Effective Health Care Program

Diagnosis of Celiac Disease
Executive Summary

Archived: This report is greater than 3 years old. Findings may be used for research purposes, but should not be considered current.

Background

Condition

Celiac disease (CD) is an immune-mediated disorder triggered in genetically susceptible individuals by ingestion of foods containing gluten, a family of proteins found in wheat, rye, barley, and related grains.[1] The prevalence of CD in the United States has been estimated at approximately 1 percent[2] but appears to be increasing for reasons that are not clear.[3] Risk factors for CD include family history, trisomy 21, Turner syndrome, and Williams syndrome, as well as several autoimmune diseases.

Clinical signs of CD include weight loss, iron deficiency anemia, aphthous ulcers, osteomalacia, dermatitis herpetiformis (a rash due to gluten sensitivity), and gastrointestinal (GI) symptoms, including diarrhea and abdominal bloating. The diagnosis of CD can be challenging because the clinical spectrum of the disease varies, and some individuals present with mild symptoms.[4]

CD causes enteropathy of the small intestine, resulting in poor absorption of nutrients. Malabsorption may result in several of the clinical signs, including iron deficiency anemia, osteomalacia, and weight loss. Young children, in particular, are susceptible to failure to thrive, stunted growth, and delayed puberty.[5] In women,

folate deficiency secondary to CD may lead to poor birth outcomes, including developmental disorders. In the long term, untreated CD increases the risk for non-Hodgkin's lymphoma, certain GI cancers, and all-cause mortality.[4]

> ## Effective Health Care Program
>
> The Effective Health Care Program was initiated in 2005 to provide valid evidence about the comparative effectiveness of different medical interventions. The object is to help consumers, health care providers, and others in making informed choices among treatment alternatives. Through its Comparative Effectiveness Reviews, the program supports systematic appraisals of existing scientific evidence regarding treatments for high-priority health conditions. It also promotes and generates new scientific evidence by identifying gaps in existing scientific evidence and supporting new research. The program puts special emphasis on translating findings into a variety of useful formats for different stakeholders, including consumers.
>
> The full report and this summary are available at **www.effectivehealthcare. ahrq.gov/reports/final.cfm**.

AHRQ
Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • www.ahrq.gov

Effective
Health Care

The only effective treatment for CD is avoidance of gluten in the diet. Timely diagnosis may be the most important component in the management of CD.

Diagnostic Strategies

A number of diagnostic methods have been developed; the validity and acceptability of some of these methods, particularly newer tests, which include combination tests and algorithms, remain controversial. These methods include various serology tests—anti-gliadin antibodies (AGA), anti-tissue transglutaminase (tTG), endomysial antibodies (EmA), and deamidated gliadin peptide (DGP) antibodies—as well as human leukocyte antigen (HLA) typing, video capsule endoscopy (VCE), and endoscopic duodenal biopsy (often considered the gold standard). Providers may use these tests sequentially in order to increase specificity and prevent false positives, or to increase sensitivity and prevent false negatives. All methods other than HLA typing require the patient to maintain a gluten-containing diet during the diagnostic process.

AGA, immunoglobulin A (IgA) and immunoglobulin G (IgG). Gliadin is one of the two groups of proteins that constitute gluten. AGA determination was used as a diagnostic tool in the 1990s, as it has high sensitivity for CD,[6] although the test has low specificity. As AGA tests are no longer recommended,[7,8] they are not addressed in this systematic review.

TTG, IgA. Tissue transglutaminase is an enzyme that causes the crosslinking of certain proteins. Anti-tTG IgA is the single test preferred by the American College of Gastroenterology (ACG) for the detection of CD in those 2 years of age and over[5] and is included in the algorithms of all recent guidelines. However, as IgA deficiency is more prevalent in CD patients than in the general population, other tests may be ordered as an alternative in those who are IgA deficient.

EmA, IgA. When the intestinal lining is damaged, endomysial antibodies develop. Most patients with active CD and many with dermatitis herpetiformis have the IgA class of anti-EmA antibodies. This test is included in some algorithms of recent guidelines for diagnosis, although it is not as widely used in the United States as in other countries. This test is less useful in IgA-deficient individuals.

DGP antibodies. This is a newer test that may give a positive result in some individuals with CD who are anti-tTG negative, including children under age 2.

HLA typing. Susceptibility to CD is linked to certain HLA class II alleles, especially in the HLA-DQ region. Approximately 95 percent of patients with CD have the HLA-DQ2 heterodimer, while the remaining 5 percent have the HLA-DQ8 heterodimer.[9] Lack of these heterodimers all but rules out CD and genetic susceptibility for the disorder. These genetic tests are part of the diagnostic algorithms recommended by the European Society for Pediatric Gastroenterology, Hepatology, and Nutrition (ESPGHAN) and the ACG.[10]

VCE. For this test, the patient ingests a capsule containing a tiny camera, providing high-quality visual evidence of the villous atrophy associated with CD. While not a traditional means of detecting CD, VCE is used in adults who seek to avoid biopsy. During the topic refinement phase of this project, Key Informants suggested that assessment of the evidence for this method be included in this report.

Endoscopic duodenal biopsy. Villous atrophy present on a duodenal biopsy and clinical remission when a gluten-free diet is followed represent the internationally accepted gold standard for CD diagnosis. However, this procedure may be difficult to execute effectively, and some patients and parents of small children are concerned about the possibility of adverse events, including perforations, bleeding, pain, and discomfort.

Scope and Key Questions

Scope of the Review

The purpose of this review is to assess the evidence on the comparative accuracy and possible harms of methods used for the diagnosis of CD, including serological tests, HLA typing, VCE, and endoscopic duodenal biopsy. The review compares the effectiveness of these diagnostic tests singly and in combination in various populations of special interest to the CD community. A protocol for the review was posted online by the Agency for Healthcare Research and Quality (AHRQ) Effective Health Care Program.

Key Questions

Figure A shows an analytic framework to illustrate the populations, interventions, outcomes, and possible adverse effects that guided the literature search and synthesis for this project.

