STOP!

The attached article has 262 pages – Don’t print it!
The answers to the questions can be found on pages 92-96
The other pages are for those inquisitive fellows who wish to know the data source

The questions are found at the end of the document (page 264)

DO NOT PRINT THIS ARTICLE!
Diagnosis of Celiac Disease
Diagnosis of Celiac Disease

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
5600 Fishers Lane
Rockville, MD 20857
www.ahrq.gov

Contract No. 290-2012-00006-I

Prepared by:
Southern California Evidence-Based Practice Center
Santa Monica, CA

Investigators:
Margaret A. Maglione, M.P.P.
Adeyemi Okunogbe, M.B.Ch.B.
Brett Ewing, M.S.
Sean Grant, Ph.D.
Sydne J. Newberry, Ph.D.
Aneesa Motala, B.A.
Roberta Shanman, M.L.S.
Nelly Mejia, M.Phil.
Aziza Arifkhanova, M.S.
Paul Shekelle, M.D., Ph.D.
Gregory Harmon, M.D.

AHRQ Publication No. 15(16)-EHC032-EF
January 2016
This report is based on research conducted by the Southern California Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-2012-00006-I). The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

The information in this report is intended to help health care decisionmakers—patients and clinicians, health system leaders, and policymakers, among others—make well-informed decisions and thereby improve the quality of health care services. This report is not intended to be a substitute for the application of clinical judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical reference and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances presented by individual patients.

This report is made available to the public under the terms of a licensing agreement between the author and the Agency for Healthcare Research and Quality. This report may be used and reprinted without permission except those copyrighted materials that are clearly noted in the report. Further reproduction of those copyrighted materials is prohibited without the express permission of copyright holders.

AHRQ or U.S. Department of Health and Human Services endorsement of any derivative products that may be developed from this report, such as clinical practice guidelines, other quality enhancement tools, or reimbursement or coverage policies, may not be stated or implied.

This report may periodically be assessed for the currency of conclusions. If an assessment is done, the resulting surveillance report describing the methodology and findings will be found on the Effective Health Care Program Web site at www.effectivehealthcare.ahrq.gov. Search on the title of the report.

Persons using assistive technology may not be able to fully access information in this report. For assistance contact EffectiveHealthCare@ahrq.hhs.gov.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of systematic reviews to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. These reviews provide comprehensive, science-based information on common, costly medical conditions, and new health care technologies and strategies.

Systematic reviews are the building blocks underlying evidence-based practice; they focus attention on the strength and limits of evidence from research studies about the effectiveness and safety of a clinical intervention. In the context of developing recommendations for practice, systematic reviews can help clarify whether assertions about the value of the intervention are based on strong evidence from clinical studies. For more information about AHRQ EPC systematic reviews, see www.effectivehealthcare.ahrq.gov/reference/purpose.cfm.

AHRQ expects that these systematic reviews will be helpful to health plans, providers, purchasers, government programs, and the health care system as a whole. Transparency and stakeholder input are essential to the Effective Health Care Program. Please visit the Web site (www.effectivehealthcare.ahrq.gov) to see draft research questions and reports or to join an email list to learn about new program products and opportunities for input.

We welcome comments on this systematic review. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 5600 Fishers Lane, Rockville, MD 20857, or by email to epc@ahrq.hhs.gov.

Richard G. Kronick, Ph.D.
Director
Agency for Healthcare Research and Quality

Stephanie Chang, M.D., M.P.H.
Director
Evidence-based Practice Center Program
Center for Evidence and Practice Improvement
Agency for Healthcare Research and Quality

Arlene S. Bierman, M.D., M.S.
Director, Center for Evidence and Practice Improvement
Agency for Healthcare Research and Quality

Karen C. Lee, M.D., M.P.H.
Medical Officer, U.S. Preventive Services Task Force Program
Task Order Officer
Center for Evidence and Practice Improvement
Agency for Healthcare Research and Quality

Nahed El-Kassar, M.D., Ph.D.
Former Task Order Officer
Center for Evidence and Practice Improvement
Agency for Healthcare Research and Quality
Acknowledgments
The authors would like to thank Sean Rubin, B.S., and Patricia Smith for their assistance on the project.

Key Informants
In designing the study questions, the EPC consulted several Key Informants who represent the end-users of research. The EPC sought the Key Informant input on the priority areas for research and synthesis. Key Informants are not involved in the analysis of the evidence or the writing of the report. Therefore, in the end, study questions, design, methodological approaches, and/or conclusions do not necessarily represent the views of individual Key Informants.

Key Informants must disclose any financial conflicts of interest greater than $10,000 and any other relevant business or professional conflicts of interest. Because of their role as end-users, individuals with potential conflicts may be retained. The TOO and the EPC work to balance, manage, or mitigate any conflicts of interest.

The list of Key Informants who provided input to this report follows:
Stefano Guandalini, M.D.
University of Chicago Celiac Disease Center
Chicago, IL
Marilyn Geller, M.S.P.H.
Celiac Disease Foundation
Woodland Hills, CA
Nancee Jaffee, M.S., R.D.
Dietitian, University of California Los Angeles
Los Angeles, CA
Martin F. Kagnoff, M.D.
University of California San Diego
Laboratory of Mucosal Immunology
San Diego, CA

Danna Korn
Raising Our Celiac Kids (ROCK)
Carlsbad, CA
Stephen Levinson, M.D.
Gastroenterologist, Community Practice
Burbank, CA
Joseph Murray, M.D.
Gastroenterologist, Mayo Clinic
Rochester, MN
Mary Schluckebier, M.A.
Celiac Sprue Association
Omaha, NE
John Whitney, M.D.
Wellpoint, Office of Medical Policy
Albany, NY

Technical Expert Panel
In designing the study questions and methodology at the outset of this report, the EPC consulted several technical and content experts. Broad expertise and perspectives were sought. Divergent and conflicted opinions are common and perceived as healthy scientific discourse that results in a thoughtful, relevant systematic review. Therefore, in the end, study questions, design, methodologic approaches, and/or conclusions do not necessarily represent the views of individual technical and content experts.
Technical Experts must disclose any financial conflicts of interest greater than $10,000 and any other relevant business or professional conflicts of interest. Because of their unique clinical or content expertise, individuals with potential conflicts may be retained. The TOO and the EPC work to balance, manage, or mitigate any potential conflicts of interest identified.

The list of Technical Experts who provided input to this report follows:
Alessio Fasano, M.D.
Director, Center for Celiac Research & Treatment
MassGeneral Hospital for Children
Boston, MA
Stefano Guandalini, M.D.
University of Chicago Celiac Disease Center
Chicago, IL
Martin F. Kagnoff, M.D.
University of California San Diego
Laboratory of Mucosal Immunology
San Diego, CA
Daniel A. Leffler, M.D., M.S.
Director of Research, The Celiac Disease Center at BIDMC
Director of Quality Assurance, Division of Gastroenterology
Beth Israel Deaconess Medical Center
Boston, MA
Joseph Murray, M.D.
Gastroenterologist, Mayo Clinic
Rochester, MN
Michelle Pietzak, M.D.
Children's Hospital Los Angeles
Los Angeles, CA

Peer Reviewers
Prior to publication of the final evidence report, EPCs sought input from independent Peer Reviewers without financial conflicts of interest. However, the conclusions and synthesis of the scientific literature presented in this report do not necessarily represent the views of individual reviewers.

Peer Reviewers must disclose any financial conflicts of interest greater than $10,000 and any other relevant business or professional conflicts of interest. Because of their unique clinical or content expertise, individuals with potential nonfinancial conflicts may be retained. The TOO and the EPC work to balance, manage, or mitigate any potential nonfinancial conflicts of interest identified.

The list of Peer Reviewers follows:
Manish J. Gandhi, M.D.
Mayo Clinic
Rochester, MN
Benjamin Lebwohl, M.D., M.S.
Celiac Disease Center at Columbia University,
New York, NY
Melissa Snyder, Ph.D.
Mayo Clinic
Rochester, MN
Ritu Verma, M.D.
Attending Physician/Director of Celiac Center at the Children’s Hospital of Philadelphia
Philadelphia
Philadelphia, PA
Diagnosis of Celiac Disease

Structured Abstract

Objectives. To report the evidence on comparative accuracy and safety of methods used in current clinical practice to diagnose celiac disease, including serological tests, human leukocyte antigen (HLA) typing, and video capsule endoscopy. Diagnostic tests used singly and in combination in various populations were compared against the reference standard of endoscopic duodenal biopsy. In addition, factors affecting biopsy accuracy were reviewed.

Data sources. Electronic searches of PubMed®, Embase®, the Cochrane Library, and Web of Science from 1990 through March 2015. Reference lists of included publications were searched for additional relevant studies, and experts were asked to suggest studies.

Review methods. Studies of diagnostic accuracy were included if all participants underwent the index test and endoscopy with duodenal biopsy as the reference standard. Systematic reviews on accuracy and studies on adverse events associated with testing were included. Standard assessment tools were used to evaluate study risk of bias. Where possible, results of accuracy studies were pooled using meta-analysis. When pooling was not possible, findings were described narratively and presented in tables and figures.

Results. A total of 7,254 titles were identified, from which 60 individual studies and 13 prior systematic reviews were included. The majority of studies were conducted in participants with symptoms. New meta-analyses found high-strength evidence to support excellent accuracy of anti-tissue transglutaminase (tTG) immunoglobulin A (IgA) tests (sensitivity = 92.5%; specificity = 97.9%) and excellent specificity of endomysial antibodies (EmA) IgA tests (sensitivity = 79.0%; specificity = 99.0%), as reported in previous systematic reviews. Promising results were reported for deamidated gliadin peptide antibodies (DGP) IgA tests (sensitivity = 87.8%; specificity = 94.1%) in a recent meta-analysis. Evidence for algorithms using multiple tests was insufficient because of diverse results, low number of studies, and heterogeneity of populations. Evidence was also insufficient for accuracy in asymptomatic general population screening and special populations such as children and patients with type 1 diabetes, anemia, and IgA deficiency.

Conclusions. New evidence on accuracy of tests used to diagnose celiac disease supports the excellent sensitivity of tTG IgA tests and excellent specificity of both tTG IgA and EmA IgA tests. Sensitivity of DGP IgA and immunoglobulin G tests is slightly less than for tTG IgA. Additional studies are needed to confirm the accuracy of diagnostic tests in special populations and to validate promising algorithms.
Contents

Executive Summary ....................................................................................................................... ES-1
Introduction ................................................................................................................................... 1
  Background ................................................................................................................................. 1
  Condition
  Diagnostic Strategies
Scope and Key Questions ........................................................................................................... 3
  Scope of the Review
  Key Questions
Organization of This Report

Methods .......................................................................................................................................... 2
  Topic Refinement and Review Protocol ..................................................................................... 2
  Literature Search Strategy ......................................................................................................... 2
  Inclusion and Exclusion Criteria ............................................................................................... 3
  Study Selection ......................................................................................................................... 4
  Data Extraction .......................................................................................................................... 4
  Quality (Risk of Bias) Assessment of Individual Studies .......................................................... 4
  Statistical Analyses ..................................................................................................................... 6
  Strength of the Body of Evidence ............................................................................................. 7
  Applicability .............................................................................................................................. 8
  Peer Review and Public Commentary ....................................................................................... 8

Results .......................................................................................................................................... 10
  Results of Literature Searches .................................................................................................. 10
  Key Question 1. Comparative Effectiveness ............................................................................ 12
    Description of Included Studies
    Key Points .............................................................................................................................. 26
    Detailed Synthesis ................................................................................................................... 24
  Key Question 2. Duodenal Biopsy Issues .................................................................................. 40
    Key Points .............................................................................................................................. 40
    Detailed Synthesis ................................................................................................................... 41
  Key Question 3. Specific Populations ...................................................................................... 46
    Key Points .............................................................................................................................. 46
    Detailed Synthesis ................................................................................................................... 46
  Key Question 4. Adverse Events .............................................................................................. 50
    Key Points .............................................................................................................................. 50
    Detailed Synthesis ................................................................................................................... 51

Discussion ..................................................................................................................................... 56
  Key Findings and Strength of Evidence
  Findings in Relationship to What Is Already Known
  Applicability
  Implications for Clinical and Policy Decisionmaking
  Limitations of the Comparative Effectiveness Review Process
  Limitations of the Evidence Base
    Volume
    Design
Appendixes
Appendix A. Search Strategy
Appendix B. List of Excluded Studies
Appendix C. Evidence Table
Appendix D. Data Abstraction Tools
Appendix E. AMSTAR Criteria
Appendix F. Strength of Evidence for Accuracy of Serology Tests
Executive Summary

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. The prevalence of CD in the United States has been estimated at approximately 1 percent but appears to be increasing for reasons that are not clear. 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.

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. 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.

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, although the test has low specificity. As AGA tests are no longer recommended, 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 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. 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.

**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

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—

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

b. **Intermediate outcomes,** such as clinical decisionmaking and dietary compliance?

c. **Clinical outcomes and complications related to CD?**

d. **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 syndrome
  - 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 instrument (revised Quality Assessment of Diagnostic Accuracy Studies instrument) was used to assess the risk of bias of accuracy studies; the McHarm instrument was used to assess the quality of studies on adverse events; and the AMSTAR 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. 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. 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.
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. 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.

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. 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.
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**

<table>
<thead>
<tr>
<th>Topic</th>
<th>EPC Conclusions and Strength of Evidence</th>
<th>Prior Systematic Reviews</th>
<th>Additional Findings From EPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Key Question 1: Accuracy of IgA tTG</td>
<td>High: IgA tTG tests have excellent sensitivity and specificity.</td>
<td>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, 95.7% to 99.0%), respectively.</td>
<td>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%.</td>
</tr>
<tr>
<td>Key Question 1: Accuracy of IgA EmA</td>
<td>High: IgA EmA tests have lower sensitivity but equal specificity to IgA tTG tests.</td>
<td>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%).</td>
<td>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 villos 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%.</td>
</tr>
<tr>
<td>Key Question 1: Accuracy of IgA DGP</td>
<td>High: IgA DGP tests are not as accurate as IgA tTG tests.</td>
<td>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%).</td>
<td>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.</td>
</tr>
<tr>
<td>Key Question 1: Accuracy of IgG DGP</td>
<td>Moderate: IgG DGP tests are not as sensitive as IgA tTG tests in non–IgA-deficient patients.</td>
<td>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.</td>
<td>One study reported sensitivity of 95.0% and specificity of 99.0% in 200 non–IgA-deficient subjects of all ages.</td>
</tr>
<tr>
<td>Key Question 1: Accuracy of HLA-DQ2 or DQ8</td>
<td>High: HLA tests can be used to rule out CD with close to 100% sensitivity.</td>
<td>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</td>
<td>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%.</td>
</tr>
</tbody>
</table>
the NPV of the HLA-DQ2/DQ8 combination test at over 99%.

A 1999 cohort also reported sensitivity of 100%, while specificity was 33.3%.

| 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. |
| 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. |
| 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 | 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. |

| 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 | Insufficient: Strength of evidence is insufficient regarding the impact of misdiagnosis. | No SRs on the impact of misdiagnosis of CD were identified. | 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 |

ES-14
to the availability of other serological tests, so applicability is limited. 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 follow-up.

| positives | to the availability of other serological tests, so applicability is limited. 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 follow-up. |

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


Introduction

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 one 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 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 aforementioned 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 gastrointestinal cancers, and all-cause mortality.\(^4\)

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. Methods include various serology tests, HLA typing, video capsule endoscopy, and endoscopic duodenal biopsy (often considered the gold standard). Serology tests include anti-gliadin antibodies (AGA), IgA & IgG; anti-tissue transglutaminase (tTG), IgA & IgG; Endomysial antibodies (EmA), IgA; and the deamidated gliadin peptide (DGP) Antibodies, IgA & IgG. These tests are often used by providers as a panel in order to increase specificity and prevent false positives or increase sensitivity and prevent false negatives. All methods other than HLA typing require that the patient maintain a gluten containing diet during the diagnostic process. Commonly used diagnostic methods are described below.

Anti-gliadin antibodies (AGA), IgA & 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.\(^6\) However, the test has low specificity, because anti-gliadin IgG is found in both acute and chronic common intestinal childhood diseases. In 2007, the World Gastroenterology Organization recommended against using these tests.\(^7\) In 2009, the UK National Institute for Clinical Excellence (NICE) also recommended against using the tests.\(^8\) As AGA tests are no longer recommended, they are not addressed in this systematic review.
Anti-tissue transglutaminase (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 celiac disease in those over the age of 2 years.5 These tests are also included in the algorithms of all recent guidelines. It is important to note that IgA deficiency is more prevalent in CD patients than in the general population; therefore, other tests may be ordered as an alternative in those who are IgA deficient.

Endomysial antibodies (EmA), IgA. The thin connective tissue layer that covers individual muscle fibers is called endomysium. When the intestinal lining is damaged, endomysial antibodies (EmA) develop. Most patients with active celiac disease and many with dermatitis herpetiformis have the IgA class of anti-EmA antibodies. Although this test is included in the algorithms of recent guidelines for diagnosis, it is not as widely used in the U.S. as in other countries, and many providers simply order a biopsy if the tTG levels are high. In addition, this test is less useful in individuals with low IgA.

Deamidated gliadin peptide (DGP) Antibodies. Elevated DGP antibodies are often seen in patients with celiac disease on a gluten-containing diet; this newer test may give a positive result in some individuals with CD who are anti-tTG negative, including children younger than 2 years old. Testing both DGP IgG and anti-tTG IgG is recommended by the ACG for those who have low IgA or IgA deficiency.5

Human leukocyte antigen (HLA) typing. Susceptibility to CD is linked to certain human leukocyte antigen (HLA) class II alleles, especially in the HLA-DQ region. HLA molecules are hypothesized to present gluten antigens to T-cells, which in turn induce tissue damage.9 Approximately 95 percent of patients with CD have the HLA-DQ2 heterodimer, while the remaining 5 percent of persons with CD have the HLA-DQ8 heterodimer.10 Since 25 percent to 40 percent of the U.S. population carries either the DQ2 or DQ8 gene, the presence of either is not pathognomonic for CD. However, lack of these heterodimers all but rules out CD and genetic susceptibility for the disorder. Thus, these genetic tests are routinely used to rule out CD and are part of the diagnostic algorithms recommended by the European Society for Pediatric Gastroenterology, Hepatology, and Nutrition (ESPGHAN) and the ACG.11

Video capsule endoscopy. In this procedure, a capsule containing a tiny camera is ingested by the patient, providing high quality visual evidence of CD. While not a traditional means of detecting CD, it is used in adults who seek to avoid biopsy. During Topic Refinement, Key Informants (KI) requested assessment of the evidence for this method 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. The Modified Marsh criteria are utilized by most pathologists in evaluating histology findings from duodenal biopsy specimens for celiac disease diagnosis and progression of treatment during follow up.12 The criteria are graded from 0-3 with grade 3 further subdivided to 3a, 3b, and 3c.13, 14 Patients with Marsh grade 0 have normal histologic findings and are very unlikely to have celiac disease. In Marsh grade 1 and 2, biopsy specimens demonstrate raised Intraepithelial lymphocytes (IELs>30 per enterocytes) alone and raised IELs with crypt hyperplasia respectively. These histologic outcomes may be found in celiac patients on treatment or in patients with dermatitis herpetiformis. However, grade 1 or 2 lesions alone in the absence of clinical or serology evidence are nonspecific and are suggestive, but not confirmatory, of celiac disease.15 Patients with Grade 3 lesions have raised IELs with crypt hyperplasia and a measure of villous atrophy. Grade 3a, 3b, and 3c, in addition to raised IELs with crypt hyperplasia, also have findings of partial villous atrophy, subtotal villous atrophy, and
Marsh grade 3 lesion is the classic celiac lesion and is characteristic, but not diagnostic, of celiac disease. Of note, some community physicians use a simple qualitative assessment of villous atrophy or elevation of intraepithelial lymphocytes to make a diagnosis rather than relying on Marsh criteria.

Obtaining properly oriented tissue samples can be difficult, patchy mucosal lesions can be missed, and limiting the portion of gut examined may risk missing the diagnosis of CD-related complications such as lymphoma and ulcerative jejunoileitis. Some patients and parents are concerned about the risk of adverse events such as perforations and bleeding. Patients may feel pain and discomfort, which is especially problematic for small children.

Combinations of the above. Many providers use a serology panel or sequential approach in order to prevent false positives that are associated with tests that don’t work well under varying circumstances. The current systematic review compares the effectiveness of diagnostic tests, singularly and in combination.

**Scope and Key Questions**

**Scope of the Review**

Several systematic reviews and guidelines on diagnosis of CD have been published in the past decade, often with contradictory findings and recommendations. At least five recent guidelines for the diagnosis of CD have been published by recognized research/academic/medical bodies such as the North American Society for Pediatric Gastroenterology, Hepatology, and Nutrition (NASPGHAN) and ESPGHAN.

These clinical practice guidelines are complex and recommend different approaches to diagnosis. For example, some guidelines propose different sequences of tests for diagnosing population groups such as children versus adults, and symptomatic versus asymptomatic patients at increased risk (e.g. ESPGHAN). In addition, some guidelines (ACG and World Gastroenterology Organization [WGO]) uphold endoscopic biopsy as the gold standard for confirming diagnosis, whereas other guidelines (ESPGHAN) explore the use of other tests to serve as substitutes for biopsy. The diagnosis of celiac disease is further complicated by lack of provider knowledge and variability in laboratory cut-off levels to indicate “positive” results. It is also unknown whether the same diagnostic criteria apply to different racial, ethnic, or other demographic subgroups, or if they may be incorrectly diagnosed or underdiagnosed. In addition, false positives and false negatives may have significant consequences: Positive diagnosis requires huge lifestyle changes, and undiagnosed CD can result in potential health harm (nutrient malabsorption, osteoporosis, and lymphoma).

This report compares the accuracy of the diagnostic methods listed above in children, adults, and sub-populations of interest to clinicians and patient groups. Diagnostic methods that are no longer included in guidelines or still in development (not approved in the U.S.) are beyond the scope of this project. Accuracy of serological tests and VCE are based on biopsy results as the reference standard. We also assess how biopsy results may vary by provider characteristics, technique, and the length of time the patient has consumed gluten. Finally, we report adverse events associated with invasive diagnostic methods (biopsy, VCE) and sequelae of false or indeterminate results of diagnosis. We provide below an analytic framework to illustrate the populations, interventions, outcomes, and adverse effects that guided the literature search and synthesis for this project (Figure 1).
Key Questions

Four Key Questions guided this systematic review, as follows.

**Key Question 1.** What is the comparative effectiveness of the different diagnostic methods (various serological tests, 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—

a. Accuracy: sensitivity, specificity, positive likelihood ratio (LR+), negative likelihood ratio (LR-), summary receiver operating characteristics (ROCs)?

b. Intermediate outcomes, such as clinical decisionmaking and dietary compliance?

c. Clinical outcomes and complications related to CD?

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

**Key Question 2.** Does 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, (sensitivity, specificity, LR+, LR-, summary ROCs) and outcomes differ among specific populations (subgroups of Key Question 1), 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 (i.e., bleeding from biopsy) or harms (related to false positives, false negatives, indeterminate results) associated with testing for CD?

In addition, we identify the following PICOTS (Populations, Interventions, Comparators, Outcomes, and Timing) for the Key Questions:

**Population(s):**
- For KQ 1, 2, and 4:
  - All populations tested for CD
- For KQ 3:
  - Patients with signs and symptoms of celiac disease, for example:
    - Diarrhea
    - Constipation
    - Dermatitis
    - Malabsorption (anemia, folate deficiency)
  - Asymptomatic individuals at risk of celiac disease
    - Family history
    - Type 1 diabetes
    - Auto-immune disease
    - Turner’s syndrome
    - Trisomy 21
  - Children, under age 24 months vs older children & adolescents
  - Adults (aged 18+)
  - Ethnic and geographic populations
  - Low socioeconomic status (SES)
  - Patients with IgA deficiency
  - Patients previously testing negative for CD

**Interventions:**
- For KQ 1, 3, 4:
  - Endomysial antibodies (EmA) IgA test
  - Anti-tissue transglutaminase (tTG) IgA test
  - Deamidated gliadin poeptide (DGP) IgA antibodies
  - EmA IgG, tTG IgG, and DGP IgG tests for IgA deficient individuals
  - HLA typing
  - Video capsule endoscopy
  - Combinations of the above
- For KQ 2:
  - Endoscopy with biopsy
Comparators:
- For KQ 1, 3:
  o Endoscopy with duodenal biopsy
- For KQ 2:
  o Repeat biopsy

Outcomes:
- For KQ 1a, KQ2 and 3a-f, for Accuracy
  o Sensitivity
  o Specificity
  o PPV, NPV, FP, FN
  o Positive and negative likelihood ratios
- For KQ 1b, for Clinical decisionmaking
  o Additional testing for CD
  o Nutritionist advice on gluten-free diet
  o Follow up and monitoring by MD
- For KQ 1c, for Clinical outcomes and complications
  o Nutritional deficits
  o Persistence of villous atrophy on biopsy
  o Lymphomas
- For KQ 1d, for Patient-centered outcomes
  o Quality of life
  o Discomfort
  o Bloating
  o Abdominal pain
  o Depression
- For KQ 4, for Harms
  o Immediate AEs from biopsy
  o Psychological stress related to false positive results
  o Sequelae of false negatives or indeterminate results

Timing:
- For KQ 2
  o Length of time ingesting gluten before biopsy

Setting:
- For all KQs
  o Outpatient: Academic
  o Outpatient: Community

Organization of This Report
The remainder of this report presents the methods used to conduct the literature searches, data abstraction, and analyses; the results of the literature searches and analyses; the conclusions; and a discussion of the limitations as well as suggestions for future research.
Methods

The methods for this comparative effectiveness review (CER) follow the methods suggested in the Agency for Healthcare Research and Quality (AHRQ) “Methods Guide for Effectiveness and Comparative Effectiveness Reviews” (available at http://www.effectivehealthcare.ahrq.gov/methodsguide.cfm). The main sections in this chapter reflect the elements of the protocol established for the CER.

Topic Refinement and Review Protocol

KIs representing a variety of end-user perspectives were recruited to assist in refining the KQs and topics to address in this systematic review. Conference calls were held with community and academic-affiliated gastroenterologists, celiac disease researchers, a celiac disease centric nutritionist, national patient organizations, and a payer representative.

AHRQ posted the KQs on the Effective Health Care Web site for public comment in February, 2014. The EPC refined and finalized the KQs after review of the public comments, taking into consideration the prior input from KIs.

A study protocol was drafted and a technical expert panel (TEP) was recruited to provide high-level content and methodological expertise throughout the development of the review.

The final protocol for the project was posted on the AHRQ Effective Health Care Web site on June 11, 2014.

Literature Search Strategy

The literature search methods are summarized in Table 1 below. The full search strategy is presented as Appendix A.

<table>
<thead>
<tr>
<th>Table 1. Literature search methods</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Publication dates</strong></td>
</tr>
<tr>
<td><strong>Search terms</strong></td>
</tr>
<tr>
<td><strong>Electronic databases</strong></td>
</tr>
<tr>
<td><strong>Scientific Information</strong></td>
</tr>
<tr>
<td><strong>Inclusion and Exclusion Criteria</strong></td>
</tr>
<tr>
<td>------------------------------------</td>
</tr>
<tr>
<td>The topic refinement phase involved a preliminary environmental scan. Several recent high quality systematic reviews (SR) on the accuracy of serological tests and VCE for diagnosis of celiac disease were identified. Electronic searches were conducted to identify additional primary studies that were not included in these SRs. Studies of diagnostic accuracy used designs such as controlled trials, prospective and retrospective cohorts, case-control studies, and case series. Studies were included if they met the following criteria:</td>
</tr>
<tr>
<td>• Diagnostic method is recommended in a current guideline (see list in Introduction.) For example, visualization via endoscopy (without biopsy) and AGA tests were excluded. Methods in development (genotyping other than HLA, ultrasound, mucosal immunohistochemistry, etc.) were also excluded.</td>
</tr>
<tr>
<td>• The tests were used to diagnose celiac disease, rather than for management of existing CD. For example, studies where serology was used to monitor adherence to diet or biopsy was used to monitor improvement of intestinal atrophy were excluded.</td>
</tr>
<tr>
<td>• All participants underwent both the “index test” and the reference standard (biopsy). If only subjects testing positive via serology underwent biopsy, the study was excluded.</td>
</tr>
<tr>
<td>• The study reports sensitivity, specificity, or data that allow their calculation.</td>
</tr>
<tr>
<td>• Study must be published in English</td>
</tr>
<tr>
<td>• Study must enroll consecutive or random sample</td>
</tr>
<tr>
<td>• Sample size is 300 or greater, unless one of the following is the focus</td>
</tr>
<tr>
<td>o Low SES</td>
</tr>
<tr>
<td>o Previously negative for celiac disease via serology or biopsy</td>
</tr>
<tr>
<td>o IgA deficient</td>
</tr>
<tr>
<td>o Type 1 diabetes</td>
</tr>
<tr>
<td>o Turner’s syndrome</td>
</tr>
<tr>
<td>o Trisomy 21 / Downs</td>
</tr>
<tr>
<td>o Iron deficiency anemia</td>
</tr>
<tr>
<td>o Family history</td>
</tr>
<tr>
<td>o Accuracy results are stratified by race / ethnicity</td>
</tr>
<tr>
<td>The following were excluded from this systematic review.</td>
</tr>
<tr>
<td>• Animal studies</td>
</tr>
<tr>
<td>• Individual case reports</td>
</tr>
<tr>
<td>• Studies not published in English</td>
</tr>
<tr>
<td>• Documents with no original data (commentary, editorial)</td>
</tr>
<tr>
<td>• Studies that measured only prevalence</td>
</tr>
</tbody>
</table>
Study Selection

Each title and abstract identified by the searches was screened by two researchers independently 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 and their reasons for exclusion is presented as Appendix B.

Data Extraction

The DistillerSR software package was used to manage the search output, screening, and data abstraction. The database can be used to calculate inter-rater reliability statistics of agreement and agreement adjusted for chance (kappa statistic) before resolution of disagreements. Assessment of inter-rater reliability can be used to guard against selection bias in choosing the articles for further review. Forms are displayed in Appendix D.

Data collection forms were designed by the project team in Distiller SR, piloted by the reviewers, further modified, and then the final forms piloted with a random selection of included studies to ensure agreement of interpretation. Study-level data abstracted included sample size; subjects’ demographic characteristics, symptoms, and risk factors; study design; type of diagnostic test including cut-off level; and any other potential confounders. A statistician abstracted all accuracy data (sensitivity, specificity, and data needed to calculate). At the project’s end, all abstracted data were uploaded to the federally-fund Systematic Review Data Repository.

Quality (Risk of Bias) Assessment of Individual Studies

Bias might be introduced at many points during the conduct of a study, affecting validity, reliability, and applicability of the results. Risk of bias of accuracy studies was assessed using the QUADAS-2 instrument; domains and items are described in Table 2 below. The AMSTAR instrument, described in Table 3, was used to assess the quality of prior systematic reviews. Finally, the McHarm instrument, presented in Table 4, was used to assess the quality of studies on adverse events. Each study was scored individually by two EPC researchers who met to reconcile any differences; discrepancies were resolved by an experienced methodologist.

The Evidence Tables (Appendix C) display each QUADAS-2 item score for each accuracy study included in this review. Appendix E displays the AMSTAR scores. The McHarm scores are presented in the body of the report as there were few studies on adverse events. The strengths and weaknesses of the studies are discussed throughout the report and are reflected in the final Strength of Evidence ratings and Discussion.

