessment such as QUADAS should usually be conducted to assess for potential sources of bias. Leeflang et al[6] and the Cochrane Collaboration have provided guidelines for systematic reviews of diagnostic accuracy studies. Nonsystematic reviews are called narrative reviews.
Verification bias	Results from differences in the verification of positive cases as determined by the index test. In FNAC studies, positive FNAC results are usually followed up by surgery, whereas negative FNAC results are usually not verified. This pattern is known as partial verification (negative results are not verified at the same rate as positive results) and can lead to bias in the observed sensitivity and specificity. Differential verification bias occurs when one method (a "gold standard") is used to verify positive cases and another method (a "brass standard") is used to verify negative cases. For example, in FNAC studies, positive results may be referred to surgery, whereas negative results are followed up clinically, which can lead to biased estimates of accuracy of the index test (FNAC) if the 2 methods of verification (histologic examination vs clinical observation) differ in accuracy.

Downloaded from https://academic.oup.com/ajcp/article/136/5/729/1766053 by guest on 09 April 2024

© American Society for Clinical Pathology

DOI: 10.1309/AJCP2SD8RFQEUZJW

11. Harbord RM, Deeks JJ, Egger M, et al. A unification of models for meta-analysis of diagnostic accuracy studies [published correction appears in *Biostatistics*. 2008;9:779]. *Biostatistics*. 2007;8:239-251.
12. Harbord RM, Whiting P. Metandi: meta-analysis of diagnostic accuracy using hierarchical logistic regression. *Stata J*. 2009;9:211-229.
13. Badoual C, Rousseau A, Heudes D, et al. Evaluation of frozen section diagnosis in 721 parotid gland lesions. *Histopathology*. 2006;49:538-540.
14. Carvalho MB, Soares JM, Rapoport A, et al. Perioperative frozen section examination in parotid gland tumors. *Sao Paulo Med J*. 1999;117:233-237.
15. Hwang SY, Brett RH. An audit of parotidectomy in Singapore: a review of 31 cases. *Med J Malaysia*. 2003;58:273-278.
16. Ishida R, Yamada H, Nishii S. Clinical analysis of parotid tumor [in Japanese]. *Practica Otologica (Kyoto)*. 2003;96:1095-1098.
17. Iwai H, Yamashita T, Izumikawa M, et al. Evaluation of frozen section diagnosis of parotid gland tumors [in Japanese]. *Nippon Jibiinkoka Gakkai Kaiho*. 1999;102:1227-1233.
18. Longuet M, Nallet E, Guedon C, et al. Diagnostic value of fine-needle aspiration biopsy and frozen section examination in the surgery of parotid gland lesions [in French]. *Rev Laryngol Otol Rhinol (Bord)*. 2001;122:51-55.
19. Tew S, Poole AG, Philips J. Fine-needle aspiration biopsy of parotid lesions: comparison with frozen section. *Aust N Z J Surg*. 1997;67:438-441.
20. Zbaren P, Guelat D, Loosli H, et al. Parotid tumors: fine-needle aspiration and/or frozen section. *Otolaryngol Head Neck Surg*. 2008;139:811-815.
21. Zheng T, Holford TR, Chen Y, et al. Are cancers of the salivary gland increasing? experience from Connecticut, USA. *Int J Epidemiol*. 1997;26:264-271.
22. Renshaw AA, Cartagena N, Granter SR, et al. Agreement and error rates using blinded review to evaluate surgical pathology of biopsy material. *Am J Clin Pathol*. 2003;119:797-800.
23. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*. 2002;21:1539-1558.
24. Rea JL. Partial parotidectomies: morbidity and benign tumor recurrence rates in a series of 94 cases. *Laryngoscope*. 2000;110:924-927.
25. Roh JL, Kim HS, Park CI. Randomized clinical trial comparing partial parotidectomy versus superficial or total parotidectomy. *Br J Surg*. 2007;94:1081-1087.
26. Pohar S, Gay H, Rosenbaum P, et al. Malignant parotid tumors: presentation, clinical/pathologic prognostic factors, and treatment outcomes. *Int J Radiat Oncol Biol Phys*. 2005;61:112-118.
27. Bell RB, Dierks EJ, Homer L, et al. Management and outcome of patients with malignant salivary gland tumors. *J Oral Maxillofac Surg*. 2005;63:917-928.
28. Guzzo M, Andreola S, Sirizzotti G, et al. Mucoepidermoid carcinoma of the salivary glands: clinicopathologic review of 108 patients treated at the National Cancer Institute of Milan. *Ann Surg Oncol*. 2002;9:688-695.
29. Lima RA, Tavares MR, Dias FL, et al. Clinical prognostic factors in malignant parotid gland tumors. *Otolaryngol Head Neck Surg*. 2005;133:702-708.

Downloaded from https://academic.oup.com/ajcp/article/136/5/729/1766053 by guest on 09 April 2024

DOI: 10.1309/AJCP2SD8RFQEUZJW
© American Society for Clinical Pathology

HOLOGIC®

First and Only FDA Cleared
Digital Cytology System

Genius™ Cervical AI

Genius™ Review Station

Genius™ Digital Imager

Empower Your Genius With Ours

Make a Greater Impact on Cervical Cancer
with the Advanced Technology of the
Genius™ Digital Diagnostics System

Click or Scan
to discover more

ADS-04159-001 Rev 001 © 2024 Hologic, Inc. All rights reserved. Hologic, Genius, and associated logos are trademarks and/or registered trademarks of Hologic, Inc. and/or its subsidiaries in the United States and/or other countries. This information is intended for medical professionals in the U.S. and other markets and is not intended as a product solicitation or promotion where such activities are prohibited. Because Hologic materials are distributed through websites, podcasts and tradeshows, it is not always possible to control where such materials appear. For specific information on what products are available for sale in a particular country, please contact your Hologic representative or write to **diagnostic.solutions@hologic.com**.

genius™
DIGITAL DIAGNOSTICS