Figure A. Analytic framework, diagnosis of celiac disease
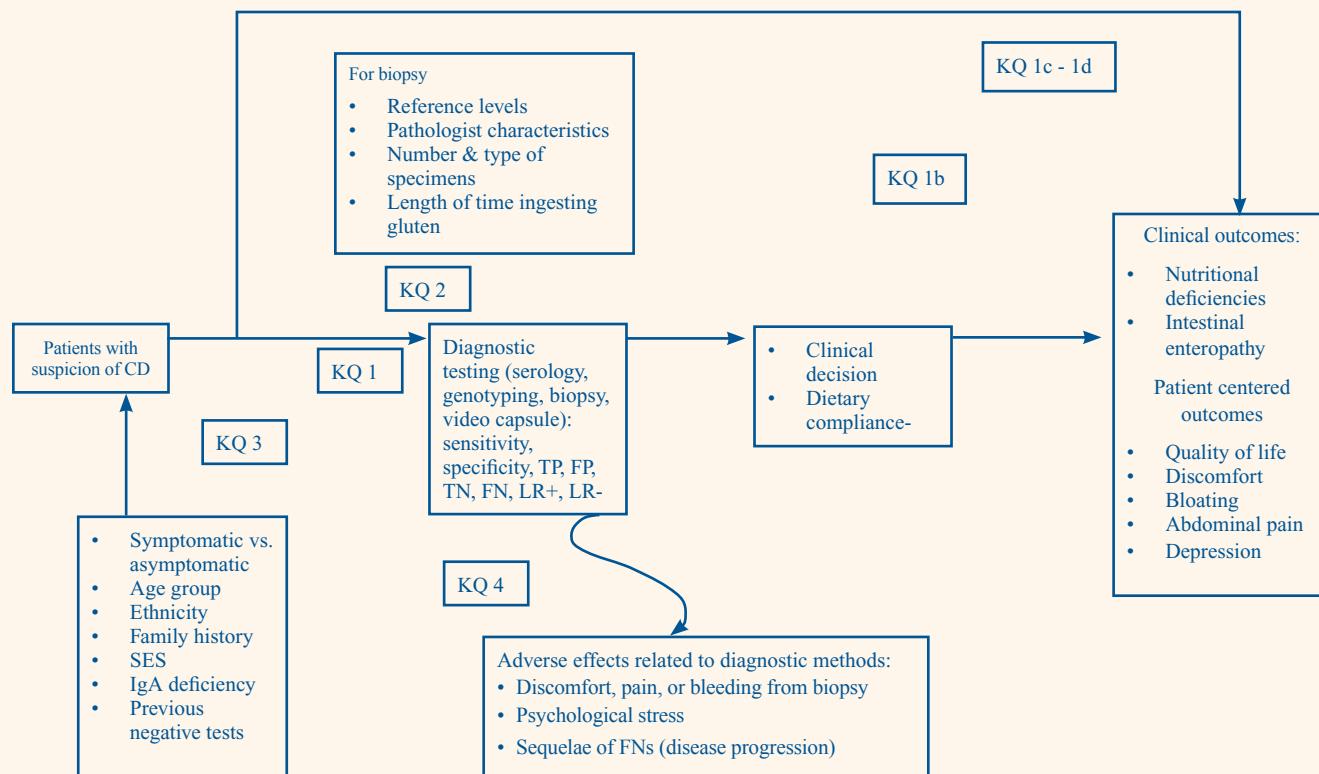
CD = celiac disease; FN = false negative; FP = false positive; IgA = immunoglobulin A; KQ = Key Question; LR+ = positive likelihood ratio; LR- = negative likelihood ratio; SES = socioeconomic status; TN = true negative; TP = true positive.

The Key Questions addressed in this review are as follows:

Key Question 1. What is the comparative effectiveness of the different diagnostic methods (various serological tests, human leukocyte antigen [HLA] typing, video capsule endoscopy, used individually and in combination) compared with endoscopy with biopsy as the reference standard, to diagnose celiac disease (CD) in terms of—

- Accuracy: sensitivity, specificity, positive likelihood ratio, negative likelihood ratio, and summary receiver-operating characteristics?

- Intermediate outcomes, such as clinical decisionmaking and dietary compliance?

- Clinical outcomes and complications related to CD?

- Patient-centered outcomes, such as quality of life (QOL) and symptoms?

Key Question 2. Do accuracy/reliability of endoscopy with duodenal biopsy vary by—

a. Pathologist characteristics (i.e., level of experience or specific training)?

b. Method (i.e., type or number of specimens)?

c. Length of time ingesting gluten before diagnostic testing?

Key Question 3. How do accuracy and outcomes differ among specific populations, such as—

a. Symptomatic patients versus nonsymptomatic individuals at risk?

b. Adults (age 18 and over) versus children and adolescents?

c. Children under age 24 months versus older children?

d. Demographics, including race, genetics, geography, and socioeconomic status?

e. Patients with IgA deficiency?

f. Patients previously testing negative for CD?

Key Question 4. What are the direct adverse effects (e.g., bleeding from biopsy) or harms (related to false positives, false negatives, indeterminate results) associated with testing for CD?

Methods

Topic Refinement and Review Protocol

Key Informants from professional associations, research centers, payers, and patient organizations were engaged to assist in refining the Key Questions (KQs) and issues to cover in this systematic review. The authors then refined and finalized the KQs after review of public comments collected on the AHRQ Effective Health Care Web site in February 2014. The final protocol was posted on the Web site in June 2014 after input from a Technical Expert Panel representing various areas of expertise in CD.

Literature Search Strategy

An experienced reference librarian designed the search strategies in collaboration with an expert on CD and project staff experienced in systematic review methods. The search strategy included search terms for CD, combined with general terms for diagnosis or terms representing each diagnostic method, plus terms representing all outcomes listed in the PICOTs (populations, interventions, comparators, outcomes, timing, and setting). The full search strategy is presented in Appendix A of the full report.

For KQ 1a, we searched for publications starting from January 1990 but did not abstract studies that were already included in recent high-quality systematic reviews. For KQ 2, on duodenal biopsy, and KQ 3, on specific populations, our search also started at January 1990. For KQ 4, on direct and indirect harms of the diagnostic procedures, our search started at January 2003, as this KQ was covered by an AHRQ-funded systematic review published in 2004.[11]

PubMed®, Embase®, the Cochrane Library, and Web of Science were searched. The AHRQ-funded Scientific Resource Center requested unpublished data from manufacturers of all serological tests. Key Informants, project clinicians, and members of the Technical Expert Panel also suggested studies. Reference lists of included articles were reviewed for identification of additional relevant studies.

Inclusion and Exclusion Criteria

Eligible studies of diagnostic accuracy included controlled trials, prospective and retrospective cohorts, case-control studies, and case series. Studies were included if they met the following criteria:

- Diagnostic method must be currently used in clinical practice, as listed in the PICOTS. Diagnostic methods no longer recommended or still in development were excluded.

- Study was about diagnosis of CD rather than management of existing CD.

- All participants underwent both the "index test" and the reference standard (biopsy).

- The study reported sensitivity, specificity, or data that allowed calculation.

- Study was published in English.

- Study enrolled a consecutive or random sample.

