A Systematic Review and Meta-Analysis of the Diagnostic Accuracy of Fine-Needle Aspiration Cytology for Parotid Gland Lesions

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

Key Words: Parotid gland; Fine-needle aspiration; Sensitivity and specificity; Meta-analysis

DOI: 10.1309/AJCPOIE0CZNAT6SQ

Upon completion of this activity you will be able to:
- describe the role of quality assessment in a meta-analysis.
- define the following terms: verification bias, review bias, misclassification bias, timing bias, and bias due to handling of indeterminate results.
- discuss the current state of knowledge regarding the factors affecting the diagnostic performance of fine-needle aspiration cytology for the assessment of salivary gland lesions.
- list the general categories of factors that lead to variation in diagnostic performance.

The ASCP is accredited by the Accreditation Council for Continuing Medical Education to provide continuing medical education for physicians. The ASCP designates this journal-based CME activity for a maximum of 1 *AMA PRA Category 1 Credit*™ per article. Physicians should claim only the credit commensurate with the extent of their participation in the activity. This activity qualifies as an American Board of Pathology Maintenance of Certification Part II Self-Assessment Module.

The authors of this article and the planning committee members and staff have no relevant financial relationships with commercial interests to disclose.

Questions appear on p 152. Exam is located at www.ascp.org/ajcpcme.

Abstract

The clinical usefulness of fine-needle aspiration cytology (FNAC) for the diagnosis of parotid gland lesions is controversial. Many accuracy studies have been published, but the literature has not been adequately summarized.

We identified 64 studies on the diagnosis of malignancy (6,169 cases) and 7 studies on the diagnosis of neoplasia (795 cases).The diagnosis of neoplasia (area under the summary receiver operating characteristic [AUSROC] curve, 0.99; 95% confidence interval [CI], 0.97-1.00) had higher accuracy than the diagnosis of malignancy (AUSROC, 0.96; 95% CI, 0.94-0.97). Several sources of bias were identified that could affect study estimates. Studies on the diagnosis of malignancy showed significant heterogeneity (P < .001). The subgroups of American, French, and Turkish studies showed greater homogeneity, but the accuracy of these subgroups was not significantly different from that of the remaining subgroup.

It is not possible to provide a general guideline on the clinical usefulness of FNAC for parotid gland lesions owing to the variability in study results. There is a need to improve the quality of reporting and to improve study designs to remove or assess the impact of bias.

The value of fine-needle aspiration cytology (FNAC) for diagnosis of parotid gland lesions is controversial. FNAC obviates the need for surgery in up to 33% of patients[1-4] and can provide useful information for surgical planning[5,6]; however, the clinical usefulness of FNAC is questioned because of low sensitivity, variation in reported results, and the belief that most parotid masses require surgery. While FNAC is now a commonplace procedure, some authors, such as Batsakis et al,[7] have suggested that FNAC is only cost-effective in limited circumstances.

FNAC provides information that informs 2 key decisions in patient management. First, FNAC differentiates between neoplastic and nonneoplastic lesions. Neoplastic lesions usually are managed surgically, whereas nonneoplastic lesions are managed conservatively. Second, given a neoplastic lesion, FNAC determines whether the lesion is malignant or benign, which determines the extent of surgery and, in particular, whether the facial nerve can be spared during surgery. The 2 central research questions regarding the clinical usefulness of FNAC for parotid lesions are: (1) Is FNAC sufficiently sensitive to exclude neoplasia and avoid surgery? (2) Is FNAC sufficiently sensitive for malignancy to allow for facial nerve–sparing surgery? The resolution of these questions requires an accurate assessment of the diagnostic performance of FNAC and an understanding of the causes of variation in performance.

Systematic reviews are the cornerstone of evidence-based medicine and provide the basis for the development of guidelines for patient management. Numerous studies on the accuracy of FNAC for the diagnosis of parotid tumors have been published; however, this body of literature has never been

adequately reviewed. While some studies have summarized the results of selected articles,[8-12] the literature has not been subject to a comprehensive systematic review. In addition, the statistical techniques for meta-analysis of diagnostic studies have been developed relatively recently,[13,14] and, as a result, previous summaries have not used modern meta-analytic methods to obtain estimates of diagnostic performance or they have used older methods that have been shown to have deficiencies.[15,16] No previous reviews have included a quality assessment of the literature and examined the potential for bias in study estimates. Finally, previous reviews have focused on the diagnosis of malignancy, and no reviews have examined the accuracy of the diagnosis of neoplasia.

Our objective was to summarize the evidence on the diagnostic accuracy of FNAC for parotid gland tumors using current guidelines for systematic review and meta-analysis of diagnostic studies.[15,17] To that end, we conducted a comprehensive systematic review of the literature and used meta-analytic methods to develop a summary receiver operating characteristic (SROC) curve of diagnostic performance. We also conducted a quality assessment of included articles to explore potential sources of bias and to provide recommendations to improve the reporting of future studies.

Materials and Methods

Literature Search

MEDLINE, EMBASE, and the bibliographies of retrieved articles were searched for studies evaluating diagnostic accuracy of FNAC for parotid lesions published between January 1, 1985, and December 31, 2010, using a sensitive search strategy developed in consultation with an experienced medical reference librarian. Language was not restricted. Scopus was used to perform a "forward search" to obtain articles citing the set of retrieved articles. Our search strategy was broad and included articles on FNAC of head and neck lesions in addition to salivary glands.

Eligibility

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

Inclusion

Eligible studies were independently evaluated by 2 authors (R.L.S. and B.J.H.), and discrepancies were resolved by consensus. Studies were included if they contained extractable data on histologically verified cases involving parotid tumors and provided data that enabled lesions to be classified into broad categories (malignant vs benign and nonneoplastic vs neoplastic).

Because our study was concerned with broad categories of disease, we excluded studies that included only data on the diagnosis of a particular disease entity. We excluded studies using needle core biopsy and included only studies in which the needle size was 20 gauge or smaller (0.60 mm inner diameter). 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). All studies except 1 reported only cases with histologic verification. In that study,[18] there were 67 cases with histologic verification and 4 benign cases with clinical follow-up. We excluded the 4 cases that had only clinical follow-up.

Data Extraction

Data extraction was completed independently by 2 authors (R.L.S. and B.J.H.), and discrepancies were resolved by consensus or by correspondence with study authors. 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 1 author (R.L.S.). Inadequate or indeterminate biopsy results were not counted in the calculation of accuracy. FNAC 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.

Quality Assessment

Quality assessment of articles written in English was conducted using the QUADAS tool.[19] Assessment was completed independently by 2 authors (R.L.S. and A.R.W.). A pilot scoring form was developing and tested on a subset of 10 studies. The scores were compared and used to clarify definitions and identify deficiencies in the form. The revised form was then used to evaluate the full set of studies. The degree of agreement was assessed by using the Cohen κ. The items with discrepant scores were reviewed. Discrepancies due to errors and misinterpretations were corrected. Discrepancies sometimes arose owing to differences in judgment. These items were discussed until a consensus was reached. The consensus approach was required for relatively few items because the initial level of agreement was high (see the "Results" section).

We debated whether to use a consensus approach or use a third evaluator as a tie-breaker. The literature on group decision making provides no clear guidance as to which procedure provides better decisions. In our opinion, the consensus process worked quite well because it required each evaluator to revisit the criteria, recheck the data used to make a judgment,

DOI: 10.1309/AJCPOIE0CZNAT6SQ

and reevaluate the decision in the light of the other evaluator's comments. Thus, the consensus process provided stringent quality control on discrepant votes. In contrast, a tie-breaker would be determined by the third vote with no guarantee of quality (the tie-breaker could assign scores randomly).

Statistical Analysis

SROC curves were developed by using the hierarchical method[13,14] to construct the curves. Computations were done using Stata 10 (StataCorp, College Station, TX) and the metandi procedure for SROC curve analysis.[20]

Results

Literature Search

We screened 3,848 titles and abstracts to obtain a set of 551 eligible articles. The reports of the eligible studies were screened and resulted in 64 studies that met our inclusion criteria. These 64 studies contained 64 data sets on the diagnostic accuracy of FNAC for the assessment of malignancy vs benignancy **Table 1**[5,6,9-12,18,21-77] with a total of 6,169 cases. There were 7 data sets for the assessment of neoplastic vs nonneoplastic lesions **Table 2** with a total of 795 cases.

Study Characteristics

We collected data on the setting (academic vs community), period of the study, location (country), study design (prospective vs retrospective), method (experience of pathologist and whether samples were immediately assessed by a pathologist), population characteristics (age, sex, unusual referral patterns and comorbidities), and potential sources of bias (blinding, percent verification, inadequacy rates, and indeterminate diagnoses). No studies were performed in a community setting. The publication rate increased during the period with half of the studies published in the last 8 years **Figure 1**. The locations are summarized in Table 1. The largest number of studies (9 of 64) were conducted in the United States. All studies were retrospective with the exception of 2 prospective studies.[22,34] No studies reported the experience of the pathologist. Most studies (46 of 64) did not specify who obtained the sample; 7 studies specified that samples were obtained by a pathologist, and 10 studies indicated that specimens were obtained by nonpathologists (clinicians, surgeons, or radiologists), and 1 study specified that samples were taken by both pathologists and nonpathologists. Approximately two thirds of the studies reported summary statistics on age and sex distributions of patients, and all those reporting were similar. Only 2 studies[45] were blinded.

