Interpretive Diagnostic Error Reduction in Surgical Pathology and Cytology:
Joint Guideline from the College of American Pathologists Pathology and Laboratory Quality Center and the Association of Directors of Anatomic and Surgical Pathology

Corresponding Author:
Raouf E. Nakhleh, MD

Authors:
Vania Nosé, MD, PhD
Lisa Fatheree, SCT(ASCP)
Christina Ventura, MT(ASCP)
Douglas McCrory, MD
Carol Colasacco SCT(ASCP), MLIS, AHIP


* The Supplemental Digital Content was not copyedited by Archives of Pathology
METHODS USED TO PRODUCE THE GUIDELINE

Panel Composition
The College of American Pathologists (CAP) Pathology and Laboratory Quality Center (the Center) and Association of Directors of Anatomic and Surgical Pathology (ADASP) convened an expert and advisory panel consisting of practicing pathologists with expertise in surgical pathology and cytology. CAP and ADASP approved the appointment of the project co-chairs (RN and VN) and panel members. These panel members and the methodologist served as the Expert Panel (EP) for the systematic evidence review.

Conflict of Interest (COI) Policy
Prior to acceptance on the expert or advisory panel, potential members completed the CAP conflict of interest (COI) disclosure process, whose policy and form (in effect April 2010) require disclosure of material financial interest in, or potential for benefit of significant value from, the guideline’s development or its recommendations 12 months prior through the time of publication. The potential members completed the COI disclosure form, listing any relationship that could be interpreted as constituting an actual, potential, or apparent conflict. The CAP Center uses the following criteria:

Nominees who have the following conflicts may be excused from the panel:
- Stock or equity interest in a commercial entity that would likely be affected by the guideline or white paper
- Royalties or licensing fees from products that would likely be affected by the guideline or white paper
- Employee of a commercial entity that would likely be affected by the guideline or white paper

Nominees who have the following potentially manageable direct conflicts may be appointed to the panel:
- Patents for products covered by the guideline or white paper
- Member of an advisory board of a commercial entity that would be affected by the guideline or white paper
- Payments to cover costs of clinical trials, including travel expenses associated directly with the trial
- Reimbursement from commercial entity for travel to scientific or educational meetings

Everyone was required to disclose conflicts prior to beginning and continuously throughout the project’s timeline. Expert panel members’ disclosed conflicts are listed in the appendix of the manuscript. The CAP and ADASP provided funding for the administration of the project; no industry funds were used in the development of the guideline. All panel members volunteered their time and were not compensated for their involvement.
CAP/ADASP Expert Panel Literature Review and Analysis

The expert panel met 26 times through teleconference webinars from December 2011 through June 12, 2014. Additional work was completed via electronic mail and the panel met in person October 12, 2013 to review evidence to date and draft recommendations.

All expert panelists participated in the systematic evidence review (SER) level of title-abstract, full-text review, and data extraction. The co-chairs (RN and VN) and methodologist (DM) performed the audit of data extraction. All articles were available as discussion or background references. All members of the expert panel participated in developing draft recommendations, reviewing open comment feedback, finalizing and approving recommendations and writing/editing of the manuscript.

Peer Review

An open comment period was held from December 2, 2013 through January 21, 2014. Five draft recommendations and three methodology questions were posted online on the CAP Web site www.cap.org. An announcement was sent to the following societies deemed to have interest:

- CAP Board of Governors, Councils, Committees and Membership
- Association of Directors of Anatomic and Surgical Pathology (ADASP)
- American Society for Clinical Pathology (ASCP)
- American Society of Cytopathology (ASC)
- Papanicolaou Society of Cytology (PSC)
- Arthur Purdy Stout Society (APSS)
- Association of Pathology Chairs (APC)
- Canadian Association of Pathologists (CAP-APC)
- United States & Canadian Academy of Pathology (USCAP)
- Quality Initiative in Interpretive Pathology (QIIP) Canadian Partnership Against Cancer
- Society to Improve Diagnoses in Medicine (SIDM)
- American Society of Clinical Oncologists (ASCO)
- Veteran’s Affairs (VA) and Department of Defense (DOD)
- Centers for Disease Control and Prevention (CDC)
- Centers for Medicare and Medicaid Services (CMS)