Table 2. Quality Assessment of Diagnostic Accuracy Studies (QUADAS) -2 questions for assessing risk of bias in diagnostic accuracy studies

<table>
<thead>
<tr>
<th>Domain 1: Patient Selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Was a consecutive or random sample of patients enrolled? (Yes/No/Unclear)</td>
</tr>
<tr>
<td>Was a case-control design avoided? (Yes/No/Unclear)</td>
</tr>
<tr>
<td>Did the study avoid inappropriate exclusions? (Yes/No/Unclear)</td>
</tr>
<tr>
<td>Could the selection of patients have introduced bias? Risk: Low/High/Unclear</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Domain 2: Index Test(s) (complete for each index test used)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Were the index test results interpreted without knowledge of the reference standard? (Yes/No/Unclear)</td>
</tr>
</tbody>
</table>
If a threshold was used, was it pre-specified? (Yes/No/Unclear)

Could the conduct or interpretation of the index test have introduced bias? Risk: Low/High/Unclear

**Domain 3: Reference Standard**

Is the reference standard likely to correctly classify the target condition? (Yes/No/Unclear)

Were the reference standard results interpreted without knowledge of the results of the index test? (Yes/No/Unclear)

Could the reference standard, its conduct, or its interpretation have introduced bias? Risk: Low/High/Unclear

**Domain 4: Flow and Timing**

Was there an appropriate interval between index test(s) and reference standard? (Yes/No/Unclear)

Did all patients receive a reference standard? (Yes/No/Unclear)

Did all patients receive the same reference standard? (Yes/No/Unclear)

Were all patients included in the analysis? (Yes/No/Unclear)

Could the patient flow have introduced bias? Risk: Low/High/Unclear

**Table 3. AMSTAR (A Measurement Tool to Assess Systematic Reviews) criteria for assessing quality of systematic reviews**

<table>
<thead>
<tr>
<th>Item</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>01. Was an a priori study design provided?</td>
<td>~ Yes ~No ~Unsure</td>
</tr>
<tr>
<td>02. Was there duplicate study selection and data extraction?</td>
<td>~ Yes ~No ~Unsure</td>
</tr>
<tr>
<td>03. Was a comprehensive literature search performed?</td>
<td>~ Yes ~No ~Unsure</td>
</tr>
<tr>
<td>04. Was the status of publication (gray literature) used as an inclusion criterion?</td>
<td>~ Yes ~No ~Unsure</td>
</tr>
<tr>
<td>05. Was a list of studies (included/excluded) provided?</td>
<td>~ Yes ~No ~Unsure</td>
</tr>
<tr>
<td>06. Were the characteristics of the included studies provided?</td>
<td>~ Yes ~No ~Unsure</td>
</tr>
<tr>
<td>07. Was the scientific quality of the included studies assessed and documented?</td>
<td>~ Yes ~No ~Unsure</td>
</tr>
<tr>
<td>08. Was the scientific quality of the included studies used appropriately in formulating conclusions?</td>
<td>~ Yes ~No ~Unsure</td>
</tr>
<tr>
<td>09. Were the methods used to combine the findings of studies appropriate?</td>
<td>~ Yes ~No ~Unsure</td>
</tr>
<tr>
<td>10. Was the likelihood of publication bias assessed?</td>
<td>~ Yes ~No ~Unsure</td>
</tr>
<tr>
<td>11. Was the conflict of interest stated?</td>
<td>~ Yes ~No ~Unsure</td>
</tr>
</tbody>
</table>

**Table 4. McMaster Quality Assessment Scale for Harms (McHarm)**

<table>
<thead>
<tr>
<th>Item</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Were the harms PRE-DEFINED using standardized or precise definitions?</td>
<td>~ Yes ~No ~Unsure</td>
</tr>
<tr>
<td>2. Were SERIOUS events precisely defined?</td>
<td>~ Yes ~No ~Not Applicable ~Unsure</td>
</tr>
<tr>
<td>3. Were SEVERE events precisely defined?</td>
<td>~ Yes ~No ~Not applicable ~Unsure</td>
</tr>
<tr>
<td>4. Were the number of DEATHS in each study group specified OR were the reason(s) for not specifying them given?</td>
<td>~ Yes ~No ~Unsure</td>
</tr>
<tr>
<td>5. Was the mode of harms collection specified as ACTIVE?</td>
<td>~ Yes ~No ~Unsure</td>
</tr>
<tr>
<td>6. Was the mode of harms collection specified as PASSIVE?</td>
<td>~ Yes ~No ~Unsure</td>
</tr>
</tbody>
</table>
Statistical Analyses

Studies that reported sensitivity, specificity, or receiver-operator characteristics, or provided the data to perform such calculations were abstracted for potential inclusion in a synthesis. Sensitivity is the ability of a test to correctly classify an individual as ‘diseased’ or in this case, having celiac disease, as confirmed by the “gold” standard (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. Sensitivity = (true positive) / (true positive + false negative). Specificity is the ability of a test to correctly classify an individual as ‘disease-free’ or in this case, as having symptoms or serologic results that indicate not having celiac disease when the individual doesn’t have celiac disease, as determined by the gold standard (biopsy). 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. Specificity = (true negative) / (true negative + false positive). 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.

Some studies of the accuracy of diagnostic tests report Likelihood Ratios (LR): 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. Diagnostic findings with LRss close to 0 indicate a decrease in the likelihood of disease. An LR of 1 indicates no change in the likelihood of disease. As the LR increases from 1, the likelihood of disease increases. Positive LRss are a measure of how the probability of the disease increases in the presence of a positive test finding: LR+ = sensitivity / (1 - specificity). Negative LRss are a measure of how the probability of the disease decreases if the test is negative: LR- = (1 - sensitivity) / specificity. An LR+ of more than 10 is considered good, as is an LR- of less than 0.1.

Finally, positive predictive value (PPV) is the probability that an individual who tests positive actually has the disease. Similarly, negative predictive value 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 sensitivity and specificity, we considered pooling their result. In such cases, studies were weighted by sample size. Sensitivity and specificity were pooled using the ‘mada’ package in R which runs a bivariate diagnostic random-effects meta-analysis. The random effects model estimates a pooled sensitivity and false positive rate (used to calculate specificity) along with the associated confidence limits. This approach accounts for the interrelated sensitivity and specificity estimates.27, 28 The bivariate modeling approach that was used accounts for between study variability by allowing for the nonindependence of sensitivity and specificity across the studies.29

Sensitivity analyses were conducted by stratifying by test type, cut-off level (titer), and population characteristics of interest. When pooling was not possible, study results were described narratively, according to comparisons of interest and study design, and presented either in tables or on figures.

**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. This method classifies (grades) the evidence based on the following domains: study limitations (risk of bias), consistency, directness, and precision. The grades and their definitions are presented below in Table 5.

<table>
<thead>
<tr>
<th>Table 5. Strength of evidence definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High</strong></td>
</tr>
<tr>
<td><strong>Moderate</strong></td>
</tr>
<tr>
<td><strong>Low</strong></td>
</tr>
<tr>
<td><strong>Insufficient</strong></td>
</tr>
</tbody>
</table>

Table 6 below, taken from the AHRQ Methods Guide for Diagnostic Tests,30 briefly describes the methods used to rate each domain. The rating system was originally designed to assess the body of evidence on health care interventions rather than diagnostic tests; thus, assessing these domains presents unique challenges. For example, in assessing the precision of estimates of test performance, it may be difficult to judge whether a particular confidence interval has any practical clinical implications. In addition, there may be no direct evidence to link a specific test with clinical outcomes.

<table>
<thead>
<tr>
<th>Table 6. Domains and their definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Domain</strong></td>
</tr>
<tr>
<td><strong>Study Limitations / Risk of Bias</strong></td>
</tr>
<tr>
<td>Domain</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>Good internal validity</td>
</tr>
<tr>
<td>Consistency</td>
</tr>
<tr>
<td>Directness</td>
</tr>
<tr>
<td>Precision</td>
</tr>
</tbody>
</table>

**Applicability**

Applicability assessment was based on the similarity of the populations in terms of characteristics listed in the PICOTs. These include age, gender, ethnicity, geographic location, SES, co-morbidities such as Type 1 diabetes, and symptoms such as iron deficiency. For example, a test might have high sensitivity and specificity in adults but not in small children, due to biological changes during the life course. These issues are addressed by KQ3.

**Peer Review and Public Commentary**

A draft version of this report was reviewed by several celiac disease experts. Reviewer names and affiliations are listed in the Acknowledgements section of the front matter. All peer
reviewers completed conflict of interest disclosures; none reported ties to any test manufacturer. The draft report was also posted on the AHRQ Effective Health Care Web site for public comment in February, 2015. Feedback from these sources was incorporated into the current version.
Results

Results of Literature Searches

As seen in Figure 2, the literature search identified a total of 7,254 titles as possibly related to the aims of this project. Of these, 68 were recommended by Key Informants, Technical Expert Panel (TEP) members, and a local content expert, 7,178 were identified by the electronic database searches, and an additional 8 were found through review of references of included studies. The titles of all 7,254 articles were screened, and 1,905 were rejected. Abstracts for the remaining 5,349 articles were reviewed, and an additional 3,649 articles were excluded, primarily because they were not about celiac disease (837) or were about celiac disease management rather than diagnosis (1,443). Other reasons for exclusion at abstract were the following: diagnostic methods out of our scope (endoscopic view without biopsy, experimental methods, etc.), focus on test processing (ELISA, PCR, etc.), serology with no biopsy comparison, simple prevalence report, no human subjects, no original data (commentary, editorial), data duplicated in other article, or individual case report. An additional 95 studies were rejected after abstract review, as they were already included in the identified relevant systematic reviews; this project was funded as a small systematic review (SR), which incorporated the aggregate results of prior SRs.

In total, 1,700 studies went on to the next phase of review. Of those, 345 did not have an abstract; upon review of a random 20-percent sample, we found none relevant and thus felt confident to exclude all. (Details are available upon request.) Four hundred and seventy were background articles, such as nonsystematic reviews, histories of diagnostic tests, or detailed descriptions of the biological processes involved. Excluding those studies left 1,230 research studies to retrieve for full-text screening. Thirteen prior systematic reviews were accepted. Based on full-text screening, 1,155 articles were excluded for the following reasons: not in English (1), not human (1), not about celiac disease (12), not about diagnostic tests (150), no original data (20), individual case reports (19), prevalence only (30), diagnostic methods beyond the scope of study (155), test processing issue (20), included in prior systematic reviews on topic (6), did not assess the accuracy or effectiveness of diagnostic tests (150), index test not compared to reference standard (biopsy) (62), not all subjects underwent both index test and reference standard (410), accuracy studies without consecutive or random sample (43), and accuracy studies where the sample size was less than 300 and not a special population (83). References for these excluded articles along with reasons for exclusion can be found in Appendix B.

Therefore, 60 individual studies and 13 SRs were included for evidence synthesis. Eleven SRs and 27 studies answered KQ1 and 3, one SR and 25 additional studies answered KQ2, and one SR and four studies answered KQ4.
Figure 2. Literature flow

Identified from EPC electronic searches (12/24/15) \(N = 7,438\)

Hand searching \(N = 8\)

Expert recommended \(N = 68\)

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

Abstracts rejected \(N = 3,649\)
- Not human: \(N = 183\)
- Not about CD: \(N = 837\)
- Not about diagnostic tests: \(N = 1,443\)
- No original data (letter, commentary, editorial, etc.): \(N = 92\)
- Individual case report: \(N = 271\)
- Prevalence only: \(N = 105\)
- 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 about CD: \(N = 12\)
- Not about diagnostic tests: \(N = 150\)
- No original data (letter, commentary, editorial, etc.): \(N = 20\)
- Individual case report: \(N = 1\)
- Prevalence only: \(N = 41\)
- Diagnostic methods outside the scope of study: \(N = 155\)
- Test processing issue: \(N = 7\)
- 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 360 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
- Sample size \(>=300\): \(N = 15\)
- Special pop with a sample size \(<300\): \(N = 12\)
- Sequelae by diagnostic method: \(N = 4\)
- SR: \(N = 11\)

KQ 2 (accuracy of biopsy)
- \(N = 25\)
  - Pathological characteristics: \(N = 5\)
  - Technical method: \(N = 19\)
  - Length of gluten challenge: \(N = 1\)
  - SR: \(N = 1\)

KQ 4 (adverse events)
- \(N = 4\)
  - SR: \(N = 1\)

CD = celiac disease; KQ = Key Question; SR = systematic review(s).
Key Question 1. Comparative Effectiveness

Description of Included Studies

We identified eleven systematic reviews (SRs) on the accuracy of the diagnostic methods within the scope of this review. Thirty-one studies published after the SRs reported sensitivity and specificity of these tests (or data that allowed their calculation) and met our inclusion criteria (consecutive or random sample, all subjects received index test and biopsy, sample size 300 or more, or special population of interest). The study characteristics are displayed in Table 7. With the exception of one study of over 12,000 subjects, studies ranged in size from 17 to 1,071 subjects. Notably, only two studies were conducted in the U.S. Four studies were conducted in the UK, five in the Middle East, one in India, and the rest in Western Europe. Ethnicity and race of participants were rarely described. All but one study included individuals presenting for testing due to symptoms, risk factors, or family history. Only one study using general population screening met our inclusion criterion that sero-negative subjects undergo biopsy. Of the accepted studies, very few included asymptomatic individuals; only one compared accuracy results in symptomatic individuals with those of asymptomatic individuals. Many of the newer studies focused on algorithms using more than one serologic test.

The far right column in Table 7 summarizes the risk of bias of the new studies with regard to their ability to determine test accuracy. Bias in patient selection was rated high in almost half of the studies; eight studies were rated so because they used a case-control design. No studies had high risk of bias due to conduct of the index test, although almost half of the studies were missing some information on this issue. No studies had high risk of bias due to the conduct of the reference test (biopsy), although eight studies were missing information to make a determination. Finally, patient flow may have introduced a high risk of bias in four studies.

In sum, the risk or bias and applicability of these studies varied widely. These issues will be discussed throughout the report and summarized in the Discussion section.
<table>
<thead>
<tr>
<th>Author, Year, Country</th>
<th>Study Design, Sample Size</th>
<th>Percent Female, Percent of Each Ethnicity, Population</th>
<th>QUADAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barada et al., 2014[77] Lebanon Prospective cohort Sample Size: 998</td>
<td>Percentage female: 55.2% Middle eastern 100% Population: Adults aged 18-70 (mean 43 years) with GI symptoms of CD. 2.6% had family history. 20% with anemia</td>
<td>QUADAS Domain 1 Biased patient selection: Low QUADAS Domain 2 Bias due to testing: Low QUADAS Domain 3 Bias due to reference test: Low QUADAS Domain 4 Could patient flow have introduced bias: Not Applicable</td>
<td></td>
</tr>
<tr>
<td>Basso et al., 2011[72] Italy Retrospective cohort Sample Size: 703</td>
<td>Percentage female: 62.4% Population: 100 U/ml. Children and adolescents with CD, latent CD, or controls. Mean age 8.</td>
<td>QUADAS Domain 1 Biased patient selection: Low QUADAS Domain 2 Bias due to testing: Unclear QUADAS Domain 3 Bias due to reference test: Unclear QUADAS Domain 4 Could patient flow have introduced bias: Not Applicable</td>
<td></td>
</tr>
<tr>
<td>Bienvenu et al., 2014[73] France Retrospective cohort Sample Size: 45</td>
<td>Percentage female: 40% Population: IgA deficient children tested with CD-LFIA (detects both human IgA and IgG anti DGP)</td>
<td>QUADAS Domain 1 Biased patient selection: Low QUADAS Domain 2 Bias due to testing: Low QUADAS Domain 3 Bias due to reference test: Low QUADAS Domain 4 Could patient flow have introduced bias: High</td>
<td></td>
</tr>
<tr>
<td>Author, Year, Country</td>
<td>Study Design, Sample Size</td>
<td>Percent Female, Percent of Each Ethnicity, Population</td>
<td>QUADAS</td>
</tr>
<tr>
<td>------------------------</td>
<td>----------------------------</td>
<td>-------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Cekin et al., 2012&lt;sup&gt;37&lt;/sup&gt; Turkey</td>
<td>Prospective cohort Sample Size: 84</td>
<td>Percentage female: 70.2% Population: Patients with Iron Deficiency Anemia of obscure origin aged between 16 and 80 years were evaluated for anti EmA IgA.</td>
<td>QUADAS Domain 1 Biased patient selection: Low QUADAS Domain 2 Bias due to testing: Low QUADAS Domain 3 Bias due to reference test: Unclear QUADAS Domain 4 Could patient flow have introduced bias: Unclear</td>
</tr>
<tr>
<td>Dahlbom et al., 2010&lt;sup&gt;35&lt;/sup&gt; Sweden</td>
<td>Case control Sample size: 301</td>
<td>Percentage female: Not reported Population: 52 children with severe malabsorption (group I) median age of 1.6 yrs, 59 children with mild symptoms (group II) median age of 8.1 yrs, 59 adults (group III) median age of 39.5 yrs and 131 disease controls (adult and children-median age of 10.8 yrs).</td>
<td>QUADAS Domain 1 Biased patient selection: High QUADAS Domain 2 Bias due to testing: Unclear QUADAS Domain 3 Bias due to reference test: Unclear QUADAS Domain 4 Could patient flow have introduced bias: Not Applicable</td>
</tr>
<tr>
<td>Dahle et al., 2010&lt;sup&gt;36&lt;/sup&gt; Sweden</td>
<td>Retrospective cohort Sample size: 176</td>
<td>Percentage female: 54.5% Population: Patients with symptoms who underwent endoscopy and biopsy without previous serological testing for anti-tTG or EmA</td>
<td>QUADAS Domain 1 Biased patient selection: Low QUADAS Domain 2 Bias due to testing: Unclear QUADAS Domain 3 Bias due to reference test: Low QUADAS Domain 4 Could patient flow have introduced bias: Unclear</td>
</tr>
<tr>
<td>Author, Year, Country</td>
<td>Study Design, Sample Size</td>
<td>Percent Female, Percent of Each Ethnicity, Population</td>
<td>QUADAS</td>
</tr>
<tr>
<td>-----------------------</td>
<td>--------------------------</td>
<td>------------------------------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>DeGaetani et al., 2013</td>
<td>Retrospective cohort</td>
<td>Percentage female: 51.4% Population: Adult patients (mean age 59yr) with villous atrophy on biopsy and negative celiac serologies.</td>
<td>QUADAS Domain 1: Biased patient selection: High</td>
</tr>
<tr>
<td>US</td>
<td>Sample size: 72</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dutta et al., 2010</td>
<td>Retrospective cohort</td>
<td>Percentage female: 40.2% Population: Symptomatic patients (age: 32.8 ± 17.4 years).</td>
<td>QUADAS Domain 1: Biased patient selection: Low</td>
</tr>
<tr>
<td>India</td>
<td>Sample size: 92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emami et al., 2012</td>
<td>Retrospective cohort</td>
<td>Percentage female: 67.7% Middle eastern 100% Population: Adult patients (mean age of 35.5 +/-13.7) with iron deficiency anemia</td>
<td>QUADAS Domain 1: Biased patient selection: Low</td>
</tr>
<tr>
<td>Iran</td>
<td>Sample size: 130</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Author, Year, Country</td>
<td>Study Design, Sample Size</td>
<td>Percent Female, Percent of Each Ethnicity, Population</td>
<td>QUADAS</td>
</tr>
<tr>
<td>-----------------------</td>
<td>---------------------------</td>
<td>------------------------------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>Harrison et al., 2013³⁷ UK</td>
<td>Retrospective cohort Sample size: 12289</td>
<td>Percentage female: Not reported Population: Patients tested for celiac disease using IgA tTG. 4 were IgA deficient (had error reading for IgA tTG, but elevated levels of IgG tTG)</td>
<td>QUADAS Domain 1 Biased patient selection: Low QUADAS Domain 2 Bias due to testing: Low QUADAS Domain 3 Bias due to reference test: Low QUADAS Domain 4 Could patient flow have introduced bias: Not Applicable</td>
</tr>
<tr>
<td>Kaukinen et al., 1999⁴¹ Finland</td>
<td>Case control Sample size: 26</td>
<td>Percentage female: 66.1% Population: Patients with endocrine disorders with median age of 37.7 years</td>
<td>QUADAS Domain 1 Biased patient selection: High QUADAS Domain 2 Bias due to testing: Low QUADAS Domain 3 Bias due to reference test: Low QUADAS Domain 4 Could patient flow have introduced bias: High</td>
</tr>
<tr>
<td>Mansour et al., 2011¹² Iraq</td>
<td>Prospective cohort Sample size: 62</td>
<td>Percentage female: 40.3% Population: Children and adults with diabetes 1. Mean age 23.4 years</td>
<td>QUADAS Domain 1 Biased patient selection: Low QUADAS Domain 2 Bias due to testing: Low QUADAS Domain 3 Bias due to reference test: Low QUADAS Domain 4 Could patient flow have introduced bias: Unclear</td>
</tr>
<tr>
<td>Author, Year, Country</td>
<td>Study Design, Sample Size</td>
<td>Percent Female, Percent of Each Ethnicity, Population</td>
<td>QUADAS</td>
</tr>
<tr>
<td>-----------------------</td>
<td>--------------------------</td>
<td>-------------------------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>Mozo et al., 2012&lt;sup&gt;33&lt;/sup&gt; Spain</td>
<td>Case control Sample size: 200</td>
<td>Percentage female: 59.5% Population: Children and adults diagnosed with CD or with various digestive pathologies. CD children mean age 2, control children mean age 2.8. CD adults mean age 39.1, and control adults mean age 43.0.</td>
<td>QUADAS Domain 1 Biased patient selection: High QUADAS Domain 2 Bias due to testing: Low QUADAS Domain 3 Bias due to reference test: Unclear QUADAS Domain 4 Could patient flow have introduced bias: Unclear</td>
</tr>
<tr>
<td>Nevoral et al., 2013&lt;sup&gt;34&lt;/sup&gt; Czech Republic</td>
<td>Retrospective cohort Sample size: 345</td>
<td>Percentage female: Not reported Population: Children and adolescents. 32 had first degree relatives with CD, 60 had Type 1 diabetes, and 187 were symptomatic</td>
<td>QUADAS Domain 1 Biased patient selection: Low QUADAS Domain 2 Bias due to testing: Low QUADAS Domain 3 Bias due to reference test: Low QUADAS Domain 4 Could patient flow have introduced bias: Not Applicable</td>
</tr>
<tr>
<td>Olen et al., 2012&lt;sup&gt;35&lt;/sup&gt; Sweden</td>
<td>Retrospective cohort Sample size: 537</td>
<td>Percentage female: 57.4% Population: Children and adolescents under 18 years of age. 71 children were under 2 years old. 16 were IgA deficient.</td>
<td>QUADAS Domain 1 Biased patient selection: High QUADAS Domain 2 Bias due to testing: Low QUADAS Domain 3 Bias due to reference test: Low QUADAS Domain 4 Could patient flow have introduced bias: Not Applicable</td>
</tr>
<tr>
<td>Author, Year, Country</td>
<td>Study Design, Sample Size</td>
<td>Percent Female, Percent of Each Ethnicity, Population</td>
<td>QUADAS</td>
</tr>
<tr>
<td>-----------------------</td>
<td>---------------------------</td>
<td>------------------------------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>Sakly et al., 2012&lt;sup&gt;Tunisia&lt;/sup&gt;</td>
<td>Case control, Sample size: 297</td>
<td>Percentage female: Not reported, Population: Adult and pediatric patients with CD as well as controls tested for IgA DGP</td>
<td>QUADAS Domain 1: Biased patient selection: High, QUADAS Domain 2: Bias due to testing: Unclear, QUADAS Domain 3: Bias due to reference test: Unclear, QUADAS Domain 4: Could patient flow have introduced bias: Unclear</td>
</tr>
<tr>
<td>Srinivas et al., 2014&lt;sup&gt;UK&lt;/sup&gt;</td>
<td>Retrospective cohort, Sample Size: 752</td>
<td>Percentage female: 66%, Population: CD vs. without CD in each of these groups: with clinical features of CD, serology profile, and duodenal macroscopic appearance</td>
<td>QUADAS Domain 1: Biased patient selection: Low, QUADAS Domain 2: Bias due to testing: Unclear, QUADAS Domain 3: Bias due to reference test: Unclear, QUADAS Domain 4: Could patient flow have introduced bias: High</td>
</tr>
<tr>
<td>Srinivas et al., 2013&lt;sup&gt;UK&lt;/sup&gt;</td>
<td>Retrospective cohort, Sample size: 752</td>
<td>Percentage female: 66%, Population: Symptomatic patients classified into 4 clinical groups: high risk of CD, low risk of CD, nutrient deficiencies and screening for diabetes.</td>
<td>QUADAS Domain 1: Biased patient selection: Low, QUADAS Domain 2: Bias due to testing: Low, QUADAS Domain 3: Bias due to reference test: Low, QUADAS Domain 4: Could patient flow have introduced bias: Not Applicable</td>
</tr>
<tr>
<td>Author, Year, Country</td>
<td>Study Design, Sample Size</td>
<td>Percent Female, Percent of Each Ethnicity, Population</td>
<td>QUADAS</td>
</tr>
<tr>
<td>-----------------------</td>
<td>---------------------------</td>
<td>------------------------------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>Sugai et al., 2010&lt;sup&gt;11&lt;/sup&gt;</td>
<td>Prospective cohort Sample size: 17</td>
<td>Percentage female: Not reported Population: IgA tTG negative adults with villous atrophy</td>
<td>QUADAS Domain 1 Biased patient selection: Low</td>
</tr>
<tr>
<td>Country Not Reported</td>
<td></td>
<td></td>
<td>QUADAS Domain 2 Bias due to testing: Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 3 Bias due to reference test: Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 4 Could patient flow have introduced bias: High</td>
</tr>
<tr>
<td>Swallow et al., 2013&lt;sup&gt;30&lt;/sup&gt;</td>
<td>Retrospective cohort Sample size: 756</td>
<td>Percentage female: Not reported Population: Adults</td>
<td>QUADAS Domain 1 Biased patient selection: High</td>
</tr>
<tr>
<td>UK</td>
<td></td>
<td></td>
<td>QUADAS Domain 2 Bias due to testing: Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 3 Bias due to reference test: Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 4 Could patient flow have introduced bias: Low</td>
</tr>
<tr>
<td>Van Meensel et al., 2004&lt;sup&gt;41&lt;/sup&gt;</td>
<td>Retrospective cohort Sample size: 175</td>
<td>Percentage female: 59.4% Population: Children and adult patients with biopsy-confirmed CD, with GFD, 5 were IgA deficient.</td>
<td>QUADAS Domain 1 Biased patient selection: High</td>
</tr>
<tr>
<td>Belgium</td>
<td></td>
<td></td>
<td>QUADAS Domain 2 Bias due to testing: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 3 Bias due to reference test: Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 4 Could patient flow have introduced bias: Unclear</td>
</tr>
<tr>
<td>Author, Year, Country</td>
<td>Study Design, Sample Size</td>
<td>Percent Female, Percent of Each Ethnicity, Population</td>
<td>QUADAS</td>
</tr>
<tr>
<td>-----------------------</td>
<td>---------------------------</td>
<td>-----------------------------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>Vermeersch et al., 2010&lt;sup&gt;53&lt;/sup&gt; Belgium</td>
<td>Case control Sample size: 827</td>
<td>Percentage female: 59.9% Population: 86 consecutive CD patients (recruited between August 1st 2000 and November 31st 2008) and 741 consecutive disease control patients (recruited between May 1st 2004 and October 31st 2007). The study population consisted of 599 adults and 228 children with a mean age of 29 years</td>
<td>QUADAS Domain 1 Biased patient selection: High QUADAS Domain 2 Bias due to testing: Unclear QUADAS Domain 3 Bias due to reference test: Low QUADAS Domain 4 Could patient flow have introduced bias: Unclear</td>
</tr>
<tr>
<td>Vermeersch et al., 2010&lt;sup&gt;53&lt;/sup&gt; Belgium</td>
<td>Case control Sample size: 588</td>
<td>Percentage female: 64.1% Population: 43 CD (mean age of 43.7) and 545 non-CD (mean age of 39.8)</td>
<td>QUADAS Domain 1 Biased patient selection: High QUADAS Domain 2 Bias due to testing: Unclear QUADAS Domain 3 Bias due to reference test: Low QUADAS Domain 4 Could patient flow have introduced bias: Unclear</td>
</tr>
<tr>
<td>Vermeersch et al., 2012&lt;sup&gt;54&lt;/sup&gt; Belgium</td>
<td>Case control Sample size: 649</td>
<td>Percentage female: 59.2% Population: Children and adults. Average age of celiac patients is 30.0 and that of controls is 26.4.</td>
<td>QUADAS Domain 1 Biased patient selection: High QUADAS Domain 2 Bias due to testing: Unclear QUADAS Domain 3 Bias due to reference test: Low QUADAS Domain 4 Could patient flow have introduced bias: Unclear</td>
</tr>
<tr>
<td>Author, Year, Country</td>
<td>Study Design, Sample Size</td>
<td>Percent Female, Percent of Each Ethnicity, Population</td>
<td>QUADAS</td>
</tr>
<tr>
<td>-----------------------</td>
<td>---------------------------</td>
<td>------------------------------------------------------</td>
<td>--------</td>
</tr>
</tbody>
</table>
| Wakim-Fleming et al., 2014<sup>15</sup> | Prospective cohort | Percentage female: 46.1%  
White: 82.8%  
Black: 13.3%  
Asian: 0.49%  
Latino: 2.9%  
American Indian: 0.49%  
Population:  
Patients with biopsy proven cirrhosis | QUADAS Domain 1  
Biased patient selection: Low  
QUADAS Domain 2  
Bias due to testing: Low  
QUADAS Domain 3  
Bias due to reference test: Low  
QUADAS Domain 4  
Could patient flow have introduced bias: Low |
| Walker, 2010<sup>36</sup> | Prospective cohort | Percentage female: 51.2%  
Population:  
Random sample of general population, adults | QUADAS Domain 1  
Biased patient selection: Low  
QUADAS Domain 2  
Bias due to testing: Low  
QUADAS Domain 3  
Bias due to reference test: Low  
QUADAS Domain 4  
Could patient flow have introduced bias: Low |
| Wolf et al., 2014<sup>37</sup> | Case control | Percentage female: 55%  
Population:  
Selective IgA deficiency (sIgAD) in 27 patients | QUADAS Domain 1  
Biased patient selection: High  
QUADAS Domain 2  
Bias due to testing: Low  
QUADAS Domain 3  
Bias due to reference test: Low  
QUADAS Domain 4  
Could patient flow have introduced bias: Low |
<table>
<thead>
<tr>
<th>Author, Year, Country</th>
<th>Study Design, Sample Size</th>
<th>Percent Female, Percent of Each Ethnicity, Population</th>
<th>QUADAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zanini et al., 2012</td>
<td>Retrospective cohort</td>
<td>Percentage female: 76%</td>
<td>QUADAS Domain 1</td>
</tr>
<tr>
<td>Italy</td>
<td>Sample size: 945</td>
<td>Population:</td>
<td>Biased patient selection: High</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Adult patients, aged 16 to 82 years, with symptoms,</td>
<td>QUADAS Domain 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>familiarity or presence of associated diseases. Mean age</td>
<td>Bias due to testing: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td>36.5</td>
<td>QUADAS Domain 3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Bias due to reference test: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Could patient flow have introduced bias: Not Applicable</td>
</tr>
</tbody>
</table>

Au/ml = absorbance units per milliliter; CD = celiac disease; CI = confidence interval; DGP = deamidated gliadin peptide (DGP); HLA = human leukocyte antigen; IgA = immunoglobulin A; IgG = immunoglobulin G; L = liter; NR = not reported; QUADAS = Quality Assessment of Diagnostic Studies; tTG = anti-tissue transglutaminase; U = units; U/mL = units per milliliter.
Key Points

**IgA tTG.** A 2010 meta-analysis pooled 12 studies and found a sensitivity of 93.0% (95% CI: 91.2%, 94.5%) and specificity of 96.5% (95% CI: 95.2%, 97.5%). A 2012 meta-analysis restricted to point of care tests in children reported sensitivity and specificity of 96.4% (95% CI: 94.3%, 97.9%) and 97.7% (95% CI: 95.8%, 99.0%) respectively when five studies were pooled. A 2013 systematic review did not pool data and echoed these results. We sixteen studies published after the prior SR search dates reported data needed for pooling; sensitivity was 92.6% (95% CI: 90.2%, 94.5%) and specificity 97.6% (95% CI: 96.3%, 98.5%). LR+ was 40.19% (95% CI: 25.29, 62.22) and LR- was 0.08 (95% CI: 0.06, 0.10). Significant heterogeneity was detected, as studies used a wide range of thresholds. Lower threshold levels increased sensitivity while higher threshold levels increased specificity. Our sensitivity analysis excluding data for threshold levels not used in clinical practice found sensitivity of 92.5% (95% CI: 89.7%, 94.6%) and specificity of 97.9% (95% CI: 96.5%, 98.7%). Positive predictive value (PPV) was 89.4% (95% CI: 88.3%, 90.5%) and negative predictive value (NPV) was 99.0% (95% CI: 98.8%, 99.1%).

**IgA EmA.** These tests have lower sensitivity than—and similar specificity to—IgA tTG tests, as confirmed by three SRs and several subsequent studies. LR+ was 65.98 (95% CI: 29.64, 126.33) and LR- was 0.21 (95% CI: 0.14, 0.30) in our pooled analysis of seven studies published after the SRs, after dropping data points where Marsh 1 and 2 level atrophy was classified as CD (not standard practice). Pooled sensitivity was 79.0% (95% CI: 71.0%, 86.0%); specificity was 99.0% (95% CI: 98.4%, 99.4%). PPV was 78.9% (95% CI: 71.0%, 85.5%) and NPV was 99.1% (95% CI: 98.6%, 99.5%). Significant heterogeneity was detected, likely due to variation in patient populations.

**IgA DGP.** A 2010 SR pooled eleven studies; sensitivity was estimated at 87.8% (95% CI: 85.6%, 89.9%) and specificity at 94.1% (95% CI: 95.2%, 97.5%) for all age groups combined. LR+ was 13.33 (95% CI: 9.64, 18.42) and LR- was 0.12 (95% CI: 0.08, 0.18). A 2012 SR reviewed the three 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%, 93.1%). One new study reported sensitivity of 97.0% and specificity of 90.7% in symptomatic adults and children at one clinic, while another reported both sensitivity and specificity of 96% in a similar population.

**IgG DGP.** A 2013 SR of seven studies of non-IgA-deficient adults reported sensitivities of 75.4% to 96.7% and specificities of 98.5% to 100%. A 2012 SR of three studies in non-IgA-deficient children reported sensitivities of 80.1% to 98.6% and specificities of 86.0% to 96.9%. Authors of these reviews did not pool data. A new study reported sensitivity of 95.0% and specificity of 99.0% in 200 participants of all ages. In children under age seven, both sensitivity and specificity were 100.0%.

**HLA-DQ2 and HLA-DQ8.** No SRs of the accuracy of testing for HLA-DQ2 or DQ8 were identified. These tests are used to rule out CD; the ACG estimates the negative predictive value of the combination at over 99%.

**Algorithms.** 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, more studies are needed to confirm the findings due to the wide range of values reported and populations studied.
**Video capsule endoscopy.** A previous SR of moderate quality pooled six studies and reported sensitivity of 89% (95% CI: 82.0%, 94.0%) and specificity of 95% (95% CI: 89.0%, 99.0%). LR+ was 12.90 (95% CI: 2.89, 57.58) and LR- was 0.16 (95% CI: 0.10, 0.25).

**Adherence to gluten free diet.** A previous SR of low quality indicated that 42% to 91% of adult CD patients in research studies strictly adhere to gluten-free diets; estimates varied primarily due to the definition and measurement of “strict.” The SR included three studies that reported no statistical difference in adherence levels between patients whose celiac disease was detected via screening and those whose celiac disease was “symptom- detected.” Association between specific diagnostic method and adherence was not addressed. We found only one study on this topic. Of ten blood donors in Israel who tested positive by tTG, only four adhered to gluten free diet; the other six patients did not believe they had CD, and four of those were told by physicians that asymptomatic patients did not need to modify their diets.

