Chapter 4. Discussion

Celiac 1: Sensitivity and Specificity of Tests for CD

Serology

Systematic reviews of studies of diagnostic accuracy are similar in many ways to reviews of other study types, such as randomized controlled trials. However, important differences exist in large part because of the weaknesses inherent to the diagnostic-accuracy study design and its potential sources of bias.[24] In addition to these considerations, the topic of CD introduces further difficulties, and bias because of the nature of how the disease itself is defined, and the methods of patient selection for inclusion in the study. Ideally, a diagnostic-accuracy study should include a consecutive or randomly selected sample of patients from a clinically relevant patient population. That is to say, a study population who's characteristics match those of the population in which the test will ultimately be used, and both patients and controls are selected from this population. Unfortunately, selection spectrum bias is common in studies of diagnostic tests in general, and in practice it is easier for investigators to select cases and controls as separate groups in a case-control design. The practice of choosing cases that have previously been identified as having the disease, especially if more severe, introduces bias in the estimates of sensitivity (artificially raising it), while choosing completely healthy individuals as controls introduces bias in the estimates of specificity—artificially raising it as well.[24] The importance of these biases comes back to the issue of the relevant clinical population. If the test is to be used in screening healthy individuals, then the estimate of the reported sensitivity is higher than it should, but the specificity estimate is likely valid. On the other hand, if the test is to be applied to suspected cases of the disease, then the reported estimate of sensitivity may not be that far off, but the specificity estimate would be higher than it should. Other important sources of bias also exist in relation to the study population, such as the mix of other diseases present in the population with similar features as the disease in question, and ensuring an appropriate mix of disease severity in the tested population. This last point regarding disease severity is especially important for this report, and is discussed at length below.

Lijmer et al.[423] reviewed 11 meta-analyses of diagnostic tests, and assessed the characteristics of the included studies using multivariate regression analysis. The authors identified several threats to the validity of a diagnostic study's results. Case-control designs overestimated diagnositic odds ratios (DORs) by three-fold compared with studies using a clinical cohort (relevant clinical population). As well, studies that applied different reference tests to those with and without disease (in case control) or to those testing positive or negative (in relevant clinical populations) overestimated the DOR by 2.2-fold. Interpreting the reference test, with knowledge of the results of the test under study, overestimated the DOR by 1.3-fold. DORs from studies without adequate descriptions of the test or study population were 70% and 40% higher, respectively, than in studies reporting these details. Inadequate descriptions of the reference test were also identified as sources of bias.

With this information at hand we tried to minimize bias in this report, by using what some may consider fairly strict inclusion criteria which also eliminated many poor quality studies. We included both case-control studies and cohort (relevant clinical population) designs but grouped

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

them separately. Studies were only included if an adequate description of the test under study and the reference test (biopsy, and a statement of the criteria defining CD) were provided, and both the cases and controls had to have had the same reference test (i.e., biopsy) applied at the same definition or level (i.e., biopsy grade).

The results of the systematic review demonstrate that in the studied populations IgA-EMA and IgA-tTG have sensitivities and specificities each in excess of 90% in both children and adults. In fact, the pooled specificity of EMA was 100% in adults using either EMA-ME or EMA-HU. In studies of children, the specificity of EMA using these two substrates was 97% and 95%, respectively, with overlapping 95% CIs, suggesting no statistical difference between these values. In adults, the pooled specificity of tTG-GP and tTG-HR were 95% and 98%, respectively, with overlapping CIs. Similarly, in children the specificities were 96% and 99%, again with overlapping CIs. Among the three studies in adults,[32,45,70] and four studies in children[35,52,70,79] that assessed both EMA and tTG, the specificities were nearly identical. Overall, these results suggest that EMA and tTG antibodies demonstrate extremely high specificities in both adults and children.

We identified a tendency towards greater variability in sensitivity between studies and between antibodies, compared with specificity. IgA-EMA-ME demonstrated sensitivities of 97% and 96% in adults and children, respectively. EMA-HU demonstrated a similar sensitivity of 97% in children, although the pooled estimate in adults was somewhat lower at 90%. Among two studies that assessed both EMA-ME and EMA-HU in adults, one demonstrated identical sensitivities of 95%,[81] whereas, the other[57] showed a lower sensitivity of HU compared with ME (90% vs 100%). This last study only included 20 untreated patients with CD, all of whom were ME positive, but two of whom were HU negative. None of the included mixed-age studies assessed both of these antibodies. Heterogeneity existed in the analyses of sensitivity of tTG-GP in the adult, but it is likely close to 90%. In children, the pooled estimate was 93%. The sensitivity of tTG-HR was 98% in adults and 96% in children, although in both cases the CIs included a low of 90%. In studies of mixed-age populations the sensitivity was 90%.

Estimates of the sensitivity of the IgG class antibodies of EMA and tTg suggest that these tests have poor sensitivities around 40%, although the specificities were quite high at around 98%. These finding suggest that this class of antibody would be inappropriate as a single test for CD, but may be useful in IgA deficient patients, or in combination with an IgA class antibody. One study that assessed the use of IgA-tTG-HR with IgG-tTG-HR found a sensitivity of 99% and a specificity of 100% for the combination.[72]

The analyses of all the AGA subgroups demonstrated significant heterogeneity, making pooled estimates impossible. Be that as it may, the sensitivity of IgA-AGA in adults is likely not much higher than 80%, but seems somewhat higher in children. The specificity likely lies between 80% and 90%, in adults and children, although the studies of serial testing of AGA followed by EMA or tTG in the prevalence section of this report suggest that the specificity is low as well. Even if one considers an optimistic range, the performance of IgA-AGA in both adults and children is inferior to that of the other antibodies discussed above.

The analyses of IgG-AGA suffered from significant clinical and statistical heterogeneity, making even general summary statements difficult. With this in mind, the typical sensitivity of this test likely lies below 80% in adults, and between 80% and 90% in children. The specificities are likely close to 80% in adults and between 80% and 90% in children with the same warning coming from the prevalence studies, suggesting that in the era of EMA and tTG, testing for CD with AGA has a limited role.

In assessing the PPV and NPV of these tests it is important to keep in mind the prevalence of CD in the tested population. In all the included studies, the prevalence of CD would be considered quite high, the minimum study prevalence was 9%, and many studies demonstrated prevalences in excess of 40%. In comparison, Fasano et al.[15] found the prevalence of CD in at-risk first-degree relatives of CD patients to be 4.55%. In general, based on our report, the prevalence of CD in high-risk groups such as suspected CD patients, and first-degree relatives was less than 20% (in non-tertiary centers), and the prevalence in patients with anemia and diabetes was generally less than 10% (Celiac 2 section). As expected, overall the included studies demonstrated the classic relationship between prevalence and the PPV and NPVs. At the relatively high prevalence of CD in these studies, the PPV (the chance that a positive test represents a true positive test) was quite high (>90%), but started dropping at a prevalence below 35% to values generally below 80%. Figures 21 and 22 represent the actual unweighted individual study data. It is therefore not surprising that the studies maintaining a high PPV at a low prevalence were all studies of small sample sizes. In the expected reverse relationship, at a prevalence above 45% the included studies showed a drop in the NPVs. However, in contrast to the situation with the PPV, the NPV would be expected to be between 95% and 100%, if not actually close to 100%, at the expected prevalence of CD in most clinical situations. The same relationship was seen when the pooled estimates of the sensitivity and specificity for each analysis group was used to calculate the PPV over a range of prevalences (Figure 23). Therefore, the potential problem with EMA and tTG serological testing lies in their performance in situations of "low" prevalence of CD (i.e., less than 20%, a value that is still higher than the prevalence of CD in most at-risk groups). Unfortunately, it was difficult to directly estimate the PPV of EMA and tTG based on the prevalence studies, such as the one by Fasano et al., since many of the studies only performed serology testing, or there was incomplete biopsy confirmation. However, in studies where it could be estimated using the best performing EMA or tTG serological test, the PPV ranged from 66.7% to 95.0%,[209,211,212,214,215,220,223,323] with all but one study having a PPV of less than 88.9%. Most of the studies had PPVs in the range of 70% to 80%. In this same group of studies that assessed the prevalence of CD in a general population, five studies showed 100% PPV, however, in all these studies there was less than ten confirmed CD cases,[213,217,222,225,231,269] and in three studies there were three or fewer confirmed cases.[217,222,231] The PPV of IgA/IgG AGA screening alone was considerably worse, and it was not uncommon in serial testing studies to see a ten-fold drop in potential cases when moving from AGA to subsequent EMA and tTG confirmation.

From the preceding discussion it is clear that in the diagnostic studies of the serological tests, the sensitivities of EMA and tTG antibodies for the detection of CD are quite high. Furthermore the specificities and NPVs are nearly perfect, making these antibodies appealing candidates for screening, as well as for the diagnosis of suspected CD patients. However, the pressing question is whether the reported high sensitivities and PPVs in these studies, and the enthusiasm surrounding these antibody tests, will hold true when these tests are applied to different clinically relevant populations. Of concern, is the true PPV of these tests when they are applied in populations with a relatively "low" prevalence (<10%-20%) of CD. This is an important issue, since the proportion of patients who would undergo unnecessary further testing will rise as the PPV falls. For example, if the PPV falls to a value of 80% (based on the examination of Figure 21), then 20% of screen-positive individuals would undergo unnecessary testing and/or treatments. From the estimates discussed above derived from the population screening studies,

and from the plots of PPV versus prevalence, it would appear that the PPV of these tests is potentially lower than the diagnositic test studies suggest it is.

The vast majority of studies, as well as our own TEP, required that the small intestinal mucosa show at least partial villous atrophy histologically for the diagnosis of CD to be made. In fact, most of the studies used patients with subtotal or total villous atrophy. Furthermore, inherent to the clinical definitions of classic, atypical, and silent CD described in the methods, is the requirement of having a "fully developed" villous atrophy. However, Fasano et al.,[15] in a large American prevalence study, found that only 34% of biopsied EMA-positive subjects had subtotal or total villous atrophy (modified Marsh IIIb or IIIc). In this study, no EMA-positive patient had a Marsh I lesion, 26% had a Marsh II lesion and 40% had a Marsh IIIa lesion. It is clear from this study, and from the discussion about biopsy later in this section, that true CD exists in patients with histologic grades less severe than classic Marsh III lesions, and that patients with silent CD do not have to have fully developed villous atrophy. The problem that then arises is whether the reported sensitivities of these antibodies holds in the majority of patients who have CD, yet with less severe histology. As well, if the sensitivity is not as high as reported then, by definition, the nearly perfect NPV of IgA EMA and tTG would also be expected to suffer.

This question has been answered in several studies that have correlated histology with the sensitivity of these serological markers, and also mirrors to some extent the antibody response that occurs once patients with CD are placed on a GFD. A description of results of these studies follows below, while a full narrative with tables is located in the Appendix H.

Rostami et al.[16] evaluated the diagnostic value of IgA EMA and AGA in 101 untreated patients with CD. The combination of the two tests showed an overall sensitivity of 76%. But, alarmingly, the sensitivity of EMA in these patients dropped precipitously with milder histological grades. EMA demonstrated a sensitivity of 100% in Marsh IIIc, 70% in Marsh IIIb and only 30% in Marsh IIIa. The authors did not consider patients with Marsh I or II lesions as having CD.

Tursi et al.[424] assessed the relationship of the histologic grade to tTG positivity in 119 consecutive adult CD patients defined by characteristic duodenal biopsy and "permanent gluten sensitive enteropathy." In this study, the frequency of tTG-positivity (sensitivity) and mean tTG levels, were greatest with the highest modified Marsh grade, and dropped steadily with milder histologic grades reaching a low of only 8% positivity in CD patients with Marsh I lesions. The sensitivities of tTG in Marsh IIIc, IIIb, IIIa, and II were 96%, 84%, 56%, and 33%, repectively. In another publication, likely using the same population of "permanent gluten-sensitive enteropathy," Tursi et al.[425] demonstrated similar results with AGA and EMA in a population of atypical CD (defined in methods). The sensitivities of EMA in Marsh IIIc, IIIb, IIIa, II, and I, were 97%, 92%, 89%, 40%, and 0%, respectively. The results with AGA showed a similar pattern, with the sensitivity dropping from 90% to 30% in March IIIc to Marsh II.

Furthermore, in likely the same population of "permanent gluten-sensitive enteropathy," Tursi et al.[426] found a relationship between clinical manifestation of CD and EMA sensitivity. EMA was positive in 77 of 96 (80.8%) patients with atypical CD and in 17 of 27 (63.0%) patients with silent CD. EMA was negative in patients with Marsh I lesions. Once again, assuming that all these patients with "permanent gluten-sensitive enteropathy" are truly CD patients, then EMA would miss 19% of atypical CD, and 37% of silent CD that were picked up on the basis of biopsy.

Demir et al.[427] studied the presentation and clinical features of 104 newly diagnosed Turkish children. EMA and biopsy correlation was available for 72 children. Similar to what was described above, EMA was positive in 92% of patients with Marsh III lesions versus 66.6% of patients with Marsh I-II lesions. Kotze et al.[428] assessed 47 symptomatic subjects with CD with intestinal biopsy, tTG and EMA antibodies. The authors found a statistically significant correlation between antibody titres of EMA and tTG, and histologic grades.

Hoffenberg et al.[317] studied a group of children at risk of CD who were part of a large prospective study of the genetic and environmental factors associated with autoimmune diseases. No relationship was found between Marsh grade and the genetic risk factor leading to screening, but a significant correlation was found between Marsh grade and tTG ($r=0.57$, $p<0.01$).

In a small case-control study assessing the diagnostic value of EMA, Sategna-Guidetti et al. also found that in patients with documented CD, EMA positivity correlated with the severity of the histologic grade.[429] In this study, EMA was falsely negative in 50% of CD patients without villous atrophy.

The findings of the large prevalence study by Fasano et al.,[15] however, require further discussion within this context. This study demonstrated a very high prevalence of CD of 0.95% (1:105) in asymptomatic not-at-risk adults using IgA-EMA. Additionally, 34% of biopsied EMA positive subjects had subtotal or total villous atrophy (modified Marsh IIIb or IIIc), 40% had a Marsh IIIa lesion, and 26% had a Marsh II lesion. No CD patient in this study had a Marsh I lesion, although this is in part likely due to how they defined CD. In any case, there are at least two ways to interpret these results. The first is that EMA testing does pick up the mild Marsh grades, given the high prevalence of CD in this study. While the second interpretation is that based on the preceding discussion and the serology monitoring data, this study has missed an unknown number of CD patients with milder histological grades. Unfortunately, since we do not have follow-up data on the screen-negative patients in this study, this question will be difficult to answer and arguments can be made on both sides.

The question that remains, however, is whether subjects with low grade histologic lesions are at the same risk of long-term complications as those with more advanced histologic grades. On the one hand, it is apparent that symptoms may not correlate with histologic grade but rather with the length of affected small bowel. When the distribution of histological grades is compared among patients with CD who are clinically asymptomatic versus symptomatic, the same distribution of grades is seen. For practical reasons, few of the studies we identified assessed length of small bowel involvement with CD. But another question arises: are patients with early March lesions who test positive for serology the ones who have more extensive small bowel disease?[430] These questions add to the uncertainty regarding the true performance of serological testing, and whether missing early grade histologic lesions is important. Although we could not find direct evidence comparing outcomes in patients based on their histologic grades, it is not unreasonable to think that a patient with Marsh I-II lesions would still have an increased risk of CD complications (see Celiac 4 and 5 for some data regarding this point).

In summary, it is clear that from our pooled estimates of the included studies that IgA-EMA and IgA-tTG antibodies provide excellent specificity for the diagnosis of CD. However, the high reported sensitivities may only apply to the selected group of patients with villous atrophy. Furthermore, if the sensitivity is in fact lower when the entire biopsy spectrum of CD is considered, then the nearly perfect NPV of these tests, particularly in low prevalence populations, would also be expected to suffer. Finally, the PPV of these tests may not be as high as suggested when the tests are applied in low-prevalence populations, as demonstrated by our

estimates of PPV from the population screening studies. These potential limitations of serological testing can have profound implications for population screening initiatives, and verification of the sensitivity of these antibodies in a large population of CD patients showing the full histological spectrum is urgently required.

HLA DQ2/DQ8

The HLA DQ2 haplotype represents the occurrence of the 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. HLA DQ2 occurs in about 20% to 40% of the general population,[9,10,15,100,135,136,138-141,143,146,147,150-157,159,167] 48% to 65% of healthy relatives of patients with CD,[158,161,164,166,167,169,172,177] and in up to 73% of non-CD patients with type I diabetes.[97,165] In one study, 100% of patients with enteropathy associated T-cell lymphoma (EATCL) were HLA DQ2 positive.[151] Non-CD patients with Down Syndrome appeared to have the same frequency of HLA DQ2 as the general population.[109,134,160]

Populations of non-Western European descent demonstrated very wide variations in the frequencies of HLA DQ2 both in CD patients and controls.[120,137,142,148,159,163]

Overall, it can be seen that HLA DQ2 alone offers a sensitivity in excess of 90%, which can be improved to close to 100% if a strategy of testing for both HLA DQ2 and HLA DQ8 is utilized (either test being positive). The specificity of both tests together, or either test alone, is not as good as the sensitivity, falling in the range of 55% to 80%. The specificity becomes considerably worse if a population with a higher expected frequency of HLA DQ2 or HLA DQ8, such as first-degree relatives of patients with CD or patients with type 1 diabetes, is tested. The PPV, (the probability that a positive test represents a true positive result) of testing for HLA DQ2/8 in an average population is generally low. One, however, needs to keep in mind the dependence of predictive values on the prevalence of CD in the population to be tested. Therefore, in high-risk groups, such as first-degree relatives or patients with type I diabetes, the PPV tends to be higher. Conversely, it appears that the value of testing for HLA DQ2/8 is highest when a negative test is found. Given the high NPV of this test, average-risk patients can have the diagnosis of CD excluded based on a negative test. The situation is more complex in high-risk groups, since the NPV decreases with increasing prevalence, and with the recognition that there are HLA DQ2/DQ8-negative patients with CD. These findings, along with the cost of HLA testing, make routine use of this modality for screening or diagnosis inappropriate. However, the use of this test is most useful in cases of diagnositic uncertainty or as part of a multi-test gold standard in clinical studies.

Biopsy

Unfortunately, we could not identify any studies that assessed the sensitivity or specificity of biopsy for the diagnosis of CD. This is perhaps not surprising considering that CD has historically been, and for the most part continues to be, diagnosed based on characteristic histological features. These histologic features have been classified and categorized by Marsh and others,[1,16] and criteria for the diagnosis of CD have been proposed,[2] and modified[4] (Appendix A). A biopsy showing characteristic features that improves with a GFD and recurs with gluten challenge is by definition the gold standard for the diagnosis of CD and therefore would be expected to be highly specific (some patients such as those with refractory sprue will not improve on a GFD but are still considered to have CD, so the specificity of this definition is not absolute nor perhaps completely valid). Although we do not have actual numbers, it would appear from the qualitative assessment of the identified articles that a biopsy classified as a Mash IIIa or higher is likely to have a high specificity for the diagnosis of CD. However, as seen in the study by Fasano et al.,[15] such criteria would be expected to have a low sensitivity. Alternatively, one would expect that biopsy could have a very high sensitivity if a Marsh I lesion was used to define CD, though clearly given the wide differential of mild histologic changes (Table 1, Appendix A), the specificity would be expected to drop. Therefore, to try to estimate the sensitivity and specificity of biopsy, and particularly the lower histology grades, we have compiled some articles below that provide "uncontrolled indirect information" on this subject.

Inter-observer agreement in the histologic assessment of small bowel pathology. As previously described, there are several potential criteria for the diagnosis of CD. The original and modified ESPGAN criteria[2,4] appear direct. Most of these criteria, as well as the assembled TEP, felt that some degree of villous abnormality is required for the diagnosis of CD. In practical terms, even distinguishing between a Marsh II (no villous abnormality) and a Marsh IIIa (minimal villous changes) can be difficult.[431]. This concern is further confounded by potential problems with the biopsy specimens themselves such as size, orientation, quality, and proper biopsy sampling. Hence, agreement between different pathologists and between the same pathologist at different times becomes important. The biopsy literature search identified a few articles that addressed pathologist aggreement.

Weile et al.[432] assessed inter and intra-observer agreement among three experienced Swedish and Danish pathologists reading the small bowel histology of patients suspected of having CD. Ninety small-bowel biopsies taken by capsule near the ligament of Treitz from 73 children were selected at random from a larger sample taken from 1987 to 1994. The final diagnosis was made on the basis of evaluation of specimens by dissecting microscopy, formalin-fixed H&E-stained slides, intestinal disaccaridases, serology and clinical presentation. The initial biopsy reports from patient files were sorted into normal (66; normal or minor nonspecific abnormalities—85% were on a gluten-containing diet [GCD]), pathological (17; total and severe villous atrophy, all on GCD), and inconclusive (seven; because of poor orientation, small sample, or autolysis). Several years later (1997) the same three pathologists who read the initial biopsies, performed a second reading of the slides given to them in random order. In comparison with the first reading, the number of inconclusive readings rose from seven to 22, there was a corresponding fall in the number biopsies read as normal and pathological. Considering the overall biopsy reading and diagnosis, the Kappa statistics (a statistical measure of agreement "correcting" for chance[433])

were (0.57, 0.63, and 0.75) for the three pair-wise comparisons of the three pathologists. These kappa values were reported to be "moderate" (for two out of the three agreement kappa scores) to "substantial" in terms of agreement, and suggest that agreement is far from perfect even when the same pathologist reads the same slide twice

Vilela et al.[431] also assessed inter-observer agreement among Brazilian pathologists in the diagnosis of CD. Three experienced masked pathologists independently read the slides of 34 patients with CD based on ESPGAN criteria. Agreement differed among the three possible pair-wise comparisons, with the best agreement occurring between pathologists A and C. Good to excellent agreement (kappa 0.61-0.85) was obtained for the assessment of villous structure. Reasonable to good agreement was observed for increased number of crypt mitosis (kappa 0.63), and decrease in the overall number of villi (kappa 0.47-0.53). However, agreement about the number of IELs using standard staining was weak (kappa 0.39). Interestingly, the agreement regarding overall histologic grade was also weak between two pathologist pairs, and reasonable to good for the last pair. As with the above study, it is difficult to comment on the generalizability of these results. The authors suggest that the number of CD cases seen was fewer than expected, and qualitative rather than quantitative measures of such parameters as villous height and IELs were used. Still, the findings suggest that agreement regarding the histologic grades should not be taken for granted.

Several authors have suggested that quantitating various histologic features, such as the number of IELs per 100 or more enterocytes, results in greater reproducibility of biopsy readings.[434] Authors that used quantitiative criteria during studies of inter-observer agreement likewise showed better agreement than reported above.[435-437] These studies suggest that the use of quantative methods in the reading and reporting of small bowel histology, by pathologists experienced in the reading of CD biopsy specimens, leads to greater agreement among pathologists and presumably more uniform and standardized reporting.

Latent CD. The presence of latent CD is a threat to the diagnostic accuracy of biopsy, since these patients truly have normal intestinal histology.

Stenhammar et al.[438] conducted an initial study of 100 first-degree relatives of 32 patients with CD. All 100 relatives were biopsied and two cases of CD were identified. In a 20-year follow-up study, Hogberg and Stenhammar[247] performed serological evaluation (AGA, EMA, tTg) on these same 100 relatives and their offspring, with positive results prompting intestinal biopsy. All relatives with initial "mild or moderate mucosal" abnormalities remained unchanged and were not considered to have CD. Eight new CD cases were identified, two of these were relatives of the two cases diagnosed in the first study. One of these, a parent of an affected child, had a grade II-III lesion in the first study that normalized on a GFD, and remained normal after 3 years of a GCD; she was not classified as CD, though in retrospect she likely represents a late relapser rather than transient gluten intolerance or a true latent CD. The other patient had a grade II lesion, but initially was not regarded as having CD because of the absence of symptoms. She was also found to be DQ2 positive. The remaining six newly diagnosed subjects were offspring of index CD cases and were not part of the initial cohort. In all, only two subjects of the initial biopsied cohort were "missed" in the first study. In retrospect, these subjects should have been included. This suggests that biopsy has the potential of high sensitivity and specificity for CD. Unfortunately, in the follow-up study, the number and HLA status of those with mild-to-moderate mucosal abnormalities (serology negative) was not reported, and since not all subjects

were rebiopsied it is also unclear if there is a group of serology-negative, initially normal biopsy relatives that have developed higher grade histology at follow-up, suggesting latent CD.

Maki et al.[62] likewise after an initial biopsy screen of 113 first-degree relatives of CD patients, discovered 13 relatives with villous atrophy and crypt hyperplasia. During a 3-year follow-up period another three relatives, with previously "normal biopsies" who were AGA positive, were found to have CD. Unfortunately, the authors do not report on the number of relatives with low-grade histologic lesions, and whether the new cases were in patients with completely normal (Marsh 0) lesions or normal in terms of absence of villous atrophy.

Troncone et al.[439] searched the medical records of 25 centres in Italy over a 10-year period to identify children with latent CD defined as either individuals with initial normal biopsies who later developed villous atrophy and responded to a GFD (Group 1), or people who were previously diagnosed with CD by ESPGAN criteria and who were subsequently found to have normal histology on a GCD for 2 years (Group 2). Nineteen such cases were found. All these patients had normal morphometric analysis and IEL counts on the initial biopsy. Four of the 14 GFD responders were considered at risk of CD (first degree, diabetes). The authors suggested that the five Group 2 patients could either represent true transient gluten-intolerance, or, in their opinion, more likely be late relapsers. These results of apparent post-pubertal recovery from CD are similar to those reported by Maki et al.[440] and by Schmitz.[441] Although the authors do not report on the number of charts or children screened, the findings of this study suggest that latent CD is very rare and unlikely to impact on the diagnostic accuracy of biopsy. It, however, underscores the importance of a time dimension in studies of CD, to accurately assess the true false positive and negative rates of diagnostic tests for CD.

IELs with normal villous structure. CD exists in patients with normal villous structure. The biopsy can pick up these patients on the basis of crypt changes and/or changes in the number and type of IELs.

Ferguson et al.[442] assessed the relationship of raised levels of IELs to the final diagnosis among children with diarrhea. The authors found a lack of correlation between IEL counts and morphologic grading of the biopsy. However, among seven children ultimately found to have no organic disease, all had normal IEL counts in the range of 14-25/100 epithelial cells (ECs). Two of three children with CD on a GFD also had normal IEL counts. In contrast, the values were elevated to greater than 38 IEL/100 ECs in untreated CD patients. High counts were also found in three children with failure to thrive or diarrhea of unknown etiology, and in three of nine children with giardiasis. Though in these cases, the mean values were lower than in the untreated CD cases. Interestingly, among 14 children with gastroenteritis, ten had abnormalities of the villi, crypts or lamina propria, but all but one had IEL counts within the normal range. Although, the differential of mild mucosal changes is large, this study suggests that one of the histologic features of CD can distinguish between CD and other mild enteropathies, and could potentially allow for a relatively high sensitivity by allowing CD to be defined by a low-grade Marsh lesion, while maintaining some of the specificity. This theme will be revisited in studies that follow.

Iltanen et al.[136] assessed the γδ+ IELs in patients with and without CD. One hundred and seven patients were evaluated for possible CD. Twenty seven were found to have CD (25%) on the basis of ESPGAN criteria. As well, 28 biopsy-negative adults who underwent endoscopy for dyspepsia were used as controls. Table 46 details the main study findings.

Table 46: Results of study assessing γδ+ IELs in patients with and without CD[136]

Test	Celiac (n=27)	CD excluded on biopsy (n=79)	Biopsy-negative controls (n=28)
Mean # of γδ+ IELs	40.4 (95%CI: 32.7-48.2)	6.7 (95%CI: 4.8-8.5)	1.6 (95% CI: 1.1-2.1)
Elevated γδ+ IELs (> 4.4 cells/mm)	27 (100%)	39 (49%)	n/a
AGA positive	21/26 (81%)	33/66 (50%)	n/a
Reticulin antibodies	27/27 (100%)	18/78 (23%)	n/a
HLA DQ2	19/21 (90%)	20/67 (30%)	

The mean density of γδ+ IELs was significantly greater in CD patients compared with those patients where CD was excluded on biopsy, and compared with biopsy-negative controls. The density of these IELs was also significantly higher in patients with CD excluded on biopsy compared with controls. Because the authors used the ESPGAN criteria, which requires some degree of villous atrophy, the 50% of subjects with CD excluded based on this criteria who were AGA positive begs the question of how many of these were actually CD patients. However, based on the reported data, elevated γδ+ IELs were calculated to have a sensitivity of 100%, but a specificity of only 50.6%, although the true specificity is likely higher. In the biopsy-negative suspected CD group, 66 out of the 79 underwent testing for HLA DQ2. Out of these patients, 46 tested negative for HLA DQ2. Given the high NPV of this test, it is likely that most of those patients do not have CD. Recalculating the specificity based on this assumption would raise its value, but unfortunately a breakdown of the number of patients with normal and elevated IEL in relation to HLA DQ2 was not reported. In any case, a better comparison would have been with the biopsy-negative control subjects, but the number of control subjects with raised IELs is not reported. Based on the mean density of IELs in this group, the number of patients with elevated IELs is likely to be low. During follow-up of the children suspected of having CD, but with normal mucosal biopsy and positive serology, four patients developed CD and responded to a GFD, further suggesting that this "control" group of patients with CD "excluded" on biopsy likely contained true CD patients who did not have villous atrophy. The results also suggest that the measurement of γδ+ IELs can be valuable in the diagnosis of CD, and hints at the fact that the requirement of villous atrophy on biopsy may miss some subjects with CD, particularly if they have raised IEL levels , positive serology and are HLA DQ2 positive.

Kutlu et al.[443] also studied the density of γδ+ IELs in untreated CD, treated CD and control patients (Table 47). The study population was made up of five children with classic CD with total villous atrophy and improvement on a GFD (Group A), seven patients studied after 1 to 11 years of a GFD with mucosal recovery (Group B), and 22 patients with CD by ESPGAN criteria who were left on a normal diet for 1 month to 10 years (Group C). The control group consisted of 15 children with various GI disorders other than CD, and 15 adults undergoing intestinal surgery for gastric and pancreatic disorders. The report aggregated data from groups A and C.

Table 47: Results of study assessing density of γδ+ IELs in patients with untreated CD, treated CD and control patients[443]

	Sub-total/total villous atrophy (n=18)	Moderate villous atrophy (n=7)	Normal mucosa (n=9)	Pediatric controls (n=15)	Adult controls (n=15)
Diet	normal		GFD	n/a	
γδ+ IELs/100 ECs	14.8	17.5	14.5	3.1	3.6

The density of γδ+ IELs/100 enterocytes was significantly higher in CD patients (15.4, n=34) compared with pediatric and adult control patients (3.1 and 3.6, respectively). However, the density did not correlate with histologic grade or with a GFD. Unfortunately, this study has several methodological flaws, and estimates of the sensitivity and or specificity of IEL in CD could not be derived. However, the study does indicate the potential usefulness of measuring γδ+ IELs in the overall evaluation of biopsy specimens for possible CD, and again demonstrates that CD patients can have a biopsy with normal villous structure which can be distinguished from normals by assessing the number of IELs.

In an interesting comparative study of the correlation of IELs with AGA positivity by ELISA, O'Farrelly et al.[444] studied 25 patients who had typical histologic features of CD and who were subsequently placed on a GFD. Ten of these were AGA positive, whereas 15 were negative. The second group consisted of 28 subjects suspected of CD but with "normal" small bowel histology. Twelve were AGA positive and 16 were negative. Increased levels of IELs were seen in both AGA positive (82.5) and negative (74.3) CD patients (difference not significant). On the other hand, among those with "normal" histology, AGA positive subjects had a significantly higher density of IELs than those who were AGA negative (42.4 vs 17, $p<0.001$). This data suggests that subjects suspected of CD with normal villous atrophy who have raised IEL densities should be further evaluated for CD, especially if serology is positive. These are also the types of patients where response to a GFD may be invaluable to firmly establish the diagnosis and help clarify the diagnostic value of low-grade histologic lesions.

Saputo et al.[445] compared the density of IELs between patients with confirmed CD, those undergoing investigation for CD, and control subjects (Table 48). The normal IEL range was determined to be between 4.68 and 17.60 based on the control group mean +/- 2 SD.

Table 48: Results of study comparing density of γδ+ IELs in patients with confirmed CD, those undergoing investigation for CD, and control subjects[445]

	Confirmed CD (n=9)	CD under investigation (n=40)	Controls (n=143)
IELs/50 ECs	68.55	51.21	11.14
# with raised IELs (estimated from figure)	9	40	2

These results again suggest the usefulness of IELs in the evaluation of histology of patients being assessed for CD, and suggest a sensitivity of raised IELs of 100%, and a specificity of 98.6%. Unfortunately, the authors do not report the number of individuals under investigation for CD who actually ended up having CD, so as to estimate the diagnostic parameters in this group.

Similarly, Jarvinen[436] studied IEL density and villous/crypt ratio in 928 Finnish patients with a suspicion of CD, and 59 biopsy-negative controls with dyspepsia (Table 49). CD was diagnosed on the basis of a suggestive small intestinal biopsy showing some degree of villous atrophy with subsequent later improvement on GFD. The main results excluding DH patients are presented below.

Table 49: Results of study comparing IEL density and villous/crypt ratio in patients with a suspicion of CD, and 59 biopsy-negative controls with dyspepsia[436]

	Untreated CD (n=138)	Treated CD (n=198)	Suspicion of CD with normal villi (n=545)	Controls (n=59)
CD3 + IELs	68*	40*	26	30
γδ+ IELs	19.8*	12*	3.2	2.3
Villous/crypt ratio	0.6*	1.9*	2.8	3.0
*statistically different from control				

The authors noted that using a cut off of 37 cells/mm for CD3+ and 4.3 cells/mm for γδ+ IELs, the sensitivities and specificities were 93% and 73% for CD3+, and 93% and 88% for raised γδ+ IELs, respectively. The PPVs and NPVs for raised γδ+ IELs were 95% and 85%, respectively, in this population. However, these results are based on the well-documented clear-cut CD group, and did not take into consideration the CD patients that might be in the suspicious but normal villi group. