Chapter 2. Methods

Overview

The UO-EPC's evidence report on CD is based on a systematic review of the scientific-medical literature to identify, and synthesize the results from studies addressing the key questions put forth by the AHRQ. The Celiac Review Team, together with content experts, identified specific issues integral to the review. A Technical Expert Panel (TEP) refined the research questions, as well as highlighted key variables requiring consideration in the evidence synthesis. Evidence tables presenting the key study characteristics and results were developed. Summary tables were derived from the evidence tables. The methodological quality of reports of the included studies was appraised, and individual study results were summarized. For some objectives a narrative interpretation of the literature was provided.

Key Questions Addressed in This Report

The AHRQ task order requested answers to the questions outlined below:

1) Objective 1 – Sensitivity and specificity of tests for CD (Celiac 1)

a) What is the sensitivity and specificity of the following tests for CD:
   i) AGA;
   ii) EMA;
   iii) human tTG lgA antibodies;
   iv) HLA (DQ2/DQ8);
   v) duodenal/jejunal biopsy (see section below on celiac definition)
b) Do sensitivity and specificity vary in different target populations (e.g., symptomatic vs. asymptomatic; geographic populations)?

2) Objective 2 – Prevalence and incidence of CD (Celiac 2)

a) What is the prevalence and incidence of symptomatic and "clinically silent" CD in:
   i) the general population;
   ii) high-risk populations:
      (1) family member of patient with CD;
      (2) type 1 diabetes mellitus;
      (3) iron deficiency anemia (IDA);
      (4) osteoporosis?
b) How does prevalence and incidence in the general population vary in different geographic and racial/ethnic populations?

3) Objective 3 – Celiac associated lymphoma (Celiac 3)

Note: Appendixes and Evidence Tables are provided electronically at http://www.ahrq.gov/clinic/tp/celiactp.htm

a) What is the association between CD and GI lymphoma?
   i) What is the cumulative risk of developing GI lymphoma in patients with CD?
   ii) Does the cumulative risk vary with clinical presentation?

4) Objective 4 – Expected consequences of testing for CD (Celiac 4)

a) What are the expected consequences of testing for CD in the following populations:
   i) patients with symptoms suggestive of CD;
   ii) asymptomatic, at-risk populations (affected family members, patients with type 1 diabetes);
   iii) the general population?

b) "Consequences" include:
   i) false-positive results;
   ii) follow-up testing;
   iii) invasive procedures (biopsies);
   iv) cases diagnosed;
   v) patients complying with treatment; and
   vi) response to treatment.

5) Objective 5 – Promoting or monitoring adherence to a GFD (Celiac 5)

a) What interventions are effective for promoting or monitoring adherence to a GFD?

Study Criteria Used in this Review

Histological

From the preceding discussion in the methodological consideration section it is clear that current histological criteria using a cut-off grade to define CD have important shortcomings. We therefore adopted an open histological definition of CD when selecting a study for inclusion, as long as the authors' explicitly stated or described the criteria used to define CD (see inclusion criteria below)**.** ***However, with the help of the TEP, we defined a "standard" histological definition of CD as a biopsy grade showing a modified Marsh IIIa or greater***. This definition was NOT used as an inclusion/exclusion criterion, but simply to frame our results and to allow for the evaluation of the effect of different histological criteria on the performance of the various CD tests.

The choice of biopsy criteria and/or histological grade "cut-off" used to define CD has important implications for the interpretation of the studies of serology, HLA, and biopsy. It is recognized that some patients with CD may have Marsh I or II lesions, and by definition patients with latent CD have Marsh 0 lesions. However, as emphasized by Marsh,[1] and as is discussed further below, in order to correctly interpret these early lesions, prospective follow-up studies are required, and an individual patient follow-up and documented response to gluten withdrawal would be required to firmly establish the diagnosis of CD.

The practical importance of the histological definition is evident from our preliminary review of articles that demonstrated considerable heterogeneity in the histological criteria used within the studies to define CD. Some used strict definitions, whereas, others accepted milder grade lesions. Furthermore, since the existence of latent CD and some silent CD without fully developed histology is now recognized, a study that aims to assess the sensitivity and specificity of biopsy itself in CD needs to use a design that incorporates the most sensitive and specific serologic and HLA tests available. The biopsy and serology should be performed simultaneously, with patients having discordant test results being further evaluated. Those with normal biopsy and positive serology would have to be followed over time to see if they have a latent form of CD. Conversely, patients with positive biopsies and normal serology would have to demonstrate improvement in histology on a GFD, and ideally, certification of relapse by biopsy with reintroduction of gluten. This type of study design was sought in order to address the objective of the sensitivity and specificity of biopsy.