**Clinical outcomes, complications, or patient-centered outcomes.** We identified no studies of how these outcomes differ by initial diagnostic method.

**Detailed Synthesis**

**Key Question 1a. Accuracy of Diagnostic Methods and Algorithms**

**Anti-tissue transglutaminase (tTG) tests.** Four recent SRs assessed the accuracy of IgA tTG tests in symptomatic adults and children. They are discussed below; corresponding data are presented in Table 8. In 2010, Lewis published a meta-analysis that pooled 12 studies and found a sensitivity of 93.0% (95% CI: 91.2%, 94.5%) and specificity of 96.5% (95% CI: 95.2%, 97.5%). LR+ was 25.62 (95% CI: 15.64, 41.99) and LR- was 0.07 (95% CI: 0.05, 0.12). Significant heterogeneity was detected. The previous year, the NICE clinical guidelines reported that based on 19 studies, sensitivities ranged from 38% to 100% and specificities ranged from 25% to 100%; pooling was not performed.

In another systematic review, Giersiepen, 2012 compared the accuracy of IgA tTG tests by assay method in children. Fifteen studies using ELISA reported sensitivities from 73.9% to 100% and specificities from 77.89% to 100%. Data were not pooled due to heterogeneity. Three studies using receptor binding assay (RBA) reported sensitivities ranging from 89.0% to 100%; pooled specificity was 95.6% (95% CI: 91.3%, 98.5%). Pooled sensitivity and specificity were 96.4% (95% CI: 94.3%, 97.9%) and 97.7% (95% CI: 95.8%, 99.0%) respectively in five point-of-care tests. Finally, Collatz-Schyum and colleagues (2013) reported in their systematic review that sensitivities ranged from 76.0% to 96.8% and specificities from 90.9% to 98.0% in eight studies of adults.

Collatz-Schyum and NICE each reported accuracy results from two studies of IgG tTG tests. The populations were heterogeneous, with some including IgA deficient subjects. Evidence is insufficient to estimate accuracy of this test in the non-IgA-deficient population.

Three of the SRs were of moderate quality. The NICE review did not present a table of characteristics of the included studies, Giersiepen did not report whether dual selection or abstraction was used, and Lewis did not report an a priori design or protocol. Schyum's review was poor quality; the quality of the included studies was not assessed, and it was unclear whether studies were screened and abstracted in duplicate. A full assessment of the quality of each SR using the AMSTAR criteria is presented as Appendix E.
<table>
<thead>
<tr>
<th>Reference</th>
<th>Serologic Test</th>
<th># of Studies</th>
<th># Participants</th>
<th>Baseline Prevalence</th>
<th>Threshold for Positive</th>
<th>Method for Pooling</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Additional Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lewis, Scott, 2010 ⁹⁷</td>
<td>IgA tTG</td>
<td>13 (5 adult, 2 children, 6 child/adult; 1 study did not report information to calculate sensitivity)</td>
<td>NR</td>
<td>NR</td>
<td>The Spearman correlation coefficient (calculated between the log odds of sensitivity and 1-specificity) gave no indication of a threshold effect (0.34, p=0.28).</td>
<td>DerSimonian Laird method in a random effects model</td>
<td>93.0% (95% CI: 91.2-94.5)</td>
<td>96.5% (95% CI: 95.2-97.5)</td>
<td>LR⁺ = 25.62 (95% CI: 15.64, 41.99) LR⁻ = 0.07 (95% CI: 0.05, 0.12)</td>
</tr>
<tr>
<td>NICE Clinical Guidelines, 2009⁸</td>
<td>IgA tTG</td>
<td>19 (6 child, 9 adult, 4 child/adult)</td>
<td>4,799</td>
<td>NR</td>
<td>None</td>
<td>None</td>
<td>Range 38 to 100% (adults 71 to 100%) (children 89 to 100%)</td>
<td>Range 25 to 100% (adults 65 to 100%) (children 25 to 100%)</td>
<td>One study that compared children &lt;=2 years old vs. &gt;2 years old found IgA tTG and IgA EmA to be similarly accurate.</td>
</tr>
<tr>
<td>Giersiepen et al., 2012 ⁶⁰</td>
<td>IgA tTG (ELISA only)</td>
<td>15 child</td>
<td>1,694 CD; 1,138 non-CD</td>
<td>59.82%</td>
<td>cut-off given by manufacturer</td>
<td>None</td>
<td>73.9 to 100% (not pooled due to heterogeneity)</td>
<td>77.8 to 100% (not pooled due to heterogeneity)</td>
<td></td>
</tr>
<tr>
<td>Giersiepen et al., 2012 ⁶⁰</td>
<td>IgA tTG(RB As only)</td>
<td>3 child</td>
<td>255 CD; 146 non-CD</td>
<td>63.59%</td>
<td>cut-off given by manufacturer</td>
<td>MetaDiSc Software; weighted based on sample size</td>
<td>89.0 to 100% (not pooled due to heterogeneity)</td>
<td>95.9% (95% CI: 91.3-98.5)</td>
<td></td>
</tr>
<tr>
<td>Giersiepen et al., 2012 ⁶⁰</td>
<td>IgA TG2 POC (point of care)</td>
<td>5 child</td>
<td>470 CD; 399 non-CD</td>
<td>54.09%</td>
<td>cut-off given by manufacturer</td>
<td>MetaDiSc Software; weighted based on sample size</td>
<td>96.4% (95% CI: 94.3-97.9)</td>
<td>97.7% (95% CI: 95.8-99.0)</td>
<td></td>
</tr>
<tr>
<td>NICE Clinical Guidelines, 2009⁸</td>
<td>IgG tTG</td>
<td>2 (1 adult, 1 child/adult)</td>
<td>365</td>
<td>NR</td>
<td>None</td>
<td>None</td>
<td>23 to 85%</td>
<td>89 to 98%</td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>Serologic Test</td>
<td># of Studies</td>
<td># Participants</td>
<td>Baseline Prevalence</td>
<td>Threshold for Positive</td>
<td>Method for Pooling</td>
<td>Sensitivity</td>
<td>Specificity</td>
<td>Additional Information</td>
</tr>
<tr>
<td>-----------------</td>
<td>----------------</td>
<td>--------------</td>
<td>----------------</td>
<td>---------------------</td>
<td>------------------------</td>
<td>--------------------</td>
<td>-------------</td>
<td>-------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>Schyum, 2013</td>
<td>IgA tTG</td>
<td>8 adult</td>
<td>3,871</td>
<td>39.1% to 44.9% in cohorts</td>
<td>20 U/ml</td>
<td>None</td>
<td>76.0% to 96.8%</td>
<td>90.9% to 98.0%</td>
<td></td>
</tr>
<tr>
<td>Schyum, 2013</td>
<td>IgG tTG</td>
<td>2 adult</td>
<td>Unclear</td>
<td>NR</td>
<td>NR</td>
<td>None</td>
<td>41.4% to 84.2%</td>
<td>98.8%</td>
<td></td>
</tr>
</tbody>
</table>

tTG = anti-tissue transglutaminase; IgA = immunoglobulin A; ELISA = enzyme-linked immunosorbent assay; IgG = immunoglobulin G; NICE = National Institute for Clinical Excellence; NR = not reported; CD = celiac disease; CI = confidence interval RBA= radio blinding assays.
We identified 19 additional original studies reporting accuracy of tTG IgA tests. Study characteristics and risk of bias are displayed in Table 7 at the beginning of this chapter. Three included only IgA-deficient individuals43, 45, 52 and will be discussed in the results for Key Question 3 on populations of special interest. The pooled sensitivity and specificity results for the sixteen other studies47, 55, 57 31, 40 32, 39, 48, 50, 54, 58 35, 36, 53 42, 51 are displayed in Figure 3 and Table 9 below. (Data from the 16 new studies could not be pooled with the results of the prior published meta-analyses because the prior reviews did not report the number of true and false positives and negatives by arm for each study.) Pooled sensitivity was 92.6% (95% CI: 90.2%, 94.5%) and pooled specificity was 97.6% (95% CI: 96.3%, 98.5%). The LR+ was 40.19 (95% CI: 25.29, 62.22) and the LR- was 0.08 (95% CI: 0.06, 0.10). We did not calculate a summary ROC because the data reported high accuracy at a wide variety of threshold levels. The sensitivity and specificity obtained in our pooled analysis are not statistically different from those of the recent Lewis meta-analysis and confirm the high accuracy rate of tTG tests in non-IgA deficient individuals. The I-squared value in our analysis was 96.9%, indicating evidence of heterogeneity, likely due to the wide range of threshold levels used in the studies.

Figure 3. Sensitivity and specificity results for tissue transglutaminase immunoglobulin A tests

Table 9. Accuracy of tTG IgA tests

<table>
<thead>
<tr>
<th>tTG IgA</th>
<th>Threshold</th>
<th>True Positive</th>
<th>False Negative</th>
<th>False Positive</th>
<th>True Negative</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Van Meensel,</td>
<td>2.64 kilounits/L</td>
<td>101</td>
<td>4</td>
<td>1</td>
<td>69</td>
<td>0.96</td>
<td>0.99</td>
</tr>
<tr>
<td>200441</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dahlbom,</td>
<td>3 U m/L</td>
<td>170</td>
<td>0</td>
<td>1</td>
<td>130</td>
<td>1.00</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tTG IgA</td>
<td>Threshold</td>
<td>True Positive</td>
<td>False Negative</td>
<td>False Positive</td>
<td>True Negative</td>
<td>Sensitivity</td>
<td>Specificity</td>
</tr>
<tr>
<td>----------------------</td>
<td>---------------</td>
<td>---------------</td>
<td>----------------</td>
<td>----------------</td>
<td>---------------</td>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td>2010&lt;sup&gt;13&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Van Meensel, 2004&lt;sup&gt;21&lt;/sup&gt;</td>
<td>3.13 kilounits/L</td>
<td>101</td>
<td>4</td>
<td>1</td>
<td>69</td>
<td>0.96</td>
<td>0.99</td>
</tr>
<tr>
<td>Van Meensel, 2004&lt;sup&gt;21&lt;/sup&gt;</td>
<td>3.69 kilounits/L</td>
<td>101</td>
<td>4</td>
<td>0</td>
<td>70</td>
<td>0.96</td>
<td>1.00</td>
</tr>
<tr>
<td>Van Meensel, 2004&lt;sup&gt;21&lt;/sup&gt;</td>
<td>4 kilounits/L</td>
<td>98</td>
<td>7</td>
<td>1</td>
<td>69</td>
<td>0.93</td>
<td>0.99</td>
</tr>
<tr>
<td>Van Meensel, 2004&lt;sup&gt;21&lt;/sup&gt;</td>
<td>4.43 kilounits/L</td>
<td>104</td>
<td>1</td>
<td>1</td>
<td>69</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>Van Meensel, 2004&lt;sup&gt;21&lt;/sup&gt;</td>
<td>5 kilounits/L</td>
<td>98</td>
<td>7</td>
<td>1</td>
<td>69</td>
<td>0.93</td>
<td>0.99</td>
</tr>
<tr>
<td>Dahle, 2010&lt;sup&gt;16&lt;/sup&gt;</td>
<td>5 U/mL</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.76</td>
<td>0.95</td>
</tr>
<tr>
<td>Van Meensel, 2004&lt;sup&gt;21&lt;/sup&gt;</td>
<td>7 kilounits/L</td>
<td>96</td>
<td>9</td>
<td>0</td>
<td>70</td>
<td>0.91</td>
<td>1.00</td>
</tr>
<tr>
<td>Van Meensel, 2004&lt;sup&gt;21&lt;/sup&gt;</td>
<td>7 kilounits/L</td>
<td>102</td>
<td>3</td>
<td>0</td>
<td>70</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>Vermeersch, 2010&lt;sup&gt;13&lt;/sup&gt;</td>
<td>7 U/mL</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.95</td>
<td>0.93</td>
</tr>
<tr>
<td>Vermeersch, 2012&lt;sup&gt;24&lt;/sup&gt;</td>
<td>7 U/mL</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.81</td>
<td>0.99</td>
</tr>
<tr>
<td>Vermeersch, 2012&lt;sup&gt;24&lt;/sup&gt;</td>
<td>7 U/mL</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.95</td>
<td>0.76</td>
</tr>
<tr>
<td>Zanini, 2012&lt;sup&gt;28&lt;/sup&gt;</td>
<td>7 U/mL</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.95</td>
<td>0.76</td>
</tr>
<tr>
<td>Van Meensel, 2004&lt;sup&gt;21&lt;/sup&gt;</td>
<td>7.16 kilounits/L</td>
<td>102</td>
<td>3</td>
<td>0</td>
<td>70</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>Van Meensel, 2004&lt;sup&gt;21&lt;/sup&gt;</td>
<td>7.98 kilounits</td>
<td>101</td>
<td>4</td>
<td>0</td>
<td>70</td>
<td>0.96</td>
<td>1.00</td>
</tr>
<tr>
<td>Zanini, 2012&lt;sup&gt;23&lt;/sup&gt;</td>
<td>8 U/mL</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.88</td>
<td>0.92</td>
</tr>
<tr>
<td>Van Meensel, 2004&lt;sup&gt;21&lt;/sup&gt;</td>
<td>9.73 kilounits/L</td>
<td>99</td>
<td>6</td>
<td>0</td>
<td>70</td>
<td>0.94</td>
<td>1.00</td>
</tr>
<tr>
<td>Emami, 2012&lt;sup&gt;19&lt;/sup&gt;</td>
<td>10 AU/mL</td>
<td>5</td>
<td>8</td>
<td>4</td>
<td>113</td>
<td>0.38</td>
<td>0.97</td>
</tr>
<tr>
<td>Srinivas, 2013&lt;sup&gt;48&lt;/sup&gt;</td>
<td>10 IU/mL</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.84</td>
<td>0.96</td>
</tr>
<tr>
<td>Wolf, 2014&lt;sup&gt;37&lt;/sup&gt;</td>
<td>10 ULN</td>
<td>310</td>
<td>10 (32 grey zone)</td>
<td>2 (17 grey zone)</td>
<td>673</td>
<td>0.88</td>
<td>0.97</td>
</tr>
<tr>
<td>Srinivas, 2014&lt;sup&gt;44&lt;/sup&gt;</td>
<td>10 IU/mL</td>
<td>73</td>
<td>15</td>
<td>29</td>
<td>635</td>
<td>0.83</td>
<td>0.96</td>
</tr>
<tr>
<td>Van Meensel, 2004&lt;sup&gt;21&lt;/sup&gt;</td>
<td>15 kilounits</td>
<td>99</td>
<td>6</td>
<td>0</td>
<td>70</td>
<td>0.94</td>
<td>1.00</td>
</tr>
<tr>
<td>Vermeersch, 2010&lt;sup&gt;13&lt;/sup&gt;</td>
<td>15 U/mL</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.86</td>
<td>0.95</td>
</tr>
<tr>
<td>Mansour, 2011&lt;sup&gt;22&lt;/sup&gt;</td>
<td>15 U/mL</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>51</td>
<td>0.71</td>
<td>0.93</td>
</tr>
<tr>
<td>Zanini, 2012&lt;sup&gt;28&lt;/sup&gt;</td>
<td>16 U/mL</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.89</td>
<td>0.88</td>
</tr>
<tr>
<td>Basso, 2011&lt;sup&gt;22&lt;/sup&gt;</td>
<td>17.5 U/mL</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.95</td>
<td>0.97</td>
</tr>
<tr>
<td>Van Meensel, 2004&lt;sup&gt;21&lt;/sup&gt;</td>
<td>19.05 kilounits/L</td>
<td>98</td>
<td>7</td>
<td>0</td>
<td>70</td>
<td>0.93</td>
<td>1.00</td>
</tr>
<tr>
<td>Vermeersch, 2012&lt;sup&gt;24&lt;/sup&gt;</td>
<td>20 U/mL</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.84</td>
<td>0.96</td>
</tr>
<tr>
<td>Basso, 2011&lt;sup&gt;22&lt;/sup&gt;</td>
<td>20 U/mL</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.94</td>
<td>0.97</td>
</tr>
<tr>
<td>Srinivas, 2013&lt;sup&gt;48&lt;/sup&gt;</td>
<td>20 IU/mL</td>
<td>5</td>
<td>0</td>
<td>7</td>
<td>192</td>
<td>1.00</td>
<td>0.96</td>
</tr>
</tbody>
</table>
Different laboratories or test manufacturers may define “positive” celiac disease diagnosis using different threshold levels of IgA tTG. Due to the wide range of thresholds used in the research studies and the heterogeneity identified in our meta-analysis, we conducted a sensitivity analysis by omitting accuracy results for thresholds not used in clinical practice. This pooled analysis of tests that used threshold levels of less than 40 U/mL resulted in sensitivity of 92.5% (95% CI: 89.7%, 94.6%) and specificity of 97.9% (95% CI: 96.5%, 98.7%). This result does not differ significantly from the results when all thresholds were pooled. Positive predictive value (PPV) was 89.4% (95% CI: 88.3%, 90.5%) and negative predictive value was 99.0% (95% CI: 98.8%, 99.1%).

Figure 4 displays sensitivity and specificity by threshold or “cut-off” where reported. Data were insufficient to pool by threshold level, but diagnostic thresholds of 31 to 40 U/ml reported low sensitivity, and thresholds of 40 U/ml or higher reported very low sensitivity. The higher thresholds are not currently recommended; they were reported by studies that aimed to define a cut-off value of tTG antibody with 100% specificity or high positive likelihood ratio for duodenal atrophy in patients with suspected celiac disease. One study concluded that a cutoff

<table>
<thead>
<tr>
<th>tTG IgA</th>
<th>Threshold</th>
<th>True Positive</th>
<th>False Negative</th>
<th>False Positive</th>
<th>True Negative</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wolf, 2014</td>
<td>20 U/mL</td>
<td>342</td>
<td>10</td>
<td>19</td>
<td>673</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>Van Meensel, 2004</td>
<td>20.47 kilounits</td>
<td>102</td>
<td>3</td>
<td>0</td>
<td>70</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>Zanini, 2012</td>
<td>21 U/mL</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.38</td>
<td>0.97</td>
</tr>
<tr>
<td>Basso, 2011</td>
<td>24 U/mL</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.96</td>
<td>0.81</td>
</tr>
<tr>
<td>Zanini, 2012</td>
<td>24 U/mL</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.59</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>35 U/mL</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.10</td>
<td>1.00</td>
</tr>
<tr>
<td>Van Meensel, 2004</td>
<td>40 kilounits/L</td>
<td>101</td>
<td>4</td>
<td>3</td>
<td>67</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>Zanini, 2012</td>
<td>40 U/mL</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.43</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>48 U/mL</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.70</td>
<td>0.59</td>
</tr>
<tr>
<td>Van Meensel, 2004</td>
<td>50 kilounits/L</td>
<td>98</td>
<td>7</td>
<td>5</td>
<td>65</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>56.9 kilounits/L</td>
<td>96</td>
<td>9</td>
<td>1</td>
<td>69</td>
<td>0.91</td>
<td>0.99</td>
</tr>
<tr>
<td>Basso, 2011</td>
<td>75.6 U/mL</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.91</td>
<td>0.97</td>
</tr>
<tr>
<td>Zanini, 2012</td>
<td>80 U/mL</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.59</td>
<td>0.43</td>
</tr>
<tr>
<td>Basso, 2011</td>
<td>100 U/mL</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.76</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>909.3 U/mL</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.63</td>
<td>1.00</td>
</tr>
<tr>
<td>Barada, 2014</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.72</td>
<td>0.98</td>
</tr>
<tr>
<td>Harrison, 2013</td>
<td>NR</td>
<td>66</td>
<td>10</td>
<td>11</td>
<td>12,202</td>
<td>0.87</td>
<td>1.00</td>
</tr>
<tr>
<td>Swallow, 2013</td>
<td>NR</td>
<td>5</td>
<td>2</td>
<td>72</td>
<td>654</td>
<td>0.75</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>4</td>
<td>72</td>
<td>654</td>
<td>0.87</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>3</td>
<td>77</td>
<td>656</td>
<td>0.88</td>
<td>0.90</td>
<td></td>
</tr>
</tbody>
</table>

Pooled Results*  Sensitivity: 0.93 (0.90, 0.95)  Specificity: 0.98 (0.96, 0.99)
* pooled results include only studies with all information (true and false negatives and positives) reported

tTG = anti-tissue transglutaminase; IgA = immunoglobulin A; U = unit; mL = milliliter; grey zone = indeterminate results; NR = not reported.

Table 4: Sensitivity and specificity by threshold or “cut-off” where reported. Data were insufficient to pool by threshold level, but diagnostic thresholds of 31 to 40 U/ml reported low sensitivity, and thresholds of 40 U/ml or higher reported very low sensitivity. The higher thresholds are not currently recommended; they were reported by studies that aimed to define a cut-off value of tTG antibody with 100% specificity or high positive likelihood ratio for duodenal atrophy in patients with suspected celiac disease. One study concluded that a cutoff
level five-fold higher than the upper limit of normal is 100% specific for duodenal atrophy and using this cut-off could prevent the need for biopsy in one-third of patients.\textsuperscript{58}

**Figure 4. Accuracy by threshold level for tissue transglutaminase immunoglobulin A**

---

**Endomyosial antibodies (EmA) tests.** Three recent SRs of IgA EmA tests were identified. A review conducted for the 2009 NICE clinical guidelines\textsuperscript{8} included 23 studies: 10 of children, nine of adults, and four of both adults and children. Sensitivity ranged from 68% to 100%, while specificity ranged from 77% to 100%. Pooling was not performed. Giersiepen\textsuperscript{60} included 11 studies of IgA EmA accuracy in children. Sensitivity ranged from 82.6% to 100%. Pooled specificity was 98.2\% (95\% CI: 96.7\%, 99.1\%). Finally, Schyum, 2013\textsuperscript{61} described four studies of adults that were included in the previous reviews; pooling was not conducted. As mentioned above, quality of the NICE and Giersiepen SRs was rated as moderate. Data are presented in Table 10.
**Table 10. Systematic reviews of EmA IgA tests**

<table>
<thead>
<tr>
<th>Reference</th>
<th>Serologic Test</th>
<th># of Studies</th>
<th># Participants</th>
<th>Baseline Prevalence</th>
<th>Threshold for Positive</th>
<th>Method for Pooling</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Additional Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Giersiepen et al., 2012</td>
<td>IgA EmA</td>
<td>11 child</td>
<td>1,034 CD; 558 non-CD</td>
<td>64.95%</td>
<td>cut-off given by manufacturer</td>
<td>Weighted based on sample size</td>
<td>82.6 to 100% (not pooled due to heterogeneity)</td>
<td>98.2% (95% CI: 96.7-99.1)</td>
<td></td>
</tr>
<tr>
<td>NICE Clinical Guidelines, 2009</td>
<td>IgA EmA</td>
<td>23 (10 child, 9 adult, 4 child/adult)</td>
<td>5,529</td>
<td>NR</td>
<td>NR</td>
<td>None</td>
<td>Range 68 to 100% (adults 68 to 100%) (children 46 to 100%)</td>
<td>Range 89 to 100% (adults 94 to 100%) (children 77 to 100%)</td>
<td>Reported results on 1 study that compared children &lt;=2 years old vs. &gt;2 years old and found IgA tTG and IgA EMA to have similar accuracy.</td>
</tr>
<tr>
<td>Schyum, 2013</td>
<td>IgA EmA</td>
<td>4 adult</td>
<td>2,537</td>
<td>Unclear</td>
<td>NR</td>
<td>None</td>
<td>61.0% to 93.7%</td>
<td>98.0% to 100%</td>
<td></td>
</tr>
</tbody>
</table>

EmA = endomysial antibodies; IgA = immunoglobulin A; CD = celiac disease; CI = confidence interval NICE = National Institute for Clinical Excellence; NR = not reported.
Seven additional studies that reported sensitivity and specificity of IgA EmA tests\textsuperscript{31, 34, 36, 42, 47, 50, 55} were identified. Accuracy results are displayed in Figure 5 and Table 11. Swallow (2013)\textsuperscript{50} reported results for three different biopsy reference standards (definitions of celiac disease): a) Marsh 1-2 villous atrophy, b) Marsh 1-3 villous atrophy, and c) Marsh 3 villous atrophy. Sensitivities were 42.9%, 73.3%, and 82.6%, respectively.

Two studies\textsuperscript{31, 36} reported inadequate data for inclusion in pooling. Our pooling of the seven available data points resulted in sensitivity of 76.6% (95% CI: 68.7%, 82.9%) and specificity of 99.0% (95% CI: 98.4%, 99.4%). After excluding the two data points where, contrary to standard practice, Marsh 1 and 2 level villous atrophy were classified as CD, pooled sensitivity was 79.0% (95% CI: 71.0%, 86.0%) and pooled specificity was 99.0% (95% CI: 98.4%, 99.4%). LR+ was 65.98 (95% CI: 29.64, 126.33) and LR- was 0.21 (95% CI: 0.14, 0.30). Positive predictive value (PPV) was 78.9% (95% CI: 71.0%, 85.5%) and negative predictive value was 99.1% (95% CI: 98.6%, 99.5%). I-squared value was 81.8%, indicating evidence of heterogeneity. These accuracy results support findings of prior systematic reviews.

**Figure 5. Accuracy of endomysial antibodies immunoglobulin A studies published after NICE and ESPGHAN systematic reviews**

![Figure 5](image)

**Table 11. Accuracy of EmA IgA tests in studies published after NICE and ESPGHAN systematic reviews**

<table>
<thead>
<tr>
<th>Threshold</th>
<th>True Positive</th>
<th>False Negative</th>
<th>False Positive</th>
<th>True Negative</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mansour, 2011\textsuperscript{42}</td>
<td>20 U/mL</td>
<td>5</td>
<td>2</td>
<td>53</td>
<td>0.71</td>
<td>0.96</td>
</tr>
<tr>
<td>Study, Year</td>
<td>Threshold</td>
<td>True Positive</td>
<td>False Negative</td>
<td>False Positive</td>
<td>True Negative</td>
<td>Sensitivity</td>
</tr>
<tr>
<td>------------</td>
<td>-----------</td>
<td>---------------</td>
<td>----------------</td>
<td>----------------</td>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Barada, 2014</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.72</td>
<td>1.00</td>
</tr>
<tr>
<td>Swallow, 2013</td>
<td>NR, Marsh 1-3 classified as celiac</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>722</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>NR, Marsh 2 &amp; 3 classified as celiac</td>
<td>22</td>
<td>8</td>
<td>4</td>
<td>722</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>NR, March 3 classified as celiac</td>
<td>19</td>
<td>4</td>
<td>7</td>
<td>726</td>
<td>0.83</td>
</tr>
<tr>
<td>Wakim-Fleming, 2014</td>
<td>Serum dilution ≥ 1/10</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>199</td>
<td>1.00</td>
</tr>
<tr>
<td>Srinivas, 2014</td>
<td>NR</td>
<td>70</td>
<td>18</td>
<td>5</td>
<td>659</td>
<td>0.80</td>
</tr>
<tr>
<td>Cekin, 2012</td>
<td>NR</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>77</td>
<td>1.00</td>
</tr>
<tr>
<td>Dahle, 2010</td>
<td>Serum dilution 1/5</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.61</td>
</tr>
</tbody>
</table>

Pooled Results* Sensitivity: 0.77 (0.69,0.83) Specificity: 0.99 (0.98,0.99)

*Pooled results include only studies with all information (false and true negatives and positives) reported.

EmA = endomyosal antibodies; IgA = immunoglobulin A; NICE = National Institute for Clinical Excellence; ESPGHAN = European Society for Paediatric Gastroenterology Hepatology and Nutrition; U = unit; mL = milliliter; NR = not reported.

**Deamidated gliadin peptide (DGP) antibodies test –IgA.** Four recent SRs of DGP tests were identified; details are presented in Table 12. The 2009 NICE guideline report identified only two studies of IgA DGP. Both studies were included in a 2010 review by Lewis and colleagues of 12 studies; we were able to pool eleven studies and sensitivity was estimated at 87.8% (95% CI: 85.6%, 89.9%) while pooled specificity was 94.1% (95% CI: 92.5%, 95.5%). LR+ was 13.33 (95% CI: 9.64, 18.42) and LR− was 0.12 (95% CI: 0.08, 0.18). Significant heterogeneity was detected. Giersiepen reviewed three studies of IgA DGP accuracy in children that were included in the Lewis SR. Sensitivity ranged from 80.7% to 95.1% (not pooled) and specificity was estimated at 90.7% (95% CI: 87.8%, 93.1%). Schyum identified seven studies in adults; sensitivity ranged from 69.0% to 98.4% while specificity ranged from 90.3% to 98.0%.

We identified three studies not included in prior SRs that reported sensitivity and specificity of IgA DGP tests in non-IgA-deficient individuals. Sugai assessed accuracy in 17 IgA tTG-negative patients with villous atrophy; results will be discussed under Key Question 3, which considers populations of special interest (previously seronegative subjects). Sakly reported sensitivity of 97.0% and specificity of 90.7% in 297 symptomatic adults and children at one clinic. Mozo (2012) conducted a case-control study in Spain that included 100 newly diagnosed adults and children and 100 age-matched controls. (Six patients were IgA deficient.) Both sensitivity and specificity of IgA DGP were 96.0% and AUC was 98.8%.

**Deamidated gliadin peptide (DGP) antibodies test –IgG.** Two recent SRs reported accuracy of IgG DGP tests in non-IgA deficient subjects. Details are displayed in Table 12. Giersiepen, 2012 reported sensitivity ranging from 80.1% to 98.6% and specificity from 86.0% to 96.9% in three studies of children. In a SR of seven studies of adults compiled by Schyum, sensitivity ranged from 75.4% to 96.7%, while specificity ranged from 98.5% to 100%.

We identified one study of IgG DGP tests published after these systematic reviews. In the case control study noted above, Mozo (2012) reported sensitivity and specificity of IgG DGP test.
were 95.0% and 99.0% respectively, with AUC of 99.5%. In children age seven or younger, both sensitivity and specificity of IgG DGP were 100.0%.
## Table 12. Systematic reviews of DGP tests

<table>
<thead>
<tr>
<th>Reference</th>
<th>Serologic Test</th>
<th># of Studies</th>
<th># Participants</th>
<th>Baseline Prevalence</th>
<th>Threshold for Positive</th>
<th>Method for Pooling</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Additional Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>NICE Clinical Guidelines, 2009</td>
<td>DGP-based assays</td>
<td>2 (1 children, 1 adult)</td>
<td>317</td>
<td>NR</td>
<td>NR</td>
<td>none</td>
<td>Range 90.8 to 100% (adults 96.7 to 100%) (children 90.8 to 100%)</td>
<td>Range 93.8 to 100% (adults 93.8 to 100%) (children 94.7 to 98.2%)</td>
<td>Studies described narratively</td>
</tr>
<tr>
<td>Giersiepen et al., 2012</td>
<td>IgA DGP</td>
<td>3 children</td>
<td>422 CD; 346 non-CD</td>
<td>54.95%</td>
<td>cut-off given by manufacturer</td>
<td>MetaDiSc Software; weighted based on sample size</td>
<td>80.7 to 95.1% (not pooled due to heterogeneity)</td>
<td>90.7% (95% CI: 87.8-93.1)</td>
<td></td>
</tr>
<tr>
<td>Lewis, Scott 2010</td>
<td>IgA DGP</td>
<td>12 (1 study did not report information to calculate sensitivity; 4 adult, 2 children, 6 child/adult)</td>
<td>NR</td>
<td>NR</td>
<td>The Spearman correlation coefficient (calculated between the log odds of sensitivity and 1-specificity) gave no indication of a threshold effect (-0.23, p=0.50).</td>
<td>MetaDiSc and STATA; DerSimonian Laird method in a random effects model</td>
<td>87.8% (95% CI: 85.6-89.9)</td>
<td>94.1% (95% CI: 92.5-95.5)</td>
<td>LR+ = 13.33 (95% CI: 9.64, 18.42) LR- = 0.12 (% CI: 0.08, 0.18)</td>
</tr>
<tr>
<td>Schyum, 2013</td>
<td>IgA DGP</td>
<td>7 adult</td>
<td>2,555</td>
<td>Unclear</td>
<td>NR</td>
<td>None</td>
<td>69.0% to 98.4%</td>
<td>90.3% to 98.0%</td>
<td></td>
</tr>
<tr>
<td>Schyum, 2013</td>
<td>IgG DGP</td>
<td>7 adult</td>
<td>2,322</td>
<td>Unclear</td>
<td>NR</td>
<td>None</td>
<td>75.4% to 96.7%</td>
<td>98.5% to 100%</td>
<td></td>
</tr>
<tr>
<td>Giersiepen, 2012</td>
<td>IgG DGP</td>
<td>3 children</td>
<td>422 CD, 346 non-CD</td>
<td>54.95%</td>
<td>NR</td>
<td>None</td>
<td>80.1% to 98.6%</td>
<td>86.0% to 96.9%</td>
<td></td>
</tr>
</tbody>
</table>

IgA = immunoglobulin A; DGP = deamidated gliadin peptide; NICE = National Institute for Clinical Excellence; NR = not reported; CD = celiac disease; CI: confidence interval.
**Human leukocyte antigen (HLA) DQ2 and DQ8.** These tests are used to rule out CD, as the ACG\(^5\) estimates the negative predictive value of the combination at over 99%. It is estimated that 95% of CD patients are positive for HLA-DQ2, and the remainder are positive for HLA-DQ8. Patients who test negative for HLA-DQ2 and HLA-DQ8 simultaneously will rarely undergo further testing for CD. One SR\(^{61}\) discussed studies of HLA typing but did not report sensitivity or specificity.