- For representativeness and generalizability, the sample size was 300 or more unless one of the following populations of interest was the focus:

    – Low socioeconomic status
    – Previously negative for CD via serology or biopsy
    – IgA deficient
    – Type 1 diabetes
    – Turner yndrome
    – Trisomy 21/Down syndrome
    – Iron deficiency anemia
    – Family history

- Accuracy results were stratified by race/ethnicity.

The following were excluded from this systematic review:

- Animal studies

- Individual case reports

- Studies not published in English

- Documents with no original data (commentary, editorial)

- Studies that reported only prevalence

The PICOTS considered in this review are as follows.

Population(s):

For KQs 1, 2, and 4—

    All populations tested for CD

For KQ 3—

- Patients with signs and symptoms of CD; for example—

    – Diarrhea
    – Constipation
    – Dermatitis
    – Malabsorption (anemia, folate deficiency)

- Asymptomatic individuals at risk of CD because of—

    – Family history

- – Type 1 diabetes
- – Autoimmune disease
- – Turner syndrome
- – Trisomy 21
- Children under age 24 months versus older children and adolescents
- Adults (aged 18 and over)
- Ethnic and geographic populations
- Patients with low socioeconomic status
- Patients with IgA deficiency
- Patients previously testing negative for CD

Interventions:

For KQs 1, 3, 4—

- Test for EmA IgA
- Test for tTG IgA
- Test for DGP IgA antibodies
- EmA IgG, tTG IgG, and DGP IgG tests for IgA-deficient individuals
- HLA typing
- VCE
- Combinations of the above

For KQ 2—

- Endoscopy with biopsy

Comparators:

For KQs 1 and 3—

- Endoscopy with duodenal biopsy

For KQ 2—

- Repeat biopsy

Outcomes:

For KQ 1a, KQ 2, and KQs 3a–f, for accuracy—

- Sensitivity
- Specificity
- Positive predictive value, negative predictive value, false positive, false negative
- Positive and negative likelihood ratios

For KQ 1b, for clinical decisionmaking—

- Additional testing for CD
- Nutritionist advice on gluten-free diet

- Followup and monitoring by physician

For KQ 1c, for clinical outcomes and complications—

- Nutritional deficits
- Persistence of villous atrophy on biopsy
- Lymphomas

For KQ 1d, for patient-centered outcomes—

- QOL
- Discomfort
- Bloating
- Abdominal pain
- Depression

For KQ 4, for harms—

- Immediate adverse events from biopsy
- Psychological stress related to false positive results
- Sequelae of false negatives or indeterminate results

Timing:

For KQ 2—

- Length of time ingesting gluten before biopsy

Setting:

For all KQs—

- Outpatient: academic
- Outpatient: community

Study Selection

Each title and abstract identified by the searches was screened independently by two researchers, and the combination of their selections was retrieved for full-text review. Two researchers independently screened each full-text article for inclusion in the project, with a senior researcher resolving discrepancies. A list of excluded studies with reasons for exclusion is presented as Appendix B of the full report.

Data Extraction

The DistillerSR software package was used to manage the search output, screening, and data abstraction. Data collection forms were designed by the project team in DistillerSR, piloted by the reviewers, and further modified; then the final forms were piloted with a random selection of included studies to ensure agreement of interpretation. Articles accepted for inclusion were abstracted in DistillerSR; a statistical analyst abstracted accuracy data

in Excel. The project leader reviewed data for all included studies for accuracy and made revisions accordingly. Forms are displayed in Appendix D of the full report.

Quality (Risk-of-Bias) Assessment of Individual Studies

The QUADAS-2[12] instrument (revised Quality Assessment of Diagnostic Accuracy Studies instrument) was used to assess the risk of bias of accuracy studies; the McHarm instrument[13] was used to assess the quality of studies on adverse events; and the AMSTAR[14] instrument (a measurement tool for the assessment of multiple systematic reviews) was used to assess the quality of prior systematic reviews. These instruments are described in detail in the Methods chapter of the full report. Each study was scored individually by two Evidence-based Practice Center researchers, who met to reconcile any differences; the project leader resolved discrepancies.

Diagnostic Accuracy—Statistical Analyses

Studies that reported sensitivity, specificity, or ROCs, or provided the data to calculate these values, were abstracted for potential inclusion in a synthesis. Sensitivity is also known as the "true positive rate," the ability of a test to correctly classify an individual as having a condition—in this case, having CD as confirmed by biopsy. Sensitivity ranges from 0 to 100, with values closer to 100 indicating a greater probability of a test being positive when the disease is present.[15] Specificity, also known as the "true negative rate," is the ability of a test to correctly classify an individual as not having a condition—in this case, when the individual is determined by biopsy not to have CD. Specificity ranges from 0 to 100, with values closer to 100 indicating a greater probability of a test being negative when the disease is not present.[15] A perfect diagnostic test would have both sensitivity and specificity of 100 percent. In general, sensitivity and specificity are considered good if at least 70.0 percent, very good from 80.0 percent to 89.9 percent, and excellent if 90.0 percent or greater.[15]

Some studies of the accuracy of diagnostic tests report likelihood ratios (LRs), the probability of a positive finding in patients with a disease divided by the probability of the same finding in patients without the disease. Likelihood ratios can range from 0 to infinity. An LR of 1 indicates no change in the likelihood of disease.[16] As the LR increases from 1, the likelihood of disease increases. LR+ (positive likelihood ratio) is a measure of how the probability of the disease increases in the presence of a positive test finding, while LR- (negative likelihood ratio) is a measure of how the probability of the disease decreases if the test is negative. An LR+ of greater than 10 is considered good, as is an LR- of less than 0.1.[17]

Finally, positive predictive value (PPV) is the probability that an individual who tests positive actually has the disease. Similarly, negative predictive value (NPV) is the probability of not having a disease when an individual tests negative. Unlike sensitivity and specificity, predictive values (PPV, NPV) are largely dependent on the prevalence of a disease in a study population. With increased prevalence in a population, PPV increases while NPV decreases.

If three or more studies of the same diagnostic method and comparator reported the number of true positives, false positives, true negatives, and false negatives by arm, their results were pooled in order to estimate overall sensitivity, specificity, LRs, and predictive values. Additional analyses were conducted by stratifying by test type, threshold (titer), and population characteristics of interest. When pooling was not possible, study results were described narratively according to comparisons of interest and presented in tables and figures in the full report.

Strength of the Body of Evidence

The overall strength of evidence for accuracy outcomes was assessed using guidance developed by experts in systematic reviews for the AHRQ Effective Health Care Program.[18] This method classifies the strength of evidence based on the following domains: study limitations (risk of bias), consistency, directness, and precision. The domains are described in the Methods chapter of the full report. In this Executive Summary, we report the strength of evidence for each KQ and subquestion. Appendix F in the full report displays the results for each domain for the evidence on accuracy of serological tests in each population.

Applicability

Applicability assessment was based on the similarity of the populations in terms of characteristics listed in the PICOTs.

Peer Review and Public Commentary

A draft version of this report was reviewed by several CD experts; names and affiliations are listed in the front matter of the report. All Peer Reviewers completed conflict-of-interest disclosure forms; none reported ties to any test manufacturers. A draft version of this report was posted on the AHRQ Effective Health Care Web site in February 2015 for public comment. The authors reviewed the comments and incorporated the feedback into the final version.
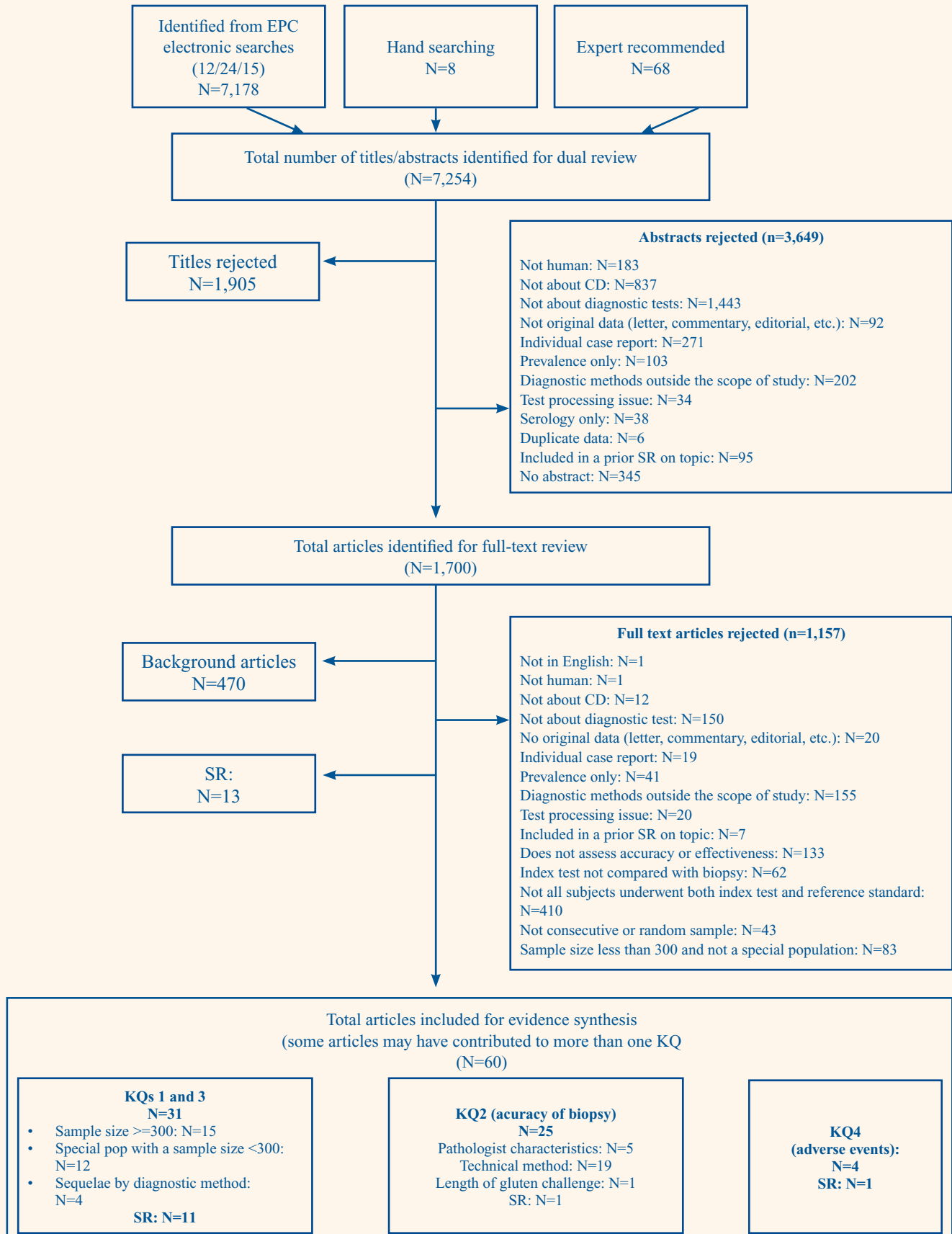
Results

Overview

Figure B is a literature flow diagram that displays the number of studies identified through electronic searches and contact with experts. It shows the number of studies accepted at each stage of screening and reasons for excluding the others. Table A presents the key findings from prior systematic reviews, results reported in newly identified studies, summary conclusions by KQ and subquestion, and strength of evidence. The applicability and limitations of the evidence are discussed, followed by overall conclusions.

Results of Literature Searches

As displayed in Figure B, of a total of 7,254 titles from the literature search, 60 individual studies and 13 prior systematic reviews (SRs) were included for evidence synthesis. References for the excluded articles, along with reasons for exclusion, can be found in Appendix B of the full report. Thirty-one articles reporting original data and 11 SRs addressed KQ 1 and KQ 3, 25 articles and 1 SR addressed KQ 2, and 4 articles and 1 SR addressed KQ 4.

Figure B. Literature flow

Identified from EPC electronic searches (12/24/15) N=7,178

Hand searching N=8

Expert recommended N=68

Total number of titles/abstracts identified for dual review (N=7,254)

Titles rejected N=1,905

Abstracts rejected (n=3,649)
Not human: N=183
Not about CD: N=837
Not about diagnostic tests: N=1,443
Not original data (letter, commentary, editorial, etc.): N=92
Individual case report: N=271
Prevalence only: N=103
Diagnostic methods outside the scope of study: N=202
Test processing issue: N=34
Serology only: N=38
Duplicate data: N=6
Included in a prior SR on topic: N=95
No abstract: N=345

Total articles identified for full-text review (N=1,700)

Background articles N=470

SR: N=13

Full text articles rejected (n=1,157)
Not in English: N=1
Not human: N=1
Not about CD: N=12
Not about diagnostic test: N=150
No original data (letter, commentary, editorial, etc.): N=20
Individual case report: N=19
Prevalence only: N=41
Diagnostic methods outside the scope of study: N=155
Test processing issue: N=20
Included in a prior SR on topic: N=7
Does not assess accuracy or effectiveness: N=133
Index test not compared with biopsy: N=62
Not all subjects underwent both index test and reference standard: N=410
Not consecutive or random sample: N=43
Sample size less than 300 and not a special population: N=83

Total articles included for evidence synthesis (some articles may have contributed to more than one KQ (N=60)

KQs 1 and 3
N=31
- Sample size >=300: N=15
- Special pop with a sample size <300: N=12
- Sequelae by diagnostic method: N=4
SR: N=11

KQ2 (acuracy of biopsy)
N=25
Pathologist characteristics: N=5
Technical method: N=19
Length of gluten challenge: N=1
SR: N=1

KQ4 (adverse events): N=4
SR: N=1

CD = celiac disease; EPC = Evidence-based Practice Center; KQ = Key Question; SR = systematic review.