Populations

1) **Unselected general population**. The unselected general population implies a representative sample of a given population, such as a random sample of healthy blood donors or healthy school children. Some unselected populations are better than others for determining the true prevalence or incidence of CD. For example, blood donors are required to have normal hemoglobin and no iron deficiency, and therefore may underestimate the true numbers of patients with CD.

2) **Suspected CD.** Patients with suspected CD include patients with GI symptoms, such as diarrhea or symptomatic malabsorption, who are being investigated for the possibility of CD. These patients are typically undergoing other investigations in addition to being worked-up for CD.

3) **High-risk populations.** High-risk populations include populations with an expectedly higher prevalence of CD. Such populations include asymptomatic family members of patients with CD, patients with type I diabetes where identified CD would likely be silent or latent, and populations such as those with iron deficiency or osteoporosis where identified CD would be in the atypical CD classification.

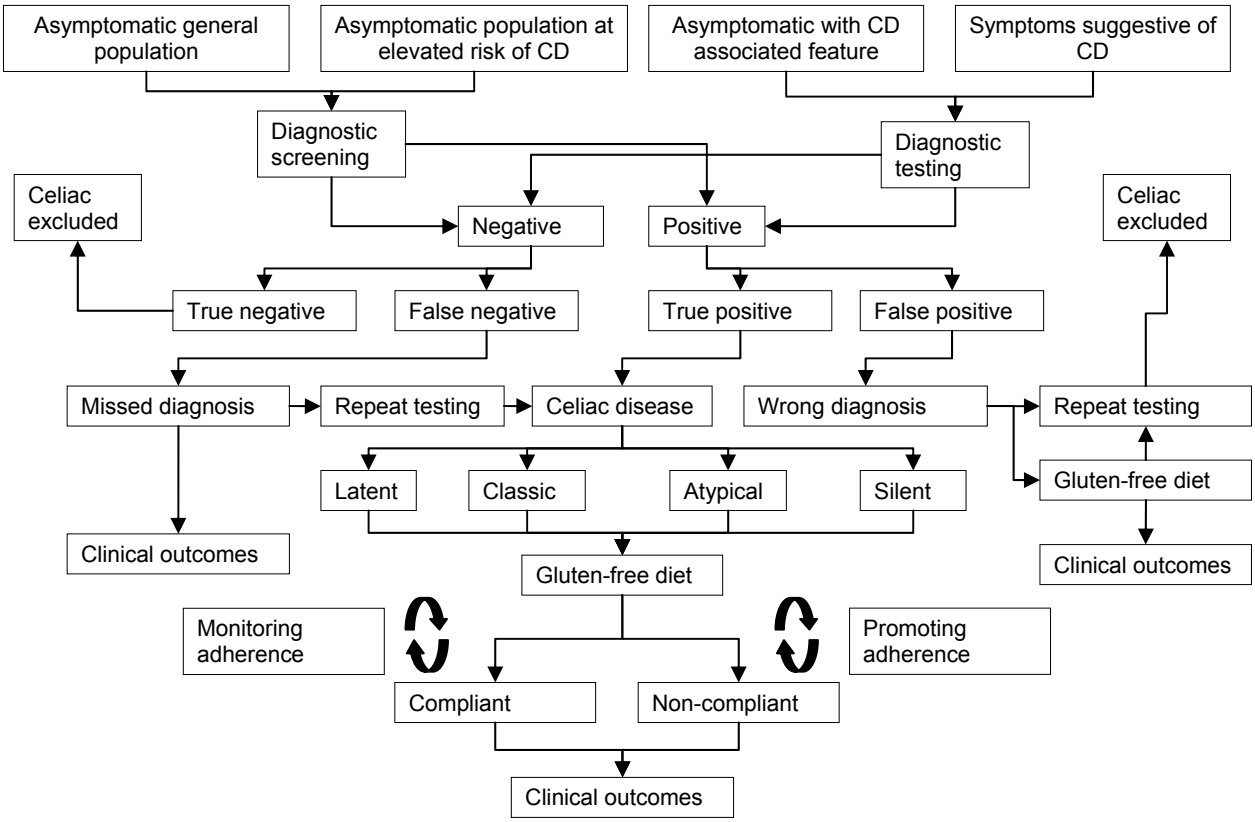
HLA DQ2/DQ8

The HLA DQ2 haplotype represents the occurrence of HLA class II heterodimer alleles DQA1*0501 and DQB1*0201. These typically occur in a cis position as HLA DR3-DQ2 or in a trans position as HLA DR5/DR7-DQ2. The HLA DQ8 haplotype DQA1*0301/DQB1*302 typically occurs in association with DR4.

