COLLEGE of AMERICAN PATHOLOGISTS

Supplemental Digital Content* | Methodology | May 2021

Validating Whole Slide Imaging for Diagnostic Purposes in Pathology: Guideline Update

Authors:
Andrew J. Evans, MD, PhD, FCAP
Christina Lacchetti, MHSc
Kearin Reid, MLIS, AHIP, MLS(ASCP)
Nicole E. Thomas, MPH, CT(ASCP)

Evans A, Brown RW, Bui MM, et al. Validating whole slide imaging systems for diagnostic purposes in pathology: guideline update from the College of American Pathologists in Collaboration with the American Society for Clinical Pathology and the Association for Pathology Informatics. *Arch Pathol Lab Med*. 2021;146(4):440-450. doi: 10.5858/arpa.2020-0723-CP

*The Supplemental Digital Content was not copyedited by *Archives of Pathology & Laboratory Medicine*.

Supplemental Digital Content. The Supplemental Digital Content was not copyedited by *Archives of Pathology & Laboratory Medicine*.

METHODS USED TO PRODUCE THE GUIDELINE

Panel Composition

The College of American Pathologists (CAP) in collaboration with the American Society for Clinical Pathology (ASCP) and the Association for Pathology Informatics (API) convened an expert panel (EP) consisting of nine pathologists, two laboratory professionals, and a methodologist consultant to update the 2013 Validating Whole Slide Imaging for Diagnostic Purposes guideline. The CAP approved the appointment of the project chair and panel members. The EP members performed the systematic evidence review, drafted the recommendations, evaluated the public comments, revised the recommendations and contributed to the manuscripts.

An advisory panel (AP) of seven pathologists also helped in the development of the guideline. The role of the AP members was to provide guidance and feedback on the scope and key questions for the literature search, vet the draft recommendations prior to the public comment period, and to review and provide feedback for the manuscript and supplemental digital content (SDC).

Conflict of Interest (COI) Policy

Prior to acceptance on the expert or advisory panel, potential members completed the collaborative conflict of interest (COI) disclosure process, whose policy and form (in effect November 2017) require disclosure of material financial interest in, or potential for benefit of significant value from, the guideline's development or its recommendations 24 months prior through 12 months post-publication. The majority of expert panel members cannot have relevant conflicts of interest. Potential members completed the COI disclosure form, listing any relationship that could be interpreted as constituting an actual, potential, or apparent conflict. Each potential expert panel member's disclosures were assessed by three COI review teams (CAP staff and member groups) and categorized as:

No Relevant Conflicts of Interest**:** Individuals with no relevant COI are approved for full participation including determining the scope and questions to be addressed, reviewing and discussing the evidence, formulating and grading recommendations, voting on recommendations, and writing the document. Research funding that is free of direct or indirect industry funding or control, such as that provided by a government program or a non-profit organization that does not receive industry funding and uses an award mechanism and oversight that is independent of industry, is not regarded to be a conflict of interest. Service on a data and safety monitoring board for such research is also not regarded as a conflict of interest. Finally, industry funded research unrelated to the content of the *Recommendations* is not regarded as a conflict of interest.

Manageable Conflicts of Interest: Individuals with manageable conflicts must disclose their conflicts to the whole guideline panel (done via report at every meeting). They may participate in discussions about the evidence, but must excuse themselves or be recused from decision-making, including formulating, voting on, writing, and grading recommendations related to their COI (i.e., recommendations addressing a product of the

commercial entity with which they have a relationship or addressing a product of a competitor of the commercial entity with which they have a relationship). COI that require management include:

A. Research funding from an industry grant that is paid to the participant's institution and related to the content of the *Recommendations*;
B. Research funding from a government program or non-profit organization that receives funding from industry with business interests in the content of the *Recommendations*;
C. Participation on a data and safety monitoring board concerned with research that is relevant to the content of the *Recommendations* and is funded by an industry with business interests in the content of the *Recommendations*, or by a government program or non-profit organization that receives funding from industry with business interests in the content of the *Recommendations*.
D. Participation in scientific advisory board or consultant activities that are exclusively scientific in nature (i.e., does not involve any activities that could be perceived as promotional) related to the subject matter of the *Recommendations.*
E. Participation in industry-funded research, scientific advisory committees, consulting roles, non-promotional speaking engagements, or expert testimony on matters that are unrelated to the subject matter of the *Recommendations*, but the company involved is known to have business interest in the subject matter;
F. Delivery of non-promotional talks in which the speaker has full control of the content and is either unpaid or paid by a third party that is responsible for ensuring that the event is free of influence of relevant industry (i.e. if the event has industry financial support, all planning and content must be free of industry influence, and any payment of expenses and honoraria must occur through a third party, such as the medical society or institution sponsoring the event, or an event manager acceptable to them, rather than directly by a commercial entity with an interest in guideline subject matter or its agent);
G. Professional roles or activities (i.e., roles and activities performed as part of an individual's profession, whether reimbursed or not) that place an individual in a position to personally gain or lose depending upon the recommendations.

<u>Disqualifying Conflicts of Interest</u>: Disqualifying conflicts of interest include the following:

A. A direct financial relationship with a relevant commercial entity that has an interest in the content of the *Recommendations*, exclusive of the research, data safety monitoring board activities, and scientific advisory board and consultant activities noted above. Such direct financial relationships include the following, whether paid to or held by the individual directly or issued to another entity at the direction of the individual (such as to a panelist's institution):
    i. Payment of wages, consulting fees, honoraria, or other payments (in cash, in stock or stock options, or in kind) by a relevant company as compensation for the individual's services or expertise, exclusive of the research and data safety monitoring board activities noted above. Examples of such services are: participation on scientific advisory committees or consulting that is, in full or in part, promotional in nature; non-continuing medical education speaking engagements and inclusion in speaker bureaus where control of material is held by industry; expert testimony on matters related to guideline content provided on behalf of a relevant company or a law firm representing a relevant company; employment by a relevant commercial entity (such as a relevant pharmaceutical or medical device company or a third party payer exclusive of commercial laboratory employment that has financial interests in the content of the *Recommendations*).

ii. Investments in relevant companies by the panelist or the panelist's spouse or life partner (exclusive of general mutual funds).

B. A patent or other intellectual property that is relevant to the *Recommendations'* subject matter and has resulted or could result in payments to the panelist or the panelist's institution.

All panel members were required to disclose conflicts prior to beginning and continuously throughout the project's timeline.

Funding

The CAP provided funding for the administration of the project; no industry funds were used in the development of the guideline.

Disclosures of interest judged by the oversight group as manageable conflicts are listed in the manuscript. Appendix 1 in the manuscript also includes a table of all disclosed interest of the expert panel members during the development of the guideline for complete transparency.

Systematic Evidence Review (SER)

The objective of the SER was to identify articles of sufficient quality that would provide data to inform the recommendations. The scope of the SER and the key questions (KQs) were established by the EP and AP in consultation with the methodologist prior to beginning the literature search.

Search and Selection

A comprehensive literature search was performed in Ovid on 6/26/2018. The search was limited to 1/1/2012 - 6/26/2018. An additional search was performed in Embase with the same search date parameters. The databases searches used indexed terms and keywords for the concepts of whole slide imaging for diagnostic purposes; concordance; validation; and methodological, analytical, and procedural variables. Search results were limited to English language and the Cochrane search filter for humans was applied. A publication filter to exclude letters, commentaries, editorials, case reports, and conference abstracts was added. Results of both searches were combined, and duplicate references were removed. A literature search refresh was completed in Ovid and Embase on 6/14/2019 and 07/15/2020. Both search strategies can be found in Supplemental Figure 1.

A search for grey (unindexed) literature included a review of the ClinicalTrials.gov, Cochrane Library, Guidelines International Network, Trip search engine, and applicable U.S. and international organizational websites.

Study Selection Criteria

Eligible Study Designs

Included study types: clinical practice guidelines, systematic reviews, randomized controlled trials, observational studies, non-comparative studies. Excluded studies: follow-up studies, qualitative studies, mixed methods studies, time series, narrative reviews, consensus documents, letters, comments, editorials, meeting abstracts.