**Algorithms.** We identified nine studies that used multiple serological tests simultaneously in diagnostic algorithms. Each algorithm combined a tTG screen with an additional serological test. Accuracy results are displayed in Table 13 below. Data could not be pooled due to study heterogeneity. Algorithms that combined a tTG test and EmA test had sensitivity ranging from 57% to 93% and specificity from 64% to 99%. One study\(^{50}\) reported low sensitivity when the definition of celiac disease included patients with Marsh 1-2 level atrophy. In the same study, when only patients with atrophy level Marsh 3 or higher were considered to have celiac disease, sensitivity was 87% and specificity was 97%.

Algorithms combining tTG and DGP screens reported sensitivity ranging from 65% to 97% and specificity from 80% to 100%. Low sensitivity (65%) was reported when a threshold of 145 U/ml was used for tTG test. However, this is not common clinical practice: The authors used this high threshold level to achieve specificity of 100%. A high threshold was also used by Sugai\(^{49}\) for the same purpose. Of note, two new combined tTG IgA + DGP IgG tests reported high sensitivity and specificity.\(^{54}\)

While this report underwent peer review, we identified a SR that included three studies of combination assays.\(^{61}\) Two were already included in this report; the third did not meet our inclusion criteria.

In general, the sensitivity and specificity of algorithms were not significantly higher than that of the individual tests used alone, and it is unclear whether the small increases in accuracy are clinically meaningful.
Table 13. Accuracy of algorithms

<table>
<thead>
<tr>
<th>Test Type</th>
<th>Notes</th>
<th>True Positive</th>
<th>False Negative</th>
<th>False Positive</th>
<th>True Negative</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barada, 2014&lt;sup&gt;14&lt;/sup&gt;</td>
<td>tTG IgA, DGP IgA</td>
<td>Symptomatic adults in Lebanon</td>
<td>13</td>
<td>5</td>
<td>25</td>
<td>955</td>
<td>0.722</td>
</tr>
<tr>
<td></td>
<td>tTG IgA</td>
<td>Symptomatic adults in Lebanon</td>
<td>13</td>
<td>5</td>
<td>16</td>
<td>965</td>
<td>0.722</td>
</tr>
<tr>
<td></td>
<td>EmA IgA</td>
<td>Symptomatic adults in Lebanon</td>
<td>13</td>
<td>5</td>
<td>3</td>
<td>978</td>
<td>0.722</td>
</tr>
<tr>
<td>Basso, 2011&lt;sup&gt;15&lt;/sup&gt;</td>
<td>tTG IgA, DGP IgA</td>
<td>Children, threshold 145 U/ml</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.653</td>
</tr>
<tr>
<td></td>
<td>tTG IgA</td>
<td>Children, threshold 100 U/ml</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.757</td>
</tr>
<tr>
<td></td>
<td>tTG IgA, DGP IgA</td>
<td>Children, threshold 20 U/ml</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.967</td>
</tr>
<tr>
<td></td>
<td>tTG IgA</td>
<td>Children, threshold 20 U/ml</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.942</td>
</tr>
<tr>
<td></td>
<td>tTG IgA, DGP IgA</td>
<td>Children, Threshold 32 U/ml</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.945</td>
</tr>
<tr>
<td></td>
<td>tTG IgA</td>
<td>Children, Threshold 17.5 U/ml</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.945</td>
</tr>
<tr>
<td>Dahle, 2010&lt;sup&gt;16&lt;/sup&gt;</td>
<td>tTG IgG or IgA combined w/ DGP IgG or IgA</td>
<td>Symptomatic adults, Threshold 20 Au/mL</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>tTG IgG or IgA combined w/ DGP IgG or IgA</td>
<td>Symptomatic adults, Threshold 35 Au/mL</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>tTG IgA</td>
<td>Symptomatic adults, Threshold 5 U/mL</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>DGP IgG or IgA</td>
<td>Symptomatic adults, Threshold 20 Au/mL</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>EmA IgA</td>
<td>Symptomatic adults</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.61</td>
</tr>
<tr>
<td>Nevoral, 2013&lt;sup&gt;17&lt;/sup&gt;</td>
<td>tTG IgA, EmA IgA</td>
<td>Marsh 2 or 3</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>tTG IgA, EmA IgA</td>
<td>First degree relatives</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>tTG IgA, EmA IgA</td>
<td>Asymptomatic Marsh 2 or 3</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>tTG IgA, EmA IgA</td>
<td>Type 1 diabetes</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.93</td>
</tr>
<tr>
<td>Srinivas, 2013&lt;sup&gt;18&lt;/sup&gt;</td>
<td>tTG IgA, EmA IgA</td>
<td>Symptomatic or Type 1 Diabetes</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>tTG IgA</td>
<td>Symptomatic or Type 1 Diabetes</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>EmA IgA</td>
<td>Symptomatic or Type 1 Diabetes</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.83</td>
</tr>
</tbody>
</table>

42
<table>
<thead>
<tr>
<th>Test Type</th>
<th>Notes</th>
<th>True Positive</th>
<th>False Negative</th>
<th>False Positive</th>
<th>True Negative</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sugai, 2010&lt;sup&gt;39&lt;/sup&gt;</td>
<td>tTG IgA, DGP IgA or IgG</td>
<td>tTG positive and negative patients with enteropathy</td>
<td>6</td>
<td>8</td>
<td>0</td>
<td>3</td>
<td>0.43</td>
</tr>
<tr>
<td>Swallow, 2013&lt;sup&gt;30&lt;/sup&gt;</td>
<td>tTG IgA, EmA IgA - NICE two step strategy</td>
<td>Adults, Marsh 1-2 considered celiac</td>
<td>4</td>
<td>3</td>
<td>20</td>
<td>706</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>tTG IgA, EmA IgA - NICE two step strategy</td>
<td>Adults, Marsh 1-3 considered celiac</td>
<td>24</td>
<td>6</td>
<td>20</td>
<td>706</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>tTG IgA, EmA IgA - NICE two step strategy</td>
<td>Adults, Marsh 3 considered celiac</td>
<td>20</td>
<td>3</td>
<td>23</td>
<td>710</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>tTG IgA, 2008-2009 data</td>
<td>Adults, Marsh 3 considered celiac</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.875</td>
</tr>
<tr>
<td></td>
<td>EmA IgA, 2008-2009 data</td>
<td>Adults, Marsh 3 considered celiac</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.826</td>
</tr>
<tr>
<td>Vermeersch, 2012&lt;sup&gt;24&lt;/sup&gt;</td>
<td>tTG IgA, DGP IgG combined screen</td>
<td>Adults, Brand 1</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.897</td>
</tr>
<tr>
<td></td>
<td>tTG IgA, DGP IgG combined screen</td>
<td>Adults, Brand 2</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.888</td>
</tr>
<tr>
<td></td>
<td>tTG IgA</td>
<td>Adults, Brand 1</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.841</td>
</tr>
<tr>
<td></td>
<td>tTG IgA</td>
<td>Adults, Brand 2</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.813</td>
</tr>
<tr>
<td></td>
<td>DGP IgG</td>
<td>Adults, Brand 1</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.850</td>
</tr>
<tr>
<td></td>
<td>DGP IgG</td>
<td>Adults, Brand 2</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>0.869</td>
</tr>
<tr>
<td>Wolf, 2014&lt;sup&gt;57&lt;/sup&gt;</td>
<td>tTG IgA, DGP IgG</td>
<td>Children without IgA deficiency</td>
<td>314</td>
<td>8 (30 grey zone)</td>
<td>2 (31 grey zone)</td>
<td>659</td>
<td>0.892</td>
</tr>
<tr>
<td></td>
<td>tTG IgA, DGP IgG</td>
<td>Children with IgA deficiency</td>
<td>7</td>
<td>6 (11 grey zone)</td>
<td>0</td>
<td>3</td>
<td>0.292</td>
</tr>
</tbody>
</table>

tTG = anti-tissue transglutaminase; IgA = immunoglobulin A; DGP = deamidated gliadin peptide antibodies; IgG = immunoglobulin G; EmA = endomysial antibodies; NICE = National Institute for Clinical Excellence; CI = confidence interval; grey zone = indeterminate results; NR = not reported.
**Video capsule endoscopy.** We identified two systematic reviews (SR) on the accuracy of video capsule endoscopy (VCE). A 2012 SR by Rokkas and colleagues included six studies, three of which represented all studies included in the prior SR by El-Matary on the topic. The Rokkas SR pooled five studies with a total of 166 subjects; sensitivity of 89% (95% CI: 82.0%, 94.0%) and specificity of 95% (95% CI: 89.0%, 99.0%) were reported. LR+ was 12.90 (95% CI: 2.89, 57.58), and LR- was 0.16 (95% CI: 0.10, 0.25).

Data are presented in Table 14 below. The Area under the Curve (AUC) of the weighted symmetric summary ROC curve was 0.95. The quality of this SR was rated moderate, as the publication made no mention of assessment of the quality of the included studies. In addition, patient ages and symptomatology were not described. No additional studies of the accuracy of VCE that met our inclusion criteria were identified.

<table>
<thead>
<tr>
<th>Reference</th>
<th># of Studies</th>
<th># of Participants</th>
<th>Baseline Prevalence</th>
<th>Method for Pooling</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>El-Matary, 2009&lt;sup&gt;64&lt;/sup&gt;</td>
<td>3</td>
<td>107</td>
<td>58.88%</td>
<td>Simple pooling method with CIs computed by using the modified-Wald approach.</td>
<td>83% (95% CI: 71-90%)</td>
<td>95% (95% CI: 88-99.6%)</td>
</tr>
<tr>
<td>Rokkas &amp; Niv, 2012&lt;sup&gt;63&lt;/sup&gt;</td>
<td>6</td>
<td>166</td>
<td>NR</td>
<td>Mantel Haenszel method</td>
<td>89% (95% CI: 82-94%)</td>
<td>95% (95% CI: 89-98%)</td>
</tr>
</tbody>
</table>

CI = confidence interval; NR = not reported.

**Key Question 1b. Intermediate Outcomes Such as Clinical Decisionmaking and Dietary Compliance**

We identified a 2009 SR on factors associated with adherence to gluten-free diet in adult celiac disease patients<sup>65</sup>. The authors identified 38 studies published between January 1980, and November 2007. Studies were included regardless of whether adherence was a primary or secondary focus. The quality of the review was rated low according to AMSTAR criteria (see appendix). The rate of “strict” adherence ranged from 42% to 91% in the included studies and varied by definition of “strict” and whether adherence was measured by self-report or estimated via biological markers. Most factors investigated were socio-demographic; none of the 38 studies compared adherence by type of diagnostic test. Three included studies reported no statistical difference in adherence levels between patients whose celiac disease was detected via screening and those whose celiac disease was “symptom- detected.” These three studies were conducted in adults in the U.S. and Western Europe after 2001, and each enrolled 100 or fewer CD patients.

We identified one additional relevant study, whose results conflict with those reported in the 2009 SR. This study, conducted in Israel, sheds light on clinician and patient decisionmaking after asymptomatic individuals screened positive. In 2007, researchers conducted a study on prevalence of CD by screening 1,571 healthy adult blood donors using the IgA-tTG (or IgG-tTG for IgA deficient).<sup>66</sup> The fifty-nine patients who tested positive were given their results and
counseled; 51 participated in a telephone interview about three years later. Only 30 had undergone biopsy; of these 30, ten were diagnosed with CD. Four of the ten strictly adhered to a gluten-free diet; the other six did not believe they had celiac disease, because they considered themselves asymptomatic. Four of these six reported having been told by their physicians (two gastroenterologists, two primary care physicians) to ignore the test results because there was no need for asymptomatic patients to modify their diet. Of the 29 who did not undergo biopsy, twelve consulted a physician, and nine of these twelve were advised against biopsy due to lack of symptoms. While this study provides interesting information, risk of bias is high (the sample size is small and the information on clinical and patient decisionmaking is self-reported rather than assessed via medical record) and applicability to the current U.S. situation is not certain.

We identified no other studies of how clinical decisions differ by method of patient diagnosis. The official guidance on clinical management of celiac disease does not differ after the initial diagnosis is finalized.\(^5\)

**Key Question 1c. Clinical Outcomes and Complications Related to CD**

No studies reporting how clinical outcomes or complications differ by diagnostic method were identified. Differences due to false negative results are discussed in the results section for Key Question 4.

**Key Question 1d. Patient-Centered Outcomes Such as QOL and Symptoms**

No studies reporting how patient-centered outcomes differ by diagnostic method were identified. Differences due to false negative results are discussed in the results section for Key Question 4.

**Key Question 2. Duodenal Biopsy Issues**

**Key Points**

One very large retrospective national study found that in the U.S., adherence to American Gastroenterological Association duodenal biopsy protocol (4+ specimens) was worse at endoscopy suites with a higher volume of endoscopies with duodenal biopsy, while adherence was better at endoscopy suites with a higher number of gastroenterologists.

Three retrospective studies evaluating inter-observer variability in histological diagnosis of CD between different pathologists and clinical settings indicate that CD-related histological findings are underdiagnosed in community-based hospital- and practice settings when compared to academic settings.

Two previous SRs and several additional primary studies indicate that the number and location of biopsy specimens influence diagnostic findings of biopsy, and they recommend taking multiple specimens from different sites of the duodenum.

A 2013 SR of high quality on clinical response to gluten challenge indicates that a 3-month gluten challenge with a moderate-to-high dose (e.g., 15g daily) should be sufficient to diagnose the majority of CD patients; however, based on more recent data, the ACG recommends three grams daily for two weeks and six additional weeks if tolerable for adults.
Detailed Synthesis

Key Question 2a. Characteristics of Pathologists and Health Care Providers

One of the Key Questions for this systematic review is whether the likelihood of positive diagnosis with endoscopy with duodenal biopsy varies depending on the characteristic of pathologists or other medical staff (e.g. level of training or experience). Four studies addressing this issue were identified; results are displayed in Table 15. In their study of the influence of provider characteristics (e.g. procedure volume [defined by the number of endoscopies with duodenal biopsy performed] and the number of physicians in each endoscopy suite) on adherence to the American Gastroenterological Association’s protocol (submissions of four or more specimens during duodenal biopsy), Lebwohl and colleagues (2013) found that adherence reduced with increasing procedure volume, but increased with increasing number of gastroenterologists working in an endoscopy suite.67

Three studies investigated the inter-observer variability in the histological diagnosis of CD between different pathologists and clinical settings. In a study from the Netherlands in which the histological diagnoses of suspected pediatric CD patients were reviewed,68 Mubarak et al. (2011) found a moderate inter-observer agreement (k=0.486) in the Marsh classification of the villous atrophy of biopsy specimens between the referring/originating pathologist and the study pathologist as well as a high inter-observer agreement (k=0.850) for CD diagnosis. A similar study by Arguelles-Grande et al. from a US center in 2012 found the inter-observer agreement in pathologists’ diagnosis—based on biopsy interpretation—between the pathologist in the study hospital and the pathologist in the referring or originating hospital to be moderate (k=0.529, p<0.0001), leading to a 20% increase in CD diagnosis.69 The same study examined inter-observer agreement by type of pathology practice setting and found it ranged between ‘very good’ when the study hospital was compared with other university hospitals (k=0.888) and ‘moderate’ when the study hospital was compared with community hospitals or commercial pathology laboratories (k=0.465 and k=0.419, respectively) (Arguelles-Grande et al. 2012). A study in Argentina that evaluated the accuracy of the histologic diagnosis of CD performed in a community clinical setting compared with that of an experienced academic center found a divergence of 46.3% in diagnosis and a poor agreement (k=0.16) between settings (Sanchez et al. 2007).70 These studies support that CD-related histological findings are underdiagnosed in community-based hospital and practice settings.
Table 15. Diagnosis by duodenal biopsy: Variation by pathologist and setting characteristics

<table>
<thead>
<tr>
<th>Study (Author, Year)</th>
<th>Study Objective</th>
<th>Methodology/Data Source</th>
<th>Sample Size/Pop.</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Picarelli et al., 2014</td>
<td>Verified the correct and uniform application of Marsh–Oberhuber criteria, observing their reliability and diagnostic accuracy in CD diagnosis by testing the repeatability of histological evaluation of the same histological samples carried out by 5 different operators (histologists).</td>
<td>Evaluation of histological findings of duodenal biopsies by five different histologists not aware of patients' clinical data. (the most experienced histologist in CD, was used as standard for comparison)</td>
<td>66 active CD patients and 48 controls with no CD.</td>
<td>The strength of agreement was good/very good for Marsh–Oberhuber classification (Kappa statistic: 0.54-0.78) as well as CD diagnosis (Kappa statistic: 0.78).</td>
</tr>
<tr>
<td>Lebwohl et al., 2013</td>
<td>Studied the influence of provider characteristics (procedure volume, defined by the number of endoscopies with duodenal biopsy performed, the number of physicians in each endoscopy suite, and the regional physician density) on adherence to standard of care (four or more specimens during duodenal biopsy)</td>
<td>National pathology database/ Multivariate analysis.</td>
<td>92,580 adults with potential CD</td>
<td>Reduced adherence was observed with higher procedure volume [odds ratio (OR) for each additional 100 procedures, 0.92; 95% CI, 0.88–0.97; P = 0.002]. An increased adherence was reported for gastroenterologists working at suites with higher numbers of gastroenterologists (OR for each additional gastroenterologist, 1.08; 95% CI, 1.04–1.13; P &lt; 0.001).</td>
</tr>
<tr>
<td>Arguelles-Grande et al., 2012</td>
<td>Evaluated the agreement in biopsy interpretation (characteristic histological alterations of small bowel mucosa) between different pathology practice types (university hospitals, community hospitals, commercial pathology laboratories)</td>
<td>Retrospective review of biopsy slides of patients from referring centers.</td>
<td>102 suspected adult/pediatric CD patients</td>
<td>Inter-observer agreement in the diagnosis of CD between study pathologist and the referring pathologist was moderate (k=0.529, p&lt;0.0001). In addition, agreement ranged between 'very good' with other university hospitals (k=0.888) and 'moderate' with community hospitals or commercial laboratories (k=0.465 and k=0.419, respectively).</td>
</tr>
<tr>
<td>Mubarak et al., 2011</td>
<td>Determination of the inter-observer variability in the histological diagnosis of CD.</td>
<td>Retrospective review of histology slides of biopsy specimens by a single experienced pathologist</td>
<td>297 consecutive pediatric patients with suspected CD.</td>
<td>The inter-observer variability for the Marsh classification was found to be moderate with a Kappa value of 0.486 while the Kappa value for CD diagnosis was high (0.850) A total of 160 (53.9%) patients were original diagnosed with CD while 172 (57.9%) patients were diagnosed according to the second pathologist.</td>
</tr>
<tr>
<td>Sanchez et al., 2007</td>
<td>Evaluated the accuracy of the diagnosis of CD performed in the community clinical setting compared with that of an academic experienced center.</td>
<td>Retrospective review of original biopsy slides and reports (from community clinical setting).</td>
<td>70 consecutive adult CD patients</td>
<td>25 patients (46.3%) had a divergent diagnosis between the two practice settings (23 patients originally identified as CD and 2 diagnosed as non CD) (Kappa statistic: 0.16 -denoting poor agreement).</td>
</tr>
</tbody>
</table>

CD = celiac disease; CI = confidence interval.
Key Question 2b. Specimens: Number and Location

Two nonsystematic reviews provided background information on biopsy strategy, i.e. specimen location and number. After reference-mining these two reviews and conducting electronic searches, we identified 19 studies relevant to this key sub-question. These studies are described narratively due to heterogeneity. Sensitivity and specificity were rarely discussed, as there was no “reference” test per se; the studies generally compared the number of positive diagnoses by strategy.

Several studies reported that the number of biopsy specimens influenced the likelihood of positive diagnosis, suggesting that multiple biopsy specimens should be taken from the duodenum in order to optimize diagnostic yield in both pediatric and adult populations. One study in an adult population showed that the probability of a new diagnosis was increased when at least 4 specimens were submitted compared to when fewer than 4 specimens were taken (1.8% vs 0.7%; p < .0001). Another retrospective study found that obtaining two duodenal biopsy specimens led to a confirmed diagnosis in 90% of cases, whereas obtaining at least four duodenal specimens led to confirmed diagnosis in all cases. Another study suggests a minimum of three biopsies incorporating a duodenal bulb biopsy as essential to detect villous atrophy, with a five-biopsy regime useful for detecting the most severe lesion.

Most of the studies recommend obtaining biopsies from multiple sites in the duodenum, as the site/location from which duodenal specimens are taken could affect the likelihood of CD diagnosis. The evidence suggests that duodenal specimens should include the duodenal bulb and the distal duodenum for optimal diagnostic yield in adult/general and pediatric populations. In a suspected pediatric CD population, a study suggested taking biopsies from both the bulb and the second part of the duodenum, because mucosal changes may occur in only one site.

Key Question 2c. Length of Gluten Challenge

Table 16 presents data on the length of time a patient needs to remain on a gluten-containing diet for accurate diagnosis. A SR (Bruins, 2013) addressed clinical response to gluten challenge among adult or pediatric patients with suspected or diagnosed CD previously consuming a gluten-free diet. This review was rated as high quality based on AMSTAR criteria (see appendix). The main focus of the review is serology; here we present results regarding intestinal histology. According to the authors, 51 to 100 percent of children developed moderate to severe mucosal histological abnormalities within 2 to 3 months of gluten challenge, with data from two trials indicating that 59 to 78 percent of children may experience an increase in villous atrophy within 3 months of gluten challenge. Evidence from a small body of trials found no more than 50 percent of adult patients to be positive for EmA-IgA, tTG-IgA, or DGP-IgA/IgG antibodies within 6 weeks to 3 months of gluten challenge. Mucosal tTG-IgA deposits can appear in the majority of adult patients within 2 weeks of gluten challenge; however, two weeks or more of high-dose gluten challenge may be needed to detect small intestinal mucosal morphology changes for most patients. The ACG recommends an additional six weeks for patients who can tolerate gluten exposure. Bruins and colleagues concluded that a 3-month gluten challenge with a moderate-to-high dose (such as 15g daily, in accord with ESPGHAN guidelines) should be sufficient to diagnose the majority of CD patients, with combination testing of antibodies and mucosal histology potentially accelerating diagnoses. More recently, Leffler, 2013 conducted small gluten challenge study of adults with biopsy-proven celiac...
disease. Patients were assigned a dose of either 3 or 7.5 grams of gluten daily. At 14 days, 89.5% of patients had sufficient villous atrophy to diagnose celiac disease via duodenal biopsy.
<table>
<thead>
<tr>
<th>Diagnostic Method</th>
<th>Population</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>EmA-IgA Antibodies</td>
<td>Children</td>
<td>Majority show positive levels within 3 months.</td>
</tr>
<tr>
<td></td>
<td>Adults</td>
<td>Few to none develop positive levels within 2 months.</td>
</tr>
<tr>
<td>tTG-IgA Antibodies</td>
<td>Children</td>
<td>Majority show positive levels within 12 weeks.</td>
</tr>
<tr>
<td></td>
<td>Adults</td>
<td>At most half have positive levels within 3 months.</td>
</tr>
<tr>
<td>DGP-IgA/IgG Antibodies</td>
<td>Adults</td>
<td>At most half have positive levels within 4 weeks.</td>
</tr>
<tr>
<td>HLA-DQ2 or -DQ8</td>
<td>Adults</td>
<td>Not affected by gluten intake</td>
</tr>
<tr>
<td></td>
<td>Children</td>
<td>Not affected by gluten intake</td>
</tr>
<tr>
<td>Histology</td>
<td>Children</td>
<td>Most developed moderate-to-severe abnormalities within 2 to 3 months.</td>
</tr>
<tr>
<td></td>
<td>Adults</td>
<td>At least 2 weeks needed to detect changes for most patients; 6 additional weeks recommended for those who are able to continue without severe discomfort</td>
</tr>
</tbody>
</table>

EmA = endomysial antibodies; IgA = immunoglobulin A; tTG = anti-tissue transglutaminase; IgG = immunoglobulin G.
**Key Question 3. Specific Populations**

**Key Points**

A 2010 SR limited to studies of patients with GI symptoms reported pooled sensitivity of 90% (95% CI: 80.0%, 95.0%) and specificity of 99% (95% CI: 98.0%, 100.0%) for IgA EmA tests (8 studies) and pooled sensitivity of 89% (95% CI: 82.0%, 94.0%) and specificity of 98% (95% CI: 95.0%, 99.0%) for IgA tTG tests. These results are similar to those presented for Key Question 1, which included patients with various symptoms of CD.

Only one study of general population screening met the inclusion criteria. Sensitivities were 100% and 85.7% for tTG IgA and EmA IgA, respectively. Specificities were 97.4% and 99.0%, respectively, in this high quality study.

Two low quality studies provided data that allowed calculation of accuracy of serology in patients with iron deficiency. Studies were conducted in the Middle East; applicability to the U.S. is uncertain.

Two high quality studies reported accuracy in children with Type 1 diabetes. These studies were conducted in Iraq and the Czech Republic with small samples. Applicability to the U.S. is uncertain.

No studies provided test accuracy data on patients with other auto-immune diseases, Turner’s syndrome, or Trisomy 21.

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 presenting with symptoms.

Two large moderate quality studies found both tTG and DGP tests less sensitive in adults than in children.

No studies reported accuracy by race, ethnicity, or SES.

Two small studies of the accuracy of new combination tests in IgA-deficient patients were published in 2014; results were inconsistent.

**Detailed Synthesis**

**Key Question 3a. Symptomatic Patients Versus Nonsymptomatic Individuals at Risk**

Two SRs reported on diagnostic test accuracy in individuals with symptoms of CD. In 2009, Ford and colleagues published an SR on yield of diagnostic tests in patients with Irritable Bowel Syndrome (IBS). The goal was to estimate prevalence of celiac disease in unselected adults with IBS; test accuracy was not a primary outcome. The authors included 14 studies; 54 percent of the 4,204 individuals met diagnostic criteria for IBS. Although sensitivity and specificity were not reported, the authors computed odds ratios for positive diagnosis in IBS patients, compared to non-IBS controls, and the results indicate differences by diagnostic method. Pooled odds ratios for celiac disease in subjects with IBS versus controls were 3.40, (95% CI [1.62, 7.13]) for IgA EmA; 2.94, 95% CI (1.36, 6.35) for IgA tTG; and 4.34, (95% CI [1.78, 10.60]) for biopsy. Prevalence was estimated at 4.0 percent, 1.65 percent, and 4.34 percent by IgA EmA, IgA tTG, and biopsy, respectively.
In 2010, van der Windt and colleagues published an SR on performance of diagnostic tests in patients with abdominal symptoms typical of celiac disease. Studies where 50 percent or more of participants reported gastrointestinal (GI) symptoms were included. Eight such studies that used IgA EmA were pooled, resulting in estimates of 90.0% (95% CI: 80.0%, 95.0%) for sensitivity and 99.0% (95% CI: 98.0%, 100%) for specificity. Pooled results of seven studies on IgA tTG estimated sensitivity at 89.0% (95% CI: 82.0%, 94.0%) and specificity at 98.0% (95% CI: 95.0%, 99.0%).

The quality of both SRs was moderate according to AMSTAR criteria. The Ford (2009) SR did not report quality assessment of included studies. Neither the Ford nor the van der Windt SRs included lists of excluded studies. (Full AMSTAR criteria for all SRs are presented in Appendix E.)

Only one study of general population screening met the inclusion criteria. This study presented data for 1,000 randomly selected Swedes. Sensitivity was 100% and 85.7% for tTG IgA and EmA IgA, respectively. Specificity was 97.4% and 99.0%, respectively.

Only one study compared accuracy in symptomatic versus asymptomatic individuals at risk for celiac disease. Nevoral, 2013 studied the accuracy of the new ESPGHAN guideline (combining tTG IgA and EmA IgA tests) in 32 first degree relatives, 60 patients with Type 1 diabetes, and 187 subjects presenting with symptoms of CD in the Czech Republic. Specificity was lower in the asymptomatic subjects at risk: 0.70 for those with family history and 0.64 for those with Type 1 diabetes, compared to 0.85 for subjects presenting with symptoms. Sensitivity was 0.81 for those with family history, 0.93 for those with diabetes, and 0.76 for symptomatic subjects. This is the only study that provided accuracy results specifically for subjects with a family history.

Two individual studies provided test accuracy data for patients with iron deficiency, a common symptom of CD. Data are presented in Table 17. One study conducted in Iran reported sensitivity of 0.38 and specificity of 0.97 for IgA tTG. The other study, conducted in Turkey, reported sensitivity of 1.00 and specificity of 0.99 for IgA EmA and sensitivity of 0.33 and specificity of 0.96 for IgG EmA. The primary goal of the first study was to assess accuracy, while the goal of the latter study was prevalence estimation. In both publications, it was unclear whether interpretation of the index test and reference test (biopsy) were blinded; the interval between these tests was also unclear.

Table 17. Accuracy data for persons with iron deficiency

<table>
<thead>
<tr>
<th>Test Type</th>
<th>True Positive</th>
<th>False Negative</th>
<th>False Positive</th>
<th>True Negative</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emami, 2012</td>
<td>tTG IgA</td>
<td>5</td>
<td>8</td>
<td>4</td>
<td>113</td>
<td>0.38</td>
</tr>
<tr>
<td>Cekin, 2012</td>
<td>EmA IgA</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>77</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>EmA IgG</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>75</td>
<td>0.33</td>
</tr>
</tbody>
</table>

* tTG = anti-tissue transglutaminase; IgA = immunoglobulin A; EmA = endomysial antibodies; IgG = immunoglobulin G.

Two studies provided accuracy data in patients with Type 1 diabetes. Data are displayed in Table 18 below. Mansour, 2011 reported on use of IgA tTG, IgG tTG, and IgA EmA tests in 62 adults and children with Type 1 diabetes in Iraq. The study’s primary goal was to assess prevalence of asymptomatic CD in persons with Type 1 diabetes. IgA tests had higher accuracy than the IgG test, as displayed below. As noted above, Nevoral (2013) followed the recent ESPGHAN guidelines (tTG and EmA test) with 60 children and adolescents with Type 1
diabetes in the Czech Republic and reported sensitivity of 0.93 and specificity of 0.64 for the two tests combined. Both studies were rated high quality according to QUADAS 2 criteria.

Table 18. Accuracy data for persons with type 1 diabetes

<table>
<thead>
<tr>
<th>Test Type</th>
<th>True Positive</th>
<th>False Negative</th>
<th>False Positive</th>
<th>True Negative</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mansour, 2011</td>
<td>tTG IgA</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>51</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>tTG IgG</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>51</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>EmA IgA</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>53</td>
<td>0.71</td>
</tr>
<tr>
<td>Nevoral, 2013</td>
<td>tTG IgA, EmA IgA</td>
<td>0.93</td>
<td>0.64</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

tTG = anti-tissue transglutaminase; IgA = immunoglobulin A; IgG = immunoglobulin G; EmA = endomysial antibodies.

We found no studies that provided accuracy data on patients with other auto-immune diseases, Turner’s syndrome, or Trisomy 21. Although cirrhosis of the liver is not a risk factor for CD, one study of prevalence in patients with biopsy proven cirrhosis allowed calculation of sensitivity and specificity. Wakin-Fleming, 2014 conducted a high quality prospective cohort study that allowed comparison of IgA EmA (serum dilution >=1/10) to IgA tTG (> 20 U) in 204 cirrhosis patients. Based on biopsy results, five patients were diagnosed with celiac disease. IgA EmA had both sensitivity and specificity of 100%, while sensitivity was 100% and specificity was 96% for IgA tTG. This study had low risk of bias despite not being designed as an accuracy study.

Key Question 3b. Adults (Age 18 and Over) Versus Children and Adolescents

Studies that address this comparison are discussed in section c, below.

Key Question 3c. Children Under Age 24 Months Versus Older Children

Most studies reported mean age or age range of subjects, whereas a few simply described their patients as “adults” or “children.” Accuracy data could not be pooled by age group due to heterogeneity of populations, varying definitions of adult, and missing age data in some studies.

Three studies compared the accuracy of tests among age groups. These studies are important to highlight because in each study the populations are fairly homogenous, and each patient received the same diagnostic tests in terms of assay and threshold level. Data from these three studies are displayed in Table 19 below. As described earlier, Mozo (2012) compared the accuracy of IgA and IgG DGP tests in patients aged six months to 74 years. In children age seven or younger, the IgG tests had the same accuracy as IgA tests. In patients over age seven, the IgG tests had higher specificity and positive predictive value but lower sensitivity and negative predictive value. Olen (2012) compared accuracy of tTG and DGP tests in 537 children and adolescents. In children with normal IgA who were younger than two years old (N = 71), sensitivity and specificity were 96.0% and 98.0%, respectively. Sensitivity of a combined IgA/IgG DGP test was 100%, and specificity was 31% in children younger than two. As presented below, these tests were less accurate in the entire sample, which ranged in age from one to eighteen years. (Accuracy results were not presented separately for patients over 24 months old.) Finally, Vermeersch (2010) compared accuracy in adults (over age 16) versus
children as part of a study comparing IgG DGP tests with IgA tTG screens. Using a case-control design, they compared accuracy of three IgG DGP tests, three IgA and two IgG anti-tTG assays, and one IgA DGP screen in 827 patients. Results are presented below; all tests reported lower sensitivity in adults than in children.