Key Findings and Strength of Evidence

The key findings and strength of evidence are summarized in Table A. Additional details on strength-of-evidence ratings are provided as Appendix F of the full report.

Table A. Summary of findings and strength of evidence

Topic	EPC Conclusions and Strength of Evidence	Prior Systematic Reviews	Additional Findings From EPC
Key Question 1: Accuracy of IgA tTG	High: IgA tTG tests have excellent sensitivity and specificity.	A 2010 meta-analysis that pooled 12 studies found a sensitivity of 93.0% (95% CI, 91.2% to 94.5%) and specificity of 96.5% (95% CI, 95.2% to 97.5%). A 2012 meta-analysis restricted to 5 studies of point-of-care tests in children reported sensitivity and specificity of 96.4% (95% CI, 94.3% to 97.9%) and 97.7% (95% CI, 95.8% to 99.0%), respectively.	Sixteen studies were published after the SRs were pooled. Excluding data for threshold levels higher than used in clinical practice, sensitivity was 92.5% (95% CI, 89.7% to 94.6%) and specificity was 97.9% (95% CI, 96.5% to 98.7%). LR+ was 40.19 and LR- was 0.08. PPV was 89.4%, while NPV was 99.0%.
Key Question 1: Accuracy of IgA EmA	High: IgA EmA tests have lower sensitivity but equal specificity to IgA tTG tests.	A 2009 SR including 23 studies found sensitivity ranging from 68% to 100%, while specificity ranged from 77% to 100%; pooling was not performed. A 2012 SR included 11 studies in children; sensitivity ranged from 82.6% to 100% and pooled specificity was 98.2% (95% CI, 96.7% to 99.1%).	Seven studies were published after the SRs were pooled. Sensitivity was 79.0% (95% CI, 71.0% to 86.0%) and specificity was 99.0% (95% CI, 98.4% to 99.4%) after excluding data points where Marsh Grade I and II villous atrophy was classified as CD (not standard practice). LR+ was 65.98 and LR- was 0.21. PPV was 78.9%; NPV was 99.1%.
Key Question 1: Accuracy of IgA DGP	High: IgA DGP tests are not as accurate as IgA tTG tests.	A 2010 SR pooled 11 studies on accuracy in all ages; sensitivity was 87.8% (95% CI, 85.6% to 89.9%), while specificity was 94.1% (95% CI, 95.2% to 97.5%). LR+ was 13.33, while LR- was 0.12. A 2012 SR reviewed 3 of those studies that included only children: sensitivities ranged from 80.7% to 95.1% (not pooled) and pooled specificity was estimated at 90.7% (95% CI, 87.8% to 93.1%).	One new study reported sensitivity of 97.0% and specificity of 90.7% in symptomatic adults and children at 1 clinic, while another reported both sensitivity and specificity of 96% in a similar population.
Key Question 1: Accuracy of IgG DGP	Moderate: IgG DGP tests are not as sensitive as IgA tTG tests in non–IgA-deficient patients.	A 2013 SR of 7 studies of non–IgA-deficient adults reported sensitivity of 75.4% to 96.7% and specificity of 98.5% to 100%. A 2012 SR of 3 studies in non–IgA-deficient children reported sensitivities of 80.1% to 98.6% and specificities of 86.0% to 96.9%. Authors did not pool data.	One study reported sensitivity of 95.0% and specificity of 99.0% in 200 non–IgA-deficient subjects of all ages.
Key Question 1: Accuracy of HLA-DQ2 or DQ8	High: HLA tests can be used to rule out CD with close to 100% sensitivity.	No SRs of the accuracy of testing for HLA-DQ2 or DQ8 were identified. Based on studies from which sensitivity (but not specificity) could be calculated, the American College of Gastroenterology estimated the NPV of the HLA-DQ2/DQ8 combination test at over 99%.	Two studies were identified on the accuracy of HLA testing. A large 2013 prospective cohort found that HLA testing had a sensitivity of 100% and specificity of 18.2%. A 1999 cohort also reported sensitivity of 100%, while specificity was 33.3%.

Table A. Summary of findings and strength of evidence (continued)

Topic	EPC Conclusions and Strength of Evidence	Prior Systematic Reviews	Additional Findings From EPC
Key Question 1: Accuracy of algorithms	Insufficient: Strength of evidence is insufficient to determine comparative accuracy of different algorithms in specific populations.	No SRs of the accuracy of algorithms were identified.	Nine studies of algorithms were identified; all used tTG tests. Adding an EmA test to a tTG test resulted in increased specificity, with either no change or a slight decrease in sensitivity. Adding a DGP test to a tTG test resulted in increased sensitivity but decreased specificity. However, the increase in accuracy compared with individual tests was rarely clinically significant. The sensitivity and specificity results varied widely, populations were diverse, and the evidence base had high heterogeneity.
Key Question 1: Accuracy of VCE	Moderate: VCE has very good sensitivity and excellent specificity.	A previous SR of moderate quality on the accuracy of VCE pooled 6 studies, and estimated sensitivity at 89.0% (95% CI, 82.0% to 94.0%) and specificity at 95.0% (95% CI, 89.0% to 99.0%). LR+ was 12.90 and LR- was 0.16.	No additional studies met our inclusion criteria.
Key Question 1: Intermediate outcomes	Insufficient: Strength of evidence is insufficient regarding how method of diagnosis affects adherence.	A previous SR of low quality (3 studies) reported no statistical difference in adherence levels between patients diagnosed via screening and those diagnosed because they were symptomatic. Association between diagnostic test type and adherence was not addressed.	In 1 study on blood donors in Israel who tested positive for IgA tTG (or IgG tTG if IgA deficient), only 4 of 10 patients with asymptomatic biopsy-proven CD adhered to a gluten-free diet; the other 6 patients did not believe they had CD, and 4 of those were told by physicians that asymptomatic patients did not need to modify their diets.
Key Question 1: Clinical outcomes and complications	Insufficient: Strength of evidence is insufficient regarding how method of diagnosis affects clinical outcomes and complications.	No prior SRs on this topic were identified.	No studies on this topic were identified.
Key Question 1: Patient- centered outcomes such as quality of life	Insufficient: Strength of evidence is insufficient regarding how method of diagnosis affects patient-centered outcomes such as quality of life.	No prior SRs on this topic were identified.	No studies on this topic were identified.