A total of 451 inadequate and 79 indeterminate FNAC results were reported (Table 1). These cases, representing 8.6% of the total, were excluded from the analysis. In addition, there were 37 FNAC results reported as suspicious and 14 results reported as atypical. Suspicious and atypical results were reclassified as malignant and accounted for 4.2% of the malignant lesions.

Diagnostic Accuracy

Malignant vs Benign Lesions

The SROC curve for malignant vs benign lesions is shown in **Figure 2**. The area under the ROC curve (AUROC) was 0.96 (95% confidence interval [CI], 0.94-0.97). The summary estimates for the sensitivity and specificity were 0.80 (95% CI, 0.76-0.83) and 0.97 (95% CI, 0.96-0.98), respectively. The positive likelihood ratio was 28.6 (95% CI, 20.5-39.8), and the negative likelihood ratio was 0.21 (95% CI, 0.17-0.25). There was significant heterogeneity among studies ($P < .001$). The prevalence of malignant disease was 25%, the positive predictive value was 0.90, and the negative predictive value was 0.94.

We used the data on study characteristics (described in a preceding section) to investigate sources of heterogeneity. In general, the data were insufficient (eg, experience of the pathologist or immediate assessment) or had too little variability (eg, sex distribution, age distribution, characteristics, study design, and blinding) to allow for statistical tests. We hypothesized that study results might vary by time and compared studies completed before 2000 with those completed after 2000; however, there was no significant difference between these 2 groups (pre-2000 AUROC, 0.98; 95% CI, 0.97-0.99; post-2000 AUROC, 0.97; 95% CI, 0.95-0.98]. The results for both groups (ie, pre-2000 vs post-2000) showed significant heterogeneity ($P < .001$). We also tested whether results varied by inadequacy rates but found no significant correlation. Finally, we examined whether the heterogeneity could be caused by a small set of outliers. To that end, we sequentially dropped studies in order of their contribution to the Cochrane Q statistic until we obtained a homogeneous set. It was necessary to drop more than 20 studies before a homogeneous subset could be obtained and no underlying theme for the homogeneous subset could be identified.

We hypothesized that results might vary by location and found that the studies conducted in the United States formed a homogeneous subgroup ($P < .28$). The SROC curve for this group is given in **Figure 3**. The diagnostic accuracy of the American studies was not significantly different from the remaining group of studies **Table 3**. We found that studies from France and Turkey also formed homogeneous groups, while groups composed of studies from Australia, Japan, and Italy were heterogeneous. We tested whether we might be able to form a larger homogeneous subgroup by combining the results from high-income countries. To that end, we compared the diagnostic

DOI: 10.1309/AJCPOIE0CZNAT6SQ

Table 1
Included Studies for Diagnosis of Malignant vs Benign Lesions

Study	No.	TP	FP	FN	TN	ND	Suspicious*	Sensitivity	Specificity	Location
1/Akbas et al, 2004[21]	82	16	1	1	64	0	0	0.94	0.98	Turkey
2/Al Salamah et al, 2005[22]	33	5	0	0	28	0/4	0	1.00	1.00	Saudi Arabia
3/Al-Khafaji et al, 1998[23]	150	61	11	13	65	10/4	8	0.82	0.86	US
4/Atula et al, 1996[24]	204	23	9	21	151	3	20	0.52	0.94	Turkey
5/Aversa et al, 2006[25]	310	34	0	7	269	NR	0	0.83	1.00	Italy
6/Awan and Ahmad, 2004[26]	50	7	1	3	39	NR	U	0.70	0.98	Pakistan
7/Bartels et al, 2000[27]	43	17	3	1	22	5	U	0.94	0.88	US
8/Behbehani et al, 1990[18]	85	16	0	5	64	NR	U	0.76	1.00	Kuwait
9/Behzatoglu et al, 2004[28]	67	11	1	1	54	4	0	0.92	0.98	Turkey
10/Berrone et al, 1995[29]	352	145	20	7	180	10	0	0.95	0.90	Italy
11/Brennan et al, 2010[30]	103	16	5	7	75	17/16	0	0.70	0.94	England
12/Buhler et al, 2007[31]	58	0	3	3	52	NR	5	0.00	0.94	Brazil
13/Burgess and Serpell, 2008[32]	72	6	4	2	60	NR	0	0.75	0.94	Australia
14/Califano et al, 1992[33]	60	9	0	0	51	NR	0	1.00	1.00	Italy
15/Carrillo et al, 2009[34]	135	60	1	5	69	3	0	0.92	0.99	Mexico
16/Contucci et al, 2003[35]	139	12	0	9	118	6	0	0.57	1.00	Italy
17/Costas et al, 2000[36]	80	17	3	3	57	0	0	0.85	0.95	Spain
18/Deans et al, 1995[37]	23	2	1	1	19	2	0	0.67	0.95	Northern Ireland
19/Deneuve et al, 2010[38]	78	7	4	0	67	7	0	1.00	0.94	France
20/Filho et al, 2001[39]	174	17	3	6	148	18	U	0.73	0.98	Brazil
21/Filopoulos et al, 1998[40]	124	37	1	2	84	3	2	0.94	0.99	Greece
22/Friese, 2005[10]	164	30	3	4	127	U	U	0.88	0.98	Germany
23/Gete Garcia et al, 2006[41]	128	26	3	6	93	11	0	0.81	0.97	Spain
24/Gobić et al, 2010[42]	176	13	13	3	147	NR	0	0.81	0.92	Croatia
25/Gooden et al, 2002[43]	87	16	8	2	61	57	0	0.89	0.88	Canada
26/He et al, 2003[44]	121	31	0	9	81	U	U	0.78	1.00	China
27/Herrera Hernandez et al, 2008[45]	46	7	3	6	30	NR	0	0.54	0.91	Columbia
28/Inohara et al, 2008[46]	81	19	2	3	57	U	U	0.86	0.97	Japan
29/Jafari et al, 2009[47]	101	12	3	6	80	9	0	0.67	0.96	France
30/Kamal and Othman, 1997[48]	18	8	3	1	6	NR	0	0.89	0.67	Saudi Arabia
31/Kaur et al, 1993[49]	24	4	0	0	20	4	0	1.00	1.00	Singapore
32/Knudsen et al, 1985[50]	166	23	8	5	130	U	U	0.82	0.94	Norway
33/Kondo et al, 2007[51]	17	2	0	4	11	U	U	0.33	1.00	Japan
34/Lim et al, 2007[52]	81	8	0	2	71	10	0	0.80	1.00	Singapore
35/Lin and Bhattacharyya, 2007[5]	22	17	0	4	1	5	0	0.81	1.00	US
36/Longuet et al, 2001[53]	102	11	2	3	86	12	0	0.79	0.98	France
37/Lurie et al, 2002[54]	41	4	0	5	32	11	0	0.44	1.00	Israel
38/Malata et al, 1997[55]	16	14	0	2	0	4	0	0.88	—	England
39/Altuna Mariezkurrena et al, 2006[56]	34	4	1	1	28	U	0	0.80	0.97	Spain
40/Marrazzo et al, 1993[57]	37	3	0	1	33	U	0	0.75	1.00	Italy
41/Mianroodi et al, 2006[58]	53	17	3	5	28	9	0	0.78	0.90	Australia
42/Mohammed et al, 2008[59]	189	21	6	14	148	22	0	0.60	0.96	Canada
43/Ortega et al, 2000[11]	60	27	2	6	25	37	0	0.82	0.93	Mexico
44/Osanai et al, 2003[60]	36	4	0	6	26	U	U	0.40	1.00	Japan
45/Paris et al, 2005[61]	133	25	5	6	97	0/15	0	0.81	0.95	France
46/Pons Rocher et al, 2003[62]	118	16	6	13	83	U	U	0.55	0.93	Spain
47/Que Hee and Perry, 2001[63]	155	27	0	20	108	17	0	0.58	1.00	Australia
48/Riley et al, 2005[64]	97	31	3	2	61	NR	0	0.94	0.95	NZ
49/Rodriguez et al, 1989[65]	46	11	1	2	32	18	0	0.85	0.97	US
50/Schelkun and Grundy, 1991[66]	8	4	0	0	4	2	U	1.00	1.00	US
51/Schroder et al, 2000[12]	284	27	2	2	253	U	U	0.93	0.99	Germany
52/Seethala et al, 2005[67]	208	54	12	9	133	12	0/14	0.86	0.91	US
53/Shashinder et al, 2009[68]	70	16	2	5	47	6	0	0.76	0.95	Malaysia
54/Sonmez, 2005[9]	78	9	1	4	64	NR	0	0.69	0.98	Turkey
55/Takashima et al, 1999[69]	24	12	0	2	10	2	0	0.86	1.00	Japan
56/Tew et al, 1997[70]	129	18	0	2	109	29/37	0	0.90	1.00	Australia
57/Tsai and Hsu, 2002[71]	40	3	1	2	34	NR	0	0.60	0.97	Taiwan
58/Uğuz et al, 2007[72]	29	7	1	5	16	0	2	0.58	1.00	Turkey
59/Upton et al, 2007[73]	53	20	2	2	29	10	0	0.91	0.93	US
60/ Van Lierop et al, 2007[74]	67	8	1	3	55	45	0	0.73	0.98	South Africa
61/Weinberger et al, 1992[75]	47	11	3	3	30	2	0	0.79	0.91	US
62/Zafar et al, 1997[76]	28	8	1	2	17	0	0	0.80	0.94	Pakistan
63/Zbaren et al, 2008[6]	110	50	5	18	37	6	0	0.74	0.88	US
64/Zurrida et al, 1993[77]	223	31	0	14	178	23	0	0.69	1.00	Italy

FN, false-negative; FP, false-positive; ND, nondiagnostic (No. inadequate or No. inadequate/No. indeterminate); NR, not reported; NZ, New Zealand; TN, true-negative; TP, true-positive; U, unknown; UK, United Kingdom; US, United States.
* Given as No. "suspicious" or No. suspicious/No. inadequate.