The website received 303 comments in total (Agree and Disagree responses were also captured). All the recommendations achieved between 87% to 93% agreement. Pairs of expert panel members were assigned 1 draft recommendation for which to review all comments received and provide an overall summary to the rest of the panel. Following panel discussion, and the final quality of evidence assessment, the panel members determined whether to maintain the original draft recommendation as is, revise it with minor language change, or consider it as a major recommendation change. Resolution of all changes was obtained by majority consensus of the panel using nominal group technique (rounds of email discussion and multiple edited recommendations) amongst the panel members. The final recommendations were approved by the expert panel with a formal vote. The panel considered the risks and benefits throughout the whole process in their considered judgment process. Formal cost analysis or cost effectiveness was not performed.

An independent review panel (IRP) was assembled to review the guideline and recommend approval to the CAP. The IRP was masked to the expert panel and vetted through the COI process. Final approval was done by CAP Council on Scientific Affairs and ADASP Executive Board.

Assessing the Strength of Recommendations

Strength of recommendation is determined by the balance between desirable and undesirable consequences of alternative management strategies, quality of evidence, variability in values and preferences, resource use. Formal cost analysis or cost effectiveness was not performed.
The central question that the panel addressed in developing the guideline was “What are the most effective ways to reduce interpretive diagnostic errors in Anatomic Pathology?”

Development of recommendations requires that the panel review the identified evidence and make a series of key judgments:

1. What are the significant findings related to each KQ or outcome?
2. What is the overall quality of evidence supporting each KQ or outcome? Quality of evidence was assessed according to the GRADE framework as described below. Summary of Findings tables or when sufficient information was not available, an alternative short evidence table format, were prepared for each question. To maintain consistency with previous CAP guideline language, quality of evidence is described as Convincing, Adequate or Inadequate as shown in Table 1.
3. What is the net balance of benefits and harms? The consideration of net balance of benefits and harms will focus on the laboratory redundancy, efficiency and feasibility in comparison to the reduction of errors or potential errors and their impact on patient outcomes.
4. What is the strength of each recommendation? The implications of a strong recommendation for clinicians is that most patients should receive the recommended course of action; while the implications of a weak recommendation are that different choices may be appropriate for different patients or that different management options may be preferred by different stakeholders. We used the current CAP designations for strength of recommendations of strong recommendation, recommendation and expert consensus opinion (Table 2), as determined by a considered judgment process in which the guideline panel weighs the quality of evidence, balance of benefits and harms, variability in values preference and data on costs or resource use.

Dissemination Plans
CAP will host an Interpretive Diagnostic Error Reduction Through Targeted Case Reviews In Surgical Pathology And Cytology Resource web page which will include a link to the manuscript and supplemental digital content; summary of recommendations, teaching PowerPoint, and a frequently asked question (FAQ) document. The ADASP webpage will include a link to the CAP guideline resource page. The guideline will be promoted and presented at various professional society meetings including the College of American Pathologists, the United States and Canadian Academy of Pathology (USCAP), and the American Society of Clinical Pathology (ASCP).

SYSTEMATIC EVIDENCE REVIEW (SER)

The objectives of the SER were to investigate the most effective ways to reduce interpretive diagnostic errors in surgical pathology and cytology. If of sufficient quality, findings from this review could provide an evidence base to support development of the laboratory practice guideline. The scope of the SER and the key questions (KQs) were established by the EP in consultation with a methodologist.

Key Questions:
1. Does targeted review (either done at analytic or post-analytic phase) of surgical pathology or cytology cases (slides and/or reports) reduce the error rate (often measured as amended reports) or increase the rate of interpretive error detection compared to no review, random review or usual review procedures?
2. What methods of selecting cases for review have been shown to increase the rate of interpretive error detection compared to no review, random review or usual review procedures?

Detailed Scope Questions:
• Can a targeted review of cases lead to increased detection of errors?
• Is there a particular method (eg, prospective vs. retrospective, random reviews, etc.) that results in lower error rates? Or amended report rates? If yes, what is the method?
• Is there a particular type of diagnosis(es) (eg, malignant, benign, borderline) more prone to error? If yes, what?
• Is there a particular organ/system associated with a higher rate of error/disagreements revealed as problematic in the literature (eg, thyroid FNA, lymphoma, brain biopsy)? If yes, what?
• How does the error rate for multi-organ reviews compare with single organ reviews? (eg, internal and/or external)
• Do blinded reviews find more or less errors than non-blinded reviews?
• Can we learn anything from external review studies that can be applied to internal reviews (or vice-versa)?
• Can we establish an external error rate (benchmark) for surgical pathology or cytology?
• Can we establish an internal error rate (benchmark) for surgical pathology or cytology?
• What are the costs of conducting internal or external reviews?
• What is the effect on turn-around times (efficacy) for internal or external reviews?
• What is the effect on patient care when conducting internal or external reviews?
• Has there been improvement in reducing error reduction over time?
• What other factors (eg, standard criteria, clinical correlation, ancillary testing) are emphasized most frequently and with which organ system or diagnosis?

Search and Selection
The systematic literature search for relevant studies included a search of MEDLINE using the Ovid SP interface on November 12, 2013, with the date parameters of January 1992 through October 2012. Medical subject headings (MeSH) for the concept pathology (eg, Pathology, Surgical/Pathology, Clinical; Pathology) were combined with MeSH terms for the concept quality (eg, Quality Assurance, Health Care; Quality Control; Quality Improvement; Reproducibility of Results; Diagnostic Errors). MeSH terms were supplemented with keywords (eg, histopathology, cytopathology, histology, or cytology; and second opinion, misinterpretation, or interpretation errors). A targeted concept of slide/case review included keywords such as targeted review, peer review, or random review and the keywords slide, case, or report. Limits were set for human studies published in English. The search was not limited by study design in order to capture editorials, letters, or commentaries that might be relevant and useful for discussion purposes. A literature refresh of the OvidSP search strategy was run on November 21, 2013, to identify relevant studies published since October 2012. The full Ovid search strategy is included in the appendix.

The Ovid search strategy was modified for PubMed (1/1/92-12/31/12), and Google Scholar (1/1/12 – 1/26/13). In addition, a handsearch of relevant journals (American Journal of Clinical Pathology, American Journal of Surgical Pathology, Archives in Pathology and Laboratory Medicine, Cancer, Cancer Cytopathology, Diagnostic Cytopathology, Histopathology, Modern Pathology) was completed for issues published from January 2008, through December 2012. A search for meeting abstracts was completed utilizing Biosis Previews (Web of Science) (1/1/2008-12/31/2012) and by handsearching published abstracts from relevant meetings (American Society of Cytology, American Society for Clinical Pathology, British Society for Clinical Cytology, College of American Pathologists, European Congress of Cytopathology, International Academy of Pathology, United States and Canadian Academy of Pathology) held from January 2008 through December 2012. Reference lists from included articles were scanned for additional relevant studies.

Two reviewers were used at all levels of review (eg, title/abstract, full article) and for data/information extraction. Conflicts were resolved by discussion or referral to the panel co-chairs for a decision. When article abstracts or document summaries were not available or a conflict was not resolved, full articles were reviewed.

Selection at all levels of the review was based on predetermined inclusion/exclusion criteria.

Inclusion criteria:
• Surgical pathology or cytology studies
• Original research addressing targeted review
• English language articles
• All study types were initially included
Exclusion criteria:

- Clinical pathology studies
- Additional diagnostic techniques, issues related to competency, use of checklists, standardized language, taxonomy or formatting
- Studies focused on pre-analytic specimen processes
- Post-analytic typographic errors
- Non-English studies
- Animal studies

Outcomes of Interest

We are interested in identifying discrepancies in interpretation between a primary pathologist review and a second pathologist review as a way of estimating the error rate. To the extent that erroneous readings can be identified in excess of an expected degree of disagreement, then a method of targeted review would be said to be effective. Thus, studies with a control group are desirable; as a practical matter, however, it is necessary to examine uncontrolled series, too. Studies had to report numbers of discrepant diagnoses among a defined population of specimens examined to allow calculation of a discrepancy rate.

We are not interested in discrepancies from the pre-analytic specimen process (i.e., related to tissue collection and processing) or post-analytic errors (e.g., typographic or transcription errors, amended reports), additional diagnostic techniques (e.g., immunomarkers), issues related to competency, or the use of checklists, standardized language, taxonomy or formatting.

Various studies classify errors in different ways (e.g., major versus minor). Recognizing that all errors are not alike, we assessed the severity of interpretive errors according to the clinical impact on a patient. We considered the clinical impact of errors as follows: 1) diagnostic thinking (error results in a change in diagnosis or diagnostic category); 2) therapeutic efficacy (error results in a change in therapeutic choice); or 3) patient outcome efficacy (error results in a change in outcome (e.g., procedure avoided); demonstrating this unequivocally may require long-term follow-up). We also considered the efficiency or cost (in terms of effort or dollars) that a targeted review strategy entails.

Data Extraction and Management

The data elements from an included article/document were extracted by one reviewer into standard data formats and tables developed using systematic review database software (DistillerSR, Evidence Partners Inc., Ottawa, Canada); a second reviewer confirmed accuracy and completeness. In all cases, the methodologist acted as either the primary or secondary reviewer. Any discrepancies in data extraction were resolved by discussion with the methodologist. A bibliographic database was established in EndNote (Thomson Reuters, Carlsbad, CA) to track all literature identified and reviewed during the study.

Environmental Scan

An environmental scan for established guidelines was performed including a targeted search of pathology organizations’ web sites and a search of guideline clearinghouse websites (TRIP Database, Guidelines International Network, Agency for Healthcare Research and Quality) using the search terms “pathology or laboratory” and “guidelines or regulations.”

Quality Assessment

Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach provides a system for rating quality of evidence and strength of recommendations that is explicit, comprehensive, transparent, and pragmatic and is increasingly being adopted by organizations worldwide. The GRADE approach examines the quality of evidence as the level of individual studies and also at the review level. GRADE was used for rating the quality of evidence. At the individual study level, we assessed studies according to three criteria: 1) study design rating; 2) risk of bias rating; and 3) applicability concerns. Study design was assessed according to the following hierarchy:
1) noncomparative studies (where a single method of targeted review is considered, in determining a single error rate)

2) comparative studies (where two or more methods of targeted review are independently applied and compared in one group of patients)

3) comparative studies (where two or more methods of targeted review are compared in different groups of patients
a. historical (e.g., a before-after series)

b. concurrent (e.g., comparison of two different institution’s targeted review programs)

c. quasi-randomized (concurrent, similar or identical sites, allocated with attempt to reduce bias beyond that provided by 3.a. or 3.b.)

d. randomized

Risk of bias ratings were based on the following three criteria, rated as yes, no, or unclear:

1) Were review diagnoses independent, blinded from the primary diagnosis?

2) Was case selection done using explicit, objective and reproducible criteria?

3) Was interpretive error ascertained without confounding by pre- or post-analytic error?

Applicability concerns were assessed for three areas, rated as yes, no, or unclear:

1) Is study limited to particular organ systems, specimen types or diagnoses?

2) Is the inter-rater performance examined and acceptable?

3) Is there a distinction between major and minor errors based on impact on patient treatment or outcome?

At the review level, we examined the collection of studies according to the following domains:

1) Consistency

2) Directness

3) Precision

4) Magnitude of effect

5) Number of studies/patients

We developed a GRADE evidence table, and mapped the quality of evidence ratings to the CAP rating scheme according to table 1.

Data Analysis

Rates of discrepancy and major discrepancy were described for all studies, and subgroups based on type of specimen (surgical pathology, cytopathology or both), focusing on one organ or organ system (single-organ) versus multi-organ studies, and whether conducted within a single institution (internal) or reviews of cases diagnosed at a different institution (external). Studies were tested for homogeneity using Comprehensive Meta Analysis version 2.2.064. Nonparametric descriptive statistics including median, and 1st and 3rd quartiles were calculated using Excel.

Results

Among the 823 citations identified by electronic and hand searches, 141 were selected for inclusion. These included 130 published peer-reviewed articles, and 11 grey literature documents (Appendix 1).

Among the extracted documents, 4 articles/documents did not meet minimum quality standards, presented incomplete data or data that were not in useable formats, or included only information based on expert opinion. These articles were not included in analyses or narrative summaries.

Of 137 studies included, 128 (91.4%) were single arm clinical case series. Of 12 comparative studies, 2 reported a comparison of 2 or more methods of targeted review in one group of patients; 6 compared targeted review methods in different historical cohorts (e.g., a before-after study), and 4 compare 2 or more targeted review methods in concurrent groups of patients. No studies used random allocation or other robust measures to control for potential bias.
Risk of bias assessments of included studies showed that review diagnoses were made independently (blinded from the primary diagnosis) in 30 (21.4%) studies, unclear in 2 (1.4%) and without blinding in 108 (77.1%). Case selection was done using explicit, objective and reproducible criteria in 56 (40%), and with unclear, subjective, non-reproducible or without criteria in 84 (59.3%). Interpretive error was ascertained without confounding by pre- or post-analytic error in 120 (85.7%), unclear in 2 (1.4%) and not in 18 (12.9%) studies.

Applicability concerns were as follows: Studies were limited to particular organ systems, specimen types or diagnoses in 111 (79.3%) cases, and not limited in 29 (20.7%) cases. Inter-rater performance was examined and acceptable in 29 (20.7%), unclear in 6 (4.3%), and not in 105 (75%) of studies. There was a distinction between major and minor errors based on impact on patient treatment or outcome in 78 (55.7%), unclear in 2 (1.4%), and not in 60 (42.9%) of studies.

Discrepancy rates and major discrepancy rates by study characteristics are provided in the manuscript. The distribution of discrepancy rates shows that the variability was greater for studies of smaller sample size. The distribution of discrepancy rates by sample size is shown in figures 1 and 2, both of which show high variability in discrepancy rates for studies with smaller sample sizes, with larger studies tending to have lower discrepancy rates.

Evidence on each question was summarized in terms of study quality and effects in Evidence Profile Tables (Tables 3-5) which were used as the basis for quality of evidence determinations with the panel.
Table 1. Quality of Evidence Ratings: mapping from GRADE categories to CAP

<table>
<thead>
<tr>
<th>GRADE</th>
<th>Definition</th>
<th>CAP</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>Further research is very unlikely to change our confidence in the estimate of effect.</td>
<td>High (Convincing)</td>
<td>High confidence that available evidence reflects true effect</td>
</tr>
<tr>
<td>Moderate</td>
<td>Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.</td>
<td>Intermediate (Adequate)</td>
<td>Moderate confidence that available evidence reflects true effect</td>
</tr>
<tr>
<td>Low</td>
<td>Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.</td>
<td>Low (Inadequate)</td>
<td>Little confidence that available evidence reflects true effect</td>
</tr>
<tr>
<td>Very low</td>
<td>Any estimate of effect is very uncertain.</td>
<td>Insufficient</td>
<td>Evidence is insufficient to discern net effect</td>
</tr>
</tbody>
</table>

Adapted from Guyatt et al² with permission from BMJ Publishing Group Ltd.
Table 2. Strength of Recommendations

<table>
<thead>
<tr>
<th>CAP Designation</th>
<th>Recommendation</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Strong Recommendation</strong></td>
<td>Recommend For or Against a particular pathology review practice (Can include must or should)</td>
<td>Supported by high (convincing) or intermediate (adequate) quality of evidence and clear benefit that outweighs any harms</td>
</tr>
<tr>
<td><strong>Recommendation</strong></td>
<td>Recommend For or Against a particular pathology review practice (Can include should or may)</td>
<td>Some limitations in quality of evidence (intermediate [adequate] or low [inadequate]), balance of benefits and harms, values, or costs but panel concludes that there is sufficient evidence to inform a recommendation.</td>
</tr>
<tr>
<td><strong>Expert Consensus Opinion</strong></td>
<td>Recommend For or Against a particular pathology review practice (Can include should or may)</td>
<td>Serious limitations in quality of evidence (low [inadequate] or insufficient), balance of benefits and harms, values or costs, but panel consensus is that a guideline is necessary.</td>
</tr>
<tr>
<td><strong>No Recommendation</strong></td>
<td>No recommendation for or against a particular pathology review practice</td>
<td>Insufficient evidence, confidence, or agreement to provide a recommendation.</td>
</tr>
</tbody>
</table>

Adapted from Teutsch et al4 with permission from Macmillian Publishers Ltd. Modified by the CAP Pathology and Laboratory Quality Center.
Table 3: Evidence profile for Recommendation 1: Should case review be used for diagnostic evaluation of pathology materials?*

<table>
<thead>
<tr>
<th>No of studies</th>
<th>Design</th>
<th>Risk of bias</th>
<th>Inconsistency</th>
<th>Indirectness</th>
<th>Imprecision</th>
<th>Other considerations</th>
<th>Case review</th>
<th>Control</th>
<th>Reliability (95% Confidence Interval)</th>
<th>Absolute</th>
<th>Quality</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No serious risk of bias</td>
<td>-</td>
<td>⊕ΟΟΟΟ</td>
<td>IMPORTANT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No serious inconsistency</td>
<td>-</td>
<td>⊕ΟΟΟΟ</td>
<td>IMPORTANT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No serious indirectness</td>
<td>-</td>
<td>⊕ΟΟΟΟ</td>
<td>IMPORTANT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No serious imprecision</td>
<td>-</td>
<td>⊕ΟΟΟΟ</td>
<td>IMPORTANT</td>
</tr>
<tr>
<td>122</td>
<td>observational studies\textsuperscript{a}</td>
<td>no serious risk of bias</td>
<td>no serious inconsistency</td>
<td>no serious indirectness</td>
<td>no serious imprecision</td>
<td>none</td>
<td>27178/5 13268 (5.3%)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>⊕ΟΟΟΟ</td>
<td>IMPORTANT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Discrepancy Rate (follow-up 1 years; assessed with case review)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>79</td>
<td>observational studies\textsuperscript{a}</td>
<td>no serious risk of bias</td>
<td>no serious inconsistency \textsuperscript{b}</td>
<td>no serious indirectness \textsuperscript{c}</td>
<td>no serious imprecision</td>
<td>none</td>
<td>4416/17 9130 (2.5%)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>⊕ΟΟΟΟ</td>
<td>CRITICAL</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Major Discrepancy Rate (assessed with: case review)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>observational studies</td>
<td>no serious risk of bias</td>
<td>Serious \textsuperscript{d}</td>
<td>Serious \textsuperscript{e}</td>
<td>no serious imprecision</td>
<td>reporting bias</td>
<td>-</td>
<td>-</td>
<td>kappa ranged from 0.16 to 0.97</td>
<td>-</td>
<td>⊕ΟΟΟΟ</td>
<td>IMPORTANT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Inter-observer agreement</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\textsuperscript{a} Studies include single- and multi-institution case series that quantify diagnostic discrepancies or errors according to various definitions.

\textsuperscript{b} Although the magnitude of discrepancy rates vary considerably from study to study based on sample size, specimen type, single versus multi-organ, review type (internal versus external), and definition of discrepancy, studies are remarkably consistent in finding non-trivial discrepancy rates.

\textsuperscript{c} Few studies reported the impact of discrepant diagnoses on treatment choice, and even fewer on patient outcome.

\textsuperscript{d} Inter-rater reliability studies had high variability. Experts agreed better than non-experts.

\textsuperscript{e} Inter-rater agreement studies usually used highly selected samples and small numbers of observations.
Table 4: Evidence profile for Recommendation 2: Should prospective review versus retrospective review be used for diagnostic evaluation of pathology materials?

<table>
<thead>
<tr>
<th>Quality assessment</th>
<th>No. of patients</th>
<th>Effect</th>
<th>Quality</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No of studies</strong></td>
<td>Design</td>
<td>Risk of bias</td>
<td>Inconsistency</td>
<td>Indirectness</td>
</tr>
<tr>
<td>4 observational studies</td>
<td>no serious risk of bias</td>
<td>no serious inconsistency</td>
<td>no serious indirectness</td>
<td>no serious imprecision</td>
</tr>
<tr>
<td>2 observational studies</td>
<td>no serious risk of bias</td>
<td>no serious inconsistency</td>
<td>no serious indirectness</td>
<td>no serious imprecision</td>
</tr>
</tbody>
</table>

CI – confidence interval; RD – risk difference
Table 5: Evidence Profile for Recommendation 3: Should targeted case review versus random case review be used for diagnostic evaluation of pathology materials?

<table>
<thead>
<tr>
<th>Quality assessment</th>
<th>No. of patients</th>
<th>Effect</th>
<th>Quality</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of studies</td>
<td>Design</td>
<td>Risk of bias</td>
<td>Inconsistency</td>
<td>Indirectness</td>
</tr>
<tr>
<td>Discrepancy rate</td>
<td>1</td>
<td>observational studies</td>
<td>no serious risk of bias</td>
<td>no serious inconsistency</td>
</tr>
<tr>
<td>Major discrepancy</td>
<td>1</td>
<td>observational studies</td>
<td>no serious risk of bias</td>
<td>no serious inconsistency</td>
</tr>
</tbody>
</table>

CI – confidence interval; OR – Odds ratio
Figure 1: Discrepancy rate distribution by sample size$^6$-112
Figure 2: Major discrepancy rate distribution by sample size.

**APPENDIX**

Appendix A: Literature Review Results
Adapted with permission from Moher et al.130

Records identified through database searching (n =676)

Records after duplicates removed (n=823)

Records screened (n=823)

Full-text articles assessed for eligibility (n =294)

Studies included for extraction and grading (n = 141)

Records excluded (n =529)

Full-text articles excluded, with reasons (n =153)

Data extraction articles excluded* (n = 4)

*Excluded based on expert opinion, did not meet minimum quality standards, presented incomplete data or data that were not in useable formats

Appendix B: Ovid MEDLINE Search Strategy

1. *Pathology/
2. *Pathology, Surgical/
3. Pathology, Clinical/
4. Pathology department, Hospital/
5. Cytodiagnosis/
6. Biopsy/
7. **"Diagnostic Techniques and Procedures"*/
8. (pathology$ or cytolog$ or histolog$ or histopatholog$ or cytopatholog$).tw.
9. or/1-8
AND

1. Medical Errors/
2. "Referral and Consultation"
3. Quality Assurance, Health Care/
4. Quality Control/
5. exp Diagnostic Errors/
6. Quality Improvement/
7. exp "Peer Review"
8. "Root Cause Analysis"
9. Total Quality Management/
10. "Reproducibility of Results"
11. "Sensitivity and Specificity"
12. Medical Audit/
13. "Insurance Claim Review"
15. performance improvement.tw.
16. (corrected or amended) adj2 report$.tw.
17. (misinterpretation or misdiagnosis or medicolegal or "patient safety" or "second opinion").tw.
18. (interpret$ or diagnostic) adj2 error$.tw.
19. (error adj2 (reduction or prevention or rate)).tw.
21. (quality adj3 (improvement or control or assurance or practice$ or measure$ or process$)).tw.
22. or/1-21

AND

1. (routine or target$ or random$ or blind$ or peer$) adj4 review$.tw.
2. (case$ or slide$ or report$) adj4 review$.tw.
3. or/1-2

REFERENCES


121. Li X, Heller K, Cangiarella J, Simsir A. Interinstitutional second opinion in thyroid cytology: should second opinion be mandated prior to definitive surgery if fine needle aspiration was performed elsewhere? *CytoJournal*. 2011;8(3):S63.