Analytical Framework

The analytical framework is presented in Figure 1. In this framework, we wanted to represent the diagnostic pathways and the potential outcomes of testing various populations for CD. Each step of the pathway represents a portion of this systematic review, starting with the identification of the populations of interest, their diagnostic pathways, and ultimately the clinical outcomes, as well as consequences of testing.

Figure 1: Analytic framework



Study Identification

Although the objectives of this task order are contained within a request for a single evidence report, we conducted five separate reviews, from the literature search onwards, as the objectives of this mandate were more orthogonal than overlapping.

Search Strategy

A series of searches were performed by National Library of Medicine staff in support of the literature review for CD. Strategies were developed using the guidelines supplied by the UO-EPC, and were divided into the five questions posed by AHRQ. All searches were limited to human studies published in English language journal articles. The specific strategies used for each search are located in Appendix B.

1. What is the sensitivity and specificity of the following tests for CD:
    a. EMA
    b. human tTG IgA antibodies
    c. AGA
    d. HLA DQ2/DQ8
    e. small bowel biopsy

Searches were run in the MEDLINE® and EMBASE databases for each of the five tests. With the exception of the search for small bowel biopsy, a reference to CD or its synonyms was not a requirement for retrieval in order to obtain the widest possible information on these tests. Because of their complexity, a separate search was run for each test, then the results combined into one Pro-Cite file and duplicates eliminated. Individual case reports and letters to the editor were also removed.

The MEDLINE® searches were run in October 2003 for the year 1966 forward and yielded a total of 2885 citations, with a follow-up search for HLA DQ2 and DQ8 performed in November 2003 that yielded an additional 390 citations. The EMBASE searches were run in December 2003 for the year 1974 forward and yielded a total of 1,046 citations after duplicates to MEDLINE® were removed.

2. What is the prevalence and incidence of symptomatic and clinically silent CD in the general population and in the following identified high-risk populations:
    a. patients with an affected family member
    b. type 1 diabetes mellitus
    c. IDA
    d. osteoporosis

Searches were run in the MEDLINE® and EMBASE databases. The MEDLINE® search was performed in October 2003 for the year 1966 forward and retrieved a total of 1,584 citations. The EMBASE search was run in December 2003 for the year 1974 forward and yielded 467 citations after duplicates to the MEDLINE® retrieval were removed. Individual case reports and letters to the editor were also removed from both searches.

3. What is the association between CD and GI lymphoma?

Searches were run in the MEDLINE® and EMBASE databases. The MEDLINE® search was performed in October 2003 for the year 1966 forward and retrieved a total of 230 citations. The EMBASE search was run in December 2003 for the year 1974 forward and yielded 97 citations after duplicates to the MEDLINE® retrieval were removed. Individual case reports and letters to the editor were also removed from both searches.

4. What are the expected consequences of testing for CD in the following populations:
    a. patients with symptoms suggestive of CD
    b. asymptomatic, at-risk populations
    c. general population

Searches were run in the MEDLINE®, EMBASE, PsycINFO, AGRICOLA, CAB, and Sociological Abstracts databases. In order to obtain the widest possible retrieval, all articles on screening for celiac and its synonyms were included, not just those discussing consequences.

The MEDLINE® search was performed in October 2003 for the year 1966 forward and retrieved a total of 917 citations. The EMBASE (1974 forward), PsycINFO (1840 forward), AGRICOLA (1970 forward), CAB (1972 forward), and Sociological Abstracts (1963 forward) database searches were run in December 2003 and yielded a combined total of 204 citations after duplicates to the MEDLINE® retrieval were removed. Individual case reports and letters to the editor were also removed from both searches.

5. What interventions are effective for promoting or monitoring adherence to a GFD?

Searches were run in the MEDLINE®, EMBASE, PsycINFO, AGRICOLA, CAB, and Sociological Abstracts databases. Because of the small number of citations retrieved, a few selected articles discussing adherence to dietary limitations for other conditions were included. The MEDLINE® search was performed in October 2003 for the year 1966 forward and retrieved a total of 152 citations. The EMBASE (1974 forward), PsycINFO (1840 forward), AGRICOLA (1970 forward), CAB (1972 forward), and Sociological Abstracts (1963 forward) database searches were run in December 2003 and yielded a combined total of 168 citations after duplicates to the MEDLINE® retrieval were removed. Individual case reports and letters to the editor were also removed from both searches.