Table 19. Accuracy results by age

<table>
<thead>
<tr>
<th>Age Category</th>
<th>True Positive</th>
<th>False Negative</th>
<th>False Positive</th>
<th>True Negative</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>tTG IgA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mozo, 2012¹</td>
<td>&lt;=7 yrs old</td>
<td>40</td>
<td>3</td>
<td>2</td>
<td>37</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>&gt;7 yrs old</td>
<td>49</td>
<td>8</td>
<td>4</td>
<td>57</td>
<td>0.86</td>
</tr>
<tr>
<td>Olen, 2012²</td>
<td>&lt;=18 yrs old</td>
<td>259</td>
<td>17</td>
<td>36</td>
<td>218</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>&lt;2 yrs old</td>
<td>25</td>
<td>1</td>
<td>1</td>
<td>40</td>
<td>0.96</td>
</tr>
<tr>
<td>Vermeersch,</td>
<td>&lt;16 yrs old</td>
<td>27</td>
<td>1</td>
<td>11</td>
<td>189</td>
<td>0.96</td>
</tr>
<tr>
<td>2010³</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt;16 yrs old</td>
<td>27</td>
<td>1</td>
<td>3</td>
<td>197</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>&gt;=16 yrs old</td>
<td>49</td>
<td>9</td>
<td>27</td>
<td>514</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>&gt;=16 yrs old</td>
<td>45</td>
<td>13</td>
<td>9</td>
<td>532</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>&gt;=16 yrs old</td>
<td>46</td>
<td>12</td>
<td>39</td>
<td>502</td>
<td>0.79</td>
</tr>
<tr>
<td><strong>tTG IgG</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vermeersch,</td>
<td>&gt;=16 yrs old</td>
<td>31</td>
<td>27</td>
<td>4</td>
<td>537</td>
<td>0.53</td>
</tr>
<tr>
<td>2010³</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;=16 yrs old</td>
<td>17</td>
<td>41</td>
<td>9</td>
<td>532</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>&lt;16 yrs old</td>
<td>21</td>
<td>7</td>
<td>10</td>
<td>190</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>&lt;16 yrs old</td>
<td>16</td>
<td>12</td>
<td>2</td>
<td>198</td>
<td>0.57</td>
</tr>
<tr>
<td><strong>DGP IgA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mozo, 2012¹</td>
<td>&lt;=7 yrs old</td>
<td>41</td>
<td>2</td>
<td>1</td>
<td>38</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>&gt;7 yrs old</td>
<td>55</td>
<td>2</td>
<td>3</td>
<td>58</td>
<td>0.97</td>
</tr>
<tr>
<td>Olen, 2012²</td>
<td>&lt;=18 yrs old</td>
<td>172</td>
<td>16</td>
<td>164</td>
<td>56</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>&lt;2 yrs old</td>
<td>24</td>
<td>0</td>
<td>31</td>
<td>14</td>
<td>1.00</td>
</tr>
<tr>
<td>Vermeersch,</td>
<td>&gt;=16 yrs old</td>
<td>40</td>
<td>18</td>
<td>3</td>
<td>538</td>
<td>0.69</td>
</tr>
<tr>
<td>2010³</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt;16 yrs old</td>
<td>26</td>
<td>2</td>
<td>3</td>
<td>197</td>
<td>0.93</td>
</tr>
<tr>
<td><strong>DGP IgG</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mozo, 2012¹</td>
<td>&lt;=7 yrs old</td>
<td>43</td>
<td>0</td>
<td>0</td>
<td>39</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>&gt;7 yrs old</td>
<td>52</td>
<td>5</td>
<td>1</td>
<td>60</td>
<td>0.91</td>
</tr>
<tr>
<td>Vermeersch,</td>
<td>&gt;=16 yrs old</td>
<td>40</td>
<td>18</td>
<td>3</td>
<td>538</td>
<td>0.69</td>
</tr>
<tr>
<td>2010³</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&gt;=16 yrs old</td>
<td>46</td>
<td>12</td>
<td>2</td>
<td>539</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>&gt;=16 yrs old</td>
<td>47</td>
<td>11</td>
<td>11</td>
<td>530</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>&gt;=16 yrs old</td>
<td>42</td>
<td>16</td>
<td>12</td>
<td>529</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>&lt;16 yrs old</td>
<td>26</td>
<td>2</td>
<td>3</td>
<td>197</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>&lt;16 yrs old</td>
<td>26</td>
<td>2</td>
<td>6</td>
<td>194</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>&lt;16 yrs old</td>
<td>27</td>
<td>1</td>
<td>9</td>
<td>191</td>
<td>0.96</td>
</tr>
</tbody>
</table>
Key Question 3d. Demographics, Including Race, Genetics, Geography, SES

No studies reported accuracy by race, ethnicity, or SES. Many studies were conducted outside the U.S., most often in Europe or the Middle East. However, even though we identified studies conducted in one country only, and even if that country has a homogeneous ethnic population, results may have little applicability to persons of the same racial or ethnic group living in the U.S.

Key Question 3e. Patients With IgA Deficiency

Two studies of accuracy in IgA-deficient patients met our inclusion criteria. Beinvenu (2014) studied a multi-analytic lateral-flow immunochromatographic assay (CD-LFIA) based on both IgA DGP and IgG DGP in 45 IgA-deficient children presenting with symptoms of CD or having risk factors. The study was retrospective —the new test was used on stored blood samples of patients who previously underwent biopsy. Researchers were blinded to those results. The authors reported a sensitivity of 100% and specificity of 89.2%. Wolf (2014) conducted a case-control study of adding an IgG DGP test to an IgA tTG test; 27 of their 1,071 subjects were IgA deficient. From the data presented in their article, we calculated 100% specificity in the children with IgA deficiency. However, sensitivity in this group was only 29%.

In addition, Dutta (2010) reported on 92 consecutive patients who presented with symptoms of possible celiac disease at a clinic in Vellore, India. Eighteen patients (19.5%) were diagnosed with CD; 14 had positive serology. Sensitivity and specificity of IgG tTG tests were 77.8% and 89.1%, respectively. It was unclear why IgG tests were the only serology tests conducted and whether patients were IgA deficient. The study had no major flaws.

Key Question 3f. Patients Who Previously Tested Negative for CD

One study of test accuracy in patients with previously negative serology was identified. As part of a larger study, Sugai (2010) retrospectively conducted IgA and IgG DGP tests on 17 IgA tTG-negative serum samples from patients with indications of celiac disease (either intestinal enteropathy or dermatitis herpetiformis). In the IgA tTG-negative patients, detection of “gluten sensitivity” increased 31.6% when an IgA tTG/ IgA DGP dual screen was applied and 26.3% when a dual IgA and IgG DGP screen was used. The sensitivity and specificity of the tTG/DGP dual screen were 35.7% and 100% in this group, whereas those rates were 42.9% and 100% for dual DGP screen. However, the results may be biased: Five of the 22 originally identified IgA-deficient patients refused biopsy, leaving 17 for this small retrospective analysis.

Key Question 4. Adverse Effects

Key Points

A systematic review of 150 studies on VCE not specific for CD found a capsule retention rate of 1.4%; in three studies specific to CD, the rate ranged from 0.9% to 4.6%. No studies on safety of upper GI endoscopy and/or duodenal biopsy for diagnosis of CD were identified. According to the American Society for Gastrointestinal Endoscopy...
Infection from upper GI endoscopy is very rare, as is bleeding unless a polyp is removed during the procedure.

Two studies reporting sequelae in patients with positive EmA serology but normal biopsy indicated that 30 percent to 50 percent of these patients are diagnosed with CD after a gluten-free diet or gluten challenge.

One study of 34 children with intestinal villous atrophy and simultaneous negative EmA-IgA tests found 2 infants to have confirmed CD after 6-10 years of iterative cycles of gluten challenges and gluten-free diet.

**Detailed Synthesis**

**Direct Adverse Events Associated With Invasive Methods of Diagnosing Celiac Disease**

“Invasive” methods used to diagnose CD include upper GI endoscopy with duodenal biopsy, often referred to as the “gold standard,” as well as VCE for patients who wish to avoid biopsy. A systematic review on the general safety of these procedures was beyond the scope of this small project; however, due to the dearth of data specific to their use in CD diagnosis, we review evidence on the safety of their use to provide some context. Applicability to patients with suspected CD is uncertain.

The main adverse event from VCE is capsule retention, which could cause acute small bowel obstruction or the need for surgical removal. According to International Center for Clinical Excellence (ICCE) consensus, capsule retention may occur in patients with refractory celiac disease who may have strictures. Three studies reporting retention rates with VCE used for CD diagnosis were identified; rates ranged from 0.9% to 4.6%. Study characteristics are displayed in Table 20. Gomez and colleagues (2013), in a comparison of patients aged 80 years or older with those under age 80, found a similar frequency in occurrence of capsule retention and concluded that capsule endoscopy can be performed safely in the elderly population.

A 2010 systematic review of VCE by Liao and colleagues in populations suspected of various small intestine pathologies found a pooled retention rate of 1.4% in 104 prospective studies and 46 retrospective studies published between 2000 and 2008. The rates of capsule retention appear to depend on VCE indication; pooled retention rates of 1.2 percent, 2.6 percent, and 2.1 percent for obscure gastrointestinal bleeding, Crohn’s disease, and neoplastic lesions indications were estimated from 47, 23, and 12 studies respectively. There were no studies in this systematic review that investigated capsule retention when VCE was used specifically to evaluate patients with suspected CD. This systematic review was rated as low quality, as a listing or bibliography of the included articles was not provided.

Only two studies on biopsy to diagnose CD were identified; they estimated the amount of absorbed radiation due to X-ray fluoroscopy-guided small intestine biopsies in a pediatric population. These studies were excluded, as fluoroscopy has been replaced by video systems. No studies on the safety of upper GI endoscopy when used specifically for diagnosis of CD were identified. According to ASGE, infection from upper GI endoscopy is very rare, as is bleeding, unless a polyp is removed during the procedure.

We employed the McHarm scale to evaluate the quality of the VCE adverse event studies. Data are displayed in Table 21. All used precise definitions to define harms (generally, capsule retention was defined as a situation in which a capsule endoscope remains in the digestive tract for a minimum of two weeks, which may necessitate surgical intervention in order to retrieve the
capsule endoscope.) No studies explicitly mentioned severe or serious adverse events. Adverse events were actively collected or ascertained in all studies. All studies reported the number and full spectrum of adverse events that occurred; the total number of study participants affected by harms; and the type of analyses conducted for adverse events data.
<table>
<thead>
<tr>
<th>Study (Author, Year)</th>
<th>Study Design (Claims Data Analysis, RCT, Survey, Etc.)</th>
<th>Diagnostic Procedures Assessed (Endoscopy, Serology Test, Etc.)</th>
<th>Sample Size</th>
<th>Population</th>
<th>Event</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vere et al., 2012</td>
<td>Retrospective study</td>
<td>Video Capsule endoscopy</td>
<td>43</td>
<td>Adults</td>
<td>Capsule retention</td>
<td>Capsule retention in two patients (4.6%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Slower intestinal transit time)</td>
<td></td>
</tr>
<tr>
<td>Gomez, 2013</td>
<td>Retrospective matched cohort study (&gt;80yr old patients vs. &lt;80 yr old patients)</td>
<td>Video capsule endoscopy</td>
<td>780</td>
<td>Adults</td>
<td>Capsule retention</td>
<td>Capsule retention occurred at a similar frequency in patients age &gt;80yrs (1.0%) compared to those &lt;80 yrs (0.9%)</td>
</tr>
<tr>
<td>Atay et al., 2009</td>
<td>Retrospective chart review</td>
<td>Video capsule endoscopy</td>
<td>207</td>
<td>Children</td>
<td>Capsule retention</td>
<td>Capsule retention occurred in 3 of 207 procedures (1.4%)</td>
</tr>
</tbody>
</table>

RCT = randomized controlled trial.
Table 21. Quality of adverse events studies

<table>
<thead>
<tr>
<th>McHarm Items</th>
<th>Vere et al. 2012\textsuperscript{90}</th>
<th>Gomez 2013\textsuperscript{91}</th>
<th>Atay et al. 2009\textsuperscript{92}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Were the harms PRE-DEFINED using standardized or precise definitions?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>2 Were SERIOUS events precisely defined, if mentioned?</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>3 Were SEVERE events precisely defined, if mentioned?</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>4 Were the number of DEATHS in each study group specified OR were the reason(s) for not specifying them given?</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>5 Was the mode of harms collection specified as ACTIVE?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>6 Was the mode of harms collection specified as PASSIVE?</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>7 Did the study specify WHO collected the harms?</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>8 Did the study specify the TRAINING or BACKGROUND of who ascertained the harms?</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>9 Did the study specify the TIMING and FREQUENCY of collection of the harms?</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>10 Did the author(s) use STANDARD scale(s) or checklist(s) for harms collection?</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>11 Did the authors specify if the harms reported encompass ALL the events collected or a selected SAMPLE?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>12 Was the NUMBER of participants that withdrew or were lost to follow-up specified for each study group?</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>13 Was the TOTAL NUMBER of participants affected by harms specified for each study arm?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>14 Did the author(s) specify the NUMBER for each TYPE of harmful event for each study group?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>15 Did the author(s) specify the type of analyses undertaken for harms data?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

NA = not applicable.

Sequelae of False Positives, False Negatives, and Indeterminate Results

Three studies reported sequelae in patients whose serology and biopsy results were discordant. All were conducted in Europe and involved EmA test results. Kurppa (2012)\textsuperscript{97} evaluated 405 consecutive EmA positive children and adults at a university hospital who were referred by physicians for suspicion of CD or who participated in population-based research studies. Of these, 40 patients had low and 17 had high EmA and tTG serum antibody values without simultaneous villous atrophy at baseline. Eventually 12 (30 percent) in the low-titer group and 8 (47 percent) in the high-titer group were diagnosed with CD based on villous atrophy on a gluten-challenge, while 17 (43 percent) and 8 (47 percent) were diagnosed with CD based on positive symptom and serological responses alongside the disappearance of early mucosal changes during a gluten-free diet. Unfortunately, length of follow-up was not report, so no data are provided on the length of gluten challenge or gluten-free diet.

A 2014 study by the same authors\textsuperscript{98} randomized 40 EmA-positive yet otherwise asymptomatic adults to either a gluten-free diet or gluten-challenge.\textsuperscript{98} These 40 patients were identified from 3,031 consecutive individuals screened for CD at a university hospital in Finland. Of these 40 participants, two in the gluten-free diet group and two in the gluten-challenge group had positive EmA but no villous atrophy. Of these 4 participants, one in each group were EmA-negative and had increased villous atrophy ratios after one-year follow-up; however, the gluten-
free diet participant demonstrated improved gastrointestinal symptoms and self-perceived general health, whereas the gluten-challenge participant worsened in both of these areas.

The final study, by Kwiecien (2005), documented sequelae of discordant results via a retrospective analysis of data from 1985-2000 on 34 children with subtotal or total intestinal villous atrophy with simultaneously negative IgA EmA tests. This group included all children with discordant results from over 1,300 consecutive diagnoses. Of the 34 children, 15 completed diagnostic follow-up with three biopsies, and two of these had confirmed CD with repeated positive IgA EmA tests. One 13 month-old with subtotal villous atrophy had a second biopsy after three years of a gluten-free diet, which revealed normal intestinal mucosa. Then, after a 12-month gluten challenge, positive IgA EmA occurred simultaneously with mucosal relapse and clinical symptoms of malabsorption syndrome, leading to a CD diagnosis at 5 years, 9 months. The other child, with total villous atrophy at 11 months, had a gluten-free diet for 2.5 years, then underwent a gluten-challenge for 2 years; at this point, subtotal villous atrophy was discovered, yet EmA-IgA tests were negative. Another two-year gluten-free diet led to normalization of intestinal mucosa, while a 3-year gluten-challenge led to positive IgA EmA and subtotal villous atrophy, and thus a CD diagnosis at 10 years, 3 months. Although these examples shed light on clinical pathways, this study was conducted prior to the availability of other serological tests. Thus, the results may have little applicability to current clinical practice.
Discussion

Key Findings and Strength of Evidence

The key findings and strength of evidence are summarized in Table 22. Additional details on strength of evidence ratings are provided as Appendix F.

Table 22. Summary of findings and strength of evidence

<table>
<thead>
<tr>
<th>Topic</th>
<th>EPC Conclusions and Strength of Evidence</th>
<th>Prior Systematic Reviews</th>
<th>Additional Findings From EPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Key Question 1: Accuracy of IgA tTG</td>
<td>High: IgA tTG tests have excellent sensitivity and specificity.</td>
<td>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.</td>
<td>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%.</td>
</tr>
<tr>
<td>Key Question 1: Accuracy of IgA EmA</td>
<td>High: IgA EmA tests have lower sensitivity but equal specificity to IgA tTG tests.</td>
<td>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%).</td>
<td>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%.</td>
</tr>
<tr>
<td>Key Question 1: Accuracy of IgA DGP</td>
<td>High: IgA DGP tests are not as accurate as IgA tTG tests.</td>
<td>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%).</td>
<td>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.</td>
</tr>
<tr>
<td>Key Question 1: Accuracy of IgG DGP</td>
<td>Moderate: IgG DGP tests are not as sensitive as IgA tTG tests in non–IgA-deficient patients.</td>
<td>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 96.0% to 98.9%. Authors did not pool data.</td>
<td>One study reported sensitivity of 95.0% and specificity of 99.0% in 200 non–IgA-deficient subjects of all ages.</td>
</tr>
<tr>
<td>Key Question 1: Accuracy of HLA-DQ2 or DQ8</td>
<td>High: HLA tests can be used to rule out CD with close to</td>
<td>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,</td>
<td>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</td>
</tr>
<tr>
<td>Topic</td>
<td>EPC Conclusions and Strength of Evidence</td>
<td>Prior Systematic Reviews</td>
<td>Additional Findings From EPC</td>
</tr>
<tr>
<td>-------</td>
<td>-----------------------------------------</td>
<td>--------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td></td>
<td>100% sensitivity.</td>
<td>the American College of Gastroenterology estimated the NPV of the HLA-DQ2/DQ8 combination test at over 99%.</td>
<td>of 18.2%. A 1999 cohort also reported sensitivity of 100%, while specificity was 33.3%.</td>
</tr>
<tr>
<td>Key Question 1: Accuracy of algorithms</td>
<td>Insufficient: Strength of evidence is insufficient to determine comparative accuracy of different algorithms in specific populations.</td>
<td>No SRs of the accuracy of algorithms were identified.</td>
<td>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.</td>
</tr>
<tr>
<td>Key Question 1: Accuracy of VCE</td>
<td>Moderate: VCE has very good sensitivity and excellent specificity.</td>
<td>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.</td>
<td>No additional studies met our inclusion criteria.</td>
</tr>
<tr>
<td>Key Question 1: Intermediate outcomes</td>
<td>Insufficient: Strength of evidence is insufficient regarding how method of diagnosis affects adherence.</td>
<td>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.</td>
<td>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.</td>
</tr>
<tr>
<td>Key Question 1: Clinical outcomes and complications</td>
<td>Insufficient: Strength of evidence is insufficient regarding how method of diagnosis affects clinical outcomes and complications.</td>
<td>No prior SRs on this topic were identified.</td>
<td>No studies on this topic were identified.</td>
</tr>
<tr>
<td>Topic</td>
<td>EPC Conclusions and Strength of Evidence</td>
<td>Prior Systematic Reviews</td>
<td>Additional Findings From EPC</td>
</tr>
<tr>
<td>-------</td>
<td>------------------------------------------</td>
<td>--------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>Key Question 1: Patient-centered outcomes such as quality of life</td>
<td>Insufficient: Strength of evidence is insufficient regarding how method of diagnosis affects patient-centered outcomes such as quality of life.</td>
<td>No prior SRs on this topic were identified.</td>
<td>No studies on this topic were identified.</td>
</tr>
<tr>
<td>Key Question 2: Biopsy and provider characteristics</td>
<td>Moderate: Physician adherence to biopsy protocol decreases with volume performed per endoscopy suite and increases with number of gastroenterologists per endoscopy suite.</td>
<td>No SRs on this topic were identified.</td>
<td>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).</td>
</tr>
<tr>
<td>Key Question 2: Biopsy and pathologist characteristics</td>
<td>Moderate: CD-related histological findings are underdiagnosed in community settings when compared with academic settings.</td>
<td>No SRs on this topic were identified.</td>
<td>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.</td>
</tr>
<tr>
<td>Key Question 2: Biopsy specimens—number and location</td>
<td>High: Increasing the number and location of biopsy specimens increases diagnostic accuracy.</td>
<td>No SRs addressed how the number and location of biopsy specimens influence diagnostic findings of biopsy.</td>
<td>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.</td>
</tr>
<tr>
<td>Key Question 2: Biopsy and length of time ingesting gluten</td>
<td>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.</td>
<td>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.</td>
<td>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.</td>
</tr>
<tr>
<td>Topic</td>
<td>EPC Conclusions and Strength of Evidence</td>
<td>Prior Systematic Reviews</td>
<td>Additional Findings From EPC</td>
</tr>
<tr>
<td>-------</td>
<td>-----------------------------------------</td>
<td>--------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>Key Question 3: Symptomatic patients vs. nonsymptomatic individuals at risk</td>
<td>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.</td>
<td>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.</td>
<td>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.</td>
</tr>
<tr>
<td>Key Question 3: Children vs. adults</td>
<td>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%.</td>
<td>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.</td>
<td></td>
</tr>
<tr>
<td>Key Question 3: Demographics, including race</td>
<td>Insufficient: There was insufficient evidence to estimate the accuracy of diagnostic methods by demographic characteristics. No SRs on this topic were identified.</td>
<td>No studies reported accuracy by race, ethnicity, or socioeconomic status.</td>
<td></td>
</tr>
<tr>
<td>Key Question 3: Patients with IgA deficiency</td>
<td>Insufficient: There was insufficient evidence to estimate the accuracy of diagnostic methods in IgA-deficient patients. No SRs on this topic were identified.</td>
<td>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.</td>
<td></td>
</tr>
<tr>
<td>Key Question 3: Patients who previously tested negative for CD</td>
<td>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.</td>
<td>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.</td>
<td></td>
</tr>
<tr>
<td>Topic</td>
<td>EPC Conclusions and Strength of Evidence</td>
<td>Prior Systematic Reviews</td>
<td>Additional Findings From EPC</td>
</tr>
<tr>
<td>-------</td>
<td>----------------------------------------</td>
<td>--------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>Key Question 4: Direct adverse events—VCE</td>
<td>High: The rate of capsule retention is less than 5%.</td>
<td>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.</td>
<td>In 3 studies specific to CD, the capsule retention rate ranged from 0.9% to 4.6%.</td>
</tr>
<tr>
<td>Key Question 4: Direct adverse events—endoscopy with duodenal biopsy</td>
<td>Moderate: Adverse events during upper GI endoscopy are rare.</td>
<td>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.</td>
<td>No studies specific to diagnosis of CD were identified.</td>
</tr>
<tr>
<td>Key Question 4: Indirect adverse events—false negatives or positives</td>
<td>Insufficient: Strength of evidence is insufficient regarding the impact of misdiagnosis.</td>
<td>No SRs on the impact of misdiagnosis of CD were identified.</td>
<td>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. 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.</td>
</tr>
</tbody>
</table>

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.
Findings in Relationship to What Is Already Known

Table 22 displays findings from prior SRs along with the findings from the newly identified studies that met our inclusion criteria. We identified enough studies of the accuracy of tTG IgA tests and EmA IgA tests to conduct new meta-analyses. Our findings confirm the excellent specificity of both tests, the excellent sensitivity of tTG IgA and the good specificity of EmA IgA reported in prior SRs. A prior SR reported promising accuracy results for DGP tests; we found only one new study.

Several studies on whether adding an EmA or DGP test to a tTG test increases accuracy have recently been published. Results are insufficient to determine whether such increases are clinically meaningful.

No SRs have been conducted on the association between setting (academic vs community) and provider performance in CD diagnosis. We identified three retrospective studies evaluating inter-observer variability in histological diagnosis of CD between different pathologists and clinical settings. Results indicate that CD-related histological findings are underdiagnosed in community-based hospital and practice settings when compared to academic settings.

No SRs on how method of diagnosis affects patient adherence or clinical decisionmaking have been published. Very few studies have addressed these issues; we found insufficient evidence to answer Key Questions on this topic.

Applicability

Several factors affect the applicability of this review.

To increase generalizability, this report limited accuracy studies to those that included consecutive patients or a random sample. Several studies were excluded because we could not determine enrollment based on the information available.

Only one general population screening study met the criteria 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 makes the conduct of such studies challenging. Thus, the evidence on accuracy of diagnostic screening 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 U.S., study location was not a basis for 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 celiac disease.) In contrast, many community physicians base their diagnosis on a simple qualitative assessment of villous atrophy or elevation of intraepithelial lymphocytes.

Accuracy of serology assays may vary by both laboratory and manufacturer. For example, Li and colleagues (2009)\textsuperscript{100} used 150 samples from participants of known CD status to compare accuracy of tTG tests at 20 laboratories in the US and Europe. Sensitivity was less than 75% at...
four laboratories. Using a similar research design, Rozenberg and colleagues (2011) found differences in performance of tTG across various manufacturers.

Finally, VCE is not a first line diagnostic method—it is indicated for adults who refuse biopsy. A 2012 SR of six studies reported very good sensitivity and excellent specificity. However, patient characteristics may differ between those who refuse a biopsy and those who accept. 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 systematic reviews on the accuracy of individual diagnostic tests using immunoglobulin A (IgA). All IgA tests for celiac disease have excellent specificity; DGP IgA has slightly lower specificity than tTG IgA and EmA IgA. tTG IgA testing has a high positive predictive value 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, we were unable to determine which tests, if any, are more accurate in patients with specific symptoms or risk factors due to a dearth of studies meeting our inclusion criteria. Patients with symptoms associated with celiac disease 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, and 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 more accurate than tTG in small children; strength of evidence is low but could increase if findings are replicated. tTG IgA had greater sensitivity than EmA IgA 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 celiac disease prevalence similar to the US.

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 fold above the upper limit of normal for specific celiac tests may be accurate for diagnosing celiac disease without requiring histopathology specimens; however, the currently available evidence on comparative accuracy of algorithms is inconclusive, due to the wide range of results, heterogeneity of populations studied, and the lack of clinically significant increases in accuracy compared to 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 duodenal 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 celiac disease is underdiagnosed by pathologists in community settings compared to academic settings; continued education on diagnostic protocols may be warranted for community physicians.
Limitations of the Comparative Effectiveness Review Process

At the request of AHRQ we conducted an assessment of the evidence on comparative effectiveness of various diagnostic methods currently used in the U.S. to diagnosis celiac disease. We conducted an extensive literature search; however, our consideration of unpublished literature was limited. Although a Scientific Resource Center (SRC) funded by AHRQ requested information from test manufacturers and major laboratories, no information was provided; we did not search FDA databases for such information ourselves.

In addition, this project was funded as a “small” systematic review and budgeted to include abstraction and analysis of fewer than 50 studies. Thus, the project protocol was to assess evidence from recent applicable systematic reviews and to abstract studies published thereafter. Data were not abstracted from individual studies included in prior SRs; we assumed the data presented in the SRs were abstracted accurately.

Limitations of the Evidence Base

The literature that addresses the diagnosis of celiac disease has numerous limitations that make it difficult to draw firm conclusions. These limitations can be divided into three categories: study volume, design, and reporting quality.

Volume

We identified many studies on the accuracy of tTG and EmA screens in symptomatic adults and children, including several recent systematic reviews. There were fewer studies of DGP antibody tests, as this diagnostic method is relatively new. There were also few studies assessing the accuracy of using algorithms such as those suggested by the most recent NICE and ESPGHAN guidelines.

No studies stratified accuracy results by race, ethnicity, or SES. Several studies in non-Caucasian populations were identified; however, these were not U.S. studies, and results may not be generalizable to populations in the U.S. We identified no studies of diagnostic accuracy in persons with Turner’s syndrome of Trisomy 21. Literature was sparse on other populations of interest; several studies of accuracy in patients with Type 1 diabetes, iron deficiency anemia, or IgA deficiency were identified.

Almost no studies examined the impact of diagnostic method on decisionmaking or clinical or patient centered outcomes. Although the impact of living with undiagnosed celiac disease is well documented, very few studies report outcomes of individuals who initially receive false positive or false negative results.

Design

Diagnostic accuracy is generally assessed through case-control and cohort studies; we included both designs. In studies employing a case-control design, a group of patients with known disease and a different group known not to have the disease undergo both the “index” test and the reference standard. Researchers are blinded to initial disease status. In a cohort design, a group of patients suspected of having the disease (but without a confirmed diagnosis) undergo both diagnostic methods. In a cohort design, the group is defined based on symptoms, while in a
case-control design, the group is based on disease status. The latter design is more subject to bias.

We used the QUADAS-2 instrument to assess the quality of studies of diagnostic accuracy. The ratings for each QUADAS item for each study are presented in the Evidence Tables (Appendix C); case control studies are identified. Strengths and weaknesses of individual studies are discussed in the results section of this report and taken into consideration in rating the strength of the evidence.

To lessen bias, the decision to perform the reference standard should ideally be independent of the results of the test being studied. Thus, we included only studies where all patients underwent both tests. Many studies were identified where patients first underwent serological testing and only those who tested positive underwent biopsy; although these studies provide data on false positives, they were excluded. In addition, to increase generalizability, we included only studies that enrolled a random or consecutive sample.

The use of biopsy results as the reference standard also presents concerns. As discussed in the results for Key Question 2, inter-rater reliability of interpretation is higher at academic centers than community settings. Most of the published accuracy studies included in this review took place in an academic setting.

Regarding comparative accuracy, 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.

Finally, most of the prior SRs described in this report were of moderate quality. Strength of evidence (SOE) was not rated by the authors; we took the strengths and weaknesses of these SRs (as we assessed using AMSTAR) into consideration when we graded the SOE of the body of evidence. An additional item we considered regarding prior SRs was the method of pooling sensitivity and specificity; pooling both jointly in a bivariate model is recommended.

**Reporting Quality**

Failure to report important study design details in publications is a further limitation. Some accuracy studies were vague regarding blinding of assessors and the time lapse between implementation of the index test and reference standard. Data on these items were abstracted as part of QUADAS-2 and are displayed in the Evidence Tables. Such weaknesses are discussed in the Results section and were taken into consideration in rating the strength of evidence.

**Research Gaps**

Although there is high strength of evidence of the accuracy of various serologic tests for celiac disease in symptomatic individuals, strength of evidence on the accuracy of algorithms such as recommended by organizations such as ESPGHAN is insufficient due to the small number of studies and inconsistent results. Appendix F contains details on the test combinations, populations, and the strength of evidence domains for each algorithm studied. Further studies should be conducted.

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

Due to the inherent invasive nature of biopsy, the vast majority of studies of serologic test accuracy using biopsy as the reference standard have been conducted in patients presenting for
testing due to symptoms. The most common symptoms are gastrointestinal (diarrhea, constipation, pain, etc.) as well as 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 evidence of accuracy in the asymptomatic general population is low because the study has never been replicated. This does not mean the tests are inaccurate in asymptomatic individuals; lack of evidence does not equal evidence of inaccuracy.

No studies addressing the key subquestion “What impact does the method of initial diagnosis have on how a physician follows up with a patient?” were identified. Retrospective analyses of existing databases may shed light in this area.

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

Conclusions

New evidence on accuracy of tests used to diagnosis celiac disease supports the high sensitivity of IgA tTG tests and high specificity of both IgA tTG and IgA EmA tests reported in prior SRs. Regarding comparative accuracy, IgA EmA tests have lower sensitivity but equal 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; however, strength of evidence is high, given the large number of studies, the consistency of results, and the precision of the confidence intervals.

High strength of evidence of accuracy, particularly in children, was found for DGP tests in recent SRs. Algorithms combining tTG with either EmA or DGP tests appear to be accurate in both children and adults. Adding an EmA test to a tTG test resulted in increased specificity, with either no change or a slight decrease in sensitivity. In contrast, adding a DGP test to a tTG test resulted in increased sensitivity but decreased specificity. However, strength of evidence 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 if a patient demonstrates a tTG result greater than (10x) the normal limit, the patient should then undergo an EmA test and HLA typing; if the patient tests positive, then responds to gluten exclusion diet, a diagnosis of celiac disease can be made without use of biopsy. These guidelines have not been adopted by societies in the U.S. Evidence seems to support that a multiple-testing strategy without biopsy is accurate; however, additional studies are needed to confirm the test threshold levels that would optimize accuracy for general and specific populations.

VCE is a safe and fairly accurate means of diagnosing celiac disease in adults who wish to avoid biopsy; risk of retaining the capsule is approximately 4.6%. However, our pooled results reveal that 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 settings.