DOI: 10.1309/AJCPOIE0CZNAT6SQ

Table 2
Included Studies for Diagnosis of Neoplastic vs Nonneoplastic Lesions

Study	No.	TP	FP	FN	TN	ND	Sensitivity	Specificity	Location
1/Atula et al, 1996[24]	204	135	8	35	26	10	0.79	0.76	Turkey
2/Behzatoglu et al, 2004[28]	67	63	0	0	4	4	1.00	1.00	Turkey
3/Buhler et al, 2007[31]	58	42	0	9	7	NR	0.82	1.00	Brazil
4/He et al, 2003[44]	121	73	0	7	41	NR	0.91	1.00	China
5/Lim et al, 2007[52]	81	72	2	2	5	10	0.97	0.71	Singapore
6/Lurie et al, 2002[54]	41	31	0	5	5	11	0.86	1.00	Israel
7/Zurrida et al, 1993[77]	223	204	0	0	19	23	1.00	1.00	Italy

FN, false-negative; FP, false-positive; ND, nondiagnostic/inadequate; NR, not reported; TN, true-negative; TP, true-positive.

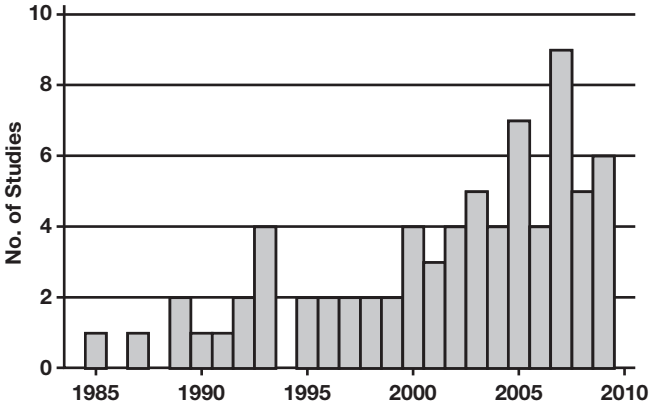

Figure 1 Number of studies by year of publication.

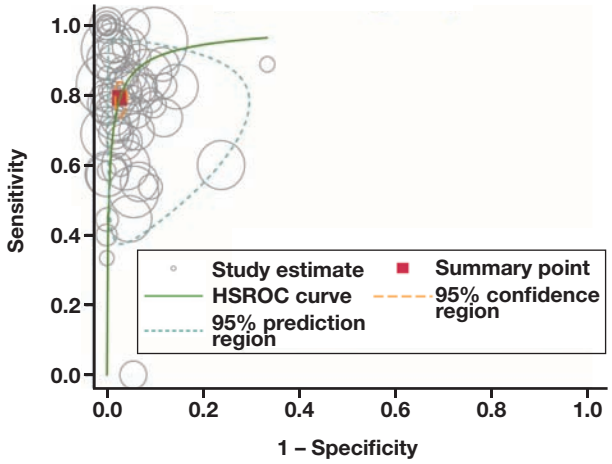

Figure 2 Hierarchical summary receiver operating characteristic (HSROC) curve for the diagnosis of malignancy. Each circle represents a study. The size of the circle is proportional to the weight given to the study in the final accuracy estimate.

accuracy of studies performed in high-income OECD (Organisation for Economic Co-operation and Development) countries[78] (AUROC, 0.097; 95% CI, 0.95-0.98) with the accuracy of those completed in non–high-income OECD countries (AUROC, 0.97; 95% CI, 0.95-0.98). The results for both groups (ie, OECD30 vs non-OECD30) showed significant heterogeneity ($P < .001$).

Neoplastic vs Nonneoplastic Lesions

The SROC curve for neoplastic vs nonneoplastic lesions is shown in **Figure 4**. The AUROC curve was 0.99 (95% CI, 0.98-1.00) and showed no significant heterogeneity ($P < .09$). The summary estimate for sensitivity was 0.96 (95% CI, 0.83-0.99). The summary estimate for specificity was 0.98 (95% CI, 0.67-1.00). The positive likelihood ratio was 58.0 (95% CI, 2.0-1,651.9), and the negative likelihood ratio was 0.04 (95% CI, 0.01-0.18). The prevalence of neoplastic disease was 85%, which gives a positive predictive value of 1.00 and a negative predictive value of 0.81.

Quality Assessment

The results of the interrater agreement study are given in **Table 4**. Using the criteria of Fleiss,[79] the degree of agreement ranged from good to excellent. There was no disagreement on

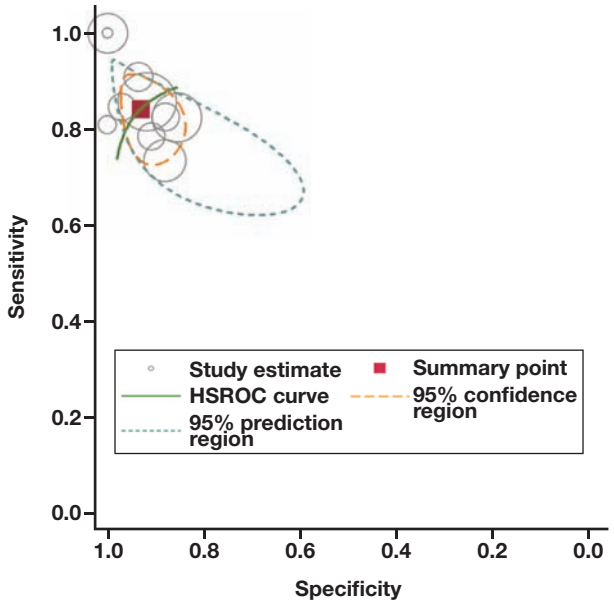

Figure 3 Hierarchical summary receiver operating characteristic (ROC) curve for diagnosis of malignancy for American studies.

DOI: 10.1309/AJCPOIE0CZNAT6SQ

Table 3
Comparison of American Studies With Non-American Studies for Malignant vs Benign Lesions*

	American Studies (n = 9)	Non-American Studies (n = 55)
Area under the curve	0.94 (0.91-0.96)	0.97 (0.95-0.98)
Inconsistency, I^2 (%)	<1 (not determined)	97 (95-99)
Log rank Q, df, *P*	0.05, 2, .49	65.9, 2, .00
Sensitivity	0.83 (0.77-0.88)	0.79 (0.74-0.83)
Specificity	0.93 (0.87-0.93)	0.97 (0.97-0.98)
Positive likelihood ratio	8.8 (6.3-12.3)	33.7 (23.4-48.4)
Negative likelihood ratio	0.19 (0.13-0.26)	0.22 (0.18-0.27)
Disease prevalence	0.43	0.23
Positive predictive value	0.90	0.89
Negative predictive value	0.89	0.94

* Values in parentheses are 95% confidence intervals.

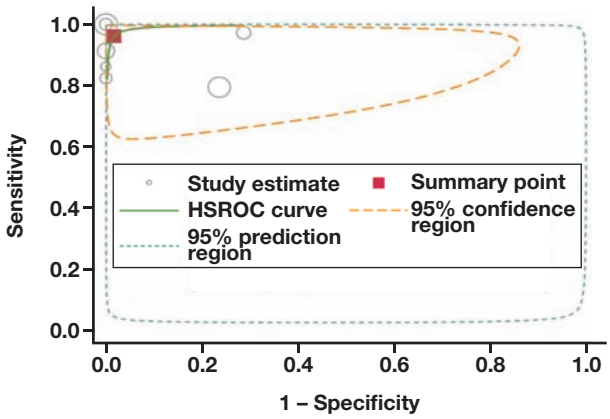

Figure 4 Hierarchical summary receiver operating characteristic (ROC) curve for diagnosis of neoplasia. Each circle represents a study. The size of the circle is proportional to the weight given to the study in the final accuracy estimate.

QUADAS items 3 through 7, 9 through 12, and 14 (which did not vary by study), and these were not counted in the rater agreement study.

The results of the quality assessment are given in **Table 5**. The QUADAS survey questions and information about our findings are as follows:

Item 1: Were the selection criteria clearly described? Most studies (36/56) clearly specified that all cases within a particular period were included. In some cases (17/56), it seemed likely that consecutive cases were included; however, it was not clearly stated. In 3 studies, the selection criteria were not clear and seemed not to include consecutive cases.

Item 2: Was the spectrum of patients representative of the patients who will receive the test in practice? We scored this item as negative if the patients were drawn from an unusual referral pattern (eg, patients receiving magnetic resonance imaging [MRI] and FNA). We scored the item as positive if it seemed likely that consecutive patients were included, no unusual referral pattern was mentioned, and if a summary of patient demographics was provided showing that the population was broadly similar to the overall population (age and sex). We scored 34 of 59 studies as positive, 18 as unclear, and 4 as negative. We found no difference in the accuracy of studies in which the spectrum of patients was representative compared with those in which the accuracy was unclear or possibly not representative.

Item 3: Is the reference standard likely to correctly classify the target condition? The reference standard (H&E histologic findings) was standard across all studies. The reference standard is imperfect and gives rise to misclassification errors, and the error rate would likely vary across studies according to the skill of pathologists. We searched the literature but were unable to find any studies on interobserver variation or accuracy studies for the diagnosis of salivary gland tumors. This item was scored unclear for all studies.