Some citations fulfilled the criteria of more than one celiac objective. Duplicates within each celiac objective were electronically removed. The obtained citations were uploaded into an internal web-based review system (SRS) for online collaborative citation screening and abstraction. Articles passing the first level screen were retrieved in full for further screening (see below).

Reference lists of included studies, book chapters, and narrative or systematic reviews retrieved after having passed the first level of relevance screening, were manually searched to identify additional unique references. Through contact with content experts, and the TEP, attempts were made to identify other studies not identified by the search.

Study Selection and Eligibility Criteria

Study selection was performed using three levels of screening with increasingly more strict criteria to ensure that all relevant articles were captured (Table 1). Each celiac objective had its own selection criteria for each level of screening and, as discussed previously, each celiac objective was treated as a separate sub-review. Following a calibration exercise, two reviewers

independently screened all studies using the SRS web-based system. This system allows automatic identification of review disagreements. Any disagreements were resolved by the two reviewers by consensus; rarely, a third reviewer was used to break an impasse. The specific screening questions for each screen level are included in Appendix C.

Level 1 broad screening. Level 1 screening was used to identify any potentially relevant citation, based on review of the title, abstract and key words. For each objective, the SRS system displayed the corresponding task order questions alongside the citation details. Reviewers answered a broad question of whether the citation potentially related to the current objective. Furthermore, the SRS system was set-up in such a way that articles which were identified in one celiac objective silo, that could also be relevant to another objective, could be identified and moved/copied to the other silo. The review team was divided up so that two members could be simultaneously reviewing each objective.

Level 2 refined screening. Potentially relevant articles identified at level 1 were obtained in full for level 2 screening. Again, using the SRS system with the actual articles on hand, reviewers selected articles that related to each of the specific objectives. The reviewers were asked to err on the side of inclusion for this level, and to classify articles as "original" or "review". Original articles meeting level 2 inclusion also had basic demographic data—such as screening test used, celiac definition, and study population identified—recorded into the SRS system.

Level 3 final screening. Level 3 screening identified articles that specifically allowed for the answering of the task order questions. These articles fulfilled the final inclusion/exclusion criteria, allowed actual extraction of the required data, and did not have fatal methodological flaws.

Table 1: Inclusion/exclusion criteria by level of screening

Objective	Level	Inclusion	Exclusion
Celiac 1	1	Any article reporting sensitivity/specificity of AGA, EMA, tTG, HLA DQ2/DQ8, or biopsy.	Clearly unrelated citation.
	2	For serology and HLA – articles where sensitivity and specificity could be extracted. For biopsy – articles were included if some measure of diagnostic utility could be obtained.	
	3	Articles that allowed determination of sensitivity or specificity for all tests were included.	• Articles with major methodological flaws excluded • Control group did not have gold standard test (biopsy) applied • No description of biopsy criteria given • Celiac group known to be positive for test under evaluation • Control group known to be negative for the test under evaluation • Control groups included patients with Marsh I or II biopsy lesions • AGA test performed without commercial ELISA kit or before 1990
Celiac 2	1	Any potential citation of prevalence or incidence of CD in general and high-risk populations or association of CD with other disorders	Clearly unrelated citation.
	2	Citations limited to those that gave evidence of the prevalence or incidence of CD in the general population or the AHRQ identified high-risk populations (e.g., diabetes, relatives, iron deficiency, osteoporosis). Countries: North America, western Europe, Australia, New Zealand.	Any studies of other CD-associated disorders not identified by the task order. Citations of the prevalence of specific disorders in patients with celiac (i.e., reverse of the inclusion). Any other country.
	3	Incidence and/or prevalence could be extracted from the article.	Serious methodological flaws: • patients identified by surveys, through solicitation of celiac societies • incidence studies without a population density denominator