Importantly, few applicable studies on the sequelae of false positive or false negative diagnoses were identified. Long-term follow-up of patients, regardless of diagnostic outcomes, should be encouraged.
References


Abbreviations/Acronyms

ACG - American College of Gastroenterology
AEs - Adverse Events
AGA - Anti-Gliadin Antibodies
AHRQ - Agency for Healthcare Research and Quality
CD - Celiac Disease
CI - Confidence Interval
COI - Conflict of Interest
DGP - Deamidated Gliadin Peptide (DGP) Antibodies
ELISA - Enzyme-Linked Immunosorbent Assay
EmA - Endomysial antibodies
EPC - Southern California Evidence-based Practice Center
ESPGHAN - European Society for Pediatric Gastroenterology, Hepatology, and Nutrition
FN - False Negative
FP - False Positive
GI - Gastrointestinal
HLA - Human Leukocyte Antigen
IBS - Irritable Bowel Syndrome
IgA - Immunoglobulin A
IgG - Immunoglobulin G
IU - International Unit
KI - Key Informants
KQ - Key Question
LR - Likelihood Ratio
NA - Not Applicable
NASPGHAN - North American Society for Pediatric Gastroenterology, Hepatology, and Nutrition
NICE - National Institute for Clinical Excellence
NPV - Negative Predictive Value
NR - Not Reported
PICOTS - Populations, Interventions, Comparators, Outcomes, and Timing
PPV - Positive Predictive Value
QOL - Quality of life
RBA - Radio blinding assay
ROC - Receiver Operating Characteristic
SES - Socioeconomic status
SR - Systematic Review
tTG - Tissue Transglutaminase
VCE - Video Capsule Endoscopy
WGO - World Gastroenterology Organization
Appendix A. Search Strategy

KQ1 (DIAGNOSTIC METHODS):

SEARCH #1 (DIAGNOSTIC ACCURACY)
DATABASE SEARCHED & TIME PERIOD COVERED:
PubMed – 1/1/2010-1/07/2015

LANGUAGE:
English

SEARCH STRATEGY:
"celiac disease"[Mesh] OR "celiac disease"[tiab] OR "coeliac disease"[tiab]
AND
diagnosis OR diagnoses OR diagnostic OR diagnose OR diagnosing OR endomysial OR
transglutaminase* OR serolog* OR antibody OR antibodies OR leucocyte* OR hla OR biopsy OR
biopsies OR test OR tests OR testing OR screen OR screening OR screened OR
"Transglutaminases"[Mesh] OR "HLA Antigens"[Mesh] OR iga OR ttg OR dgp OR IGG OR mass
screening OR diagnosis[mh] OR biopsy[mh] OR biopsies[mh] OR "deamidated gliadin peptide
antibodies" OR Human leukocyte antigen* OR "video capsule endoscopy" OR endoscop*
AND
reactions[mh] OR false positive* OR False negative reactions[mh] OR False negative* OR Predictive
value OR predictive value of tests[mh] OR Distinguish* OR Differential* OR Identif* OR Detect* OR
valid* OR reliab* OR reproducibility of results

SEARCH #2 (INTERMEDIATE OUTCOMES):
DATABASE SEARCHED & TIME PERIOD COVERED:
PubMed – 1/1/2010-1/07/2015

LANGUAGE:
English

SEARCH STRATEGY:
"celiac disease"[Mesh] OR "celiac disease"[tiab] OR "coeliac disease"[tiab]
AND
diagnosis OR diagnoses OR diagnostic OR diagnose OR diagnosing OR endomysial OR
transglutaminase* OR serolog* OR antibody OR antibodies OR leucocyte* OR hla OR biopsy OR
biopsies OR test OR tests OR testing OR screen OR screening OR screened OR
"Transglutaminases"[Mesh] OR "HLA Antigens"[Mesh] OR iga OR ttg OR dgp OR IGG OR mass
screening OR diagnosis[mh] OR biopsy[mh] OR biopsies[mh] OR "deamidated gliadin peptide
antibodies" OR Human leukocyte antigen* OR "video capsule endoscopy" OR endoscop*
AND
intermediate outcome* OR decision* OR dietary OR diet OR nutrition* OR eating OR food OR foods
OR compliance OR comply OR complying OR patient compliance OR adherence OR (("Decision
Making"[Mesh]) OR "Decision Support Systems, Clinical"[Mesh]) OR "Food Habits"[Mesh]
SEARCH #3 (CLINICAL OUTCOMES/COMPLICATIONS):
DATABASE SEARCHED & TIME PERIOD COVERED:
PubMed – 1/1/2010-1/07/2015

LANGUAGE:
English

SEARCH STRATEGY:
"celiac disease"[Mesh] OR "celiac disease"[tiab] OR "coeliac disease"[tiab]
AND
diagnosis OR diagnoses OR diagnostic OR diagnose OR diagnosing OR endomysial OR
transglutaminase* OR serolog* OR antibody OR antibodies OR leucocyte* OR hla OR biopsy OR
biopsies OR test OR tests OR testing OR screen OR screening OR screened OR
"Transglutaminases"[Mesh] OR "HLA Antigens"[Mesh] OR iga OR ttg OR dgp OR IGG OR mass
screening OR diagnosis[mh] OR biopsy[mh] OR biopsies[mh] OR "deamidated gliadin peptide
antibodies" OR Human leukocyte antigen* OR "video capsule endoscopy" OR endoscop*
AND
clinical outcome* OR complication* OR adverse event* OR adverse effect* OR harm* OR enteropathy
OR "quality of life" OR villous atrophy OR abdominal OR anemia OR anemic OR (deficien* AND (folic
acid OR folate)) OR "Outcome and Process Assessment (Health Care)[Mesh] OR "complications"
[Subheading] OR "adverse effects" [Subheading] OR "Quality of Life"[Mesh] OR "Folic Acid
Deficiency"[Mesh]

SEARCH #4 (ADD TERMS “MISDIAGNOS* OR “UNDIAGNOS*”):
DATABASE SEARCHED & TIME PERIOD COVERED:
PubMed – 1/1/2010-1/07/2015

LANGUAGE:
English

SEARCH STRATEGY:
"celiac disease"[Mesh] OR "celiac disease"[tiab] OR "coeliac disease"[tiab]
AND
misdiagnos* OR undiagnos*

NOTE – THESE RESULTS WERE INCORPORATED INTO THE PREVIOUS RESULT SETS

SEARCH #1 (DIAGNOSTIC ACCURACY):
DATABASE SEARCHED & TIME PERIOD COVERED:

LANGUAGE:
English

SEARCH STRATEGY: ("ts"= topical search)
ts=("celiac disease" OR "coeliac disease")
AND
ts=(diagnosis OR diagnoses OR diagnostic OR diagnose OR diagnosing OR endomysial OR 
transglutaminase* OR serolog* OR antibody OR antibodies OR leucocyte* OR hla OR biopsy OR 
biopsies OR test OR tests OR testing OR screen OR screening OR screened OR Transglutaminase* OR 
iga OR ttg OR dgp OR IGG OR mass screening OR "Human leukocyte antigen*" OR endoscop* OR 
misdiagnos* OR undiagnos*)
AND
ts=(Accura* OR Sensitivity OR Specificity OR false positive* OR False negative* OR Predictive value 
OR Distinguish* OR Differential* OR Identif* OR Detect* OR valid* OR reliab* OR reproducib*)

SEARCH #2 (INTERMEDIATE OUTCOMES):
DATABASE SEARCHED & TIME PERIOD COVERED:

LANGUAGE:
English

SEARCH STRATEGY:
ts=("celiac disease" OR "coeliac disease")
AND
ts=(diagnosis OR diagnoses OR diagnostic OR diagnose OR diagnosing OR endomysial OR 
transglutaminase* OR serolog* OR antibody OR antibodies OR leucocyte* OR hla OR biopsy OR 
biopsies OR test OR tests OR testing OR screen OR screening OR screened OR Transglutaminase* OR 
iga OR ttg OR dgp OR IGG OR mass screening OR "Human leukocyte antigen*" OR endoscop* OR 
missdiagnos* OR undiagnos*)
AND
ts=(intermediate outcome* OR decision* OR dietary OR diet OR nutrition* OR eating OR food OR foods 
OR compliance OR comply OR complying OR adherence)

SEARCH #3 (CLINICAL OUTCOMES/COMPLICATIONS):
DATABASE SEARCHED & TIME PERIOD COVERED:

LANGUAGE:
English

SEARCH STRATEGY:
ts=("celiac disease" OR "coeliac disease")
AND
ts=(diagnosis OR diagnoses OR diagnostic OR diagnose OR diagnosing OR endomysial OR 
transglutaminase* OR serolog* OR antibody OR antibodies OR leucocyte* OR hla OR biopsy OR 
biopsies OR test OR tests OR testing OR screen OR screening OR screened OR Transglutaminase* OR 
iga OR ttg OR dgp OR IGG OR mass screening OR "Human leukocyte antigen*" OR endoscop* OR 
misdiagnos* OR undiagnos*)
AND
ts=("adverse effect**" OR "adverse event**" OR "clinical outcome**" OR complication* OR harm* OR enteropathy OR "quality of life" OR villous atrophy OR abdominal OR anemia OR anemic OR "folic acid" OR folate)

DATABASE SEARCHED & TIME PERIOD COVERED:

Cochrane Databases – 1/1/2010-1/07/2015

LANGUAGE:

English

SEARCH STRATEGY:
"celiac disease" OR "coeliac disease" in Title, Abstract, Keywords

NUMBER OF RESULTS: 65

By database:
Cochrane Reviews (0)
Other Reviews (5)
Trials (54)
Methods Studies (0)
Technology Assessments (3)
Economic Evaluations (3)
Cochrane Groups (0)

KQ2 (ENDOSCOPY WITH DUODENAL BIOPSY)

DATABASE SEARCHED & TIME PERIOD COVERED:
PubMed – 1/1/1990-1/07/2015

LANGUAGE:

English

SEARCH STRATEGY:
"celiac disease "[Mesh] OR "celiac disease"[tiab] OR "coeliac disease"[tiab]
AND
endoscopy
AND
duodenal or duodenum
AND
biopsy OR biopsies
SEARCH STRATEGY:
TOPIC: ("celiac disease" OR "coeliac disease")
AND
TOPIC: (endoscop* AND (duodenal or duodenum) AND (biopsy OR biopsies))

KQ3 (POPULATION):
DATABASE SEARCHED & TIME PERIOD COVERED:
PubMed – 1/1/1990-1/07/2015

LANGUAGE:
English

SEARCH STRATEGY:
"celiac disease "[Mesh] OR "celiac disease"[tiab] OR "coeliac disease"[tiab]
AND
diagnosis OR diagnoses OR diagnostic OR diagnose OR diagnosing OR endomysial OR
transglutaminase* OR serolog* OR antibody OR antibodies OR leucocyte* OR hla OR biopsy OR
biopsies OR test OR tests OR testing OR screen OR screening OR screened OR
"Transglutaminases"[Mesh] OR "HLA Antigens"[Mesh] OR iga OR ttg OR dgp OR IGG OR mass
screening OR diagnosis[mh] OR biopsy[mh] OR biopsies[mh] OR "deamidated gliadin peptide
antibodies" OR Human leukocyte antigen* OR "video capsule endoscopy" OR endoscop* OR
misdiagnos* OR undiagnos*
AND
"Continental Population Groups"[Mesh] OR "Demography"[Mesh] OR population* OR symptomatic OR
 nonsymptomatic OR non-symptomatic OR child OR children OR infant OR infants OR pediatric* OR
 paediatric* OR demograph* OR race OR racial OR ethnic OR ethnicity* OR minority OR minorities OR
 genetic* OR geograph* OR region OR regions OR regional OR socioeconom* OR socio-econom* OR
economic* OR income OR (iga AND deficien*) OR negative OR country[tiab] OR countries[tiab] OR
(prevalence OR prevalen*[tiab])
AND
"outcome assessment health care"[MeSH Terms] OR Accura* OR Sensitivity and specificity[mh] OR
Sensitivity[tiab] OR Specificity[tiab] OR False positive reactions[mh] OR false positive* OR False
negative reactions[mh] OR False negative* OR Predictive value OR predictive value of tests[mh] OR
Distinguish* OR Differential* OR Identif* OR Detect* OR valid* OR reliab* OR reproducibility of
results OR outcome OR outcomes OR treatment outcome OR treatment outcomes
NOT
case report* OR case reports[pt])

DATABASE SEARCHED & TIME PERIOD COVERED:

LANGUAGE:
English
SEARCH STRATEGY:

\( ts=\("\text{celiac disease}\) OR \"\text{coeliac disease}\) \)

AND

\( ts=\) ( diagnosis OR diagnoses OR diagnostic OR diagnose OR diagnosing OR endomysial OR transglutaminase* OR serolog* OR antibody OR antibodies OR leucocyte* OR hla OR biopsy OR biopsies OR test OR tests OR testing OR screen OR screening OR screened OR Transglutaminase* OR iga OR ttg OR dgp OR IGG OR mass screening OR "Human leukocyte antigen*" OR endoscop* OR misdiagnos* OR undiagnos*)

AND

\( ts=\) (population* OR symptomatic OR nonsymptomatic OR non-symptomatic OR child OR children OR infant OR infants OR pediatric* OR paediatric* OR demograph* OR race OR racial OR ethnic OR ethnicit* OR minority OR minorities OR genetic* OR geograph* OR region OR regions OR regional OR socioeconom* OR socio-econom* OR economic* OR income OR (iga AND deficien*) OR negative OR country OR countries)

AND

\( ts=\) (Accura* OR Sensitivity OR Specificity OR false positive* OR False negative* OR "predictive value" OR Distinguish* OR Differential* OR Identif* OR Detect* OR valid* OR reliab* OR reproducib* OR outcome* OR prevalen*)

===================================================================== KQ4 (ADVERSE EVENTS):

DATABASE SEARCHED & TIME PERIOD COVERED:

PubMed – 1/1/2003-1/07/2015

LANGUAGE:

English

SEARCH STRATEGY:

"celiac disease "[Mesh] OR "celiac disease"[tiab] OR "coeliac disease"[tiab]

AND

diagnosis OR diagnoses OR diagnostic OR diagnose OR diagnosing OR endomysial OR transglutaminase* OR serolog* OR antibody OR antibodies OR leucocyte* OR hla OR biopsy OR biopsies OR test OR tests OR testing OR screen OR screening OR screened OR "Transglutaminases"[Mesh] OR "HLA Antigens"[Mesh] OR iga OR ttg OR dgp OR IGG OR mass screening OR diagnosis[mh] OR biopsy[mh] OR biopsies[mh] OR "deamidated gliadin peptide antibodies" OR Human leukocyte antigen* OR "video capsule endoscopy" OR endoscop* OR misdiagnos* OR undiagnos*

AND

adverse effect* OR adverse event* OR harm* OR bleeding OR perforat* OR danger* OR safe*[tiab] OR safety[tiab] OR patient safety OR accident*

======================================================================

DATABASE SEARCHED & TIME PERIOD COVERED:


LANGUAGE:

English

SEARCH STRATEGY:

ts="\text{celiac disease}\) OR \"\text{coeliac disease}\)"}
AND
ts=(diagnosis OR diagnoses OR diagnostic OR diagnose OR diagnosing OR endomysial OR transglutaminase* OR serolog* OR antibody OR antibodies OR leucocyte* OR hla OR biopsy OR biopsies OR test OR tests OR testing OR screen OR screening OR screened OR Transglutaminase* OR iga OR ttg OR dgp OR IGG OR mass screening OR "Human leukocyte antigen*" OR endoscop* OR misdiagnos* OR undiagnos*)

AND
ts=(adverse OR harm* OR danger* OR bleed* OR perforat* OR OR safe* OR accident*)
Appendix B. List of Excluded Studies

Not English – N=1

Not Human – N=1

Not About Celiac Disease (CD) - N=12


**Not About Diagnostic Tests - N=150**


No Original Data - N=20


**Individual Case Report - N=19**


**Prevalence Only - N=41**


**Diagnostic Method Outside the Scope of Study - N=155**


Test Processing Issue - N=20


17. Ruiz-Ortiz E, Montraveta M, Cabre E, et al. HLA-DQ2/DQ8 and HLA-DQB1*02 homozygosity typing by real-time polymerase chain reaction for the assessment of celiac disease


Included in a Prior Systematic Review on Topic - N=7


Does Not Assess Accuracy or Effectiveness - N=133

B-35


**Index Test Not Compared to Biopsy - N=62**


Not All Subjects Underwent Both Index Test and Reference Standard - N=410


115. Fabiani E, Catassi C. The serum IgA class anti-tissue transglutaminase antibodies in the
diagnosis and follow up of coeliac disease. Results of an international multi-centre study.
PMID: 11434591.


117. Ferfoglia G, Pulitano R, Sategna-Guidetti C. Do dietary antibodies still play a role in the
diagnosis and follow-up of coeliac disease? A comparison among different serological tests.

118. Fernandez-Banares F, Esteve M, Farre C, et al. Predisposing HLA-DQ2 and HLA-DQ8
haplotypes of coeliac disease and associated enteropathy in microscopic colitis. European
Journal of Gastroenterology & Hepatology. 2005 Dec;17(12):1333-8. PMID:
WOS:000233798100010.

119. Fine KD, Do K, Schulte K, et al. High prevalence of celiac sprue-like HLA-DQ genes and
enteropathy in patients with the microscopic colitis syndrome. Am J Gastroenterol. 2000


cirrhosis and of antimitochondrial antibodies in adult coeliac disease patients in Italy. Dig Liver


124. Fraser-Reynolds KA, Butzner JD, Stephure DK, et al. Use of immunoglobulin A-
antiendomysial antibody to screen for celiac disease in North American children with type 1

125. Freeman HJ. Strongly positive tissue transglutaminase antibody assays without celiac

126. Frost AR, Band MM, Conway GS. Serological screening for coeliac disease in adults with
Turner's syndrome: prevalence and clinical significance of endomysium antibody positivity. Eur


225. Mastrandrea F, Semeraro FP, Coradduzza G, et al. CD34+ hemopoietic precursor and stem cells traffic in peripheral blood of celiac patients is significantly increased but not directly related


377. Unsworth DJ, Brown DL. Serological screening suggests that adult coeliac disease is underdiagnosed in the UK and increases the incidence by up to 12%. Gut. 1994 Jan;35(1):61-4. PMID: 8307451.


No or Unclear on Consecutive or Random Sample - N=43


Sample Size Less Than 300 and Not a Special Population - N=83


# Appendix C. Evidence Table

## Table C-1. Studies of serology accuracy

<table>
<thead>
<tr>
<th>Author, Year</th>
<th>Number of Participants, Populations</th>
<th>Type of Diagnostic Test, Cut-Off Value</th>
<th>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</th>
<th>QUADAS</th>
</tr>
</thead>
</table>
| Barada et al., 2014<sup>31</sup> | Number of Participants: 999 Adults | Type of Diagnostic Test: EMA IgA Cut-off value: NR | Sensitivity: 72.2% Specificity: 99.7% Positive predictive value: 90 Negative predictive value: 99.2 | QUADAS Domain 1  
Consecutive or random sample: Yes  
Case control design avoided: Yes  
Inappropriate exclusions: Yes  
Biased patient Selection: Low |
| | | Type of Diagnostic Test: tTG IgA Cut-off value: NR | Sensitivity: 72.2% Specificity: 98.4% Positive predictive value: 44.8 Negative predictive value: 99.5 | QUADAS Domain 2  
Blinded interpretation of index test results: Yes  
Prespecified test threshold: Yes  
Bias due to testing: Low |
| | | Type of Diagnostic Test: Combined screen tTG IgA, DGP IgA Cut-off value: NR | Sensitivity: 72.2% Specificity: 97.4% Positive predictive value: 34.2 Negative predictive value: 99.5 | QUADAS Domain 3  
Valid reference standard: Yes  
Blinded analysis of reference test: Yes  
Bias due to reference test: Low |
| | | | | QUADAS Domain 4  
Appropriate interval between reference and index test: Yes  
All patients received reference test: Yes  
All patients received same test: Yes  
All patients included analysis: Yes  
Could patient flow have introduced bias: Not Applicable |

Comments: Marsh 2 & 3 were considered celiac. In addition, authors classified 1 person with Marsh 1 and positive EMA as celiac.
<table>
<thead>
<tr>
<th>Author, Year</th>
<th>Number of Participants, Populations</th>
<th>Type of Diagnostic Test, Cut-Off Value</th>
<th>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</th>
<th>QUADAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basso et al., 2011</td>
<td>Number of Participants: 703 Adults</td>
<td>Type of Diagnostic Test: tTG IgA Cut-off value: 100 U/mL</td>
<td>Sensitivity: 75.7% Specificity: 100% Positive predictive value: 100 Negative predictive value: 82.4</td>
<td>QUADAS Domain 1 Consecutive or random sample: Yes Case control design avoided: No Inappropriate exclusions: Yes Biased patient Selection: Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type of Diagnostic Test: tTG IgA Cut-off value: 17.5 U/mL</td>
<td>Sensitivity: 94.5% Specificity: 97.1% Positive predictive value: 96.6 Negative predictive value: 95.3</td>
<td>QUADAS Domain 2 Blinded interpretation of index test results: Unclear Prespecified test threshold: Yes Bias due to testing: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type of Diagnostic Test: tTG IgA Cut-off value: 20 U</td>
<td>Sensitivity: 94.2% Specificity: 97.3% Positive predictive value: 96.9 Negative predictive value: 95</td>
<td>QUADAS Domain 2 Blinded interpretation of index test results: Unclear Prespecified test threshold: Yes Bias due to testing: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type of Diagnostic Test: tTG IgA Cut-off value: 24 U/mL</td>
<td>Sensitivity: 96.3% Specificity: 81.3% Positive predictive value: 81.9 Negative predictive value: 96.2</td>
<td>QUADAS Domain 3 Valid reference standard: Yes Blinded analysis of reference test: Unclear Bias due to reference test: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type of Diagnostic Test: tTG IgA Cut-off value: 75.6 U/mL</td>
<td>Sensitivity: 90.9% Specificity: 96.5% Positive predictive value: 95.8 Negative predictive value: 92.3</td>
<td>QUADAS Domain 4 Appropriate interval between reference and index test: Yes All patients received reference test: Yes All patients received same test: Yes All patients included analysis: No Could patient flow have introduced bias: Not Applicable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type of Diagnostic Test: tTG IgA Cut-off value: 909.3 U/mL</td>
<td>Sensitivity: 62.6% Specificity: 100% Positive predictive value: 100 Negative predictive value: 75.2</td>
<td>QUADAS Domain 4 Appropriate interval between reference and index test: Yes All patients received reference test: Yes All patients received same test: Yes All patients included analysis: No Could patient flow have introduced bias: Not Applicable</td>
</tr>
<tr>
<td>Author, Year</td>
<td>Number of Participants, Populations</td>
<td>Type of Diagnostic Test, Cut-Off Value</td>
<td>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</td>
<td>QUADAS</td>
</tr>
<tr>
<td>-------------------</td>
<td>------------------------------------</td>
<td>----------------------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td>Basso et al., 2011</td>
<td>Number of Participants: 703 Adults</td>
<td>tTG IgA, DGP IgA Cut-off value: 145 U</td>
<td>Sensitivity: 65.3% Specificity: 100% Positive predictive value: 100 Negative predictive value: 76.6</td>
<td>QUADAS Domain 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>tTG IgA Cut-off value: 20 U</td>
<td>Sensitivity: 96.7% Specificity: 89.8% Positive predictive value: 89.3 Negative predictive value: 96.8</td>
<td>Consecutive or random sample: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>tTG IgA Cut-off value: 32 U</td>
<td>Sensitivity: 95.4% Specificity: 95.7% Positive predictive value: 95.2 Negative predictive value: 96</td>
<td>Case control design avoided: No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>tTG IgG Cut-off value: 20 U/mL</td>
<td>Sensitivity: 96.7% Specificity: 83.4% Positive predictive value: 83.7 Negative predictive value: 96.6</td>
<td>Inappropriate exclusions: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>tTG IgG Cut-off value: 47.6 U/mL</td>
<td>Sensitivity: 93.3% Specificity: 94.1% Positive predictive value: 93.3 Negative predictive value: 94.1</td>
<td>Biased patient Selection: Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td>tTG IgG Cut-off value: 976.8 U/mL</td>
<td>Sensitivity: 59.6% Specificity: 100% Positive predictive value: 100 Negative predictive value: 73.8</td>
<td>QUADAS Domain 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Prespecified test threshold: Yes Bias due to testing: Unclear</td>
<td>Blinded interpretation of index test results: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 3 Valid reference standard: Yes Blinded analysis of reference test: Unclear</td>
<td>QUADAS Domain 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Bias due to reference test: Unclear</td>
<td>Appropriate interval between reference and index test: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 4 All patients received reference test: Yes</td>
<td>All patients received same test: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>All patients included analysis: No</td>
<td>All patients included analysis: No</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Could patient flow have introduced bias: Not Applicable</td>
<td>Could patient flow have introduced bias: Not Applicable</td>
</tr>
<tr>
<td>Author, Year</td>
<td>Number of Participants, Populations</td>
<td>Type of Diagnostic Test, Cut-Off Value</td>
<td>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</td>
<td>QUADAS</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------------------------</td>
<td>----------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>Bienvenu et al., 2014</td>
<td>45 Number of Participants: 45 Population: Selective IgA deficient children</td>
<td>Type of Diagnostic Test: CD-LFIA (detects both human IgA and IgG anti-DGP) Cut-off value: NA</td>
<td>Sensitivity: 100.0% Specificity: 89.2% Negative predictive value: 100.0%</td>
<td>QUADAS Domain 1 Consecutive or random sample: Yes Case control design: Yes Inappropriate exclusions: Yes Biased patient Selection: Low QUADAS Domain 2 Blinded interpretation of index test results: Yes Prespecified test threshold: Yes Bias due to testing: Low QUADAS Domain 3 Valid reference standard: Yes Blinded analysis of reference test: Yes Bias due to reference test: Low QUADAS Domain 4 Appropriate interval between reference and index test: Unclear All patients received reference test: Unclear All patients received same test: Unclear All patients included analysis: Yes Could patient flow have introduced bias: High</td>
</tr>
<tr>
<td>Author, Year</td>
<td>Number of Participants, Populations</td>
<td>Type of Diagnostic Test, Cut-Off Value</td>
<td>Outcomes</td>
<td>QUADAS</td>
</tr>
<tr>
<td>-----------------</td>
<td>-----------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------</td>
<td>-----------</td>
<td>--------------------------------------------</td>
</tr>
<tr>
<td>Cekin et al., 2012</td>
<td>Number of Participants: 84 Adults with Iron Deficiency</td>
<td>Type of Diagnostic Test: EMA IgA</td>
<td>Sensitivity: 100%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type of Diagnostic Test: EMA IgG</td>
<td>Specificity: 98.72%</td>
<td>QUADAS Domain 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Positive predictive value: 85.71</td>
<td>Consecutive or random sample: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Negative predictive value: 100</td>
<td>Case control design avoided: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sensitivity: 33.33%</td>
<td>Inappropriate exclusions: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Specificity: 96.15%</td>
<td>Biased patient Selection: Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Positive predictive value: 40</td>
<td>QUADAS Domain 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Negative predictive value: 94.94</td>
<td>Blinded interpretation of index test results: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Prespecified test threshold: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bias due to testing: Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Valid reference standard: Yes</td>
<td>Blinded analysis of reference test: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bias due to reference test: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Appropriate interval between reference and index test: Unclear</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>All patients received reference test: Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>All patients received same test: Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>All patients included analysis: Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Could patient flow have introduced bias: Unclear</td>
<td></td>
</tr>
<tr>
<td>Author, Year</td>
<td>Number of Participants, Populations</td>
<td>Type of Diagnostic Test, Cut-Off Value</td>
<td>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</td>
<td>QUADAS</td>
</tr>
<tr>
<td>----------------------</td>
<td>------------------------------------</td>
<td>----------------------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>Dahlbom et al., 2010$^{55}$</td>
<td>Number of Participants: 301 Children and Adults</td>
<td>Type of Diagnostic Test: tTG IgA Cut-off value: &gt;3 U m/L</td>
<td>Sensitivity: 100% Specificity: 99.24% Positive predictive value: 99.42 Negative predictive value: 100</td>
<td>QUADAS Domain 1 Consecutive or random sample: Yes Case control design avoided: No Inappropriate exclusions avoided: Not Applicable Biased patient Selection: High</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type of Diagnostic Test: tTG IgG Cut-off value: &gt;3 U m/L</td>
<td>Sensitivity: 84.12% Specificity: 98.47% Positive predictive value: 98.62 Negative predictive value: 82.69</td>
<td>QUADAS Domain 2 Blinded interpretation of index test results: Unclear Prespecified test threshold: Yes Bias due to testing: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 4 Appropriate interval between reference and index test: Yes All patients received reference test: Yes All patients received same test: Yes All patients included analysis: Yes Could patient flow have introduced bias: Not Applicable</td>
</tr>
<tr>
<td>Author, Year</td>
<td>Number of Participants, Populations</td>
<td>Type of Diagnostic Test, Cut-Off Value</td>
<td>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</td>
<td>QUADAS</td>
</tr>
<tr>
<td>------------------------</td>
<td>------------------------------------</td>
<td>----------------------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Dahle et al., 2010</td>
<td>Number of Participants: 176 Adults</td>
<td>Type of Diagnostic Test: EMA IgA</td>
<td>Sensitivity: 61% Specificity: 100%</td>
<td>QUADAS Domain 1 Consecutive or random sample: Yes Case control design avoided: Yes Inappropriate exclusions avoided: Yes Biased patient Selection: Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cut-off value: Serum dilution 1/5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type of Diagnostic Test: tTG IgA</td>
<td>Sensitivity: 76% Specificity: 95%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cut-off value: 5 U/mL</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type of Diagnostic Test: DGP IgA or DGP IgG Cut-off value: 20 AU/mL</td>
<td>Sensitivity: 87% Specificity: 96%</td>
<td>QUADAS Domain 2 Blinded interpretation of index test results: Unclear Prespecified test threshold: Yes Bias due to testing: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type of Diagnostic Test: tTG IgG or IgA combined with DGP IgG or IgA Cut-off value: 20 AU/mL</td>
<td>Sensitivity: 91% Specificity: 80%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type of Diagnostic Test: tTG IgG or IgA combined with DGP IgG or IgA Cut-off value: 35 AU/mL</td>
<td>Sensitivity: 85% Specificity: 98%</td>
<td>QUADAS Domain 3 Valid reference standard: Yes Blinded analysis of reference test: Yes Bias due to reference test: Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

C-7
<table>
<thead>
<tr>
<th>Author, Year</th>
<th>Number of Participants, Populations</th>
<th>Type of Diagnostic Test, Cut-Off Value</th>
<th>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</th>
<th>QUADAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeGaetani et al., 2013</td>
<td>Number of Participants: 59 Adults with prior negative serology but villous atrophy. HLA test was used to rule out celiac disease.</td>
<td>Type of Diagnostic Test: HLA DQ2, HLA DQ2</td>
<td>Sensitivity: 100% Specificity: 18.18% Positive predictive value: 29.41 Negative predictive value: 100</td>
<td>QUADAS Domain 1 Consecutive or random sample: Yes Case control design avoided: Yes Inappropriate exclusions: No Biased patient Selection: High QUADAS Domain 2 Blinded interpretation of index test results: Unclear Prespecified test threshold: Unclear Bias due to testing: Unclear QUADAS Domain 3 Valid reference standard: Yes Blinded analysis of reference test: Unclear Bias due to reference test: Unclear QUADAS Domain 4 Appropriate interval between reference and index test: Unclear All patients received reference test: Yes All patients received same test: Yes All patients included analysis: Yes Could patient flow have introduced bias: Unclear</td>
</tr>
<tr>
<td>Author, Year</td>
<td>Number of Participants, Populations</td>
<td>Type of Diagnostic Test, Cut-Off Value</td>
<td>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</td>
<td>QUADAS</td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------------------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------</td>
</tr>
<tr>
<td>Dutta et al., 2010*</td>
<td>Number of Participants: 92 symptomatic adults in India</td>
<td>Type of Diagnostic Test: tTG IgG Cut-off value: &gt;15 U/mL</td>
<td>Sensitivity: 77.8% Specificity: 89.1% Positive predictive value: 63.6 Negative predictive value: 94.2</td>
<td>QUADAS Domain 1: Consecutive or random sample: Yes</td>
</tr>
<tr>
<td></td>
<td>Comment: Unclear why tTG IgG test was used</td>
<td></td>
<td></td>
<td>Case control design avoided: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Inappropriate exclusions: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Biased patient Selection: Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 2: Blinded interpretation of index test results: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Prespecified test threshold: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bias due to testing: Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 3: Valid reference standard: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Blinded analysis of reference test: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bias due to reference test: Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 4: Appropriate interval between reference and index test: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>All patients received reference test: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>All patients received same test: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>All patients included analysis: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Could patient flow have introduced bias: Not Applicable</td>
</tr>
<tr>
<td>Author, Year</td>
<td>Number of Participants, Populations</td>
<td>Type of Diagnostic Test, Cut-Off Value</td>
<td>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</td>
<td>QUADAS</td>
</tr>
<tr>
<td>----------------------</td>
<td>---------------------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| Emami et al., 2012   | Number of Participants: 130 Populations: IgA Deficient adults in Iran                              | Type of Diagnostic Test: tTG IgA Cut-off value: >10 AU/ml | Sensitivity: 38.46% Specificity: 96.58% Positive predictive value: 55.56 Negative predictive value: 93.39 | QUADAS Domain 1  
Consecutive or random sample: Yes  
Case control design avoided: Yes  
Inappropriate exclusions: Yes  
Biased patient Selection: Low  
QUADAS Domain 2  
Blinded interpretation of index test results: Unclear  
Prespecified test threshold: Yes  
Bias due to testing: Unclear  
QUADAS Domain 3  
Valid reference standard: Yes  
Blinded analysis of reference test: Unclear  
Bias due to reference test: Unclear  
QUADAS Domain 4  
Appropriate interval between reference and index test: Unclear  
All patients received reference test: Yes  
All patients received same test: Yes  
All patients included analysis: Yes  
Could patient flow have introduced bias: Unclear |
<table>
<thead>
<tr>
<th>Author, Year</th>
<th>Number of Participants, Populations</th>
<th>Type of Diagnostic Test, Cut-Off Value</th>
<th>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</th>
<th>QUADAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harrison et al., 2013&lt;sup&gt;40&lt;/sup&gt;</td>
<td>Number of Participants: 12,289, age unclear. Some IgA deficient, but number not reported</td>
<td>Type of Diagnostic Test: tTG IgA, Cut-off value: 5 U/mL</td>
<td>Sensitivity: 86.8% Specificity: 99.9%</td>
<td>QUADAS Domain 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type of Diagnostic Test: tTG IgA, tTG IgG Cut-off value: 5 U/mL</td>
<td>Sensitivity: 92.1% Specificity: 99.9%</td>
<td>Consecutive or random sample: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Case control design avoided: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Inappropriate exclusions: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Biased patient Selection: Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Blinded interpretation of index test results: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Prespecified test threshold: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bias due to testing: Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Valid reference standard: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Blinded analysis of reference test: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bias due to reference test: Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Appropriate interval between reference and index test: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>All patients received reference test: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>All patients received same test: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>All patients included analysis: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Could patient flow have introduced bias: Not Applicable</td>
</tr>
<tr>
<td>Author, Year</td>
<td>Number of Participants, Populations</td>
<td>Type of Diagnostic Test, Cut-Off Value</td>
<td>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</td>
<td>QUADAS</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------</td>
<td>---------------------------------------------------------------------------------------</td>
<td>---------------------------------------------</td>
</tr>
<tr>
<td>Kaukinen et al., 1999⁴¹</td>
<td>Number of Participants: 26 Population: Patients with endocrinologic disorders in Finland</td>
<td>Type of Diagnostic Test: HLA DQ2, HLA DQ2</td>
<td>Sensitivity: 100%                       Specificity: 33.33%                     Positive predictive value: 5.26          Negative predictive value: 100</td>
<td>QUADAS Domain 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Consecutive or random sample: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Case control design avoided: No</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Inappropriate exclusions: No</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Biased patient Selection: High</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Blinded interpretation of index test results: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Prespecified test threshold: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bias due to testing: Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Valid reference standard: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Blinded analysis of reference test: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bias due to reference test: Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Appropriate interval between reference and index test: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>All patients received reference test: No</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>All patients received same test: No</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>All patients included analysis: No</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Could patient flow have introduced bias:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>High</td>
</tr>
<tr>
<td>Author, Year</td>
<td>Number of Participants, Populations</td>
<td>Type of Diagnostic Test, Cut-Off Value</td>
<td>Outcomes</td>
<td>QUADAS</td>
</tr>
<tr>
<td>------------------------------</td>
<td>---------------------------------------------</td>
<td>---------------------------------------</td>
<td>----------</td>
<td>--------</td>
</tr>
</tbody>
</table>
| Mansour et al., 2011         | Number of Participants: 62                  | Type of Diagnostic Test: EMA IgA      | Sensitivity: 71.43% Specificity: 96.36% Positive predictive value: 71.43 Negative predictive value: 96.36 | QUADAS Domain 1  
Consecutive or random sample: Yes  
Case control design avoided: Yes  
Inappropriate exclusions: Yes  
Biased patient Selection: Low |
|                              | Population: Type 1 diabetes, Iraq           | Cut-off value: 20 U/mL                |          | QUADAS Domain 2  
Blinded interpretation of index test results: Yes  
Prespecified test threshold: Yes  
Bias due to testing: Low |
|                              |                                             | Type of Diagnostic Test: tTG IgA      | Sensitivity: 71.43% Specificity: 92.73% Positive predictive value: 55.56 Negative predictive value: 96.23 | QUADAS Domain 3  
Valid reference standard: Yes  
Blinded analysis of reference test: Yes  
Bias due to reference test: Low |
|                              |                                             | Cut-off value: 15 U/mL                |          | QUADAS Domain 4  
Appropriate interval between reference and index test: Unclear  
All patients received reference test: Yes  
All patients received same test: Yes  
All patients included analysis: Yes  
Could patient flow have introduced bias: Unclear |
<p>|                              |                                             | Type of Diagnostic Test: tTG IgG      | Sensitivity: 57.14% Specificity: 92.73% Positive predictive value: 50 Negative predictive value: 94.44 |          |</p>
<table>
<thead>
<tr>
<th>Author, Year</th>
<th>Number of Participants, Populations</th>
<th>Type of Diagnostic Test, Cut-Off Value</th>
<th>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</th>
<th>QUADAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mozo et al., 2012</td>
<td>Number of Participants: 200</td>
<td>Type of Diagnostic Test: DGP IgA Cut-off value: &gt;7 U/mL</td>
<td>Sensitivity: 96% Specifcity: 96% Positive predictive value: 96 Negative predictive value: 96</td>
<td>QUADAS Domain 1 Consecutive or random sample: Yes Case control design avoided: No Inappropriate exclusions: No Biased patient Selection: High</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type of Diagnostic Test: DGP IgG Cut-off value: &gt;7 U/mL</td>
<td>Sensitivity: 95% Specificity: 99% Positive predictive value: 98.9 Negative predictive value: 95.2</td>
<td>QUADAS Domain 2 Blinded interpretation of index test results: Yes Prespecified test threshold: Yes Bias due to testing: Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type of Diagnostic Test: tTG IgA Cut-off value: &gt;7 U/mL</td>
<td>Sensitivity: 89% Specificity: 94% Positive predictive value: 93.7 Negative predictive value: 89.5</td>
<td>QUADAS Domain 3 Valid reference standard: Yes Blinded analysis of reference test: Unclear Bias due to reference test: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 4 Appropriate interval between reference and index test: Yes All patients received reference test: No All patients received same test: Not Applicable All patients included analysis: Yes Could patient flow have introduced bias: Unclear</td>
</tr>
<tr>
<td>Author, Year</td>
<td>Number of Participants, Populations</td>
<td>Type of Diagnostic Test, Cut-Off Value</td>
<td>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</td>
<td>QUADAS</td>
</tr>
<tr>
<td>-------------</td>
<td>------------------------------------</td>
<td>----------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>Nevoral et al., 2013⁴⁴</td>
<td>Number of Participants: 345 children and adolescents Number of Participants: 32 first degree relatives Number of Participants: 263 with Marsh 2 or 3 classification Number of Participants: 40 Type 1 diabetes</td>
<td>Type of Diagnostic Test: tTG IgA, EMA IgG Cut-off value: 12 U/mL Type of Diagnostic Test: tTG IgA, EMA IgA Cut-off value: 12 U/mL Type of Diagnostic Test: tTG IgA, EMA IgA Cut-off value: 12 U/mL Type of Diagnostic Test: tTG IgA, EMA IgA Cut-off value: 12 U/mL</td>
<td>Sensitivity: 76% Specificity: 85% Positive predictive value: 94 Negative predictive value: 53 Sensitivity: 81% Specificity: 70% Sensitivity: 83% Specificity: 67% Sensitivity: 93% Specificity: 64%</td>
<td>QUADAS Domain 1 Consecutive or random sample: Yes Case control design avoided: Yes Inappropriate exclusions: Yes Biased patient Selection: Low QUADAS Domain 2 Blinded interpretation of index test results: Yes Prespecified test threshold: Yes Bias due to testing: Low QUADAS Domain 3 Valid reference standard: Yes Blinded analysis of reference test: Yes Bias due to reference test: Low QUADAS Domain 4 Appropriate interval between reference and index test: Yes All patients received reference test: Yes All patients received same test: Yes All patients included analysis: Yes Could patient flow have introduced bias: Not Applicable</td>
</tr>
<tr>
<td>Author, Year</td>
<td>Number of Participants, Populations</td>
<td>Type of Diagnostic Test, Cut-Off Value</td>
<td>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</td>
<td>QUADAS</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------------------------------------------------------------</td>
<td>---------------------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>--------</td>
</tr>
</tbody>
</table>
| Olen et al., 2012  | Number of Participants: 69  
Population: <2 years old  
Number of Participants: 408  
Population: all patients  
Number of Participants: 67  
Population: <2 years old  
Number of Participants: 530  
Population: all patients  
Comments: 93 individuals were excluded from study because the serology analyses had not been carried out at the participating immunology departments. Also, it isn’t clear why some patients did not undergo DGP tests. | Type of Diagnostic Test: DGP IgA  
Cut-off value: NR  
Type of Diagnostic Test: DGP IgA  
Cut-off value: NR  
Type of Diagnostic Test: tTG IgA  
Cut-off value: NR  
Type of Diagnostic Test: tTG IgA  
Cut-off value: NR | Sensitivity: 100%  
Specificity: 31%  
Positive predictive value: 44  
Sensitivity: 91%  
Specificity: 26%  
Positive predictive value: 51  
Sensitivity: 96%  
Specificity: 98%  
Positive predictive value: 96  
Sensitivity: 94%  
Specificity: 86%  
Positive predictive value: 88 | QUADAS Domain 1  
Consecutive or random sample: Yes  
Case control design avoided: Yes  
Inappropriate exclusions: No  
Biased patient Selection: High  
QUADAS Domain 2  
Blinded interpretation of index test results: Yes  
Prespecified test threshold: Yes  
Bias due to testing: Low  
QUADAS Domain 3  
Valid reference standard: Yes  
Blinded analysis of reference test: Yes  
Bias due to reference test: Low  
QUADAS Domain 4  
Appropriate interval between reference and index test: Yes  
All patients received reference test: Yes  
All patients received same test: Yes  
All patients included analysis: Yes  
Could patient flow have introduced bias: Not Applicable |
<table>
<thead>
<tr>
<th>Author, Year</th>
<th>Number of Participants, Populations</th>
<th>Type of Diagnostic Test, Cut-Off Value</th>
<th>Outcomes</th>
<th>QUADAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sakly et al., 2012&lt;sup&gt;16&lt;/sup&gt;</td>
<td>Number of Participants: 297 adults and children</td>
<td>Type of Diagnostic Test: DGP IgA Cut-off value: 25 IU/mL</td>
<td>Sensitivity: 97% Specificity: 90.7%</td>
<td>QUADAS Domain 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type of Diagnostic Test: DGP IgG Cut-off value: 25 IU/mL</td>
<td>Sensitivity: 94.2% Specificity: 95.4%</td>
<td>Consecutive or random sample: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Case control design avoided: No</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Inappropriate exclusions: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Biased patient Selection: High</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Blinded interpretation of index test results: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Prespecified test threshold: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bias due to testing: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Valid reference standard: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Blinded analysis of reference test: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bias due to reference test: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Appropriate interval between reference and index test: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>All patients received reference test: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>All patients received same test: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>All patients included analysis: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Could patient flow have introduced bias: Unclear</td>
</tr>
<tr>
<td>Author, Year</td>
<td>Number of Participants, Populations</td>
<td>Type of Diagnostic Test, Cut-Off Value</td>
<td>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</td>
<td>QUADAS</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------------------------</td>
<td>--------------------------------------</td>
<td>----------------------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Srinivas et al., 2014&lt;sup&gt;47&lt;/sup&gt;</td>
<td>Number of Participants: 752 Population: Clinical features of celiac disease</td>
<td>Type of Diagnostic Test: tTG IgA Cut-off value: &lt;10 IU/mL</td>
<td>Sensitivity: 0.83 Specificity: 0.96</td>
<td>QUADAS Domain 1&lt;br&gt;Consecutive or random sample: Yes&lt;br&gt;Case control design: Yes&lt;br&gt;Inappropriate exclusions: Yes&lt;br&gt;Biased patient Selection: Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type of Diagnostic Test: IgA EMA</td>
<td>Sensitivity: 0.80 Specificity: 0.99 Positive predictive value: Negative predictive value:</td>
<td>QUADAS Domain 2&lt;br&gt;Blinded interpretation of index test results: Unclear&lt;br&gt;Prespecified test threshold: Yes&lt;br&gt;Bias due to testing: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 3&lt;br&gt;Valid reference standard: Yes&lt;br&gt;Blinded analysis of reference test: Unclear&lt;br&gt;Bias due to reference test: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 4&lt;br&gt;Appropriate interval between reference and index test: No&lt;br&gt;All patients received reference test: Yes&lt;br&gt;All patients received same test: Yes&lt;br&gt;All patients included analysis: Yes&lt;br&gt;Could patient flow have introduced bias: High</td>
</tr>
</tbody>
</table>

C-18
<table>
<thead>
<tr>
<th>Author, Year</th>
<th>Number of Participants, Populations</th>
<th>Type of Diagnostic Test, Cut-Off Value</th>
<th>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</th>
<th>QUADAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Srinivas et al., 2013⁴⁸</td>
<td>Number of Participants: 75</td>
<td>Type of Diagnostic Test: EMA IgG</td>
<td>Sensitivity: 83% Specificity: 99% Positive predictive value: 93 Negative predictive value: 98</td>
<td>QUADAS Domain 1</td>
</tr>
<tr>
<td></td>
<td>Number of Participants: 102</td>
<td>Type of Diagnostic Test: tTG IgA, Cut-off value: 10 IU/mL</td>
<td>Sensitivity: 84% Specificity: 96% Positive predictive value: 72 Negative predictive value: 98</td>
<td>Consecutive or random sample: Yes</td>
</tr>
<tr>
<td></td>
<td>Number of Participants: 71</td>
<td>Type of Diagnostic Test: tTG IgA, EMA IgA</td>
<td>Sensitivity: 83% Specificity: 99% Positive predictive value: 97 Negative predictive value: 98</td>
<td>Case control design avoided: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Inappropriate exclusions: Yes Bias due to patient Selection: Low</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 2 Blinded interpretation of index test results: Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Prespecified test threshold: Yes Bias due to testing: Low</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 3 Valid reference standard: Yes Blinded analysis of reference test: Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Bias due to reference test: Low</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 4 Appropriate interval between reference and index test: Unclear</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>All patients received reference test: Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>All patients received same test: Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>All patients included analysis: Yes Could patient flow have introduced bias: Not Applicable</td>
<td></td>
</tr>
<tr>
<td>Author, Year</td>
<td>Number of Participants, Populations</td>
<td>Type of Diagnostic Test, Cut-Off Value</td>
<td>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</td>
<td>QUADAS</td>
</tr>
<tr>
<td>---------------------</td>
<td>-----------------------------------------------------------------------------------------------------</td>
<td>---------------------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| Sugai et al., 2010  | Number of Participants: 17 IgA tTG negative adults with villous atrophy                           | Type of Diagnostic Test: DGP          | Sensitivity: 35.71% Specificity: 100%                                                 | QUADAS Domain 1  
Consecutive or random sample: Yes  
Case control design avoided: Yes  
Inappropriate exclusions: Yes  
Biased patient Selection: Low                                                |
|                     | Comments: Original N = 22, five patients refused biopsy.                                             | Type of Diagnostic Test: tTG IgA,     | Sensitivity: 42.86% Specificity: 100%                                                 | QUADAS Domain 2  
Blinded interpretation of index test results: Yes  
Prespecified test threshold: Yes  
Bias due to testing: Low                                                      |
|                     |                                                                                                    | DGP IgA                                |                                                                                         | QUADAS Domain 3  
Valid reference standard: Yes  
Blinded analysis of reference test: Yes  
Bias due to reference test: Low                                                 |
|                     |                                                                                                    |                                                                                       |                                                                                         | QUADAS Domain 4  
Appropriate interval between reference and index test: Yes  
All patients received reference test: No  
All patients received same test: Yes  
All patients included analysis: No  
Could patient flow have introduced bias: High                                    |
<table>
<thead>
<tr>
<th>Author, Year</th>
<th>Number of Participants, Populations</th>
<th>Type of Diagnostic Test, Cut-Off Value</th>
<th>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</th>
<th>QUADAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swallow et al., 2013</td>
<td>Number of Participants: 733 Adults Results when Marsh 1-2 considered celiac</td>
<td>Type of Diagnostic Test: EMA IgA</td>
<td>Sensitivity: 42.9% Specificity: 99.5% Positive predictive value: 42.9 Negative predictive value: 99.5</td>
<td>QUADAS Domain 1 Consecutive or random sample: Yes Case control design avoided: Yes Inappropriate exclusions: No Biased patient Selection: High</td>
</tr>
<tr>
<td></td>
<td>Number of Participants: 756 Adults Results when Marsh 1-3 considered celiac</td>
<td>Type of Diagnostic Test: EMA IgA</td>
<td>Sensitivity: 73.3% Specificity: 99.5% Positive predictive value: 84.6 Negative predictive value: 98.9</td>
<td>QUADAS Domain 2 Blinded interpretation of index test results: Yes Prespecified test threshold: Yes Bias due to testing: Low</td>
</tr>
<tr>
<td></td>
<td>Number of Participants: 756 Adults Results when Marsh 3 considered celiac</td>
<td>Type of Diagnostic Test: EMA IgA</td>
<td>Sensitivity: 82.6% Specificity: 99.1% Positive predictive value: 73.1 Negative predictive value: 99.5</td>
<td>QUADAS Domain 3 Valid reference standard: Yes Blinded analysis of reference test: Yes Bias due to reference test: Low</td>
</tr>
<tr>
<td></td>
<td>Number of Participants: 733 Adults Results when Marsh 1-2 considered celiac</td>
<td>Type of Diagnostic Test: tTG IgA followed by EMA IgA, (NICE two step strategy)</td>
<td>Sensitivity: 57.1% Specificity: 97.3% Positive predictive value: 16.7 Negative predictive value: 99.6</td>
<td>QUADAS Domain 4 Appropriate interval between reference and index test: Yes All patients received reference test: Yes All patients received same test: Yes All patients included analysis: Yes Could patient flow have introduced bias: Not Applicable</td>
</tr>
<tr>
<td></td>
<td>Number of Participants: 756 Adults Results when Marsh 1-3 considered celiac</td>
<td>Type of Diagnostic Test: tTG IgA followed by EMA IgA, (NICE two step strategy)</td>
<td>Sensitivity: 80% Specificity: 97.3% Positive predictive value: 54.6 Negative predictive value: 99.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Number of Participants: 756 Population: Marsh 3 Comments: 473 patients were excluded because only one of the two serology tests was performed. 14 of these were diagnosed as CD via biopsy.</td>
<td>Type of Diagnostic Test: tTG IgA followed by EMA IgA, (NICE two step strategy)</td>
<td>Sensitivity: 87% Specificity: 96.9% Positive predictive value: 46.5 Negative predictive value: 99.6</td>
<td></td>
</tr>
<tr>
<td>Author, Year</td>
<td>Number of Participants, Populations</td>
<td>Type of Diagnostic Test, Cut-Off Value</td>
<td>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</td>
<td>QUADAS</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------------------------</td>
<td>----------------------------------------</td>
<td>------------------------------------------------------------------------------------</td>
<td>-------------------</td>
</tr>
</tbody>
</table>
| Swallow et al., 2013<sup>50</sup> | Number of Participants: 733 Adults Results when Marsh 1-2 considered celiac | Type of Diagnostic Test: tTG IgA | Sensitivity: 42.9%  
Specificity: 99.5%  
Positive predictive value: 42.9  
Negative predictive value: 99.5 | QUADAS Domain 1  
Consecutive or random sample: Yes  
Case control design avoided: Yes  
Inappropriate exclusions: No  
Biased patient Selection: High |
| | Number of Participants: 756 Adults Results when Marsh 1-3 considered celiac | Type of Diagnostic Test: tTG IgA | Sensitivity: 73.3%  
Specificity: 99.5%  
Positive predictive value: 84.6  
Negative predictive value: 98.9 | QUADAS Domain 2  
Blinded interpretation of index test results: Yes  
Prespecified test threshold: Yes  
Bias due to testing: Low |
| | Number of Participants: 756 Adults Results when Marsh 3 considered celiac as CD via biopsy. | Type of Diagnostic Test: tTG IgA | Sensitivity: 82.6%  
Specificity: 99.1%  
Positive predictive value: 73.1  
Negative predictive value: 99.5 | QUADAS Domain 3  
Valid reference standard: Yes  
Blinded analysis of reference test: Yes  
Bias due to reference test: Low |
| | | | | QUADAS Domain 4  
Appropriate interval between reference and index test: Yes  
All patients received reference test: Yes  
All patients received same test: Yes  
All patients included analysis: Yes  
Could patient flow have introduced bias: Not Applicable |
<table>
<thead>
<tr>
<th>Author, Year</th>
<th>Number of Participants, Populations</th>
<th>Type of Diagnostic Test, Cut-Off Value</th>
<th>Outcomes</th>
<th>QUADAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Van Meensel et al., 2004&lt;sup&gt;51&lt;/sup&gt;</td>
<td>Number of Participants: 175 Adults Comment: 5 patients were IgA deficient</td>
<td>Type of Diagnostic Test: tTG IgA Cut-off value: 10 kilounits/L Type of Diagnostic Test: tTG IgA Cut-off value: 15 kilounits Type of Diagnostic Test: tTG IgA Cut-off value: 19.05 kilounits/L Type of Diagnostic Test: tTG IgA Cut-off value: 2.64 kilounits/L Type of Diagnostic Test: tTG IgA Cut-off value: 20 kilounits Type of Diagnostic Test: tTG IgA Cut-off value: 20 kilounits/L</td>
<td>Sensitivity: 94% Specificity: 100% Sensitivity: 94% Specificity: 100% Sensitivity: 93% Specificity: 100% Sensitivity: 96% Specificity: 99% Sensitivity: 97% Specificity: 96% Sensitivity: 93% Specificity: 100%</td>
<td>QUADAS Domain 1 Consecutive or random sample: Yes Case control design avoided: No Inappropriate exclusions: Yes Biased patient Selection: High QUADAS Domain 2 Blinded interpretation of index test results: Yes Prespecified test threshold: Yes Bias due to testing: Unclear QUADAS Domain 3 Valid reference standard: Yes Blinded analysis of reference test: Yes Bias due to reference test: Low QUADAS Domain 4 Appropriate interval between reference and index test: Yes All patients received reference test: Yes All patients received same test: Yes All patients included analysis: Yes Could patient flow have introduced bias: Not Applicable</td>
</tr>
<tr>
<td>Author, Year</td>
<td>Number of Participants, Populations</td>
<td>Type of Diagnostic Test, Cut-Off Value</td>
<td>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</td>
<td>QUADAS</td>
</tr>
<tr>
<td>---------------------------</td>
<td>----------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| Van Meensel et al., 2004 | Number of Participants: 175 Adults
Comment: 5 patients were IgA deficient                                                                | Type of Diagnostic Test: tTG IgA
Cut-off value: 20.47 kilounits
Type of Diagnostic Test: tTG IgA
Cut-off value: 3.13 kilounits/L
Type of Diagnostic Test: tTG IgA
Cut-off value: 3.69 kilounits/L
Type of Diagnostic Test: tTG IgA
Cut-off value: 4 kilounits/L
Type of Diagnostic Test: tTG IgA
Cut-off value: 4.43 kilounits/L
Type of Diagnostic Test: tTG IgA
Cut-off value: 40 kilounits/L | Sensitivity: 97%
Specificity: 100%
Sensitivity: 96%
Specificity: 99%
Sensitivity: 96%
Specificity: 100%
Sensitivity: 93%
Specificity: 99%
Sensitivity: 99%
Specificity: 99%
Sensitivity: 96%
Specificity: 99% | QUADAS Domain 1
Consecutive or random sample: Yes
Case control design avoided: No
Inappropriate exclusions: Yes
Biased patient Selection: High
QUADAS Domain 2
Blinded interpretation of index test results: Yes
Prespecified test threshold: Yes
Bias due to testing: Unclear
QUADAS Domain 3
Valid reference standard: Yes
Blinded analysis of reference test: Yes
Bias due to reference test: Low
QUADAS Domain 4
Appropriate interval between reference and index test: Yes
All patients received reference test: Yes
All patients received same test: Yes
All patients included analysis: Yes
Could patient flow have introduced bias: Not Applicable |
<table>
<thead>
<tr>
<th>Author, Year</th>
<th>Number of Participants, Populations</th>
<th>Type of Diagnostic Test, Cut-Off Value</th>
<th>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</th>
<th>QUADAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Van Meensel et al., 2004&lt;sup&gt;51&lt;/sup&gt;</td>
<td>Number of Participants: 175 Adults Comment: 5 patients were IgA deficient</td>
<td>Type of Diagnostic Test: tTG IgA Cut-off value: 5 kilounits/L</td>
<td>Sensitivity: 93% Specificity: 99%</td>
<td>QUADAS Domain 1 Consecutive or random sample: Yes Case control design avoided: No Inappropriate exclusions: Yes Biased patient Selection: High</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type of Diagnostic Test: tTG IgA Cut-off value: 50 kilounits/L</td>
<td>Sensitivity: 93% Specificity: 93%</td>
<td>QUADAS Domain 2 Blinded interpretation of index test results: Yes Prespecified test threshold: Yes Bias due to testing: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type of Diagnostic Test: tTG IgA Cut-off value: 56.9 kilounits/L</td>
<td>Sensitivity: 91% Specificity: 99%</td>
<td>QUADAS Domain 3 Valid reference standard: Yes Blinded analysis of reference test: Yes Bias due to reference test: Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type of Diagnostic Test: tTG IgA Cut-off value: 7 kilounits/L</td>
<td>Sensitivity: 91% Specificity: 100%</td>
<td>QUADAS Domain 4 Appropriate interval between reference and index test: Yes All patients received reference test: Yes All patients received same test: Yes All patients included analysis: Yes Could patient flow have introduced bias: Not Applicable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type of Diagnostic Test: tTG IgA Cut-off value: 7 kilounits/L</td>
<td>Sensitivity: 97% Specificity: 100%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type of Diagnostic Test: tTG IgA Cut-off value: 7.16 kilounits/L</td>
<td>Sensitivity: 97% Specificity: 100%</td>
<td></td>
</tr>
<tr>
<td>Author, Year</td>
<td>Number of Participants, Populations</td>
<td>Type of Diagnostic Test, Cut-Off Value</td>
<td>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</td>
<td>QUADAS</td>
</tr>
<tr>
<td>----------------------</td>
<td>------------------------------------</td>
<td>---------------------------------------</td>
<td>---------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Van Meensel et al., 2004</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Number of Participants: 175  
Comment: 5 patients were IgA deficient | Type of Diagnostic Test: tTG IgA  
Cut-off value: 7.98 kilounits  
Type of Diagnostic Test: tTG IgA  
Cut-off value: 9.73 kilounits/L | Sensitivity: 96%  
Specificity: 100%  
Sensitivity: 94%  
Specificity: 100% | QUADAS Domain 1  
Consecutive or random sample: Yes  
Case control design avoided: No  
Inappropriate exclusions: Yes  
Biased patient Selection: High  
QUADAS Domain 2  
Blinded interpretation of index test results: Yes  
Prespecified test threshold: Yes  
Bias due to testing: Unclear  
QUADAS Domain 3  
Valid reference standard: Yes  
Blinded analysis of reference test: Yes  
Bias due to reference test: Low  
QUADAS Domain 4  
Appropriate interval between reference and index test: Yes  
All patients received reference test: Yes  
All patients received same test: Yes  
All patients included analysis: Yes  
Could patient flow have introduced bias: Not Applicable |
<table>
<thead>
<tr>
<th>Author, Year</th>
<th>Number of Participants, Populations</th>
<th>Type of Diagnostic Test, Cut-Off Value</th>
<th>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</th>
<th>QUADAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vermeersch et al., 2010</td>
<td>Number of Participants: 827 (599 adults, 228 children)</td>
<td>Type of Diagnostic Test: DGP IgA Cut-off value: &gt;7</td>
<td>Sensitivity: 65.1% Specificity: 99.1%</td>
<td>QUADAS Domain 1 Consecutive or random sample: Yes Case control design avoided: No Inappropriate exclusions: Yes Biased patient Selection: High</td>
</tr>
<tr>
<td></td>
<td>Number of Participants: 827</td>
<td>Type of Diagnostic Test: DGP IgG Cut-off value: 10</td>
<td>Sensitivity: 79.1% Specificity: 97.6%</td>
<td>QUADAS Domain 2 Blinded interpretation of index test results: Unclear Prespecified test threshold: Yes Bias due to testing: Unclear</td>
</tr>
<tr>
<td></td>
<td>Number of Participants: 827</td>
<td>Type of Diagnostic Test: DGP IgG Cut-off value: 20</td>
<td>Sensitivity: 83.7% Specificity: 99.3%</td>
<td>QUADAS Domain 3 Valid reference standard: Yes Blinded analysis of reference test: Yes Bias due to reference test: Low</td>
</tr>
<tr>
<td></td>
<td>Number of Participants: 827</td>
<td>Type of Diagnostic Test: DGP IgG Cut-off value: 25</td>
<td>Sensitivity: 76.7% Specificity: 99.2%</td>
<td>QUADAS Domain 4 Appropriate interval between reference and index test: Unclear All patients received reference test: Yes All patients received same test: Yes All patients included analysis: Yes Could patient flow have introduced bias: Unclear</td>
</tr>
<tr>
<td></td>
<td>Number of Participants: 827</td>
<td>Type of Diagnostic Test: DGP IgG Cut-off value: &gt;7</td>
<td>Sensitivity: 86% Specificity: 97.3%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Type of Diagnostic Test: tTG IgA Cut-off value: 7</td>
<td>Sensitivity: 84.9% Specificity: 92%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Author, Year</td>
<td>Number of Participants, Populations</td>
<td>Type of Diagnostic Test, Cut-Off Value</td>
<td>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</td>
<td>QUADAS</td>
</tr>
<tr>
<td>------------------------------------</td>
<td>-------------------------------------</td>
<td>----------------------------------------</td>
<td>---------------------------------------------------------------------------------------</td>
<td>-------------------------------------</td>
</tr>
<tr>
<td>Vermeersch et al., 2010</td>
<td>Number of Participants: 827</td>
<td>Type of Diagnostic Test: tTG IgA</td>
<td>Sensitivity: 88.4% Specificity: 94.9%</td>
<td>QUADAS Domain 1</td>
</tr>
<tr>
<td></td>
<td>Number of Participants: 827</td>
<td>Cut-off value: &gt;15</td>
<td></td>
<td>Consecutive or random sample: Yes</td>
</tr>
<tr>
<td></td>
<td>Number of Participants: 827</td>
<td>Type of Diagnostic Test: tTG IgA</td>
<td>Sensitivity: 83.7% Specificity: 98.4%</td>
<td>Case control design avoided: No</td>
</tr>
<tr>
<td></td>
<td>Number of Participants: 827</td>
<td>Cut-off value: &gt;7</td>
<td></td>
<td>Inappropriate exclusions: Yes</td>
</tr>
<tr>
<td></td>
<td>Number of Participants: 827</td>
<td>Type of Diagnostic Test: tTG IgG</td>
<td>Sensitivity: 60.5% Specificity: 98.1%</td>
<td>Biased patient Selection: High</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cut-off value: &gt;15</td>
<td></td>
<td>QUADAS Domain 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type of Diagnostic Test: tTG IgG</td>
<td>Sensitivity: 38.4% Specificity: 98.5%</td>
<td>Blinded interpretation of index test results: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cut-off value: &gt;7</td>
<td></td>
<td>Prespecified test threshold: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bias due to testing: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Valid reference standard: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Blinded analysis of reference test: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Bias due to reference test: Low</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>QUADAS Domain 4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Appropriate interval between reference and index test: Unclear</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>All patients received reference test: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>All patients received same test: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>All patients included analysis: Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Could patient flow have introduced bias: Unclear</td>
</tr>
<tr>
<td>Author, Year</td>
<td>Number of Participants, Populations</td>
<td>Type of Diagnostic Test, Cut-Off Value</td>
<td>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</td>
<td>QUADAS</td>
</tr>
<tr>
<td>---------------------------</td>
<td>------------------------------------</td>
<td>----------------------------------------</td>
<td>-----------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| Vermeersch et al., 2010⁵³ | Number of Participants: 588 Adults | Type of Diagnostic Test: tTG IgA       | Sensitivity: 86% Specificity: 95%                                                        | QUADAS Domain 1
Consecutive or random sample: Yes
Case control design avoided: No
Inappropriate exclusions: No
Biased patient Selection: High

QUADAS Domain 2
Blinded interpretation of index test results: Unclear
Prespecified test threshold: Yes
Bias due to testing: Unclear

QUADAS Domain 3
Valid reference standard: Yes
Blinded analysis of reference test: Yes
Bias due to reference test: Low

QUADAS Domain 4
Appropriate interval between reference and index test: Unclear
All patients received reference test: Yes
All patients received same test: Yes
All patients included analysis: Yes
Could patient flow have introduced bias: Unclear |
<p>|                           | Number of Participants: 588 Adults | Type of Diagnostic Test: tTG IgA       | Sensitivity: 95.3% Specificity: 92.7% Positive predictive value: 50.6                  |                                                                                                                                        |
|                           |                                    | Cut-off value: &gt;15 U/mL                |                                                                                         |                                                                                                                                        |
|                           |                                    | Type of Diagnostic Test: tTG IgA       |                                                                                         |                                                                                                                                        |
|                           |                                    | Cut-off value: &gt;=7 U/mL                |                                                                                         |                                                                                                                                        |</p>
<table>
<thead>
<tr>
<th>Author, Year</th>
<th>Number of Participants, Populations</th>
<th>Type of Diagnostic Test, Cut-Off Value</th>
<th>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</th>
<th>QUADAS</th>
</tr>
</thead>
</table>
| Vermeersch et al., 2012⁵⁴ | Number of Participants: 649 Adults and Children  
Comments: Retrospective study; the controls spanned years 2004 to 2006, while cases spanned years 2001 to 2009. | Type of Diagnostic Test: DGP IgA + tTG IgG*  
Cut-off value: 20 U/mL  
Type of Diagnostic Test: DGP IgA + tTG IgG*  
Cut-off value: 7 U/mL  
Type of Diagnostic Test: DGP IgG  
Cut-off value: 20 U/mL  
Type of Diagnostic Test: DGP IgG  
Cut-off value: 7 U/mL  
Type of Diagnostic Test: tTG IgA  
Cut-off value: 20 U/mL  
Type of Diagnostic Test: tTG IgA  
Cut-off value: 7 U/mL  
*Combined test determines whether patient has low IgA and will need IgG tests instead of IgA tests | Sensitivity: 89.7%  
Specificity: 93.3%  
Sensitivity: 88.8%  
Specificity: 95.6%  
Sensitivity: 85%  
Specificity: 99.3%  
Sensitivity: 86.9%  
Specificity: 96.7%  
Sensitivity: 84.1%  
Specificity: 95.9%  
Sensitivity: 81.3%  
Specificity: 98.5% | QUADAS Domain 1  
Consecutive or random sample: Yes  
Case control design avoided: No  
Inappropriate exclusions: Yes  
Biased patient Selection: High | QUADAS Domain 2  
Blinded interpretation of index test results: Unclear  
Prespecified test threshold: Yes  
Bias due to testing: Unclear | QUADAS Domain 3  
Valid reference standard: Yes  
Blinded analysis of reference test: Yes  
Bias due to reference test: Low | QUADAS Domain 4  
Appropriate interval between reference and index test: Unclear  
All patients received reference test: Yes  
All patients received same test: Yes  
All patients included analysis: Yes  
Could patient flow have introduced bias: Unclear |
<table>
<thead>
<tr>
<th>Author, Year</th>
<th>Number of Participants, Populations</th>
<th>Type of Diagnostic Test, Cut-Off Value</th>
<th>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</th>
<th>QUADAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wakim-Fleming et al., 2014&lt;sup&gt;15&lt;/sup&gt;</td>
<td>Number of Participants: 204 Population: Consecutive patients with biopsy proven cirrhosis</td>
<td>Type of Diagnostic Test: EMA Serum dilution &gt;= 1/10 Type of Diagnostic Test: TTG Cut-off value: above 20 U</td>
<td>Sensitivity: 1.00 Specificity: 1.00 Positive predictive value: Negative predictive value: Sensitivity: 1.00 Specificity: 0.96 Positive predictive value: Negative predictive value:</td>
<td>QUADAS Domain 1 Consecutive or random sample: Yes Case control design: Yes Inappropriate exclusions: Yes Biased patient Selection: Low QUADAS Domain 2 Blinded interpretation of index test results: Yes Prespecified test threshold: Yes Bias due to testing: Low QUADAS Domain 3 Valid reference standard: Yes Blinded analysis of reference test: Yes Bias due to reference test: Low QUADAS Domain 4 Appropriate interval between reference and index test: Yes All patients received reference test: Yes All patients received same test: Yes All patients included analysis: Yes Could patient flow have introduced bias: Low</td>
</tr>
<tr>
<td>Author, Year</td>
<td>Number of Participants, Populations</td>
<td>Type of Diagnostic Test, Cut-Off Value</td>
<td>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</td>
<td>QUADAS</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------------------------------</td>
<td>----------------------------------------</td>
<td>--------------------------------------------------------------------------------------------</td>
<td>--------</td>
</tr>
</tbody>
</table>
| Wolf et al., 2014  | Number of Participants: 1071 children Population: Selective IgA deficiency (sIgAD) was found in 27 patients | Type of Diagnostic Test: tTG IgA Cut-off value: >10 U/mL  
Type of Diagnostic Test: DGP IgG Cut-off value: >10 U/mL | Sensitivity: 0.88  
Specificity: 0.97  
Sensitivity: 0.89  
Specificity: 0.95 when added to tTG in children without IgA deficiency  
Sensitivity: 0.29  
Specificity: 1.00 when added to tTG in children WITH IgA deficiency | QUADAS Domain 1  
Consecutive or random sample: Yes  
Case control design: No  
Inappropriate exclusions: Yes  
Biased patient Selection: High |
|               |                                     |                                        | QUADAS Domain 2  
Blinded interpretation of index test results: Yes  
Prespecified test threshold: Yes  
Bias due to testing: Low | |
|               |                                     |                                        | QUADAS Domain 3  
Valid reference standard: Yes  
Blinded analysis of reference test: Yes  
Bias due to reference test: Low | |
|               |                                     |                                        | QUADAS Domain 4  
Appropriate interval between reference and index test: Yes  
All patients received reference test: Yes  
All patients received same test: Yes  
All patients included analysis: Yes  
Could patient flow have introduced bias: Low | |
<table>
<thead>
<tr>
<th>Author, Year</th>
<th>Number of Participants, Populations</th>
<th>Type of Diagnostic Test, Cut-Off Value</th>
<th>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</th>
<th>QUADAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zanini et al., 2012</td>
<td>Number of Participants: 263 Adults, (Brand B used)</td>
<td>Type of Diagnostic Test: tTG IgA Cut-off value: 16 U/mL</td>
<td>Sensitivity: 89.4% Specificity: 88.1% Positive predictive value: 90.3 Negative predictive value: 77.4</td>
<td>QUADAS Domain 1 Consecutive or random sample: Yes Case control design avoided: Yes Inappropriate exclusions: No Biased patient Selection: High</td>
</tr>
<tr>
<td></td>
<td>Number of Participants: 393 Adults, (Brand A used)</td>
<td>Type of Diagnostic Test: tTG IgA Cut-off value: 21 U/mL</td>
<td>Sensitivity: 38.2% Specificity: 97.4% Positive predictive value: 95.8 Negative predictive value: 50.7</td>
<td>QUADAS Domain 2 Blinded interpretation of index test results: Unclear Prespecified test threshold: Yes Bias due to testing: Unclear</td>
</tr>
<tr>
<td></td>
<td>Number of Participants: 289 Adults, (Brand C used)</td>
<td>Type of Diagnostic Test: tTG IgA Cut-off value: 24 U/mL</td>
<td>Sensitivity: 58.8% Specificity: 99% Positive predictive value: 99 Negative predictive value: 60.7</td>
<td>QUADAS Domain 3 Valid reference standard: Yes Blinded analysis of reference test: Unclear Bias due to reference test: Unclear</td>
</tr>
<tr>
<td></td>
<td>Number of Participants: 393</td>
<td>Type of Diagnostic Test: tTG IgA Cut-off value: 35 U/mL</td>
<td>Sensitivity: 10.1% Specificity: 100% Positive predictive value: 100 Negative predictive value: 42</td>
<td>QUADAS Domain 4 Appropriate interval between reference and index test: Yes All patients received reference test: Yes All patients received same test: Yes All patients included analysis: Yes Could patient flow have introduced bias: Not Applicable</td>
</tr>
<tr>
<td></td>
<td>Number of Participants: 289</td>
<td>Type of Diagnostic Test: tTG IgA Cut-off value: 40 U/mL</td>
<td>Sensitivity: 43.1% Specificity: 100% Positive predictive value: 100 Negative predictive value: 53.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Number of Participants: 263</td>
<td>Type of Diagnostic Test: tTG IgA Cut-off value: 48 U/mL</td>
<td>Sensitivity: 69.7% Specificity: 58.8% Positive predictive value: 100 Negative predictive value: 60.3</td>
<td></td>
</tr>
</tbody>
</table>

C-33
<table>
<thead>
<tr>
<th>Author, Year</th>
<th>Number of Participants, Populations</th>
<th>Type of Diagnostic Test, Cut-Off Value</th>
<th>Outcomes Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value</th>
<th>QUADAS</th>
</tr>
</thead>
</table>
| Zanini et al., 2012 | Number of Participants: 393 | Type of Diagnostic Test: tTG IgA Cut-off value: 7 U/mL | Sensitivity: 94.5%  
Specificity: 76.1%  
Positive predictive value: 85.9  
Negative predictive value: 90.1 | QUADAS Domain 1  
Consecutive or random sample: Yes  
Case control design avoided: Yes  
Inappropriate exclusions: No  
Biased patient Selection: High |
|               | Number of Participants: 289 | Type of Diagnostic Test: tTG IgA Cut-off value: 8 U/mL | Sensitivity: 88.1%  
Specificity: 92.2%  
Positive predictive value: 94.6  
Negative predictive value: 83.3 | QUADAS Domain 2  
Blinded interpretation of index test results: Unclear  
Prespecified test threshold: Yes  
Bias due to testing: Unclear |
|               | Number of Participants: 263 | Type of Diagnostic Test: tTG IgA Cut-off value: 80 U/mL | Sensitivity: 59.1%  
Specificity: 43.1%  
Positive predictive value: 100  
Negative predictive value: 52.9 | QUADAS Domain 3  
Valid reference standard: Yes  
Blinded analysis of reference test: Unclear  
Bias due to reference test: Unclear |

Table Notes: Au/ml – Absorbance Units per Milliliter; DGP – Deamidated Gliadin Peptide (DGP); DM – Diabetes; EMA – Endomysial Antibodies; HLA Human Leukocyte Antigen; IgA - Immunoglobulin A; IgG - Immunoglobulin G; L – Liter; NR – Not Reported; QUADAS – Quality Assessment of Diagnostic Studies; tTG - Anti-tissue Transglutaminase; U – Units; U/mL – Units per milliliter
Appendix D. Data Abstraction Tools

1. Celiac Disease Abstract Screening Form
2. Celiac Disease Full Text Screening Tool
3. Celiac Disease Data Abstraction Tool
• **Celiac Disease Abstract Screening Form**

If the title is to be excluded, please indicate by clicking the "exclude" button below. If included, please go to the abstract screening question below.

- Exclude
  Clear Response

Based on this abstract, is this an include or exclude?

- Include
- Exclude
- Needs discussion
- No abstract (exclude)

Background

- Duplicate Data [STOP] (specify ID number of which it's a duplicate)

  Clear Response

Exclude reason

- Not English language
- Not human
- Not about celiac disease (CD)
- Not about diagnosis of CD or under-diagnosis of CD
- No original data - letter, commentary, editorial, etc.
- Individual case report (Less than 10)
- Prevalence, outside U.S.
- Diagnostic method outside the scope of study. (Diagnostic methods included in the KQs are Endomysial antibodies (EmA) test, Anti-tissue Transglutaminase (tTG) test, Deamidated Gliadin Peptide (DGP) antibody, HLA typing, video capsule endoscopy, and endoscopy with biopsy)

- Test processing issue (E.g., PCR vs. other method)
- Serology Only - No comparison with biopsy
- Pre dates systematic review (SR) on topic

  Clear Response

Comment
• **Celiac Disease Full Text Screening Tool**

1. *Should the study have been rejected at abstract screening? If yes, please state the reason.*

- Yes (STOP), submit form

Specify:

- Not English language
- Not human
- Not about celiac disease (CD)
- Not about diagnosis of CD or under-diagnosis of CD
- No original data - letter, commentary, editorial, etc.
- Individual case report
- Prevalence, outside U.S.
- Diagnostic method outside the scope of study. (Diagnostic methods included in the KQs are Endomysial antibodies (EmA) test, Anti-tissue Transglutaminase (tTG) test, Deamidated Gliadin Peptide (DGP) antibody, HLA typing, video capsule endoscopy, and endoscopy with biopsy)

- Test processing issue (E.g., PCR vs. other method)
- Serology Only
- Pre dates systematic review (SR) on topic
- No

2. *Does the article address Key Question 4 - adverse effects of invasive methods (endoscopy with biopsy, video capsule endoscopy)?*

- Yes (STOP), submit form - study will be included
- No

3. *Does the article address Key Question 2 - how the accuracy of duodenal biopsy varies by MD training, method, or length of gluten ingestion?*

- Yes (STOP), submit form - study will be included

If yes, indicate which one:

- Pathologist characteristics, i.e. level of experience or specific training
- Method, i.e. type or number of specimens
- Length of time ingesting gluten before diagnostic testing
- Exclude study (indicate reason)
- No

4. *Does the article address Key Question 1 or Key Question 3 - assess the accuracy or effectiveness (e.g. outcomes) of any of the following? (check all that apply)*

- Endomysial antibodies (EmA) IgA test
Anti-tissue transglutaminase (tTG) IgA test
EmA IgG, tTG IgG, and DGP IgG tests for IgA deficient individuals
Deamidated Gliadin Peptide (DGP) IgA Antibodies
HLA-DQ2 or HLA-DQ8
Video capsule endoscopy
None of the above (STOP), submit form - study will be excluded
Does not assess accuracy or effectiveness e.g. only prevalence, test processing, etc. (STOP), submit form - study will be excluded

5. If this is an accuracy study:
Are the tests compared to biopsy?
- Yes
- No (STOP), submit form - study will be excluded

Did ALL subjects undergo both the "intervention" diagnostic method and the "reference" standard (usually biopsy)?
(i.e. everyone who received a serological test should get biopsy too, regardless of whether the test was negative or positive)
- Yes
- No (STOP), submit form - study will be excluded

6. What is the sample size?

7. Was a consecutive or random sample of patients enrolled?
- Yes
- No
- Unclear

8. Does the article present results specific to the following populations? (check all that apply)
- Children, under age 24 months vs. older children & adolescents
- Adults (aged 18+)
- Ethnic or geographic populations (Specify)
- Low socioeconomic status (SES)
- IgA deficient
- Previously negative for CD
- With signs and symptoms of CD (i.e., diarrhea, nutritional deficits, etc.)
With type 1 diabetes
With auto-immune disease
With Turner's syndrome
With trisomy 21/Down Syndrome
With family history
Other special population (Specify)
General population, no special populations

9. Does the article need further discussion?
Yes
No

Notes:

- Celiac Disease Data Abstraction Tool

Study Design
Specify the study design
- Randomized controlled trial
- Case control
- Retrospective cohort
- Prospective cohort
- Other

Clear Response

Country
Specify the country
- United States
- UK
- Other 1 (specify country)
- Other 2 (specify country)
- Other 3 (specify country)
- Not reported

Population
Specify the % of female? If not reported, please write "Not reported".
Please indicate % of race/ethnicity

- Asian (specify %)
- Black (specify %)
- Latino (specify %)
- Middle Eastern (specify %)
- White (specify %)
- Other 1 (specify ethnicity and %)
- Other 2 (specify ethnicity and %)
- Other 3 (specify ethnicity and %)
- Other 4 (specify ethnicity and %)
- Not reported

Arms

How many arms are there?

Arm 1
Please describe the population as best as possible.

Please indicate the sample size.

For tTG tests: threshold level in terms of X times upper limit of normal. If not reported, please indicate "Not reported"

Arm 2
Please describe the population as best as possible.

Please indicate the sample size.

For tTG tests: threshold level in terms of X times upper limit of normal. If not reported, please indicate "Not reported"

Arm 3
Please describe the population as best as possible.

Please indicate the sample size.

For tTG tests: threshold level in terms of X times upper limit of normal. If not reported, please indicate "Not reported"

Arm 4
Please describe the population as best as possible.

Please indicate the sample size.

For tTG tests: threshold level in terms of X times upper limit of normal. If not reported, please indicate "Not reported"

Arm 5
Please describe the population as best as possible.

Please indicate the sample size.

For tTG tests: threshold level in terms of X times upper limit of normal. If not reported, please indicate "Not reported"

Arm 6
Please describe the population as best as possible.

Please indicate the sample size.

For tTG tests: threshold level in terms of X times upper limit of normal. If not reported, please indicate "Not reported"
Arm 7
Please describe the population as best as possible.

Please indicate the sample size.

For tTG tests: threshold level in terms of X times upper limit of normal. If not reported, please indicate "Not reported"

Arm 8
Please describe the population as best as possible.

Please indicate the sample size.

For tTG tests: threshold level in terms of X times upper limit of normal. If not reported, please indicate "Not reported"

Arm 9
Please describe the population as best as possible.

Please indicate the sample size.

For tTG tests: threshold level in terms of X times upper limit of normal. If not reported, please indicate "Not reported"

Arm 10
Please describe the population as best as possible.

Please indicate the sample size.

For tTG tests: threshold level in terms of X times upper limit of normal. If not reported, please indicate "Not reported"
Arm 11
Please describe the population as best as possible.

Please indicate the sample size.

For tTG tests: threshold level in terms of X times upper limit of normal. If not reported, please indicate "Not reported"

Arm 12
Please describe the population as best as possible.

Please indicate the sample size.

For tTG tests: threshold level in terms of X times upper limit of normal. If not reported, please indicate "Not reported"

Arm 13
Please describe the population as best as possible.

Please indicate the sample size.

For tTG tests: threshold level in terms of X times upper limit of normal. If not reported, please indicate "Not reported"

Arm 14
Please describe the population as best as possible.

Please indicate the sample size.
For tTG tests: threshold level in terms of X times upper limit of normal. If not reported, please indicate "Not reported"

Arm 15
Please describe the population as best as possible.

Please indicate the sample size.

Arm 16
Please describe the population as best as possible.

Please indicate the sample size.

Arm 17
Please describe the population as best as possible.

Please indicate the sample size.

Arm 18
Please describe the population as best as possible.

Please indicate the sample size.
Arm 19
Please describe the population as best as possible.

Please indicate the sample size.

For tTG tests: threshold level in terms of X times upper limit of normal. If not reported, please indicate "Not reported"

Arm 20
Please describe the population as best as possible.

Please indicate the sample size.

For tTG tests: threshold level in terms of X times upper limit of normal. If not reported, please indicate "Not reported"
# Appendix E. AMSTAR Criteria

## Table E-1. AMSTAR criteria

<table>
<thead>
<tr>
<th>Author, Year</th>
<th>Q1: Design</th>
<th>Q2: Selection</th>
<th>Q3: Search</th>
<th>Q4: Publication Status Criterion</th>
<th>Q5: List Studies</th>
<th>Q6: Characteristics</th>
<th>Q7: Quality Assessed</th>
<th>Q8: Quality Conclusions</th>
<th>Q9: Methods Appropriate</th>
<th>Q10: Bias</th>
<th>Q11: COI</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Video Capsule Endoscopy</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rokkas et al., 2012</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>El-Matary et al., 2009</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td><strong>Serological Tests</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Giersiepen et al., 2012</td>
<td>Yes</td>
<td>NR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Lewis et al., 2010</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Rostom et al., 2005</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Rostom et al., 2004</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>NICE Guidelines, 2009</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>6</td>
</tr>
<tr>
<td>Schyum, 2013</td>
<td>Yes</td>
<td>NR</td>
<td>Yes</td>
<td>NR</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>NA</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>4</td>
</tr>
<tr>
<td><strong>Special Populations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ford et al., 2009</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>6</td>
</tr>
<tr>
<td>Author, Year</td>
<td>Q1: Design</td>
<td>Q2: Selection</td>
<td>Q3: Search</td>
<td>Q4: Publication Status Criterion</td>
<td>Q5: List Studies</td>
<td>Q6: Characteristics</td>
<td>Q7: Quality Assessed</td>
<td>Q8: Quality Conclusions</td>
<td>Q9: Methods Appropriate</td>
<td>Q10: Bias</td>
<td>Q11: COI</td>
<td>Score</td>
</tr>
<tr>
<td>-------------</td>
<td>------------</td>
<td>---------------</td>
<td>------------</td>
<td>----------------------------------</td>
<td>-----------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>-----------------------</td>
<td>------------------------</td>
<td>---------</td>
<td>--------</td>
<td>-------</td>
</tr>
<tr>
<td>van der Windt et al., 2010(^87)</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>6</td>
</tr>
<tr>
<td>Bruins, 2013(^64) on gluten challenge</td>
<td>No</td>
<td>NR</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>2</td>
</tr>
<tr>
<td>Hall et al., (^65) on adherence to GFD</td>
<td>No</td>
<td>NR</td>
<td>Yes</td>
<td>NR</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>4</td>
</tr>
<tr>
<td>Liao et al., 2010(^93) on adverse events</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>2</td>
</tr>
</tbody>
</table>

Table notes: AMSTAR = Assessment of Multiple Systematic Reviews; NR = Not reported.
AMSTAR - Assessment of Reporting Quality for Systematic Reviews
1. Was an ‘a priori’ design provided?
The research question and inclusion criteria should be established before the conduct of the review.

Note: Need to refer to a protocol, ethics approval, or pre-determined/a priori published research objectives to score a “yes.”
☐ Yes
☐ No
☐ Can't answer
☐ Not applicable

2. Was there duplicate study selection and data extraction?
There should be at least two independent data extractors and a consensus procedure for disagreements should be in place.

Note: 2 people do study selection, 2 people do data extraction, consensus process or one person checks the other’s work.
☐ Yes
☐ No
☐ Can't answer
☐ Not applicable

3. Was a comprehensive literature search performed?
At least two electronic sources should be searched. The report must include years and databases used (e.g., Central, EMBASE, and MEDLINE). Key words and/or MESH terms must be stated and where feasible the search strategy should be provided. All searches should be supplemented by consulting current contents, reviews, textbooks, specialized registers, or experts in the particular field of study, and by reviewing the references in the studies found.

Note: If at least 2 sources + one supplementary strategy used, select “yes” (Cochrane register/Central counts as 2 sources; a grey literature search counts as supplementary).
☐ Yes
☐ No
☐ Can't answer
☐ Not applicable

4. Was the status of publication (i.e. grey literature) used as an inclusion criterion?
The authors should state that they searched for reports regardless of their publication type. The authors should state whether or not they excluded any reports (from the systematic review), based on their publication status, language etc.

Note: If review indicates that there was a search for “grey literature” or “unpublished
literature,” indicate “yes.” SIGLE database, dissertations, conference proceedings, and trial registries are all considered grey for this purpose. If searching a source that contains both grey and non-grey, must specify that they were searching for grey/unpublished lit.

- Yes
- No
- Can't answer
- Not applicable

5. Was a list of studies (included and excluded) provided?
A list of included and excluded studies should be provided.

Note: Acceptable if the excluded studies are referenced. If there is an electronic link to the list but the link is dead, select “no.”

- Yes
- No
- Can't answer
- Not applicable

6. Were the characteristics of the included studies provided?
In an aggregated form such as a table, data from the original studies should be provided on the participants, interventions and outcomes. The ranges of characteristics in all the studies analyzed e.g., age, race, sex, relevant socioeconomic data, disease status, duration, severity, or other diseases should be reported.

Note: Acceptable if not in table format as long as they are described as above.

- Yes
- No
- Can't answer
- Not applicable

7. Was the scientific quality of the included studies assessed and documented?
'A priori' methods of assessment should be provided (e.g., for effectiveness studies if the author(s) chose to include only randomized, double-blind, placebo controlled studies, or allocation concealment as inclusion criteria); for other types of studies alternative items will be relevant.

Note: Can include use of a quality scoring tool or checklist, e.g., Jadad scale, risk of bias, sensitivity analysis, etc., or a description of quality items, with some kind of result for EACH study (“low” or “high” is fine, as long as it is clear which studies scored “low” and which scored “high”; a summary score/range for all studies is not acceptable).

- Yes
- No
- Can't answer
- Not applicable
8. Was the scientific quality of the included studies used appropriately in formulating conclusions?
The results of the methodological rigor and scientific quality should be considered in the analysis and the conclusions of the review, and explicitly stated in formulating recommendations.

Note: Might say something such as “the results should be interpreted with caution due to poor quality of included studies.” Cannot score “yes” for this question if scored “no” for question 7.

□ Yes
□ No
□ Can't answer
□ Not applicable

9. Were the methods used to combine the findings of studies appropriate?
For the pooled results, a test should be done to ensure the studies were combinable, to assess their homogeneity (i.e., Chi-squared test for homogeneity, I²). If heterogeneity exists a random effects model should be used and/or the clinical appropriateness of combining should be taken into consideration (i.e., is it sensible to combine?).

Note: Indicate “yes” if they mention or describe heterogeneity, i.e., if they explain that they cannot pool because of heterogeneity/variability between interventions.

□ Yes
□ No
□ Can't answer
□ Not applicable

10. Was the likelihood of publication bias assessed?
An assessment of publication bias should include a combination of graphical aids (e.g., funnel plot, other available tests) and/or statistical tests (e.g., Egger regression test, Hedges-Olken).

Note: If no test values or funnel plot included, score “no”. Score “yes” if mentions that publication bias could not be assessed because there were fewer than 10 included studies.

□ Yes
□ No
□ Can't answer
□ Not applicable

11. Was the conflict of interest included?
Potential sources of support should be clearly acknowledged in both the systematic review and the included studies.

Note: To get a “yes,” must indicate source of funding or support for the systematic
review AND for each of the included studies.
### Appendix F. Strength of Evidence for Accuracy of Serology Tests

#### Table F-1. Strength of evidence for the accuracy of serology tests

<table>
<thead>
<tr>
<th>Diagnostic Method</th>
<th>Population</th>
<th>Number of Studies</th>
<th>Risk of Bias</th>
<th>Consistency</th>
<th>Directness</th>
<th>Precision</th>
<th>SOE, Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>tTG IgA</td>
<td>Overall (Studies mixed symptomatic and high risk)</td>
<td>1 prior meta-analysis of 8 studies; 1 prior meta of 12 studies, 16 new studies</td>
<td>Moderate</td>
<td>Consistent</td>
<td>Direct</td>
<td>Precise</td>
<td>High &lt;br&gt;New meta-analysis of thresholds used in clinical practice: &lt;br&gt;Sensitivity 92.5% (95% CI: 89.7, 94.6) &lt;br&gt;Specificity 97.9% (95% CI: 96.5, 98.7)</td>
</tr>
<tr>
<td></td>
<td>Abdominal symptoms</td>
<td>1 prior meta of 7 studies</td>
<td>Moderate</td>
<td>Consistent</td>
<td>Direct</td>
<td>Precise (Specificity)</td>
<td>High &lt;br&gt;Sensitivity 89.0% (95% CI: 82.0, 94.0) &lt;br&gt;Specificity 98.0% (95% CI: 95.0, 99.0)</td>
</tr>
<tr>
<td></td>
<td>Type 1 diabetes</td>
<td>1 study (N = 62)</td>
<td>Low</td>
<td>NA</td>
<td>Direct</td>
<td>NA</td>
<td>Insufficient &lt;br&gt;Sensitivity 71.0% &lt;br&gt;Specificity 93.0%</td>
</tr>
<tr>
<td></td>
<td>Iron deficient</td>
<td>1 study (N = 130)</td>
<td>High</td>
<td>NA</td>
<td>Direct</td>
<td>NA</td>
<td>Insufficient &lt;br&gt;Sensitivity 38.0% &lt;br&gt;Specificity 97.0%</td>
</tr>
<tr>
<td></td>
<td>Asymptomatic – general population screening</td>
<td>1 study (N = 1,001)</td>
<td>Low</td>
<td>NA</td>
<td>Direct</td>
<td>NA</td>
<td>Low &lt;br&gt;Sensitivity 100.0% &lt;br&gt;Specificity 97.4%</td>
</tr>
<tr>
<td></td>
<td>Children</td>
<td>1 prior meta of 5 point-of-care tests</td>
<td>Moderate</td>
<td>Consistent</td>
<td>Direct</td>
<td>Precise</td>
<td>High &lt;br&gt;Sensitivity 96.4% (95% CI: 94.3, 97.9) &lt;br&gt;Specificity 97.7% (95% CI: 95.8, 99.0)</td>
</tr>
<tr>
<td>EmA IgA</td>
<td>Overall (Studies mixed symptomatic and high risk)</td>
<td>3 prior SRs, 7 new studies</td>
<td>Moderate</td>
<td>Consistent</td>
<td>Direct</td>
<td>Precise (Specificity)</td>
<td>High &lt;br&gt;New meta-analysis &lt;br&gt;Sensitivity 76.6% (95% CI: 68.7, 82.9) &lt;br&gt;Specificity 99.0% (95% CI: 98.4, 99.4)</td>
</tr>
<tr>
<td>Condition</td>
<td>Study Information</td>
<td>Quality Evidence</td>
<td>Specificity</td>
<td>Sensitivity</td>
<td>Insufficiency</td>
<td>Specificity</td>
<td>Sensitivity</td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>---------------------------------------------------------</td>
<td>------------------</td>
<td>-------------</td>
<td>-------------</td>
<td>---------------</td>
<td>-------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Abdominal symptoms</td>
<td>1 prior meta of 8 studies</td>
<td>Moderate</td>
<td>Consistent</td>
<td>Direct</td>
<td>High</td>
<td>Precise</td>
<td>High</td>
</tr>
<tr>
<td>Type 1 diabetes</td>
<td>1 study (N = 62)</td>
<td>Low</td>
<td>NA</td>
<td>Direct</td>
<td>Insufficient</td>
<td>Sensitivity 71.0%</td>
<td>Specificity 96.0%</td>
</tr>
<tr>
<td>Iron deficient</td>
<td>1 study (N = 84)</td>
<td>High</td>
<td>NA</td>
<td>Direct</td>
<td>Insufficient</td>
<td>Sensitivity 100.0%</td>
<td>Specificity 99.0%</td>
</tr>
<tr>
<td>Asymptomatic – general population screening</td>
<td>1 study (N = 1,001)</td>
<td>Low</td>
<td>NA</td>
<td>Direct</td>
<td>Low</td>
<td>Sensitivity 85.7%</td>
<td>Specificity 99.0%</td>
</tr>
<tr>
<td>Children</td>
<td>1 prior meta of 11 studies</td>
<td>Moderate</td>
<td>Consistent</td>
<td>Direct</td>
<td>High</td>
<td>Sensitivity 82.6% to 100% (not pooled)</td>
<td>Specificity 98.2% (95% CI: 96.7, 99.1)</td>
</tr>
<tr>
<td>DGP IgA Overall (Studies mixed symptomatic and high risk)</td>
<td>1 prior meta of 11 studies, 2 new studies</td>
<td>Moderate</td>
<td>Consistent</td>
<td>Direct</td>
<td>High</td>
<td>Sensitivity 87.8% (95% CI: 85.6, 89.9)</td>
<td>Specificity 94.1% (95% CI: 92.5, 95.5)</td>
</tr>
<tr>
<td>DGP IgG Overall (Studies mixed symptomatic and high risk)</td>
<td>1 prior SR of 7 studies, 1 new study</td>
<td>Moderate</td>
<td>Consistent</td>
<td>Direct</td>
<td>Moderate</td>
<td>Sensitivity 80.7% to 95.1% (not pooled)</td>
<td>Specificity 90.7% (95% CI: 87.8, 93.1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Antigen</th>
<th>Population</th>
<th>Sample Size</th>
<th>Quality</th>
<th>Consistency</th>
<th>Evidence</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Children</strong>&lt;br&gt;1 prior SR of 3 studies, 2 new studies</td>
<td>High</td>
<td>Direct</td>
<td>Imprecise</td>
<td>Moderate&lt;br&gt;Sensitivity 80.1% to 100%&lt;br&gt;Specificity 86.0% to 100%&lt;br&gt;New study reports sensitivity and specificity of 100% in children under 7 years of age</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DGP IgA + IgG combo</td>
<td>IgA deficient&lt;br&gt;(N = 45)</td>
<td>Moderate</td>
<td>NA</td>
<td>Direct</td>
<td>Insufficient&lt;br&gt;Sensitivity 100%&lt;br&gt;Specificity 89.2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TG IgA + DPG IgA</td>
<td>Symptomatic adults&lt;br&gt;(N=998)</td>
<td>Low</td>
<td>NA</td>
<td>Direct</td>
<td>Insufficient&lt;br&gt;Sensitivity 72.2%&lt;br&gt;Specificity 97.4%</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Children</strong>&lt;br&gt;1 study of 2 threshold levels (N= 703)</td>
<td>High</td>
<td>Consistent</td>
<td>Direct</td>
<td>NA</td>
<td>Insufficient&lt;br&gt;Sensitivity 94.5% to 96.7%&lt;br&gt;Specificity 89.8% to 95.7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TG IgA + DGP IgA/IgG combo</td>
<td>Symptomatic adults&lt;br&gt;1 study of 2 threshold levels (N = 176)</td>
<td>Moderate</td>
<td>Consistent</td>
<td>Direct</td>
<td>NA</td>
<td>Insufficient&lt;br&gt;Sensitivity 85.0% to 91.0%&lt;br&gt;Specificity 80.0% to 98.0%</td>
<td></td>
</tr>
<tr>
<td>TG IgA + EmA IgA</td>
<td>Symptomatic adults&lt;br&gt;2 studies (N = 443)</td>
<td>Moderate</td>
<td>Consistent</td>
<td>Direct</td>
<td>Imprecise</td>
<td>Insufficient&lt;br&gt;Sensitivity 76.0% to 87.0%&lt;br&gt;Specificity 85.0% to 97.0%</td>
<td></td>
</tr>
<tr>
<td>First degree relatives</td>
<td>1 study&lt;br&gt;(N = 32)</td>
<td>Low</td>
<td>NA</td>
<td>Direct</td>
<td>NA</td>
<td>Insufficient&lt;br&gt;Sensitivity 81.0%&lt;br&gt;Specificity 70.%</td>
<td></td>
</tr>
<tr>
<td>Type 1 diabetes</td>
<td>1 study&lt;br&gt;(N = 60)</td>
<td>Low</td>
<td>NA</td>
<td>Direct</td>
<td>NA</td>
<td>Insufficient&lt;br&gt;Sensitivity 93.0%&lt;br&gt;Specificity 64.0%</td>
<td></td>
</tr>
<tr>
<td>Mixed (symptomatic and / or Type 1 diabetes)</td>
<td>1 study&lt;br&gt;(N = 752)</td>
<td>High</td>
<td>NA</td>
<td>Direct</td>
<td>NA</td>
<td>Insufficient&lt;br&gt;Sensitivity 83.0%&lt;br&gt;Specificity 99.0%</td>
<td></td>
</tr>
</tbody>
</table>

[The answers to the questions are found in pages 91 to 95 of the attached 262 page article]

1. An adult patient comes to you after being on a gluten free diet for 6 months, she wants to confirm the diagnosis of celiac disease. She is HLA Q2 positive, IgA TTG negative, normal duodenal biopsies. You both agree on a gluten challenge. The **minimum** duration of gluten ingestion prior to repeating EGD with biopsy should be:
   a. one dose of 50 grams of gluten
   b. 3 days of gluten ingestion
   c. 2 weeks of gluten intake
   d. 1-2 months of gluten intake

2. A 24 year old obsessive compulsive female is convinced she has celiac sprue. You are the 6th doctor she has consulted. She wants you to use the most accurate non-invasive test to make the diagnosis. Based on her wishes you would:
   a. Order an IgA Endomysial antibody test as this is the most specific and sensitive test
   b. Order an IgA DGP as this is the most sensitive test
   c. Order HLA DQ2 & 8
   d. Order an IgA TTG test
   e. Refer her to a psychiatrist

3. A patient with IgA deficiency has signs and symptoms suggestive of celiac disease, the one non-invasive test you should obtain in an attempt to make a diagnosis should be:
   a. IgG TTG
   b. IgA anti-endomysial antibody
   c. IgG DGP
   d. Anti-gliadin antibodies IgG
   e. There is insufficient evidence to determine the best serologic test for IgA deficient patients

**True or False**

4. The accuracy of small bowel biopsies to diagnose celiac disease increases with the number of biopsies obtained, More than 4 biopsies are recommended by the AGA.

5. IgA DGP is more accurate than IgA-TTG, but is more expensive.

6. HLA testing for DQ2 and/or 8 is useful to make a diagnosis of celiac disease