Item 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? The period between FNAC and a definitive histologic diagnosis was not specified in any study; however, if indicated, the standard practice is to perform surgery within a relatively short period (eg, 1 month) following FNAC. It is unlikely that most tumors would undergo significant change during that period. This item was scored positive for all studies.

Item 5: Did the whole sample or a random selection of the sample receive verification using the reference standard of diagnosis? Complete or random verification was not used in any study. This item was scored negative for all studies.

Item 6: Did patients receive the same reference standard regardless of the index test result? By design, we included

Table 4
Rater Agreement for QUADAS Scoring[19]

QUADAS Item	Agreement (%)	Expected Agreement (%)	κ	SE (κ)	Z	*P* (κ > Z)	Qualitative Agreement*
1	98.0	73.43	0.93	0.12	7.7	.0	Excellent
2	99.0	71.4	0.96	0.12	8.2	.0	Excellent
8	74.5	54.8	0.44	0.10	4.4	.0	Good
13	82.4	55.0	0.61	0.12	5.2	.0	Good

* According to the Fleiss criteria.[79]

DOI: 10.1309/AJCPOIE0CZNAT6SQ

Table 5
Summary of QUADAS Quality Survey[19] Results*

QUADAS Item	Description	Yes	Unclear	No
1	Were the selection criteria clearly described?	35	17	3
2	Was the spectrum of patients representative of the patients who will receive the test in practice?	34	18	4
3	Is the reference standard likely to correctly classify the target condition?	0	56	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?	56	0	0
5	Did the whole sample or a random selection of the sample receive verification using the reference standard of diagnosis?	0	0	56
6	Did patients receive the same reference standard regardless of the index test result?	56	0	0
7	Was the interpretation of the reference standard independent of the index test (ie, the index test did not form part of the reference standard)?	56	0	0
8	Was the execution of the index test described in sufficient detail to permit replication of the test?	12	18	26
9	Was the execution of the reference standard described in sufficient detail to permit replication?	56	0	0
10	Were the index results interpreted without knowledge of the results of the reference standard?	56	0	0
11	Were the reference standard results interpreted without knowledge of the results of the index test?	0	0	56
12	Were the same clinical data available when test results were interpreted as would be available when the test was used in practice?	56	0	0
13	Were uninterpretable or intermediate results reported?	30	15	11
14	Were withdrawals from the study explained?	0	0	56

* Quality assessment was completed on 56 English-language articles.

only histologically verified cases, so all cases were verified by the same reference standard; however, in all studies, different proportions of cases were verified depending on the result of the index test (see item 5). This item was scored positive for all studies.

Item 7: Was the interpretation of the reference standard independent of the index test (ie, the index test did not form part of the reference standard)? The reference standard is independent of the index test. This item was scored positive for all studies.

Item 8: Was the execution of the index test described in sufficient detail to permit replication of the test? We scored this as positive if the size of needle and the number of passes were described and if it was indicated whether a pathologist or cytopathologist was immediately available to assess the adequacy of the specimen. Many studies omitted even the most basic description of the procedure (eg, needle size), and relatively few studies indicated whether a pathologist was available to assess adequacy.

Item 9: Was the execution of the reference standard described in sufficient detail to permit replication? No studies described the reference standard in detail; however, the preparation of histologic slides is standard, and there is little reason to believe there is significant variation across studies. Other items such as experience level of the pathologist could have a bearing but were not reported. We scored this item as positive for all studies.

Item 10: Were the index results interpreted without knowledge of the results of the reference standard? This was scored positive for all studies.

Item 11: Were the reference standard results interpreted without knowledge of the results of the index test? It is standard practice for pathologists to be aware of FNAC results when evaluating histologic slides. In practice, histologic findings are weighted more heavily than FNAC, and, although FNAC findings influence the final diagnosis, we view the influence to be relatively minor. This item was scored as negative for all studies.

Item 12: Were the same clinical data available when test results were interpreted as would be available when the test was used in practice? All the included studies were retrospective with cases drawn from standard practice in which clinical data are available. This item was scored positive for all studies.

Item 13: Were uninterpretable or intermediate results reported? We found the reporting of uninterpretable results to be quite variable. Some studies reported this aspect in detail, whereas other studies made no mention of such results. Thus, in such cases, it was impossible to determine whether indeterminate/uninterpretable results were found and, if so, how they were handled. We compared the diagnostic accuracy of studies in which indeterminate results were reported with those in which such results were not reported and found no difference ($P > .05$).

Item 14: Were withdrawals from the study explained? All of the included studies were retrospective with patients selected from surgery rosters. Thus, patients who were scheduled but did not undergo surgery were not included. This item was scored negative for all studies.

Discussion

Our results show that FNAC for parotid gland lesions has high specificity, and, although the sensitivity is good, the technique shows greater specificity than sensitivity. This result is

DOI: 10.1309/AJCPOIE0CZNAT6SQ

consistent with the findings of other reviews.[8,80,81] We also found that studies showed much more variability in sensitivity (SD = 0.18) than specificity (SD = 0.06).

We found that the accuracy of diagnosis for neoplasia was significantly higher than the accuracy of diagnosis for malignancy. This result was consistent for the American and non-American subgroups of studies. This is an important finding because the decision to recommend surgery generally depends on a diagnosis of neoplasia.

The American group of studies was homogeneous, and it may be possible to use the summary statistics developed herein to develop guidelines for the American setting. In contrast, the non-American group of studies on malignant lesions showed significant heterogeneity. Given the wide variation in results, it is difficult to develop general guidelines for the use of FNAC in the non-American group because the predictive value of FNAC will vary by setting. In addition, the summary statistics should be interpreted with caution owing to the heterogeneity. It is important to understand the factors that led to performance variability in order to increase consistency and improve overall performance. The knowledge gained from the study of heterogeneity in the non-American subgroup of studies can be used to improve performance in both subgroups. Indeed, the knowledge gained from the study of heterogeneity can be one of the most important benefits of a meta-analysis.

Variation in diagnostic performance can be due to 4 sources[82]: real differences in test conditions (eg, differences in population and methods), random variation, threshold effects, and bias related to study design. Our results (ie, significant heterogeneity) suggest that the variability cannot be explained by random variation.

Study Heterogeneity

Sources of Differences Between Studies

There are a large number of study factors that can give rise to differences in diagnostic performance.[83] These factors are encapsulated in the acronym PICO that, in the context of diagnostic testing, stands for population, index test, comparator, and outcome measure. Differences in any of these factors can lead to differences in diagnostic performance. In general, understanding causes of real performance variation requires one to study correlations between study-wide factors (PICO) and performance. To that end, we investigated a range of study-wide factors (described in the "Results" section).

Differences in FNAC methods are another factor that could lead to differences in study results. For example, it has been shown that having a pathologist on site for immediate evaluation of specimen adequacy improves the diagnostic yield[84-86]; however, few studies in our survey documented whether a pathologist was present. Similarly, other factors such as needle size, number of samples obtained, and the experience level of the pathologist could contribute to variability of results but are often not reported. We suggest that researchers report the following: needle size, number of passes, whether the sample was guided by imaging and the type of imaging, whether a pathologist was available for immediate evaluation of specimen adequacy, the number of pathologists involved in the study (ie, whether the accuracy statistics represent the average performance of a group of pathologists or 1 pathologist), and the staining technique used.

Population differences present another possible factor that could lead to real differences in diagnostic performance. Referral patterns are a common source of population differences because the spectrum of disease is dependent on the referral pattern. For example, the spectrum of parotid disease in patients referred for MRI and FNAC might differ from the spectrum of disease in patients who received FNAC only because the subpopulation referred for MRI may involve more complex cases. Thus, diagnostic performance could differ between studies owing to differences in the referral pattern and the resulting differences in the spectrum of disease. It is important that researchers report the referral pattern and the selection criteria, clearly state whether consecutive cases were selected within a specified period, and provide a statistical summary of the population (age and sex distribution).

We attempted to collect data on many study-wide factors; however, we found that such data were often poorly reported, and, as a consequence, it was not possible to complete many of the analyses we had planned.

Threshold Differences

Study differences can result from differences in the criteria used to assign cases to categories (benign vs malignant). Given a set of criteria, 2 pathologists may detect the same features in a case; however, one may use more stringent criteria than another in assigning malignancy. Such differences are due to threshold effects and should be distinguished from differences in accuracy. Threshold effects reflect a tradeoff between sensitivity and specificity with constant overall accuracy. Differences in accuracy reflect differences in the ability to correctly identify and interpret features. In general, differences in accuracy result in variation in the direction perpendicular to the summary average of the SROC curve, whereas differences in threshold (constant accuracy) result in variation along the SROC curve, as shown in **Figure 5**. We used a statistical procedure (midas, Stata 10) to estimate the degree of heterogeneity due to threshold effects; however, only a small percentage was estimated to be due to this factor. Threshold effects have been shown to account for a significant fraction of interrater disagreement.[87] It may be possible to reduce performance variation by more uniform application of diagnostic criteria.

DOI: 10.1309/AJCPOIE0CZNAT6SQ

Sources of Bias

Bias is a second source of potential variation that could have contributed to the heterogeneity in studies. Indeed, our quality survey highlighted several potential sources of bias in this collection of studies.