Table A. Summary of findings and strength of evidence (continued)

Topic	EPC Conclusions and Strength of Evidence	Prior Systematic Reviews	Additional Findings From EPC
Key Question 2: Biopsy and provider characteristics	Moderate: Physician adherence to biopsy protocol decreases with volume performed per endoscopy suite and increases with number of gastroenterologists per endoscopy suite.	No SRs on this topic were identified.	One very large high-quality national retrospective study found reduced physician adherence to the American Gastroenterological Association's duodenal biopsy protocol (4+ specimens) with higher procedure volume per endoscopy clinic. The OR for each 100 additional procedures was 0.92 (95% CI, 0.88 to 0.97). Adherence increase for each additional gastroenterologist per endoscopy suite was OR 1.08 (95% CI, 1.04 to 1.13).
Key Question 2: Biopsy and pathologist characteristics	Moderate: CD-related histological findings are underdiagnosed in community settings when compared with academic settings.	No SRs on this topic were identified.	Three retrospective studies reported low interobserver agreement between pathologists in community vs. academic settings, with significantly lower accuracy in community settings. Kappa statistics range from 0.16 to 0.53.
Key Question 2: Biopsy specimens— number and location	High: Increasing the number and location of biopsy specimens increases diagnostic accuracy.	No SRs addressed how the number and location of biopsy specimens influence diagnostic findings of biopsy.	Nineteen studies reported that increasing the number and location of biopsy specimens increased the likelihood of diagnosis and diagnostic yield by 25% to 50% in both pediatric and adult populations.
Key Question 2: Biopsy and length of time ingesting gluten	Moderate: A minimum 2-week gluten intake is necessary to induce intestinal changes necessary for diagnosing adults via duodenal biopsy. Low: A 2–3 month diet containing gluten may be necessary to diagnose CD in children via biopsy; strength is lower due to fewer available studies and inconsistent findings.	A previous SR of high quality on clinical response to gluten challenge indicates that 2 weeks of a moderate to high dose (e.g., 15g daily) is sufficient to cause enough intestinal changes to diagnose adults via duodenal biopsy. This same SR reports that for children, 2 to 3 months may be needed.	One small study reported that 3 grams of gluten per day for 2 weeks induces intestinal atrophy sufficient to diagnose CD in 89.5% of adults.

Table A. Summary of findings and strength of evidence (continued)

Topic	EPC Conclusions and Strength of Evidence	Prior Systematic Reviews	Additional Findings From EPC
Key Question 3: Symptomatic patients vs. nonsymptomatic individuals at risk	High: EmA and tTG tests have excellent sensitivity and specificity in patients with GI symptoms. Insufficient: How accuracy of serological tests differs between patients with risk factors such as iron deficiency or type 1 diabetes and the general symptomatic population could not be determined.	A 2010 SR including only studies of patients with GI symptoms reported pooled sensitivity of 90% (95% CI, 80.0% to 95.0%) and specificity of 99% (95% CI, 98.0% to 100.0%) for IgA EmA tests (8 studies), and pooled sensitivity of 89% (95% CI, 82.0% to 94.0%) and specificity of 98% (95% CI, 95.0% to 99.0%) for IgA tTG tests. No SRs were identified that compared test accuracy in patients with specific symptoms and asymptomatic individuals at risk.	One high-quality study compared the accuracy of the ESPGHAN algorithm (combining tTG IgA and EmA IgA) among subjects with family history, type 1 diabetes, and CD symptoms. Specificity was much higher in those with symptoms. Two small studies provided data that allowed calculation of accuracy in patients with iron deficiency, and 2 provided accuracy data for patients with type 1 diabetes. However, the studies were conducted in the Middle East and Eastern Europe; applicability to the United States is uncertain.
Key Question 3: Children vs. adults	Low: tTG and DGP tests are less sensitive in adults than children. DGP is more accurate than tTG in children under age 24 months.	No SRs assessing how test accuracy differs by age were identified. Regarding IgG DGP, one SR reported only on studies of adults, while another reported only on studies of children. A 2013 SR of 7 studies of non–IgA-deficient adults reported sensitivity of 75.4% to 96.7% and specificity of 98.5% to 100%. A 2012 SR of 3 studies in non–IgA-deficient children reported sensitivities of 80.1% to 98.6% and specificities of 86.0% to 96.9%.	Two large moderate-quality studies reported that both tTG and DGP tests were less sensitive in adults (range, 29% to 85%) than children (range, 57% to 96%). One study reported sensitivity of 96% and 100% for IgA tTG and IgA DGP, respectively, for children under age 24 months, while specificity was 98% and 31%, respectively. Accuracy was significantly lower for both tests in older children and adolescents.
Key Question 3: Demographics, including race	Insufficient: There was insufficient evidence to estimate the accuracy of diagnostic methods by demographic characteristics.	No SRs on this topic were identified.	No studies reported accuracy by race, ethnicity, or socioeconomic status.
Key Question 3: Patients with IgA deficiency	Insufficient: There was insufficient evidence to estimate the accuracy of diagnostic methods in IgA-deficient patients.	No SRs on this topic were identified.	Two small studies of the accuracy of new combination tests (IgA DGP + IgG DGP combo, IgA tTG + IgG DGP combo) in IgA-deficient patients were published in 2014; results were inconsistent.

Table A. Summary of findings and strength of evidence (continued)

Topic	EPC Conclusions and Strength of Evidence	Prior Systematic Reviews	Additional Findings From EPC
Key Question 3: Patients who previously tested negative for CD	Insufficient: There was insufficient evidence to estimate the accuracy of diagnostic methods in patients who previously tested negative for CD.	No SRs on this topic were identified.	A very small study (N = 17) found that patients with biopsy-verified CD who tested negative on IgA tested positive using IgA DGP or IgG DGP.
Key Question 4: Direct adverse events—VCE	High: The rate of capsule retention is less than 5%.	No SRs contained safety data on VCE used specifically for CD diagnosis. An SR of VCE not specific to CD found a capsule retention rate of 1.4% in 150 studies.	In 3 studies specific to CD, the capsule retention rate ranged from 0.9% to 4.6%.
Key Question 4: Direct adverse events—endoscopy with duodenal biopsy	Moderate: Adverse events during upper GI endoscopy are rare.	No SR contained safety data on upper GI endoscopy or duodenal biopsy when used specifically to diagnose CD. A review on upper endoscopy in general found infection very rare and bleeding very rare (1.6 per 1,000) unless a polyp is removed.	No studies specific to diagnosis of CD were identified.
Key Question 4: Indirect adverse events—false negatives or positives	Insufficient: Strength of evidence is insufficient regarding the impact of misdiagnosis.	No SRs on the impact of misdiagnosis of CD were identified. A study of 34 children with intestinal villous atrophy and simultaneous negative EmA IgA tests found that 2 infants were confirmed as having CD after 6–10 years of iterative cycles of gluten challenges and gluten-free diet. All 3 studies report high loss to followup.	In 2 small studies reporting sequelae in children with positive EmA serology but normal biopsy results, 30% to 50% of patients were diagnosed with CD after gluten challenge. These studies were conducted prior to the availability of other serological tests, so applicability is limited.