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

Studies included:

- Peer-reviewed full-text articles
- Referred to whole slide imaging (WSI)
- Included frozen sections, hematopathology cases, and cases submitted to anatomic pathology (biopsy, curetting, resection).
- Pertained to clinical research

Not included:

- Letters
- Commentaries
- Editorials
- Time series
- Conference abstracts
- Mixed methods studies
- Qualitative studies
- Follow-up studies
- Studies in animal models or cell lines
- Articles not in the English language
- Studies included less than 30 patients per study arm
- Studies that discussed only cytology cases
- Studies involving static and robotic digital imaging, purely technical components, only educational applications, and image analysis

Outcomes of Interest

The EP deemed the following as outcomes of interest: diagnostic concordance between WSI and glass slides/light microscopy, intra-observer agreement, major and minor discrepancies, number and types of cases for validation, time (washout period); and effects of random versus non-random order of slide review.

Data Extraction & Management

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

Literature Review and Analysis

The EP met 12 times through teleconference webinars from April 19, 2018, through March 19, 2020. Additional work was completed via electronic mail. The panel met in person July 23, 2018 to confirm the project scope and key question and February 3, 2019 to review evidence from the systematic review and draft recommendations.

The EP sought to answer what should be done to validate a whole slide digital imaging system for diagnostic purposes before it is placed in clinical service. This was the same key question addressed in the original guideline.

All EP members participated in the systematic evidence review (SER): title-abstract screening, full-text review, and data extraction. A dual review was performed for each study and in each phase of the SER; the chair adjudicated all conflicts. A literature refresh was also conducted, where studies also underwent dual review. A total of 62 studies comprised the final body of studies included in the SER. Supplemental Figure 2 displays the results of the literature review. All articles were available as discussion or background references. All members of the EP participated in developing draft recommendations, reviewing open comment feedback, finalizing and approving the final recommendations, and writing/editing of the manuscript.

Peer Review

A public, open access comment period was held from June 24 through July 15, 2019 on the CAP Web site www.cap.org for any interested stakeholder to provide feedback on the draft recommendations. Three draft recommendations, nine good practice statements, two demographic questions, three questions to assess feasibility, and five supplemental questions were posted for feedback. An announcement was sent to the following societies deemed stakeholders:

Medical Societies

- Association of Directors of Anatomic and Surgical Pathology (ADASP)
- Association of Pathology Chairs (APC)
- Association for Molecular Pathology (AMP)
- Association for Pathology Informatics (API)
- American Society of Cytopathology (ASC)
- American Society for Cytotechnology (ASCT)
- American Society for Clinical Pathology (ASCP)
- College of American Pathologists (CAP)
- Canadian Association of Pathologists (CAP-APC)
- Canadian IHC Quality Control (CIQC) program
- Digital Pathology Association (DPA)
- International Academy of Pathology
- International Society for Immunohistochemistry and Molecular Morphology (ISIMM)
- National Society for Histotechnology (NSH)

- Nordic IHC Quality Control (NordiQC) program
- Papanicolaou Society of Cytopathology (PSC)
- Quality Initiative in Interpretive Pathology (QIIP) Canadian Partnership Against Cancer
- Society to Improve Diagnosis in Medicine (SIDM)
- United States & Canadian Academy of Pathology (USCAP)
- Sociedade Brasileira de Patologia (Brazilian Society of Pathology) (SBP)

Government

- CDC, Centers for Medicare & Medicaid Services (CMS)
- CDC, Division of Laboratory Systems (DLS)
- CDC, Laboratory Medicine Best Practices (LMBP)
- National Medical Products Administration (NMPA)European Medicines Agency
- National Institute for Health and Care Excellence (NICE) (UK)
- NIH, Center for Strategic Scientific Initiatives (CSSI)
- NIH, Division of Cancer Treatment and Diagnosis (DCTD)
- US Food and Drug Administration (FDA)

Industries

- 3DHistech
- Excilone
- Huron Digital Pathology
- Leica (Aperio)
- Mikroscan Technologies, Inc.
- Motic Digital Pathology

One hundred fifty-four individuals participated in the comment period. “Agree,” Agree with modification,” and “Disagree” responses were captured for every proposed recommendation and good practice statement (GPS). One hundred forty-six written responses were also collected. All three recommendations and nine good GPSs received greater than 90% agree/agree with modification. The EP read all the comments and the chair led the members in discussion to determine if any should change. Only one GPS received a minor revision.

Decisions were obtained by majority consensus of the panel using nominal group technique (discussion at an in-person meeting, rounds of teleconference webinars, email discussion, and multiple edited recommendations) amongst the panel members. The final recommendations were agreed upon by the EP with a formal vote. The panel considered laboratory efficiency and feasibility throughout the entire considered judgment process. Seventy-six and thirty-four hundredths percent (76.34% [71 of 93]) responded that the entire guideline was feasible, 23.66% (22 of 93) responded that parts of it were feasible, and 0% (0 of 93) responded that none of it was feasible. Neither formal cost analysis nor cost effectiveness models were performed.

An independent review panel (IRP) was assembled to review and approve the guideline on behalf of the CAP Council on Scientific Affairs. The IRP was masked to the EP and to each other and were vetted through the COI process. Both the ASCP and API have approved and endorsed the guideline manuscript and supplement ahead of publication.

Quality Assessment Methods

A risk of bias assessment was performed for all retained studies following application of the inclusion and exclusion criteria. Using this method, studies deemed be of low quality would not be excluded from the systematic review, but would be retained, and their methodological strengths and weaknesses discussed where relevant. To define an overall study quality rating for each included study, validated study-type specific tools were used to assess the risk of bias, plus additional important quality features were extracted. Specific details for each study type are outlined below.

Clinical practice guidelines (CPGs)

- The following attributes were considered as per the AGREE II[1] tool using a seven-point scale:
  1. Scope and purpose
  2. Stakeholder involvement
  3. Rigor of development
  4. Clarity of presentation
  5. Applicability
  6. Editorial independence

Systematic Reviews (SRs) and Meta-analyses (MAs)

- The following questions were assessed as per the Assessing the Methodological Quality of Systematic Reviews (AMSTAR)[2] tool using yes, no, or unclear:
  1. Was an ‘a priori’ design provided?
  2. Was there duplicate study selection and data extraction?
  3. Was a comprehensive literature search performed?
  4. Was the status of publication (i.e., grey literature) used as an inclusion criterion?
  5. Was a list of studies (included and excluded) provided?
  6. Were the characteristics of the included studies provided?
  7. Was the scientific quality of the included studies assessed and documented?

8. Was the scientific quality of the included studies used appropriately in formulating conclusions?
9. Were the methods used to combine the findings of studies appropriate?
10. Was the likelihood of publication bias assessed?
11. Was the conflict of interest included?

- Additional assessed items included and were assessed as yes, no, or unclear:
  1. Reporting of funding sources.

Diagnostic accuracy studies

- The following four domains were assessed using the QUADAS-2[3] (Quality Assessment of Diagnostic Accuracy Studies) tool using low risk, high risk, or unclear risk of bias. Concerns about applicability was also assessed across the first three domains.
  1. Patient selection
  2. Index test
  3. Reference standard
  4. Flow of participants

Quality Assessment Results

A total of 221 studies were included in our systematic review with the evidentiary base supporting the recommendations consisting of 62 studies. This body of evidence comprised of two guidelines,[4, 5] one health technology assessment,[6] three systematic reviews,[7-9] and 56 diagnostic studies. Data retrieval and review occurred prospectively in 25.8% (n =16) of the studies and retrospectively in 48.4% (n=30). Nine studies[10-18] (14.5%) were non-inferiority in design. In the following sections, the quantity of the evidence as determined by the number of studies that met our inclusion criteria and were retained, the evidence type as determined by study design, the quality of that evidence as determined by the quality assessment, and its consistency are all reported, both as individual studies and in totality by outcome. The data was assessed qualitatively, and no meta-analyses nor formal testing for publication bias were performed. To help control for post- publication bias, a wide variety of databases and sources of evidence were searched, including grey literature and hand searches of reference lists of relevant articles.

Overall, the body of evidence included in this CPG represents a methodologically rigorous and representative summary of the available evidence with quality of evidence that ranges from low to high and an overall low to moderate aggregate risk of bias.

Following the quality assessment, each outcome was given a grade for quality of evidence. (Supplemental Table 1).

Assessing the Strength of Recommendations

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

1. What are the significant findings related to each KQ? This includes any regulatory requirements.
2. What is the overall quality of evidence supporting each outcome?
3. What is the strength of each recommendation?

4. What is the net balance of benefits and harms? The panel used the Evidence to Decision Framework[19] to frame, discuss, and document their decisions for each recommendation.

Strength of Recommendation, Quality of Evidence, Quality Assessment, and Summary of the Benefits and Harms by Recommendation

1. *Strong Recommendation*. – The validation process should include a sample set of at least 60 cases for one application (eg, hematoxylin and eosin [H&E] stained sections of fixed tissue, frozen sections, cytology, hematology) that reflects the spectrum and complexity of specimen types and diagnoses likely to be encountered during routine practice.
Note: The validation process should include another 20 cases for each additional application (eg, immunohistochemistry, special stains).
(Recommendation reaffirmed)
The quality of evidence is *moderate* to support this recommendation. Refer to Supplemental Tables 2-3 for the quality assessment results of the one systematic review[7] and the 32 diagnostic studies informing this recommendation.[9, 10, 12-14, 16-18, 20-43] The risk of bias concerns for applicability of included patients and reference standard was low for all but one[26] included study. While the majority of studies had a low risk of bias for the applicability of the index test, seven studies[17, 26, 28, 30, 31, 33, 36] had an unclear risk, as did one study[33] for the applicability of the reference standard. Seven studies carried an unclear risk of bias in the patient selection domain,[10, 12, 26, 27, 32, 38, 39] eleven in the index test domain,[10, 13, 17, 25, 26, 28, 30, 31, 33, 36, 37] five in the reference standard domain[12, 25, 33, 37, 39] and three in the flow and timing domain.[26, 27, 36] An assessment of consistency, directness and precision of outcomes across studies found that there were serious concerns with inconsistency in the retrospective studies, but no serious concerns for other domains across all study designs. Refer to Supplemental Table 4 for the GRADE rating.

Benefits and Harms of Implementing Recommendation 1:
Based on the available evidence and the values included in the Evidence to Decision framework,
all EP members participating in this discussion believed providing evidence-based guidance on a minimum number of cases to be a priority. All EP members also believed the recommendation was feasible to implement, the benefits of following the recommendation outweighed the harms and that the recommendation was accurate 70% (7/10) to very accurate 20% (2/10), with only 10% (1/10) who thought that accuracy varies. Seventy percent (7/10) reported the certainty of using 60 cases was high or very high. Ninety percent (9/10) of EP members believed the recommendation was acceptable to all stakeholders, while 10% (1/10) believed the recommendation would have variable acceptability. In terms of cost and resource requirements, 70% (7/10) of EP members felt the impact would be moderate to large. The EP was essentially divided on whether the balance between desirable and undesirable effects favored the comparison (ie: 60 cases) or the intervention (ie: number of cases other than 60) at 60% and 40% respectively. Refer to Supplemental Table 5 for the Evidence to Decision Summary for Recommendation 1.

2. Strong Recommendation. - The validation study should establish diagnostic concordance between digital and glass slides for the same observer (i.e., intraobserver variability). If concordance is less than 95%, laboratories should investigate and attempt to remedy the cause.

(Recommendation updated)

The quality of evidence is *moderate* to support this recommendation. Refer to Supplemental Tables 6-7 for the quality assessment results of the one systematic review[7] and the 32 diagnostic accuracy studies[7, 10, 12-14, 17, 18, 20-39, 41-44] informing this recommendation. Data collection occurred prospectively in six studies[21-23, 25, 28, 40] and retrospectively in 18.[20, 24, 26, 27, 29-32, 34-39, 41-44] One study had both prospective data collection at one site and retrospective data collection using archived specimens at two other sites.[33] Seven non-inferiority studies[10, 12-14, 16-18] were also included. Seven individual studies carried an unclear risk of bias in the patient selection domain,[10, 12, 26, 27, 32, 38, 39] eleven in the index test domain,[10, 13, 17, 25, 26, 28, 30, 31, 33, 36, 37] five in the reference standard domain[12, 25, 26, 33, 37] and three in the flow and timing domain.[26, 27, 36] The risk of bias concerns for applicability of included patients and reference standard was low for all but one[26] included study. While the majority of studies had a low risk of bias for the applicability of the index test, seven studies[17, 26, 28, 30, 31, 33, 36] had an unclear risk, as did one study[33] for the applicability of the reference standard. An assessment of consistency, directness and precision of outcomes across studies found that there were serious concerns with inconsistency in the retrospective studies, but no serious concerns for other domains across all study designs. Refer to Supplemental Table 8 for the GRADE rating.

Benefits and Harms of Implementing Recommendation 2
Based on the available evidence and the values in the Evidence to Decision framework, all EP members participating in the discussion believed providing this recommendation to be a priority, that the recommendation would be feasible to implement and the link between the results obtained by following the recommendation and decisions on WSI implementation for diagnostic purposes to be high. Ninety percent (9/10) of EP members believed the certainty of evidence of the effects of the recommendation to be high to very high. The EP members were equally divided as to whether the balance between desirable and undesirable effects would favor the intervention (WSI at 5/10) or the comparison (glass slides at 5/10). In terms of cost/resource requirements, 60% (6/10) of EP members believed the impact would be moderate to large. Ninety percent (9/10) of EP members believed the recommendation would be or probably would be acceptable to key stakeholders, with 10% (1/10) believing the recommendation to be of variable acceptability. Refer to Supplemental Table 9 for the Evidence to Decision Summary for Recommendation 2.

3. Strong Recommendation. - A washout period of at least two weeks should occur between viewing digital and glass slides.

(Recommendation reaffirmed)
The quality of evidence is *moderate* to support this recommendation. Refer to Supplemental Table 10 for the quality assessment results of the 14 studies informing this recommendation.[18, 21-24, 28, 30-33, 35, 39, 45, 46] Data collection occurred prospectively in four studies[21-23, 28] and retrospectively in eight.[24, 30-32, 35, 39, 45, 46] One study had both prospective data collection at one site and retrospective data collection using archived specimens at 2 other sites.[33] A non-inferiority study[18] was also included. Individual studies carried an unclear risk of bias in the patient selection domain,[32, 39] index test domain,[28, 30, 31, 33] reference standard domain[33] and flow and timing domain.[46] An assessment of consistency, directness and precision of outcomes across studies found that there were serious concerns with inconsistency in the retrospective studies and a serious concern with imprecision in the prospective/retrospective study, but no serious concerns for other domains across all study designs Refer to Supplemental Table 11 for the GRADE rating).

Benefits and Harms of Implementing Recommendation 3:
Based on the available evidence and the values in the Evidence to Decision framework, all EP members participating in the discussion believed this recommendation was feasible to implement and 80% (8/10) believed it would probably be or would be acceptable to key stakeholders. The

EP members were divided on many issues concerning this recommendation including its accuracy, certainty around the accuracy of the supporting evidence, certainty of the evidence of the effects of following the recommendation, certainty of the evidence for decisions on WSI implementation that are guided by the recommendation, certainty of evidence for benefits or adverse effects arising from the recommendation as well as whether the balance of its desirable or undesirable effects would favor WSI or glass slides. The presence of a divided opinion on the merits of washout periods was also apparent in responses obtained during the open comment period. Refer to Supplemental Table 12 for the Evidence to Decision Summary for Recommendation 3.

Dissemination Plans

The CAP hosts a WSI Validation Guideline Web page which will include a link to the manuscript and supplement; a summary of the recommendations, a PowerPoint slide deck (Microsoft Corporation, Redmond, WA), a frequently asked question (FAQ) document, and an infographic. The guideline will be promoted and presented at various society meetings.

Supplemental Table 1: Grades for Quality of Evidence

GRADE	Definition
High	We are very confident that the true effect lies close to that of the estimate of the effect.
Moderate	We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different.
Low	Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect.
Very Low	We have very little confidence in the effect estimate. The true effect is likely to be substantially different from the estimate of effect.

GRADE Guidelines[47]

Supplemental Table 2. Quality Assessment Results for Studies Supporting Recommendation 1 – Systematic Reviews

Reference	PICO Used	Methods and Protocols Defined in Advance and Deviations Reported	Selection of study designs for inclusion explained	Comprehensive literature search strategy	Duplicate study selection	Dual Data Extraction	Excluded study list provided	Included studies described adequately	Satisfactory RoB technique used	Funding sources reported	Account for RoB in individual studies	Satisfactory explanation/ discussion of heterogeneity observed in results	Conflicts of interest reported
Williams[7] 2017	No	Partial Yes, included review questions, search strategy, inclusion/exclusion criteria, ROB assessment	No	Yes	Yes	Yes	No	Partial Yes. Population, interventions, comparators, outcomes and research design described	No	No	No	No	Yes

PICO = Patient/Population, Intervention, Comparator, Outcome(s); RoB = Risk of bias

Supplemental Table 3. Quality Assessment Results for Studies Supporting Recommendation 1 – Diagnostic Studies

References	Risk of Bias				Applicability			OVERALL Quality*
	Selection of patients	Index test	Reference standard	Flow & timing	Selection of patients	Index test	Reference standard	
Araujo[20] 2018	Low	Low	Low	Low	Low	Low	Low	High
Rakha[44] 2018	Low	Low	Low	Low	Low	Low	Low	High
Villa[21] 2018	Low	Low	Low	Low	Low	Low	Low	High

Williams[22] 2018	Low	Low	Low	Low	Low	Low	Low	High
Tabata[23] 2017	Low	Low	Low	Low	Low	Low	Low	High
Bauer[10] 2015	Unclear	Unclear	Low	Low	Low	Low	Low	Low
Al-Janabi[24] 2014	Low	Low	Low	Low	Low	Low	Low	High
Gui[25] 2012	Low	Unclear	Unclear	Low	Low	Low	Low	Low
Fertig[26] 2017	Unclear	Unclear	Unclear	Unclear	Unclear	Unclear	Low	Low
Mukhopadhyay[12] 2018	Unclear	Low	Unclear	Low	Low	Low	Low	Low
Kent[13] 2017	Low	Unclear	Low	Low	Low	Low	Low	Moderate
Snead[14] 2016	Low	Low	Low	Low	Low	Low	Low	High
Gomez-Gelvez[27] 2015	Unclear	Low	Low	Unclear	Low	Low	Low	Low
Loughrey[28] 2015	Low	Unclear	Low	Low	Low	Unclear	Low	Low
Thrall[29] 2015	Low	Low	Low	Low	Low	Low	Low	High
Arnold[30] 2015	Low	Unclear	Low	Low	Low	Unclear	Low	Low
Ordi[16] 2015	Low	Low	Low	Low	Low	Low	Low	High
Houghton[31] 2014	Low	Unclear	Low	Low	Low	Unclear	Low	Low
Campbell[32] 2014	Unclear	Low	Low	Low	Low	Low	Low	Moderate
Gage[17] 2013	Low	Unclear	Low	Low	Low	Unclear	Low	Low
Krishnamurthy[33] 2013	Low	Unclear	Unclear	High	Low	Unclear	Unclear	Low
Bauer[18] 2013	Low	Low	Low	Low	Low	Low	Low	High
Al-Janabi[34] 2012	Low	Low	Low	Low	Low	Low	Low	High
Al-Janabi[35] 2012	Low	Low	Low	Low	Low	Low	Low	High
Campbell[36] 2012	Low	Unclear	Low	Unclear	Low	Unclear	Low	Low
Fonyad[37] 2012	Low	Unclear	Unclear	Low	Low	Low	Low	Low
Al-Janabi[38] 2012	Unclear	Low	Low	Low	Low	Low	Low	Moderate
Al-Janabi[39] 2012	Unclear	Low	Low	Low	Low	Low	Low	Moderate
Hanna[40] 2019	Low	Low	Low	Low	Low	Low	Low	High
Larghi[41] 2019	Unclear	Low	Low	Low	Low	Low	Low	Moderate

Cima[42] 2018	Low	Low	Low	Low	Low	Low	Low	High
Sturm[43] 2019	Low	Low	Low	Low	Low	Low	Low	High

*Overall quality was assessed to be high when risk bias or concern about applicability was low in all domains. A study was assessed to have moderate quality if there was no more than one unclear domain and no high risk of bias or concern about applicability. A study was of low quality if there was more than one unclear domain or any high risk of bias/concern about applicability in the any domain.

Supplemental Table 4. GRADE Rating for Studies Supporting Recommendation 1

No. of studies	Total number of cases included	Study Design	Aggregate RoB	Inconsistency	Indirectness	Imprecision	GRADE Certainty Rating
Diagnostic Concordance:							
1	8069	Systematic Review	Moderate	Not serious	Not serious	Not serious	Moderate
18	4474	Retrospective diagnostic studies	Moderate	Serious	Not serious	Not serious	Low
6	2185	Prospective diagnostic studies	Moderate	Not serious	Not serious	Not serious	Moderate
1	150	Prospective/retrospective diagnostic studies	Moderate	Not serious	Not serious	Serious	Low
7	6940	Non-inferiority	Low	Not serious	Not serious	Not serious	Moderate

No.= Number RoB = Risk of bias

Supplemental Table 5. Evidence to Decision Summary for Recommendation 1

Is the problem a priority?	No	Probably No	Probably Yes	Yes	Varies	Don't Know
	0	0	30% (3/10)	70% (7/10)	0	0
How accurate is the test?	Very Inaccurate	Inaccurate	Accurate	Very Accurate	Varies	Don't Know
	0	0	70% (7/10)	20% (2/10)	10% (1/10	0
How substantial are the desirable anticipated effects?	Trivial	Small	Moderate	Large	Varies	Don't Know
	0	0	30% (3/10)	60% (6/10)	10% (1/10)	0
How substantial are the undesirable anticipated effects?	Large	Moderate	Small	Trivial	Varies	0
	50% (5/10)	10% (1/10)	20% (2/10)	0	20% (2/10)	0
	Very Low	Low	Moderate	High	Very High	No Included Studies

What is the overall certainty of the evidence of test accuracy?	0	0	30% (3/10)	60% (6/10)	10% (1/10)	0	
What is the overall certainty of the evidence of effects of the management that is guided by the test results?	Very Low	Low	Moderate	High	Very High	No Included Studies	
	0	20% (2/10)	20% (2/10)	60% (6/10)	0	0	
How certain is the link between test results and management decisions?	Very Low	Low	Moderate	High	Very High	No Included Studies	
	0	10% (1/10)	40% (4/10)	20% (2/10)	30% (3/10)	0	
What is the overall certainty of the evidence of effects of the test?	Very Low	Low	Moderate	High	Very High	No Included Studies	
	0	0	30% (3/10)	40% (4/10)	30% (3/10)	0	
What is the overall certainty of the evidence for any critical or important direct benefits, adverse effects or burden of the test?	Very Low	Low	Moderate	High	Very High	No Included Studies	
	0	0	30% (3/10)	50% (5/10)	20% (2/10)	0	
Is there important uncertainty about or variability in how much people value the main outcomes?	Important uncertainty or variability	Possibly important uncertainty or variability	Probably no important uncertainty or variability	No important uncertainty or variability			
	10% (1/10)	50% (5/10)	30% (3/10)	10% (1/10)			
Does the balance between desirable and undesirable effects favor the intervention or the comparison?	Favors the comparison	Probably favors the comparison	Does not favor either the intervention or comparison	Probably favors the intervention	Favors the intervention	Varies	Don't Know
	30% (3/10)	30% (3/10)	0	0	40% (4/10)	0	0
How large are the resource requirements (costs)?	Large costs	Moderate costs	Negligible costs and savings	Moderate savings	Large savings	Varies	Don't Know
	10% (1/10)	60% (6/10)	30% (3/10)	0	0	0	0
What is the certainty of evidence of resource requirements (costs)?	Very Low	Low	Moderate	High	Very High	No Included Studies	

	10% (1/10)	20% (2/10)	30% (3/10)	20% (2/10)	20% (2/10)	0	
Does the cost-effectiveness of the intervention favor the intervention or the comparison	Favors the comparison	Probably favors the comparison	Does not favor either the intervention or comparison	Probably favors the intervention	Favors the intervention	Varies	No included studies
	20% (2/10)	30% (3/10)	0	0	40% (4/10)	0	10% (1/10)
What would be the impact on health equity?	Reduced	Probably reduced	Probably no impact	Probably increased	Increased	Varies	Don't Know
	0	0	20% (2/10)	20% (2/10)	30% (30/10)	10% (1/10)	0
Is the intervention acceptable to key stakeholders?	No	Probably no	Probably yes	Yes	Varies	Don't Know	
	0	0	0	90% (9/10)	10% (1/10)	0	
Is the intervention feasible to implement?	No	Probably no	Probably yes	Yes	Varies	Don't Know	
	0	0	0	100% (10/10)	0	0	

Supplemental Table 6. Quality Assessment Results for Studies Supporting Recommendation 2 – Systematic Review

Reference	PICO Used	Methods and Protocols Defined in Advance and Deviations Reported	Selection of study designs for inclusion explained	Comprehensive literature search strategy	Duplicate study selection	Dual Data Extraction	Excluded study list provided	Included Studies described adequately	Satisfactory RoB technique used	Funding sources reported	Account for RoB in individual studies	Satisfactory explanation/ discussion of heterogeneity observed in results	Conflicts of interest reported
Williams[7] 2017	No	Partial Yes, included review questions, search strategy, inclusion/exclusion criteria,	No	Yes	Yes	Yes	No	Partial Yes. Population, interventions, comparators, outcome	No	No	No	No	Yes

		ROB assessment						s and research design describe d					

PICO = Patient/Population, Intervention, Comparator, Outcome(s); RoB = Risk of bias

Supplemental Table 7. Quality Assessment Results for Studies Supporting Recommendation 2 – Diagnostic Studies

References	Risk of Bias				Applicability			OVERALL Quality*
	Selection of patients	Index test	Reference standard	Flow & timing	Selection of patients	Index test	Reference standard	
Araujo[20] 2018	Low	Low	Low	Low	Low	Low	Low	High
Rakha[44] 2018	Low	Low	Low	Low	Low	Low	Low	High
Villa[21] 2018	Low	Low	Low	Low	Low	Low	Low	High
Williams[22] 2018	Low	Low	Low	Low	Low	Low	Low	High
Tabata[23] 2017	Low	Low	Low	Low	Low	Low	Low	High
Bauer[10] 2015	Unclear	Unclear	Low	Low	Low	Low	Low	Low
Al-Janabi[24] 2014	Low	Low	Low	Low	Low	Low	Low	High
Gui[25] 2012	Low	Unclear	Unclear	Low	Low	Low	Low	Low
Fertig[26] 2017	Unclear	Unclear	Unclear	Unclear	Unclear	Unclear	Low	Low
Mukhopadhyay[12] 2018	Unclear	Low	Unclear	Low	Low	Low	Low	Low
Kent[13] 2017	Low	Unclear	Low	Low	Low	Low	Low	Moderate
Snead[14] 2016	Low	Low	Low	Low	Low	Low	Low	High
Gomez-Gelvez[27] 2015	Unclear	Low	Low	Unclear	Low	Low	Low	Low
Loughrey[28] 2015	Low	Unclear	Low	Low	Low	Unclear	Low	Low
Thrall[29] 2015	Low	Low	Low	Low	Low	Low	Low	High
Arnold[30] 2015	Low	Unclear	Low	Low	Low	Unclear	Low	Low
Ordi[16] 2015	Low	Low	Low	Low	Low	Low	Low	High
Houghton[31] 2014	Low	Unclear	Low	Low	Low	Unclear	Low	Low

Campbell[32] 2014	Unclear	Low	Low	Low	Low	Low	Low	Moderate
Gage[17] 2013	Low	Unclear	Low	Low	Low	Unclear	Low	Low
Krishnamurthy[33] 2013	Low	Unclear	Unclear	High	Low	Unclear	Unclear	Low
Bauer[18] 2013	Low	Low	Low	Low	Low	Low	Low	High
Al-Janabi[34] 2012	Low	Low	Low	Low	Low	Low	Low	High
Al-Janabi[35] 2012	Low	Low	Low	Low	Low	Low	Low	High
Campbell[36] 2012	Low	Unclear	Low	Unclear	Low	Unclear	Low	Low
Fonyad[37] 2012	Low	Unclear	Unclear	Low	Low	Low	Low	Low
Al-Janabi[38] 2012	Unclear	Low	Low	Low	Low	Low	Low	Moderate
Al-Janabi[39] 2012	Unclear	Low	Low	Low	Low	Low	Low	Moderate
Hanna[40] 2019	Low	Low	Low	Low	Low	Low	Low	High
Larghi[41] 2019	Unclear	Low	Low	Low	Low	Low	Low	Moderate
Cima[42] 2018	Low	Low	Low	Low	Low	Low	Low	High
Sturm[43] 2019	Low	Low	Low	Low	Low	Low	Low	High

*Overall quality was assessed to be high when risk bias or concern about applicability was low in all domains. A study was assessed to have moderate quality if there was no more than one unclear domain and no high risk of bias or concern about applicability. A study was of low quality if there was more than one unclear domain or any high risk of bias/concern about applicability in the any domain.

Supplemental Table 8. GRADE Rating for Studies Supporting Recommendation 2

No. of studies	Total number of cases included	Study Design	Aggregate RoB	Inconsistency	Indirectness	Imprecision	GRADE Certainty Rating
Diagnostic Concordance:							
1	8069	Systematic Review	Moderate	Not serious	Not serious	Not serious	Moderate
18	4474	Retrospective diagnostic studies	Moderate	Serious	Not serious	Not serious	Low
6	2185	Prospective diagnostic studies	Moderate	Not serious	Not serious	Not serious	Moderate
1	150	Prospective/retrospective diagnostic studies	Moderate	Not serious	Not serious	Serious	Low
7	6940	Non-inferiority	Low	Not serious	Not serious	Not serious	Moderate

No.= Number; RoB = Risk of bias

Supplemental Table 9. Evidence to Decision Summary for Recommendation 2

Is the problem a priority?	No	Probably No	Probably Yes	Yes	Varies	Don't Know
	0	0	0	100% (10/10)		
How accurate is the test?	Very Inaccurate	Inaccurate	Accurate	Very Accurate	Varies	Don't Know
	0	0	60% (6/10))	30% (3/10)	10% (1/10)	0
How substantial are the desirable anticipated effects?	Trivial	Small	Moderate	Large	Varies	Don't Know
	0	0	10% (1/10)	90% (9/10)	0	0
How substantial are the undesirable anticipated effects?	Large	Moderate	Small	Trivial	Varies	
	50% (5/10)	30% (3/10)	20% (2/10)	0	0	
What is the overall certainty of the evidence of test accuracy?	Very Low	Low	Moderate	High	Very High	No Included Studies
	0	10% (1/10)	0	70% (7/10)	20% (2/10)	0
What is the overall certainty of the evidence of effects of the management that is guided by the test results?	Very Low	Low	Moderate	High	Very High	No Included Studies
	0	10% (1/10)	10% (1/10)	50% (5/10)	30% (3/10)	0
How certain is the link between test results and management decisions?	Very Low	Low	Moderate	High	Very High	No Included Studies
	10% (1/10)	0	0	80%(8/10)	10% (1/10)	0
What is the overall certainty of the evidence of effects of the test?	Very Low	Low	Moderate	High	Very High	No Included Studies
	0	0	10% (1/10)	60% (6/10)	30% (3/10)	0
What is the overall certainty of the evidence for any critical or important direct benefits, adverse effects or burden of the test?	Very Low	Low	Moderate	High	Very High	No Included Studies
	10% (1/10)	0	10% (1/10)	70% (7/10)	10% (1/10)	0

Is there important uncertainty about or variability in how much people value the main outcomes?	Important uncertainty or variability	Possibly important uncertainty or variability	Probably no important uncertainty or variability	No important uncertainty or variability			
	20% (2/10)	30% (3/10)	40% (4/10)	10% (1/10)			
Does the balance between desirable and undesirable effects favor the intervention or the comparison?	Favors the comparison	Probably favors the comparison	Does not favor either the intervention or comparison	Probably favors the intervention	Favors the intervention	Varies	Don't Know
	20% (2/10)	30% (3/10)	0	0	50% (5/10)	0	0
How large are the resource requirements (costs)?	Large costs	Moderate costs	Negligible costs and savings	Moderate savings	Large savings	Varies	Don't Know
	10% (1/10)	50% (5/10)	30% (3/10)	0	0	0	10% (1/10)
What is the certainty of evidence of resource requirements (costs)?	Very Low	Low	Moderate	High	Very High	No Included Studies	
	0	10% (1/10)	20% (2/10)	30% (3/10)	20% (2/10)	20% (2/10)	
Does the cost-effectiveness of the intervention favor the intervention or the comparison	Favors the comparison	Probably favors the comparison	Does not favor either the intervention or comparison	Probably favors the intervention	Favors the intervention	Varies	No included studies
	30% (3/10)	10% (1/10)	0	0	50% (5/10)	0	10% (1/10)
What would be the impact on health equity?	Reduced	Probably reduced	Probably no impact	Probably increased	Increased	Varies	Don't Know
	0	0	0	20% (2/10)	60% (6/10)	0	20% (2/10)
Is the intervention acceptable to key stakeholders?	No	Probably no	Probably yes	Yes	Varies	Don't Know	
	0	0	20% (2/10)	70% (7/10)	10% (1/10)	0	
Is the intervention feasible to implement?	No	Probably no	Probably yes	Yes	Varies	Don't Know	
	0	0	0	100% (10/10)	0	0	

Supplemental Table 10. Quality Assessment Results for Studies Supporting Recommendation 3 – Diagnostic Studies

References	**Risk of Bias**	**Applicability**	

	Selection of patients	Index test	Reference standard	Flow & timing	Selection of patients	Index test	Reference standard	OVERALL Quality*
Villa[21] 2018	Low	Low	Low	Low	Low	Low	Low	High
Williams[22] 2018	Low	Low	Low	Low	Low	Low	Low	High
Tabata[23] 2017	Low	Low	Low	Low	Low	Low	Low	High
Wilbur[45] 2015	Low	Low	Low	Low	Low	Low	Low	High
Al-Janabi[24] 2014	Low	Low	Low	Low	Low	Low	Low	High
Malarkey[46] 2015	Low	Low	Low	Unclear	Low	Low	Low	Moderate
Loughrey[28] 2015	Low	Unclear	Low	Low	Low	Unclear	Low	Low
Arnold[30] 2015	Low	Unclear	Low	Low	Low	Unclear	Low	Low
Houghton[31] 2014	Low	Unclear	Low	Low	Low	Unclear	Low	Low
Campbell[32] 2014	Unclear	Low	Low	Low	Low	Low	Low	Moderate
Krishnamurthy[33] 2013	Low	Unclear	Unclear	High	Low	Unclear	Unclear	Low
Bauer[18] 2013	Low	Low	Low	Low	Low	Low	Low	High
Al-Janabi[35] 2012	Low	Low	Low	Low	Low	Low	Low	High
Al-Janabi[39] 2012	Unclear	Low	Low	Low	Low	Low	Low	Moderate

*Overall quality was assessed to be high when risk bias or concern about applicability was low in all domains. A study was assessed to have moderate quality if there was no more than one unclear domain and no high risk of bias or concern about applicability. A study was of low quality if there was more than one unclear domain or any high risk of bias/concern about applicability in the any domain.

Supplemental Table 11. GRADE Rating for Studies Supporting Recommendation 3

No. of studies	Total number of cases included	Study Design	Aggregate RoB	Inconsistency	Indirectness	Imprecision	GRADE Certainty Rating
Diagnostic Concordance by washout period:							
8	808	Retrospective diagnostic studies	Moderate	Serious	Not serious	Not serious	Low
4	2013	Prospective diagnostic studies	Moderate	Not serious	Not serious	Not serious	Moderate
1	450	Prospective/retrospective diagnostic studies	Moderate	Not serious	Not serious	Serious	Low

1	607	Non-inferiority	Low	Not serious	Not serious	Not serious	Moderate

No. = Number; RoB = Risk of bias

Supplemental Table 12. Evidence to Decision Summary of Recommendation 3

Is the problem a priority?	No	Probably No	Probably Yes	Yes	Varies	Don't Know
	0	20% (2/10)	10% (1/10)	70% (7/10)	0	0
How accurate is the test?	Very Inaccurate	Inaccurate	Accurate	Very Accurate	Varies	Don't Know
	0	10% (1/10)	30% (3/10)	30% (3/10)	20% (2/10)	10% (1/10)
How substantial are the desirable anticipated effects?	Trivial	Small	Moderate	Large	Varies	Don't Know
	0	0	20% (2/10)	40% (4/10)	20% (2/10)	20% (2/10)
How substantial are the undesirable anticipated effects?	Large	Moderate	Small	Trivial	Varies	
	30% (3/10)	60% (6/10)	0	20% (2/10)	10% (1/10)	
What is the overall certainty of the evidence of test accuracy?	Very Low	Low	Moderate	High	Very High	No Included Studies
	10% (1/10)	10% (1/10)	20% (2/10)	40% (4/10)	20% (2/10)	0
What is the overall certainty of the evidence of effects of the management that is guided by the test results?	Very Low	Low	Moderate	High	Very High	No Included Studies
	10% (1/10)	20% (2/10)	30% (3/10)	30% (3/10)	10% (1/10)	0
How certain is the link between test results and management decisions?	Very Low	Low	Moderate	High	Very High	No Included Studies
	10% (1/10)	20% (2/10)	10% (1/10)	40% (4/10)	10% (1/10)	10% (1/10)
What is the overall certainty of the evidence of effects of the test?	Very Low	Low	Moderate	High	Very High	No Included Studies
	10% (1/10)	0	40% (4/10)	40% (4/10)	10% (1/10)	0
What is the overall certainty of the evidence for any critical or important direct benefits, adverse effects or burden of the test?	Very Low	Low	Moderate	High	Very High	No Included Studies
	10% (1/10)	20% (2/10)	30% (3/10)	40% (4/10)	0	0

Is there important uncertainty about or variability in how much people value the main outcomes?	Important uncertainty or variability	Possibly important uncertainty or variability	Probably no important uncertainty or variability	No important uncertainty or variability			
	10% (1/10)	60% (6/10)	30% (3/10)	0			
Does the balance between desirable and undesirable effects favor the intervention or the comparison?	Favors the comparison	Probably favors the comparison	Does not favor either the intervention or comparison	Probably favors the intervention	Favors the intervention	Varies	Don't Know
	20% (2/10)	0	0	30% (3/10)	30% (3/10)	10% (1/10)	10% (1/10)
How large are the resource requirements (costs)?	Large costs	Moderate costs	Negligible costs and savings	Moderate savings	Large savings	Varies	Don't Know
	10% (1/10)	0	90% (9/10)	0	0	0	0
What is the certainty of evidence of resource requirements (costs)?	Very Low	Low	Moderate	High	Very High	No Included Studies	
	0	20% (2/10)	30% (3/10)	10% (1/10)	20% (2/10)	20% (2/10)	
Does the cost-effectiveness of the intervention favor the intervention or the comparison	Favors the comparison	Probably favors the comparison	Does not favor either the intervention or comparison	Probably favors the intervention	Favors the intervention	Varies	No included studies
	20% (2/20)	10% (1/10)	10% (1/10)	20% (2/10)	30% (3/10)	0	10% (1/10)
What would be the impact on health equity?	Reduced	Probably reduced	Probably no impact	Probably increased	Increased	Varies	Don't Know
	0	0	20% (2/10)	10% (1/10)	30% (3/10)	10% (1/10)	30% (3/10)
Is the intervention acceptable to key stakeholders?	No	Probably no	Probably yes	Yes	Varies	Don't Know	
	0	0	20% (2/10)	60% (6/10)	20% (2/10)	0	
Is the intervention feasible to implement?	No	Probably no	Probably yes	Yes	Varies	Don't Know	
	0	0	0	100% (10/10)	0	0	

Supplemental Figure 1:

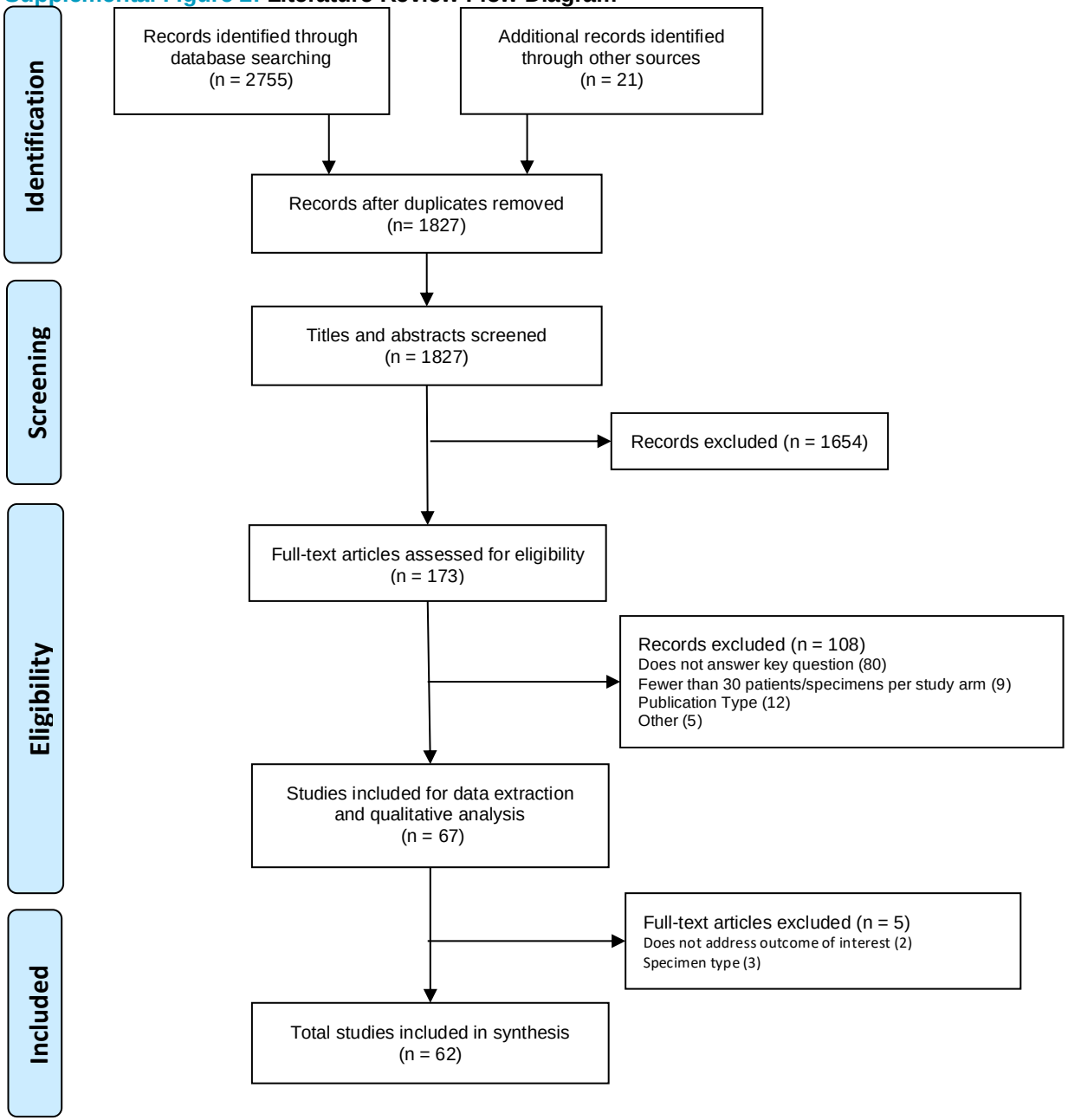
Ovid Search String:

((((((digit* or virtual or internet) adj2 (slide* or imag* or pathology or dermatopathology or hematopathology or haematopathology or cytology or neuropathology or cytopathology or histology or histopathology or reproduction or microscop*)) or (telepathology or teledermatopathology or telemicroscopy or telecytology or teleconsultation*)).ti. or whole slide?.ti,ab,kf. or telepathology/) **AND** ((((glass or light or optical or conventional or standard or traditional or analog) adj2 (slide* or imag* or microscop*)).ti,ab,kf.) **OR** ((valid* or revalid* or re-valid* or verification or concordan* or discordan* or superior* or inferior* or noninferior* or non-inferior* or reproducib* or gold standard or diagnos* or documentation or interobserver or interpathologist or intraobserver or intrapathologist or inter-observer or inter-pathologist or intra-observer or intra-pathologist or within-reader or inter-reader or intra-reader or between-reader or intra-rater or inter-rater or ((between or vari*) adj3 (rater* or reader* or observer*)) or wash-out or washout).ti,ab,kf. or validation studies/ or "reproducibility of results"/ or "feasibility studies"/ or "observer variation"/ or diagnosis/)))) **NOT** (comment/ or editorial/ or case reports/ or (letter/ not exp study characteristics/) or (exp animals/ not humans/))) **Limit to** (english language and yr="2012 -2018")

EMBASE Search String:

((((((digit* OR virtual OR internet) NEAR/2 (slide* OR imag* OR pathology OR dermatopathology OR hematopathology OR haematopathology OR cytology OR neuropathology OR cytopathology OR histology OR histopathology OR reproduction OR microscop*)):ti) OR telepathology:ti OR teledermatopathology:ti OR telemicroscopy:ti OR telecytology:ti OR teleconsultation*:ti OR "whole slide$":ti,ab,kw OR 'telepathology'/de) **AND** ((((glass OR light OR optical OR conventional OR standard OR traditional OR analog) NEAR/2 (slide* OR imag* OR reproduction OR microscop*)):ti,ab,kw) **OR** ((valid* OR 're-valid*' OR revalid* OR verification OR concordan* OR discordan* OR superior* OR inferior* OR noninferior* OR 'non-inferior*' OR reproducib* OR 'gold standard' OR diagnos* OR documentation OR interobserver OR interpathologist OR intraobserver OR intrapathologist OR 'inter-observer' OR 'inter-pathologist' OR 'intra-observer' OR 'intra-pathologist' OR 'within-reader' OR 'inter-reader' OR 'intra-reader' OR 'between-reader' OR 'intra-rater' OR 'inter-rater' OR ((between or vari*) NEAR/3 (rater* or reader* or observer*)) OR 'wash-out' OR washout):ti,ab,kw OR 'validation study'/de OR 'reproducibility'/exp OR 'feasibility study'/de OR 'observer variation'/de OR 'observer bias'/de OR 'diagnosis'/de))) **NOT** ('conference abstract'/it OR 'conference paper'/exp OR 'case report'/exp OR 'case study'/exp OR 'editorial'/exp OR 'note'/exp OR ('letter'/exp NOT 'clinical study'/exp) OR ('animal'/exp NOT 'human'/exp))) [english]/lim AND [2012-2018]/py

Supplemental Figure 2: Literature Review Flow Diagram



References

1. Brouwers MC, Kho ME, Browman GP et al. AGREE II: advancing guideline development, reporting and evaluation in health care. *CMAJ*. 2010;182(18):E839-E842.
2. Shea BJ, Reeves BC, Wells G et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ*. 2017;358:j4008.
3. Whiting PF, Rutjes AWS, Westwood ME et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529-536.
4. Pantanowitz L, Sinard JH, Henricks WH et al. Validating whole slide imaging for diagnostic purposes in pathology: guideline from the College of American Pathologists Pathology and Laboratory Quality Center. *Arch Pathol Lab Med*. 2013;137(12):1710-1722.
5. Cross S, Furness, P., Igali, L., Snead, D., Treanor, D. Best practice recommendations for implementing digital pathology: Royal College of Pathologists; January 2018. Accessed June 21, 2018. https://www.rcpath.org/uploads/assets/f465d1b3-797b-4297-b7fedc00b4d77e51/Best-practice-recommendations-for-implementing-digital-pathology.pdf
6. Lindsköld L, Samuelsson B, Carlberg I et al. Diagnostic agreement of digital whole slide imaging and routine light microscopy. Gothenburg: The Regional Health Technology Assessment Centre (HTA-centrum): HTA-rapport 2012:54. Accessed June 21, 2018. https://www.crd.york.ac.uk/crdweb/ShowRecord.asp?ID=32013000121&ID=32013000121
7. Williams BJ, DaCosta P, Goacher E, Treanor D. A Systematic Analysis of Discordant Diagnoses in Digital Pathology Compared With Light Microscopy. *Arch Pathol Lab Med*. 2017;141(12):1712-1718.
8. Goacher E, Randell R, Williams B, Treanor D. The Diagnostic Concordance of Whole Slide Imaging and Light Microscopy: A Systematic Review. *Arch Pathol Lab Med*. 2017;141(1):151-161.
9. Araujo ALD, Arboleda LPA, Palmier NR et al. The performance of digital microscopy for primary diagnosis in human pathology: a systematic review. *Virchows Arch*. 2019;474(3):269-287.
10. Bauer TW, Slaw RJ, McKenney JK, Patil DT. Validation of whole slide imaging for frozen section diagnosis in surgical pathology. *J Pathol Inform*. 2015;6(1):49.
11. Saco A, Diaz A, Hernandez M et al. Validation of whole-slide imaging in the primary diagnosis of liver biopsies in a University Hospital. *Dig Liver Dis*. 2017;49(11):1240-1246.
12. Mukhopadhyay S, Feldman MD, Abels E et al. Whole Slide Imaging Versus Microscopy for Primary Diagnosis in Surgical Pathology: A Multicenter Blinded Randomized Noninferiority Study of 1992 Cases (Pivotal Study). *Am J Surg Pathol*. 2018;42(1):39-52.
13. Kent MN, Olsen TG, Feeser TA et al. Diagnostic Accuracy of Virtual Pathology vs Traditional Microscopy in a Large Dermatopathology Study. *JAMA Dermatol*. 2017;153(12):1285-1291.
14. Snead DR, Tsang YW, Meskiri A et al. Validation of digital pathology imaging for primary histopathological diagnosis. *Histopathology*. 2016;68(7):1063-1072.
15. Jones NC, Nazarian RM, Duncan LM et al. Interinstitutional whole slide imaging teleconsultation service development: assessment using internal training and clinical consultation cases. *Arch Pathol Lab Med*. 2015;139(5):627-635.
16. Ordi J, Castillo P, Saco A et al. Validation of whole slide imaging in the primary diagnosis of gynaecological pathology in a University Hospital. *J Clin Pathol*. 2015;68(1):33-39.
17. Gage JC, Joste N, Ronnett BM et al. A comparison of cervical histopathology variability using whole slide digitized images versus glass slides: experience with a statewide registry. Hum Pathol. 2013;44(11):2542-2548.

18. Bauer TW, Schoenfield L, Slaw RJ, Yerian L, Sun Z, Henricks WH. Validation of whole slide imaging for primary diagnosis in surgical pathology. *Arch Pathol Lab Med*. 2013;137(4):518-524.
19. Alonso-Coello P, Schunemann HJ, Moberg J et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. *BMJ*. 2016;353:i2016.
20. Araujo ALD, Amaral-Silva GK, Fonseca FP et al. Validation of digital microscopy in the histopathological diagnoses of oral diseases. *Virchows Arch*. 2018;473(3):321-327.
21. Villa I, Mathieu MC, Bosq J et al. Daily Biopsy Diagnosis in Surgical Pathology: Concordance Between Light Microscopy and Whole-Slide Imaging in Real-Life Conditions. *Am J Clin Pathol*. 2018;149(4):344-351.
22. Williams BJ, Hanby A, Millican-Slater R, Nijhawan A, Verghese E, Treanor D. Digital pathology for the primary diagnosis of breast histopathological specimens: an innovative validation and concordance study on digital pathology validation and training. *Histopathology*. 2018;72(4):662-671.
23. Tabata K, Mori I, Sasaki T et al. Whole-slide imaging at primary pathological diagnosis: Validation of whole-slide imaging-based primary pathological diagnosis at twelve Japanese academic institutes. *Pathol Int*. 2017;67(11):547-554.
24. Al-Janabi S, Huisman A, Jonges GN, Ten Kate FJ, Goldschmeding R, van Diest PJ. Whole slide images for primary diagnostics of urinary system pathology: a feasibility study. *J Renal Inj Prev*. 2014;3(4):91-96.
25. Gui D, Cortina G, Naini B et al. Diagnosis of dysplasia in upper gastro-intestinal tract biopsies through digital microscopy. *J Pathol Inform*. 2012;3(1):27.
26. Fertig RM, Gaudi S, Cervantes J et al. Feasibility study in teledermatopathology: An examination of the histopathologic features of mycosis fungoides and spongiotic dermatitis. *J Cutan Pathol*. 2017;44(11):919-924.
27. Gomez-Gelvez JC, Kryvenko ON, Chabot-Richards DS, Foucar K, Inamdar KV, Karner KH. Comparative Analysis Reveals Potential Utility of Digital Microscopy in the Evaluation of Peripheral Blood Smears With Some Barriers to Implementation. *Am J Clin Pathol*. 2015;144(1):68-77.
28. Loughrey MB, Kelly PJ, Houghton OP et al. Digital slide viewing for primary reporting in gastrointestinal pathology: a validation study. *Virchows Arch*. 2015;467(2):137-144.
29. Thrall MJ, Wimmer JL, Schwartz MR. Validation of multiple whole slide imaging scanners based on the guideline from the College of American Pathologists Pathology and Laboratory Quality Center. *Arch Pathol Lab Med*. 2015;139(5):656-664.
30. Arnold MA, Chenever E, Baker PB et al. The College of American Pathologists guidelines for whole slide imaging validation are feasible for pediatric pathology: a pediatric pathology practice experience. *Pediatr Dev Pathol*. 2015;18(2):109-116.
31. Houghton JP, Ervine AJ, Kenny SL et al. Concordance between digital pathology and light microscopy in general surgical pathology: a pilot study of 100 cases. *J Clin Pathol*. 2014;67(12):1052-1055.
32. Campbell WS, Hinrichs SH, Lele SM et al. Whole slide imaging diagnostic concordance with light microscopy for breast needle biopsies. *Hum Pathol*. 2014;45(8):1713-1721.
33. Krishnamurthy S, Mathews K, McClure S et al. Multi-institutional comparison of whole slide digital imaging and optical microscopy for interpretation of hematoxylin-eosin-stained breast tissue sections. *Arch Pathol Lab Med*. 2013;137(12):1733-1739.
34. Al-Janabi S, Huisman A, Nap M, Clarijs R, van Diest PJ. Whole slide images as a platform for initial diagnostics in histopathology in a medium-sized routine laboratory. *J Clin Pathol*. 2012;65(12):1107-1111.

35. Al-Janabi S, Huisman A, Willems SM, Van Diest PJ. Digital slide images for primary diagnostics in breast pathology: a feasibility study. *Hum Pathol*. 2012;43(12):2318-2325.
36. Campbell WS, Lele SM, West WW, Lazenby AJ, Smith LM, Hinrichs SH. Concordance between whole-slide imaging and light microscopy for routine surgical pathology. *Hum Pathol*. 2012;43(10):1739-1744.
37. Fonyad L, Krenacs T, Nagy P et al. Validation of diagnostic accuracy using digital slides in routine histopathology. *Diagn Pathol*. 2012;7:35.
38. Al-Janabi S, Huisman A, Vink A et al. Whole slide images for primary diagnostics in dermatopathology: a feasibility study. *J Clin Pathol*. 2012;65(2):152-158.
39. Al-Janabi S, Huisman A, Vink A et al. Whole slide images for primary diagnostics of gastrointestinal tract pathology: a feasibility study. *Hum Pathol*. 2012;43(5):702-707.
40. Hanna MG, Reuter VE, Hameed MR et al. Whole slide imaging equivalency and efficiency study: experience at a large academic center. *Mod Pathol*. 2019;32(7):916-928.
41. Larghi A, Fornelli A, Lega S et al. Concordance, intra- and inter-observer agreements between light microscopy and whole slide imaging for samples acquired by EUS in pancreatic solid lesions. *Digest Liver Dis*. 2019;51(11):1574-1579.
42. Cima L, Brunelli M, Parwani A et al. Validation of Remote Digital Frozen Sections for Cancer and Transplant Intraoperative Services. *J Pathol Inform*. 2018;9(1):34.
43. Sturm B, Creytens D, Cook MG et al. Validation of Whole-slide Digitally Imaged Melanocytic Lesions: Does Z-Stack Scanning Improve Diagnostic Accuracy? *J Pathol Inform*. 2019;10(1):6.
44. Rakha EA, Aleskandarani M, Toss MS et al. Breast cancer histologic grading using digital microscopy: concordance and outcome association. *J Clin Pathol*. 2018;71(8):680-686.
45. Wilbur DC, Brachtel EF, Gilbertson JR, Jones NC, Vallone JG, Krishnamurthy S. Whole slide imaging for human epidermal growth factor receptor 2 immunohistochemistry interpretation: Accuracy, Precision, and reproducibility studies for digital manual and paired glass slide manual interpretation. *J Pathol Inform*. 2015;6(1):22.
46. Malarkey DE, Willson GA, Willson CJ et al. Utilizing Whole Slide Images for Pathology Peer Review and Working Groups. *Toxicol Pathol*. 2015;43(8):1149-1157.
47. Guyatt G, Oxman AD, Akl EA et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol*. 2011;64(4):383-394.