Verification Bias.—Because of their retrospective design, all studies have this source of bias; however, there is more potential for verification bias to affect the diagnosis of neoplasia more than the diagnosis of malignancy because the degree of differential verification is higher in the diagnosis of neoplasia compared with the diagnosis of malignancy. Almost all cases with a finding of neoplasia are verified by histologic examination, whereas only a small subset of the nonneoplastic cases receives histologic verification. In addition, the subset of nonneoplastic cases that proceeds to surgery is unlikely to be representative of the nonneoplastic cases. Both of these factors are likely to bias the sensitivity and specificity in the diagnosis of neoplasia. In contrast, when diagnosing malignancy, almost all cases proceed to surgery and receive histologic verification. Thus, verification bias is less of an issue for the diagnosis of malignancy than for the diagnosis of neoplasia. The effect of verification bias can be estimated as follows:

Let Sn, and Sp = the actual sensitivity and specificity; Sn' and Sp' = the apparent sensitivity and specificity; α = the sampling fraction of positive cases; β = the sampling fraction of negative cases; $r = \alpha/\beta$ = the relative sampling fraction of positive to negative cases; and TP', FN', TN', and FP' = the observed true-positives, false-negatives, true-negatives, and false-positives.

Then:

Equation 1

$$Sn' = \frac{TP'}{TP' + FN'} = \frac{\alpha Sn}{\alpha Sn + \beta(1 - Sn)} = \frac{rSn}{1 - (1 - r)Sn}$$

Equation 2

$$Sp' = \frac{TN'}{TN' + FP'} = \frac{\beta Sp}{\beta Sp + \alpha(1 - Sp)} = \frac{Sp}{r + (1 - r)Sp}$$

From which we obtain:

Equation 3

$$Sn = \frac{Sn'}{r + (1 - r)Sn'}$$

Equation 4

$$Sp = \frac{rSp'}{1 - (1 - r)Sp'}$$

Relations 1 and 2, showing the effect of verification bias on sensitivity and specificity are presented in **Figure 6** and **Figure 7**, respectively. For the diagnosis of neoplasia, we would expect the relative sampling fraction, r, to be quite high, say on the order of 5 to 10 (ie, more positive samples than negative samples receive histologic verification). Under these conditions, the apparent sensitivity would be biased upward and the apparent specificity would be biased downward. For example, using values of r = 10, Sn' = 0.96, and Sp' = 0.98 in equations 3 and 4, we obtain estimates of 0.71 and 1.00 for the

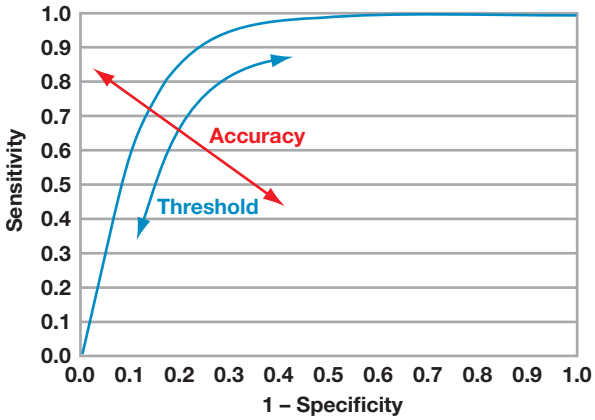

Figure 5 Comparison of accuracy vs threshold effects on the summary receiver operating characteristic (SROC) curve. Variation along the SROC curve represents studies with equal accuracy with differing thresholds for malignancy (ie, a tradeoff between sensitivity and specificity). Variation perpendicular to the SROC curve represents differences in accuracy (differences in the ability to detect or correctly interpret case features).

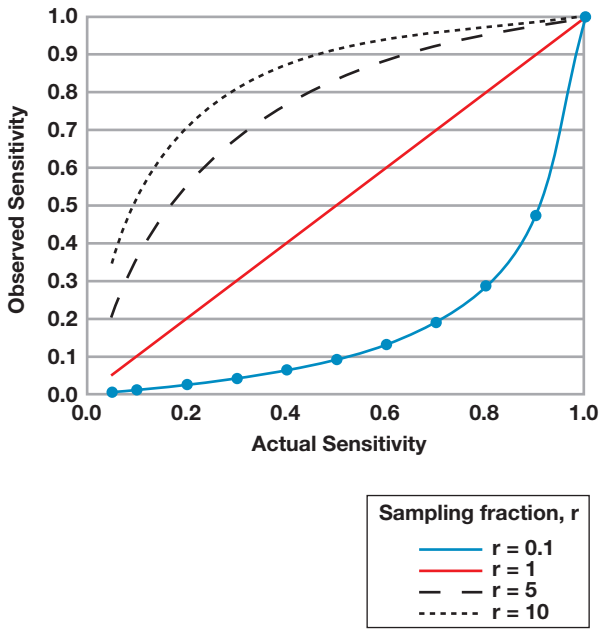

Figure 6 The effect of verification bias on the observed vs actual sensitivity for diagnosis of malignancy. The sampling fraction, r, is the relative proportion of malignant to benign cases that receive histologic verification.

DOI: 10.1309/AJCPOIE0CZNAT6SQ

actual sensitivity and specificity, respectively. We expect the sampling fraction of malignant lesions to be similar to that of nonmalignant lesions because most neoplasms are managed with surgery. For example, using our results of Sn' = 0.79, Sp' = 0.96 for diagnosis of malignant lesions and assuming the relative sampling fraction of malignant to nonmalignant lesions is r = 1.2 and using equations 3 and 4, we obtain estimates of 0.76 and 0.97 for the actual sensitivity and specificity, respectively. These results are summarized in **Table 6**, which shows that our estimate for the sensitivity of diagnosis of neoplasia may be biased upward owing to verification bias but that bias is unlikely to be a major factor in our other estimates.

We conducted a simulation to determine the effect of verification bias on our summary estimates obtained from our SROC curve. We varied the sampling fraction, r, and at each level of r recalculated the entries to Tables 1 and 2 that would have been obtained without verification bias. At each level of r, we reran our analysis and obtained the summary estimates for the 2 homogeneous groups: malignant vs benign (American group) and neoplastic vs nonneoplastic. The results **Table 7** show that the summary estimate of sensitivity is most likely to be affected by verification bias. As discussed previously, we believe that verification bias is more of an issue for the diagnosis of neoplasia than for malignancy.

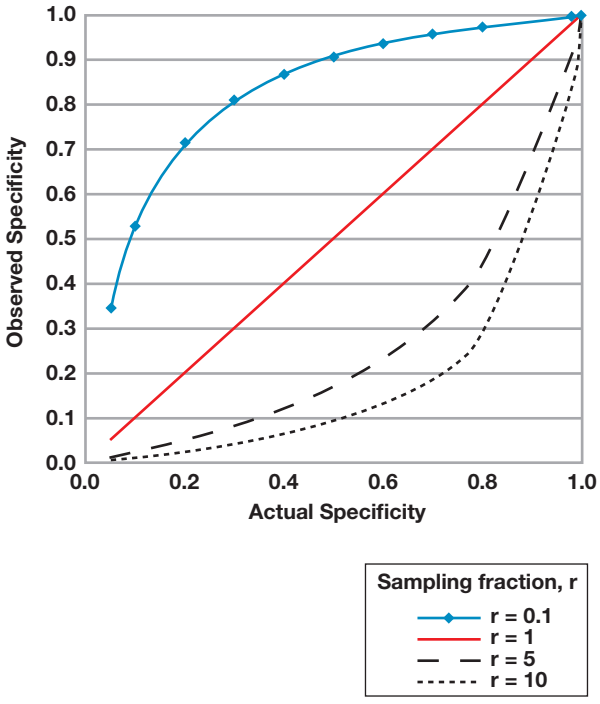

Figure 7 The effect of verification bias on the observed vs actual specificity for diagnosis of neoplasia. The sampling fraction, r, is the relative proportion of neoplastic to nonneoplastic fine-needle aspiration cytology diagnoses that receive histologic verification.

It is worth noting that sensitivity and specificity are not the only summary statistics that are affected by verification bias. The likelihood ratios and predictive values are also affected by verification bias because they are functions of sensitivity and specificity.

Resolution of verification bias would require studies with follow-up of patients who receive a diagnosis of a nonneoplastic lesion by FNAC. While it is not practical to follow up all nonneoplastic cases with histologic studies, patients can be followed up with an alternative method of verification such as clinical follow-up. In our survey, such follow-up was rarely done and, if done at all, the follow-up of negative cases is poorly documented. The study by Rapkiewicz et al[88] is an exception and provides a good example of complete documentation of negative cases. In general, studies need to document the flow of patients through the system as recommended by the STARD initiative guidelines.[89,90] Specifically, it is important to know how many patients undergo evaluation of a suspected salivary gland lesion, how many underwent FNAC and, how many underwent surgery. These data are required to provide the predictive value of FNAC that, from a clinical perspective, is the key performance measure.

Review Bias.—Only 2 of the studies in this collection were blinded. Thus, the results of FNAC were known when

Table 6
The Estimated Effect of Verification Bias in Summary Estimates of Sensitivity and Specificity*

	Diagnosis of Neoplasia (r = 10)		Diagnosis of Malignancy (r = 1.2)	
	Apparent	Estimated Actual	Apparent	Estimated Actual
Sensitivity	0.96	0.71	0.79	0.76
Specificity	0.98	1.00	0.96	0.97

* Calculations based on equations 3 and 4 (see the text). The sampling fraction, r, is the relative proportion of neoplastic to nonneoplastic or malignant to benign fine-needle aspiration cytology diagnoses that receive histologic verification.

Table 7
The Effect of Verification Bias on the Summary Receiver Operating Characteristic Curve Estimates

Sampling fraction, r*	Neoplastic vs Malignant (American Studies)		Nonneoplastic vs Neoplastic	
	Sensitivity	Specificity	Sensitivity	Specificity
1	0.83	0.93	0.96	0.98
1.5	0.72	0.96	0.95	1.00
2	0.64	0.97	0.94	1.00
4	0.45	0.99	0.90	1.00
5	—	—	0.89	1.00
10	0.29	1.00	0.84	1.00

* The sampling fraction, r, is the relative proportion of benign to malignant or nonneoplastic to fine-needle aspiration cytology diagnoses that receive histologic verification.

DOI: 10.1309/AJCPOIE0CZNAT6SQ

the reference test was conducted and the knowledge of the FNAC results could influence the interpretation of histologic slides. If such a bias exists, it would tend to increase the correlation between FNAC and histologic findings and inflate the estimates of sensitivity and specificity. Because histologic findings are weighted more heavily than FNAC findings, we do not believe that review bias is likely to have a large impact; however, future studies should use blinding to remove this source of bias.

Misclassification Bias.—The accuracy of the reference standard (definitive histologic diagnosis) is another potential source of bias. Few data are available on the levels of interrater agreement in the diagnosis of salivary gland tumors, so the level of error and the types of error (differential vs nondifferential misclassification) are unknown. Nondifferential misclassification occurs when the error rate of the reference test (definite histologic diagnosis) is independent of the result of the index test. We believe that misclassification errors are most likely nondifferential.

The potential impact of nondifferential misclassification can be seen by investigating the effect of the misclassification rate on the summary sensitivity and specificity using the totals obtained from our survey. For example, our survey found a total of 1,227 true-positives, 311 false-positives, 177 false-negatives, and 4,454 true-negatives for the diagnosis of malignancy **Table 8**. Similar totals are shown for diagnosis of neoplasia in **Table 9**. The effect of nondifferential classification on the sensitivity and specificity is shown in **Figure 8** and **Figure 9**.

The impact of misclassification depends on the distribution of cases in the 2 × 2 table, and, for that reason, the impact is different for malignancy than for neoplasia. In the case of malignancy, the proportion of true-negative cases is high, and misclassification causes a downward bias on sensitivity because the net effect of misclassification is to move cases from the true-negative category to the false-negative category. In the case of neoplasia, the proportion of true-positive cases is high, and misclassification causes true-positives to be misclassified as false-positives that, in turn, cause a downward bias in specificity. Renshaw et al[87] found a misclassification rate of approximately 3% in a range of surgical specimens, and our calculations show that relatively small misclassification rates (say 1%) can cause significant bias. It would be helpful to obtain an estimate of the misclassification rate for salivary gland lesions through future studies.

Bias Due to Handling of Indeterminate and Inadequate Results.—We found the reporting of data on inadequate and indeterminate results was often unclear. Obviously, the way these result categories are handled can have a large influence on diagnostic performance. Several articles did not mention such results, and it was unclear whether there were no results of this type, whether they were excluded, or whether they were incorporated in some other way. We are uncertain as to how much this factor contributed to the variability in study results. This source of variation could be eliminated by more standardized reporting. We recommend that results be reported in a standardized format as shown in **Table 10**. The

Table 8
Summary Totals From Included Studies for Diagnosis of Malignancy

	Histologic Diagnosis	
FNAC Diagnosis	**Malignant**	**Benign**
Malignant	1,227	177
Benign	311	4,454

FNAC, fine-needle aspiration cytology.

Table 9
Summary Totals From Included Studies for Diagnosis of Neoplasia

	Histologic Diagnosis	
FNAC Diagnosis	**Neoplastic**	**Nonneoplastic**
Neoplastic	620	10
Nonneoplastic	58	107

FNAC, fine-needle aspiration cytology.

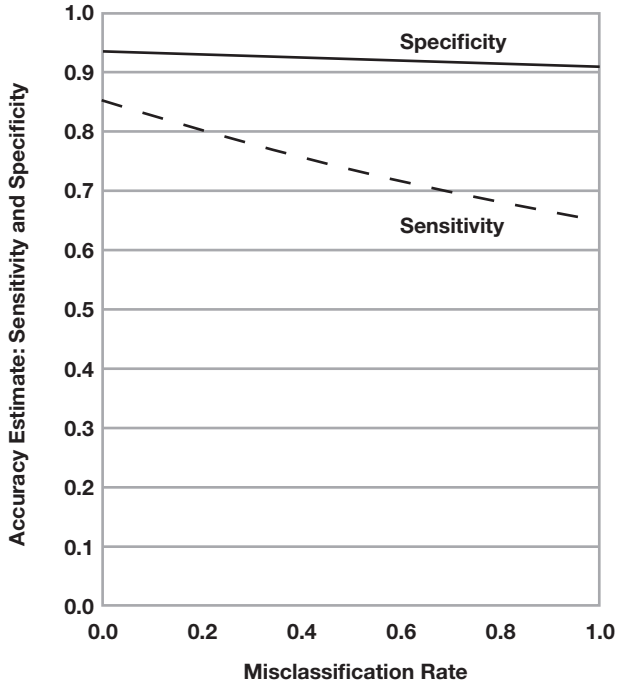

Figure 8 The effect of misclassification errors on observed sensitivity and specificity for diagnosis of malignancy. The misclassification rate refers to the error rate of the "gold standard," histologic examination.

DOI: 10.1309/AJCPOIE0CZNAT6SQ

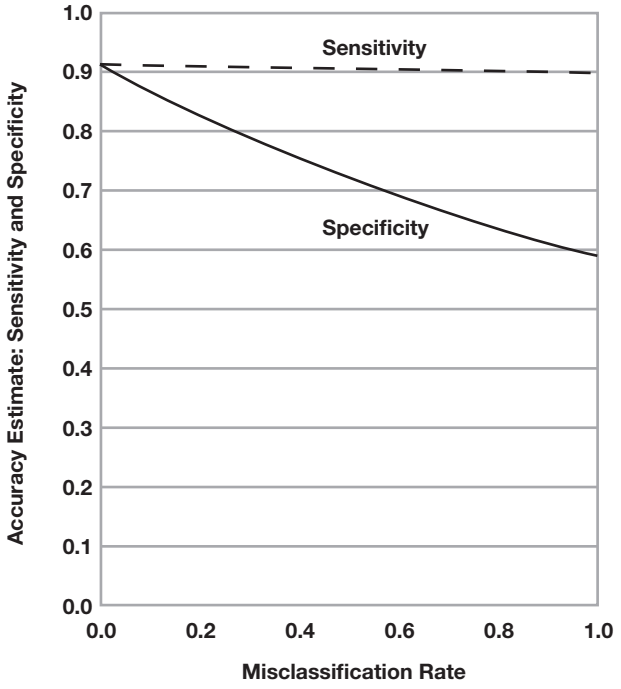

Figure 9 The effect of misclassification errors on observed sensitivity and specificity for diagnosis of neoplasia. The misclassification rate refers to the error rate of the "gold standard," histologic examination.

results for diagnosis of neoplasia and malignancy should be presented in 2 separate tables.

Timing Bias.—Timing bias occurs when there is significant disease progression during the time between performance of the index test and a reference test. While we do not believe this is likely to be a significant source of bias, it would be helpful if researchers reported summary statistics (eg, average and SD) on the time between FNAC and surgery.

The summary estimate for the sensitivity of the diagnosis of neoplasia is probably inflated owing to verification bias. Review bias probably inflates the estimates of sensitivity and specificity, but this effect is probably small. Misclassification bias probably leads to an underestimate of the sensitivity in the diagnosis of malignancy and an underestimate of the specificity for the diagnosis of neoplasia. Overall, the statistically significant difference seen between the accuracy of the diagnosis of neoplasia and malignancy may be due to bias rather than a real difference in performance.

Table 10
Suggested Format for Reporting of Accuracy Results

	Histologic Diagnosis		
FNAC Diagnosis	**Positive**	**Negative**	**Indeterminate**
Positive			
Negative			
Indeterminate			
Inadequate			

FNAC, fine-needle aspiration cytology.

Clinical Implications

FNAC provides information that informs 2 key decisions in patient management. First, the FNAC differentiates between neoplastic and nonneoplastic lesions. Neoplastic lesions generally are managed by surgery, whereas nonneoplastic lesions are managed conservatively. Second, given a neoplastic lesion, FNAC determines whether the lesion is malignant or benign, which determines the extent of surgery and, in particular, whether the facial nerve can be spared during surgery.

At present, the performance variability is too great to provide general guidelines regarding the usefulness of FNAC. Given the variability, it would not be possible to provide general guidelines based on average performance. Thus, the usefulness of FNAC can be evaluated only on a case-by-case basis depending on local diagnostic performance. Centers need to develop systems to assess local diagnostic performance for the diagnosis of neoplasia and malignancy. We found that studies grouped according to country often formed homogeneous groups with respect to diagnostic performance. Thus, it may be possible to develop guidelines that would apply to countries or regions. The impact of verification bias needs to be better understood before the estimates for neoplasia can be considered reliable.

Research Needs

In our opinion, there is room for further research including the following:

- Improved estimates of the diagnostic accuracy of the diagnosis of neoplasia. This is the most important decision because it determines whether a patient goes to surgery. A false-negative diagnosis for neoplasia can result in a delay of treatment and progression of disease, whereas a false-negative for neoplasia will usually be detected following surgery by definitive histologic findings. Thus, the diagnosis of neoplasia could be regarded as the most critical decision; however, the data on the diagnosis of neoplasia are sparse. It seems that studies have been driven by the convenience of sampling (surgery lists) and, as a result, have focused on the diagnosis of malignancy rather than the diagnosis of neoplasia. Obtaining such data will require studies with better follow-up of patients who do not undergo surgery.