Table 1 (cont'd): Inclusion/exclusion criteria by level of screening

Objective	Level	Inclusion	Exclusion
Celiac 3	1	Any potential citation of the association, prevalence or risk of lymphoma in CD, including articles on outcome of refractory sprue and ulcerative jejunoileitis.	Clearly unrelated citation.
	2	Measure of risk or prevalence/incidence of lymphoma in a population with CD.	Prevalence of CD in a population of lymphoma. Case reports and non-comparative case series.
	3	Extractable prevalence, incidence, or cumulative risk of lymphoma in CD.	Clonality of lymphocytes in ulcerative jejunoileitis-ileitis not determined or stated (as per TEP). Serious methodological flaw.
Celiac 4	1	Any potential citation of possible consequences of testing for CD.	Clearly unrelated citation.
	2	Consequences extractable from article.	
	3	Consequences limited to the AHRQ list.	Consequences obtainable from the other celiac objective sub-review – i.e., false positive and negative results, etc.
Celiac 5	1	Any potential citation of interventions for the monitoring or promotion of adherence.	Clearly unrelated citation.
	2	Studies of monitoring adherence were included if they assessed monitoring, by biopsy, serology (AGA publication date 1990 or later, EMA, tTG), or both. Any promotion intervention.	Serology prior to 1990.
	3	Data from article could be extracted. Data included follow-up by biopsy alone or serology with biopsy confirmation.	Articles assessing adherence through the measures of intestinal permeability. Studies that reported changes in mean serological titers with a GFD or gluten challenge, but did not address the potential usefulness of a serologic test to assess compliance.

Important articles answering a stated objective but not meeting inclusion criteria (i.e., containing potential threats to internal validity), were presented and discussed in the discussion section.

Data Abstraction

For each objective, a detailed and standardized data abstraction form was developed with the assistance of content experts and the TEP panel. The data abstraction forms included baseline study characteristics as well as questions allowing for the abstraction of all relevant study results and characteristics. The electronic data extraction forms began with basic study and patient

demographic questions that were common across the five sub-review forms. These included reviewer name, author name, publication year, publication type, study design type, and basic study population demographics such as race, age, gender, and type of CD population. The extraction forms then moved to specific questions geared at extracting data to answer the respective objective's questions. The individual data abstraction forms are included in Appendix C.

Celiac 1 (sensitivity and specificity) data abstraction form. Separate data abstraction forms were developed for serology, HLA, and the biopsy sub-questions. Two-by-two tables were used to abstract data on sensitivity and specificity, and to determine positive and negative predictive values and the prevalence of CD in the tested population. The biopsy studies were quite heterogeneous, and did not allow for direct numeric extraction of data.

Celiac 2 (prevalence and incidence) data abstraction form. For this objective, the data extraction form included questions for detailing the screened study population, the number of individuals screened, the number of CD cases identified and how CD was confirmed. For incidence studies, the comparison population and time period were recorded.

Celiac 3 (lymphoma) data abstraction form. In addition to the basic demographic, and study design data, the extraction form contained fields for the extraction of risk data linking GI lymphoma to CD. Types of data sought were prevalence and incidence of lymphoma in CD in the setting of comparison data from a control population. Fields for extracting standardized incidence, morbidity, and mortality ratios were included.

Celiac 4 (consequences of screening) data abstraction form. The extraction forms for this objective included text fields to detail the consequences of testing for CD. The form contained fields that identified the specific consequence of testing which was addressed by the study, as well as a data field to report the study findings. The general field approach was chosen to allow extraction of the expected varied data for this objective.

Celiac 5 (monitoring and promoting adherence) data abstraction form. For this objective, standard demographic data was collected, as well as the methods used to monitor adherence to a GFD, the response of those measures to the diet, and the correlation of serological methods with biopsy findings. Space was provided to detail the sensitivity and specificity of the monitoring method when that data was available. For the objective of promoting adherence to a GFD, a text-based form was used to allow the extractor to describe the intervention and the results of its use.

Electronic forms. The abstraction forms were developed in Microsoft Excel to allow for electronic data entry and recording, and to allow exporting the evidence table data into Microsoft Word. For each celiac objective, data abstraction was conducted by one reviewer and verified by another. The extracted data was further verified by one of the principal investigators.

Quality Assessment

The quality of reporting of diagnostic test studies was assessed using the QUADAS tool.[19] This tool is the first to be published that allows for the assessment of the quality of studies of diagnostic tests. The instrument was developed using a Delphi procedure. The Delphi panel consisted of nine experts in diagnostic research who refined an initial list of items in four rounds, after which agreement was reached on the items to be included in the tool. The QUADAS tool consists of 14 questions that are answered "yes," "no," or "unsure." The tool addresses the items individually and does not incorporate an overall quality score (Appendix D).