CD = celiac disease; CI = confidence interval; DGP = deamidated gliadin peptide; EmA = endomysial antibodies; EPC = Evidence-based Practice Center; GI = gastrointestinal; ESPGHAN = European Society for Pediatric Gastroenterology, Hepatology, and Nutrition; HLA = human leukocyte antigen; IgA = immunoglobulin A; IgG = immunoglobulin G; LR- = negative likelihood ratio; LR+ = positive likelihood ratio; NPV = negative predictive value; OR = odds ratio; PPV = positive predictive value; SR = systematic review; tTG = anti-tissue transglutaminase; VCE= video capsule endoscopy.

Applicability

Several factors affect the applicability of this review.

To increase generalizability, this report limited inclusion of accuracy studies to those that enrolled consecutive patients or a random sample. Several studies were excluded because enrollment could not be determined given the information available.

Only one study of accuracy in the asymptomatic general population met the criterion that all subjects, regardless of serology results, undergo biopsy. The cost of performing biopsies in all subjects and the low rate of acceptance of biopsy in seronegative asymptomatic individuals make the conduct of such studies challenging. Thus, the evidence on accuracy of diagnosis in the general asymptomatic population with no risk factors for CD is categorized as low strength.

Although this report is limited to diagnostic methods currently used in the United States, study location was not a basis for study exclusion. Many studies were conducted in Europe, the Middle East, and South Asia. Due to differences in genetics and disease prevalence, the applicability of these studies to the U.S. population is uncertain.

No studies stratified accuracy results by racial or ethnic group. Few studies focused on populations of special interest.

Most studies were conducted by gastroenterologists in academic settings. This report found a significant difference in interpretation of biopsy results between academic and nonacademic physicians. The majority of accuracy studies included in this report used Marsh classification to categorize biopsy results. (Marsh III or higher is classified as CD.) In contrast, many community physicians use a simple qualitative assessment of villous atrophy or elevation of intraepithelial lymphocytes to make a diagnosis.

Accuracy of serology assays may vary by both laboratory and manufacturer. For example, Li and colleagues (2009)[19] used 150 samples from subjects of known CD status to compare accuracy of tTG tests at 20 laboratories in the United States and Europe. Sensitivity was less than 75 percent at four laboratories. Rozenberg and colleagues (2012)[20] found differences in performance of tTG tests across various manufacturers by using a similar research design.

Finally, VCE is not a first-line diagnostic method: it is indicated for adults who refuse biopsy. A 2012 systematic review of six studies reported very good sensitivity and excellent specificity with VCE. However, there may be differences in patient characteristics between those who refuse and those who accept a biopsy. For example, those with more severe symptoms are hypothesized to be more likely to accept a biopsy.

Implications for Clinical and Policy Decisionmaking

The findings of this review support those of previous SRs on the accuracy of individual diagnostic tests using IgA. All IgA tests for CD have excellent specificity; DGP IgA has slightly lower specificity than tTG IgA and EmA IgA. Testing for tTG IgA has a high PPV for most clinical populations with a modest prevalence of CD. EmA IgA has good sensitivity, DGP IgA has very good sensitivity, and tTG IgA has excellent sensitivity. DGP IgG tests have very good sensitivity and excellent specificity, even in non–IgA-deficient individuals.

Unfortunately, due to a dearth of studies meeting our inclusion criteria, we were unable to determine which tests, if any, are more accurate in patients with specific symptoms or risk factors. Patients with symptoms associated with CD would impact the pretest probability and, as a result, the likelihood of disease based on a positive result. No studies of test accuracy in patients with trisomy 21, Turner syndrome, or Williams syndrome were identified. The few studies of patients with type 1 diabetes included small samples and were conducted in non-Western countries. Thus, no clinical implications for testing individuals with specific risks can be stated at this time.

New research has found DGP tests to be more accurate than tTG tests in small children; strength of evidence is low but could increase if findings are replicated. Compared with EmA IgA, tTG IgA had greater sensitivity in the one study of the general asymptomatic population identified that met our inclusion criteria that all participants undergo biopsy, regardless of serology results. The quality of this general population study was high, the sample size was large (over 1,000), and it was conducted in a Western country (Sweden) with estimated CD prevalence similar to that in the United States.

This review found insufficient evidence to determine which populations would most benefit from diagnostic algorithms that combine a tTG test with an EmA or DGP test. A combination of positive serological testing with a threshold level at or several times above the upper limit of normal for specific celiac tests may be accurate for diagnosing CD without requiring histopathology specimens. However, the

currently available evidence on comparative accuracy of algorithms is inconclusive because of the wide range of results, heterogeneity of populations studied, and lack of clinically significant increases in accuracy compared with individual tests. Future studies aimed at the diagnostic accuracy of multiple-test strategies would strengthen the evidence for this approach.

Finally, regarding biopsy, there is high-strength evidence that multiple specimens should be taken from the duodenal bulb and the distal duodenum for optimal diagnostic yield in both the adult and pediatric population. There is moderate-strength evidence that CD is underdiagnosed by pathologists in community settings compared with academic settings; continued education on diagnostic protocols may be warranted for community physicians.

Research Gaps

Although the accuracy of various serological tests for CD in symptomatic individuals has a high strength of evidence, strength of evidence on the comparative accuracy of algorithms such as those recommended by organizations such as ESPGHAN is insufficient because of the small number of studies, heterogeneity of study populations, and inconsistent results. Further studies should be conducted. Appendix F of the full report contains details on the test combinations, populations, and strength–of-evidence domains for each algorithm studied.

Evidence is insufficient to recommend specific tests for particular at-risk populations. Patient-level factors that have been hypothesized to affect test accuracy include race and ethnicity, but no studies stratified results by these characteristics.

Because of the inherently invasive nature of biopsy, the vast majority of studies of serological test accuracy using biopsy as the reference standard have been conducted in patients presenting for testing due to symptoms. The most common symptoms are GI symptoms (diarrhea, constipation, pain, etc.) as well as signs of malnutrition in children. High accuracy was found in the only general population screening study; however, despite the high scientific quality of this study, the strength of evidence for accuracy in the asymptomatic general population is low because the study has never been replicated. This lack of evidence does not mean the tests are inaccurate in asymptomatic individuals; lack of evidence does not equal evidence of inaccuracy.

No studies were identified that addressed the key issue, "What impact does the method of initial diagnosis have on how a physician follows up with a patient?" Retrospective analyses of existing databases may shed light on this question.

Finally, studies may be needed to investigate the long-term impact of misdiagnosis. False positives and false negatives may be important "harms" because of (a) huge lifestyle changes involved for positive diagnosis and (b) potential harms to health (malabsorption, intestinal damage) from undiagnosed CD.

Conclusions

New evidence on accuracy of tests used to diagnose CD supports the excellent sensitivity of IgA tTG tests and excellent specificity of both IgA tTG and IgA EmA tests reported in prior SRs. High strength of evidence of accuracy, particularly in children, was found for DGP tests in recent SRs. Regarding comparative accuracy, IgA EmA tests have lower sensitivity but similar specificity to IgA tTG tests. IgA DGP and IgG DGP tests are not as sensitive as IgA tTG tests in non–IgA-deficient adults. These conclusions are based primarily on indirect evidence—i.e., pooled results on accuracy of individual tests rather than head-to-head studies comparing accuracy of different tests in the same samples. However, strength of evidence is high given the large numbers of studies, the consistency of results, and the precision of the confidence intervals.

Algorithms combining tTG with either EmA or DGP tests appear to be accurate in both children and adults; however, strength of evidence for comparative accuracy is insufficient given the low number of studies relative to single tests, heterogeneity of populations, and wide range of results. The increase in accuracy over individual tests is not consistently clinically significant. Additional studies of algorithms are needed.

Notably, current ESPGHAN guidelines state that a patient with a tTG result greater than 10 times the normal limit should undergo an EmA test and HLA typing. If the patient tests positive and then responds to a gluten-exclusion diet, a diagnosis of CD can be made without use of biopsy. These guidelines have not been adopted by societies in the United States. Evidence seems to support the accuracy of a multiple-testing strategy without biopsy; however, additional studies are needed to confirm the threshold levels that provide the highest accuracy and population differences, if any.

VCE is a safe and fairly accurate means of diagnosing CD in adults who wish to avoid biopsy; risk of retaining the capsule is approximately 4.6 percent. However, our pooled results reveal that some serological tests have higher sensitivity and specificity. No data are available on

how VCE accuracy varies by population characteristics or setting. Endoscopy with biopsy has a very low risk of adverse events; accuracy appears to be greater in academic than community settings.

Importantly, few applicable studies on the sequelae of false positive or false negative diagnoses were identified. Long-term followup of patients, regardless of diagnosis results, should be encouraged.

References

1. Rashtak S, Murray JA. Review article: coeliac disease, new approaches to therapy. Aliment Pharmacol Ther. 2012 Apr;35(7):768-81. PMID: 22324389.

2. Rubio-Tapia A, Ludvigsson JF, Brantner TL, et al. The prevalence of celiac disease in the United States. Am J Gastroenterol. 2012 Oct;107(10):1538-44; quiz 7, 45. PMID: 22850429.

3. Ludvigsson JF, Rubio-Tapia A, van Dyke CT, et al. Increasing incidence of celiac disease in a North American population. Am J Gastroenterol. 2013 May;108(5):818-24. PMID: 23511460.

4. See J, Murray JA. Gluten-free diet: the medical and nutrition management of celiac disease. Nutr Clin Pract. 2006 Feb;21(1): 1-15. PMID: 16439765.

5. Rubio-Tapia A, Hill ID, Kelly CP, et al. ACG clinical guidelines: diagnosis and management of celiac disease. Am J Gastroenterol. 2013 May;108(5):656-76; quiz 77. PMID: 23609613.

6. Bottaro G, Rotolo N, Spina M, et al. [Evaluation of sensitivity and specificity of antigliadin antibodies for the diagnosis of celiac disease in childhood]. Minerva Pediatr. 1995 Dec;47(12):505-10. PMID: 8900559.

7. Bai J, Zeballos E, Fried M, et al. WGO-OMGE Practice Guideline: Celiac Disease. World Gastroenterology Organisation; 2007.

8. National Institute for Health and Clinical Excellence. Coeliac Disease: Recognition and Assessment of Coeliac Disease. London; National Institute for Health and Clinical Excellence; 2009 May. www.org.uk/guidance.

9. Rostom A, Murray JA, Kagnoff MF. American Gastroenterological Association (AGA) Institute technical review on the diagnosis and management of celiac disease. Gastroenterology. 2006 Dec;131(6):1981-2002. PMID: 17087937.

10. Husby S, Koletzko S, Korponay-Szabo IR, et al. European Society for Pediatric Gastroenterology, Hepatology, and Nutrition guidelines for the diagnosis of coeliac disease. J Pediatr Gastroenterol Nutr. 2012 Jan;54(1):136-60. PMID: 22197856.

11. Rostom A, Dube C, Cranney A, et al. Celiac disease. Evid Rep Technol Assess (Summ). 2004 Jun;(104):1-6. PMID: 15346868.

12. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011 Oct 18;155(8):529-36. PMID: 22007046.

13. Chou R, Aronson N, Atkins D, et al. Chapter 11. Assessing harms when comparing medical interventions. 2009. In: Methods Guide for Effectiveness and Comparative Effectiveness Reviews. AHRQ Publication No. 10(14)-EHC063-EF. Rockville, MD: Agency for Healthcare Research and Quality. January 2014. Chapters available at www.effectivehealthcare.ahrq.gov.

14. Shea BJ, Grimshaw JM, Wells GA, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol. 2007;7:10. PMID: 17302989.

15. Parikh R, Mathai A, Parikh S, et al. Understanding and using sensitivity, specificity and predictive values. Indian J Ophthalmol. 2008;56(1):45. PMID: 18158403.

16. McGee S. Simplifying likelihood ratios. J Gen Intern Med. 2002 Aug;17(8):646-9. PMID: 12213147.

17. Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. BMJ. 2004 Jul 17;329(7458):168-9. PMID: 15258077.

18. Chang SM, Matchar DB, Smetana GW, et al., eds. Methods Guide for Medical Test Reviews. AHRQ Publication No. 12-EC017. Rockville, MD: Agency for Healthcare Research and Quality; 2012. www.ncbi.nlm.nih.gov/pubmed/22834019.

19. Li M, Yu LP, Tiberti C, et al. A report on the International Transglutaminase Autoantibody Workshop for Celiac Disease. Am J Gastroenterol. 2009 Jan;104(1):154-63. PMID: 19098864.

20. Rozenberg O, Lerner A, Pacht A, et al. A novel algorithm for the diagnosis of celiac disease and a comprehensive review of celiac disease diagnostics. Clin Rev Allergy Immunol. 2012 Jun;42(3):331-41. PMID: 21279475.

Full Report