DOI: 10.1309/AJCPOIE0CZNAT6SQ

- Improved study designs and patient tracking systems to eliminate verification bias
- Studies to understand the impact of misclassification bias
- Studies to further understanding of the contribution of FNAC to the diagnostic process. FNAC is not done in isolation, so the diagnostic process often involves other tests such as a clinical examination and imaging studies. Thus, the clinical usefulness of FNAC can be assessed only by evaluating its incremental impact on the final diagnosis.
- Studies to understand and eliminate the causes of performance variability. This will require 2 things. First, there is a need for improved reporting of study factors that can be used to identify the factors that cause performance variability. Second, factors that improve performance need to be incorporated into practice to reduce variability that, in turn, makes it possible to identify more subtle causes of performance variation.

Conclusions

The specificity of FNAC is quite high for the diagnosis of neoplasia (0.98) and malignancy (0.96). Thus, a positive diagnosis is quite reliable. The sensitivity is lower and more variable for the diagnosis of neoplasia (0.96) and for malignancy (0.79). The overall accuracy of diagnosis of neoplasia is greater than the accuracy for diagnosis of malignancy; however, some of the difference may be due to verification bias. Given the wide variability in accuracy, it is not possible to give a general guideline regarding the clinical usefulness of FNAC. The usefulness of FNAC will vary by location depending on the accuracy obtained at each site. Based on our survey, there are practice locations where FNAC is sufficiently accurate that a negative result can be used to avoid surgery or to allow for nerve-sparing surgery; however, there are other sites where this is not the case. Thus, in the face of such variability, each practice location must monitor its diagnostic performance to assess the usefulness of FNAC at each individual practice location.

It is disappointing that such a large collection of cases does not contribute more to our understanding of performance variation. There is a need for more complete reporting and improved study designs to remove sources of bias. More complete reporting using, for example, the guidelines of the STARD initiative[89,90] would do much to rectify this situation. Also, adoption of the best practices (eg, having a cytopathologist available to assess slide adequacy at sample collection) would reduce variance and improve overall performance. With these improvements, future studies may provide useful data that can be used to understand causes of performance variation and to provide general guidelines regarding the clinical usefulness of FNAC for diagnosis of parotid gland lesions.

From the [1]Department of Pathology, University of Utah School of Medicine, and [2]ARUP Laboratories, Salt Lake City.

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

Acknowledgments: We thank the following people for assistance with translation: Evrim Ergodan, PhD, Larissa Furtado, MD, Yefim Lavrentyev, Anya Matynia, MD, Chie Minoda, Karen Moser, MD, Laura Parnas, PhD, Carolin Teman, MD, Reha Toydemir, MD, and Holly Zhou, MD. We also thank Mary Youngkin for assistance with the literature search.

References

1. Frable MA, Frable WJ. Fine-needle aspiration biopsy of salivary glands. *Laryngoscope*. 1991;101:245-249.
2. Qizilbash AH, Sianos J, Young JE, et al. Fine needle aspiration biopsy cytology of major salivary glands. *Acta Cytol*. 1985;29:503-512.
3. 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.
4. Nettle WJ, Orell SR. Fine needle aspiration in the diagnosis of salivary gland lesions. *Aust N Z J Surg*. 1989;59:47-51.
5. Lin AC, Bhattacharyya N. The utility of fine needle aspiration in parotid malignancy. *Otolaryngol Head Neck Surg*. 2007;136:793-798.
6. 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.
7. Batsakis JG, Sneige N, el-Naggar AK. Fine-needle aspiration of salivary glands: its utility and tissue effects. *Ann Otol Rhinol Laryngol*. 1992;101(2 pt 1):185-188.
8. Tandon S, Shahab R, Benton JI, et al. Fine-needle aspiration cytology in a regional head and neck cancer center: comparison with a systematic review and meta-analysis. *Head Neck*. 2008;30:1246-1252.
9. Sonmez B. Parotis Bezi Kitlelerinde Ince Igne Aspirasyon Sitolojisi Ve Postoperatif Histopatolojik Degerlendirmenin Karsilastirilmasi Sağlık Bakanlığı; 2005. http://www.istanbulsaglik.gov.tr/w/tez/pdf/patoloji/dr_birol_sonmez.pdf. Accessed October 15, 2010.
10. Friese A. Ergebnisse von Feinnadelpunktionszytologie, klinischer Beurteilung und endgültigem histopathologischem Befund im Vergleich bei Parotistumoren, Universität zu Würzburg; 2005.
11. Ortega PG, Perez VM, Sotelo-Regil RH. Biopsia por aspiracion can aguja delgada en el diagnostico de lesiones en region parotidea. *Rev Inst Nac Cancerol*. 2000;46:81-84.
12. Schroder U, Eckel HE, Rasche V, et al. Value of fine needle puncture cytology in neoplasms of the parotid gland [in German]. HNO. 2000;48:421-429.
13. 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.

DOI: 10.1309/AJCPOIE0CZNAT6SQ

14. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med*. 2001;20:2865-2884.
15. Leeflang MMG, Deeks JJ, Gatsonis C, et al. Systematic reviews of diagnostic test accuracy. *Ann Intern Med*. 2008;149:889-897.
16. Harbord RM, Whiting P, Sterne JAC, et al. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. *J Clin Epidemiol*. 2008;61:1095-1103.
17. Deeks JJ BP, Gatsonis C, eds. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.0*. The Cochrane Collaboration; 2009. http://srdta.cochrane.org/handbook-dta-reviews. Accessed September 15, 2010.
18. Behbehani A, Dashti H, Al-Shahawi M, et al. The value of pre-operative fine-needle aspiration biopsy in planning the management of parotid gland tumours. *Med Principles Pract*. 1990;2:27-34.
19. 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.
20. Harbord RM, Whiting P. Metandi: meta-analysis of diagnostic accuracy using hierarchical logistic regression. *Stata J*. 2009;9:211-229.
21. Akbas Y, Tuna EU, Demireller A, et al. Ultrasonography guided fine needle aspiration biopsy of parotid gland masses. *Kulak Burun Bogaz Ihtis Derg*. 2004;13:15-18.
22. Al Salamah SM, Khalid K, Khan IA, et al. Outcome of surgery for parotid tumours: 5-year experience of a general surgical unit in a teaching hospital. *ANZ J Surg*. 2005;75:948-952.
23. Al-Khafaji BM, Nestok BR, Katz RL. Fine-needle aspiration of 154 parotid masses with histologic correlation: ten-year experience at the University of Texas M. D. Anderson Cancer Center. *Cancer*. 1998;84:153-159.
24. Atula T, Greenman R, Laippala P, et al. Fine-needle aspiration biopsy in the diagnosis of parotid gland lesions: evaluation of 438 biopsies. *Diagn Cytopathol*. 1996;15:185-190.
25. Aversa S, Ondolo C, Bollito E, et al. Preoperative cytology in the management of parotid neoplasms. *Am J Otolaryngol*. 2006;27:96-100.
26. Awan MS, Ahmad Z. Diagnostic value of fine needle aspiration cytology in parotid tumors. *J Pak Med Assoc*. 2004;54:617-619.
27. Bartels S, Talbot JM, DiTomasso J, et al. The relative value of fine-needle aspiration and imaging in the preoperative evaluation of parotid masses. *Head Neck*. 2000;22:781-786.
28. Behzatoglu K, Bahadir B, Kaplan HH, et al. Fine needle aspiration biopsy of the parotid gland: diagnostic problems and 2 uncommon cases. *Acta Cytol*. 2004;48:149-154.
29. Berrone S, Cubetta M, Amasio ME, et al. Fine needle biopsy in the preoperative diagnosis of parotid tumors [in Italian]. *Minerva Stomatol*. 1995;44:515-519.
30. Brennan PA, Davies B, Poller D, et al. Fine needle aspiration cytology (FNAC) of salivary gland tumours: repeat aspiration provides further information in cases with an unclear initial cytological diagnosis. *Br J Oral Maxillofac Surg*. 2010;48:26-29.
31. Buhler RB, Mattiola LR, Pinheiro JLG, et al. Fine needle aspiration puncture in parotid gland lesions. *Int Arch Otorhinolaryngol*. 2007;11:294-299.
32. Burgess AN, Serpell JW. Parotidectomy: preoperative investigations and outcomes in a single surgeon practice. *ANZ J Surg*. 2008;78:791-793.
33. Califano L, Zupi A, Giardino C. Accuracy in the diagnosis of parotid tumours. *J Craniomaxillofac Surg*. 1992;20:354-359.
34. Carrillo JF, Ramirez R, Flores L, et al. Diagnostic accuracy of fine needle aspiration biopsy in preoperative diagnosis of patients with parotid gland masses. *J Surg Oncol*. 2009;100:133-138.
35. Contucci AM, Corina L, Sergi B, et al. Correlation between fine needle aspiration biopsy and histologic findings in parotid masses: personal experience. *Acta Otorhinolaryngol Ital*. 2003;23:314-318.
36. Costas A, Castro P, Martin-Granizo R, et al. Fine needle aspiration biopsy (FNAB) for lesions of the salivary glands. *Br J Oral Maxillofac Surg*. 2000;38:539-542.
37. Deans GT, Briggs K, Spence RA. An audit of surgery of the parotid gland. *Ann R Coll Surg Engl*. 1995;77:188-192.
38. Deneuve S, Quesnel S, Depondt J, et al. Management of parotid gland surgery in a university teaching hospital. *Eur Arch Otorhinolaryngol*. 2010;267:601-605.
39. Filho V, de Carlucci D, Sondermann A, et al. Fine needle aspirative biopsy in parotid gland diseases. *Rev Col Bras Cir*. 2001;28:189-192.
40. Filopoulos E, Angeli S, Daskalopoulou D, et al. Pre-operative evaluation of parotid tumours by fine needle biopsy. *Eur J Surg Oncol*. 1998;24:180-183.
41. Gete Garcia MP, Almodovar Alvarez C, Garcia Alvarez G, et al. Parotid tumours: correlation between fine needle aspiration biopsy and histological findings [in Spanish]. *Acta Otorrinolaringol Esp*. 2006;57:279-283.
42. Gobić MB, Pedisić D, Bekafigo IS, et al. Fine needle aspiration cytology in the evaluation of parotid gland tumors. *Coll Antropol*. 2010;34:345-348.
43. Gooden E, Witterick IJ, Hacker D, et al. Parotid gland tumours in 255 consecutive patients: Mount Sinai Hospital's quality assurance review. *J Otolaryngol*. 2002;31:351-354.
44. He Y, Zhang ZY, Tian Z. The diagnostic value of fine-needle aspiration cytology (FNAC) for lesions in the parotid gland [in Chinese]. *Shanghai Kou Qiang Yi Xue*. 2003;12:410-413.
45. Herrera Hernández AA, Diaz Perez JA, Garcia CA, et al. Evaluation of fine needle aspiration cytology in the diagnosis of cancer of the parotid gland [in Spanish]. *Acta Otorrinolaringol Esp*. 2008;59:212-216.
46. Inohara H, Akahani S, Yamamoto Y, et al. The role of fine-needle aspiration cytology and magnetic resonance imaging in the management of parotid mass lesions. *Acta Otolaryngol*. 2008;128:1152-1158.
47. Jafari A, Royer B, Lefevre M, et al. Value of the cytological diagnosis in the treatment of parotid tumors. *Otolaryngol Head Neck Surg*. 2009;140:381-385.
48. Kamal SA, Othman EO. Diagnosis and treatment of parotid tumours. *J Laryngol Otol*. 1997;111:316-321.
49. Kaur A, Chew CT, Lim-Tan SK. Fine needle aspiration of 123 head and neck masses: an initial experience. *Ann Acad Med Singapore*. 1993;22:303-306.
50. Knudsen PJ, Eriksen HE, Greisen O. Fine needle aspiration cytological study of parotid tumors [in Danish]. *Ugeskr Laeger*. 1985;147:2824-2827.
51. Kondo A, Kozawa T, Watanabe K, et al. Clinical study of parotid tumors. *Pract Otorhinolaryngol*. 2007;100:369-373.

DOI: 10.1309/AJCPOIE0CZNAT6SQ

52. Lim CM, They J, Loh KS, et al. Role of fine-needle aspiration cytology in the evaluation of parotid tumours. *ANZ J Surg*. 2007;77:742-744.
53. Longuet M, Nallet E, Guedon C, et al. Diagnostic value of needle biopsy and frozen section histological examination in the surgery of primary parotid tumors [in French]. *Rev Laryngol Otol Rhinol (Bord)*. 2001;122:51-55.
54. Lurie M, Misselevitch I, Fradis M. Diagnostic value of fine-needle aspiration from parotid gland lesions. *Isr Med Assoc J*. 2002;4:681-683.
55. Malata CM, Camilleri IG, McLean NR, et al. Malignant tumours of the parotid gland: a 12-year review. *Br J Plast Surg*. 1997;50:600-608.
56. Altuna Mariezkurrena X, Gorostiaga Aznar F, Zulueta Lizaur A, et al. Evaluation of the fine needle aspiration biopsy in the presurgical diagnosis of tumors of the parotid gland [in Spanish]. *An Otorrinolaringol Ibero Am*. 2006;33:495-503.
57. Marrazzo A, Taormina P, La Bara G, et al. The role of needle aspiration biopsy in the diagnosis of parotid masses [in Italian]. *Minerva Chir*. 1993;48:1193-1196.
58. Mianroodi AAA, Sigston EA, Vallance NA. Frozen section for parotid surgery: should it become routine? *ANZ J Surg*. 2006;76:736-739.
59. Mohammed F, Asaria J, Payne RJ, et al. Retrospective review of 242 consecutive patients treated surgically for parotid gland tumours. *J Otolaryngol Head Neck Surg*. 2008;37:340-346.
60. Osanai H, Osaki T, Nonaka S, et al. Parotid tumors: clinical study of 36 cases. *Pract Otorhinolaryngol*. 2003;96:799-804.
61. Paris J, Facon F, Pascal T, et al. Preoperative diagnostic values of fine-needle cytology and MRI in parotid gland tumors. *Eur Arch Otorhinolaryngol*. 2005;262:27-31.
62. Pons Rocher F, Estelles Ferriol E, Carrasco Llatas M, et al. Malignant tumors of the parotid gland [in Spanish]. *An Otorrinolaringol Ibero Am*. 2003;30:571-585.
63. Que Hee CG, Perry CF. Fine-needle aspiration cytology of parotid tumours: is it useful? *ANZ J Surg*. 2001;71:345-348.
64. Riley N, Allison R, Stevenson S. Fine-needle aspiration cytology in parotid masses: our experience in Canterbury, New Zealand. *ANZ J Surg*. 2005;75:144-146.
65. Rodriguez HP, Silver CE, Moisa II, et al. Fine-needle aspiration of parotid tumors. *Am J Surg*. 1989;158:342-344.
66. Schelkun PM, Grundy WG. Fine-needle aspiration biopsy of head and neck lesions. *J Oral Maxillofac Surg*. 1991;49:262-267.
67. 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.
68. Shashinder S, Tang IP, Velayutham P, et al. A review of parotid tumours and their management: a ten-year-experience. *Med J Malaysia*. 2009;64:31-33.
69. Takashima S, Takayama F, Wang Q, et al. Parotid gland lesions: diagnosis of malignancy with MRI and flow cytometric DNA analysis and cytology in fine-needle aspiration biopsy. *Head Neck*. 1999;21:43-51.
70. 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.
71. Tsai SC, Hsu HT. Parotid neoplasms: diagnosis, treatment, and intraparotid facial nerve anatomy. *J Laryngol Otol*. 2002;116:359-362.
72. Uğuz MZ, Onal HK, Eroğlu OO, et al. Sensitivity and specificity of fine needle aspiration biopsy in parotid masses [in Turkish]. *Kulak Burun Bogaz Ihtis Derg*. 2007;17:96-99.
73. 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.
74. Van Lierop AC, Fagan JJ. Parotidectomy in Cape Town: a review of pathology and management. *S Afr J Surg*. 2007;45:96-98, 100, 102-103.
75. Weinberger MS, Rosenberg WW, Meurer WT, et al. Fine-needle aspiration of parotid gland lesions. *Head Neck*. 1992;14:483-487.
76. Zafar A, Shafi M, Hassan SH, et al. Fine needle aspiration cytology in parotid lumps. *J Pak Med Assoc*. 1997;47:188-190.
77. Zurrida S, Alasio L, Tradati N, et al. Fine-needle aspiration of parotid masses. *Cancer*. 1993;72:2306-2311.
78. The World Bank. Country and Lending Groups. 2010. http://data.worldbank.org/about/country-classifications/country-and-lending-groups. Accessed November 1, 2010.
79. Fleiss JL. *Statistical Methods for Rates and Proportions*. 2nd ed. New York, NY: Wiley; 1981.
80. Layfield LJ, Tan P, Glasgow BJ. Fine-needle aspiration of salivary gland lesions: comparison with frozen sections and histologic findings. *Arch Pathol Lab Med*. 1987;111:346-353.
81. Hughes JH, Volk EE, Wilbur DC. Pitfalls in salivary gland fine-needle aspiration cytology: lessons from the College of American Pathologists Interlaboratory Comparison Program in Nongynecologic Cytology. *Arch Pathol Lab Med*. 2005;129:26-31.
82. Dinnes J, Deeks J, Kirby J, et al. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. *Health Technol Assess*. 2005;9:1-113.
83. Irwig L, Bossuyt P, Glasziou P, et al. Designing studies to ensure that estimates of test accuracy are transferable. *BMJ*. 2002;324:669-671.
84. Eisele DW, Koch WM, Richtsmeier WJ, et al. Utility of immediate on-site cytopathological procurement and evaluation in fine needle aspiration biopsy of head and neck masses. *Laryngoscope*. 1992;102:1328-1330.
85. Singh N, Ryan D, Berney D, et al. Inadequate rates are lower when FNAC samples are taken by cytopathologists. *Cytopathology*. 2003;14:327-331.
86. Wu M, Burstein DE, Yuan S, et al. A comparative study of 200 fine needle aspiration biopsies performed by clinicians and cytopathologists. *Laryngoscope*. 2006;116:1212-1215.
87. 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.
88. Rapkiewicz A, Le BT, Simsir A, et al. Spectrum of head and neck lesions diagnosed by fine-needle aspiration cytology in the pediatric population. *Cancer*. 2007;111:242-251.
89. Bossuyt PM, Reitsma JB, Bruns DE, et al; for the STARD Group. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Clin Chem*. 2003;49:1-6.
90. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin Chem*. 2003;49:7-18.

DOI: 10.1309/AJCPOIE0CZNAT6SQ