Cohort and case-control study reports were assessed using the Newcastle-Ottawa scale (NOS; Appendix D). The NOS is an ongoing collaboration between the Universities of Newcastle, Australia and Ottawa, Canada. It was developed to assess the quality of non-randomized studies with its design, content and ease-of-use directed to the task of incorporating the quality assessments in the interpretation of meta-analytic results. A "star system" has been developed in which a study is judged on three broad perspectives: the selection of the study groups; the comparability of the groups; and the ascertainment of either the exposure or outcome of interest for case-control or cohort studies, respectively. The goal of this project is to develop an instrument that provides an easy and convenient tool for quality assessment of non-randomized studies for use in a systematic review.

The inter- and intra-rater reliability of the NOS have been established. The face content validity of the NOS has been reviewed based on a critical review of the items by several experts in the field, who evaluated its clarity and completeness for the specific task of assessing the quality of studies to be used in a meta-analysis. Furthermore, the validity of the NOS criteria has been established by comparisons to more comprehensive but cumbersome scales. An assessment plan is being formulated for evaluating its construct validity, with consideration of the theoretical relationship of the NOS to external criteria and the internal structure of the NOS components.[20]

Quality assessments of cross-sectional reports were assessed using a 19-item instrument adapted from Ophthalmology (Appendix D).[21]

We did not conduct any sensitivity analysis of quality assessments on the observational studies, as there is little by way of guidance to suggest what a poor quality study score would be based on for these assessment instruments.

One reviewer assessed the quality of an entire celiac objective to maintain internal consistency. Quality assessment was not performed under masked conditions.

Data Synthesis and Analysis

The data obtained from this review fell into several broad categories, which correspond in large part to the individual study objectives. These will be addressed in turn.

Data for the sensitivity and specificity of each serological marker was considered separately. In addition, studies were subdivided by the population age group (adults, children, mixed population), and by study design (case control, relevant clinical population/cohort).

Attempts were made to identify, explain, and minimize clinical and statistical heterogeneity in the included studies. Heterogeneity was assessed graphically by plotting receiver operator (ROC) curves for each of the included studies in a given analysis. A Pearson's Chi Square with

n-1 degrees of freedom, where n represents the number of included studies in an analysis was calculated to assess statistical heterogeneity.

Pooled estimates were only calculated if clinically and statistically appropriate. In situations where pooling was not performed, a narrative systematic review was conducted.

There are several potential ways to pool the results of studies of diagnostic tests, each having both advantages and disadvantages. The simplest and most intuitive is to simply perform a weighted mean of the sensitivity and specificity for the studies in question. This method provides a pooled estimate that is easy to interpret by clinicians. Several other techniques involve the pooling of diagnostic odds ratios or likelihood ratios. These methods have the distinct disadvantage of difficulty in interpretation, and the inability to derive a pooled sensitivity or specificity from the resulting estimates. Lastly, one can use one of several methods to produce a summary ROC curve. The method described by Littenberg and Moses,[22,23] has the advantage of being able to produce a summary curve while taking into account a threshold effect. This can occur when different studies use different thresholds to define a positive test, or even from differences in labs using the same cut-off. To interpret summary ROC curves it is necessary to know the sensitivity or specificity of the test in question in the population in which it will be applied. Since neither of these values is estimable without conducting yet another diagnostic accuracy study for the given population, the clinical usefulness of using this method alone is limited.[24,25]

In order to produce clinically useful pooled statistics, we calculated a weighted mean of the sensitivity and specificity from those of the included study. For both sensitivity and specificity, this pooling relies on the assumption that the test statistic is the same in all of the included studies. For each pooled estimate, a 95% confidence interval (CI) was calculated using both a fixed and random effects model. The results of which were compared as a further test for heterogeneity. The pooled estimates for the sensitivity and specificity were also compared with a summary ROC curve calculated for the same group of studies as a second check of the estimates (summary ROC Curves are included in Appendix E).

The prevalence and incidence data from the Celiac 2 objective, and the CD-lymphoma data from the Celiac 3 objective, were anticipated to be quite heterogeneous considering the different, countries, age groups, and risk characteristics of the studied patients. Attempts were made to group studies of prevalence by age group, study population, and serological screening method. If the grouped studies did not show evidence of heterogeneity, pooled estimates of the prevalence were produced for that group of studies, otherwise a descriptive presentation of the data with a qualitative systematic review was conducted. Likewise, the outcome measures of the Celiac objectives 4 and 5 were presented in a qualitative systematic review, except in cases where it was possible to pool the sensitivity and specificity data as measures of monitoring of patients at various stages of recovery on a GFD.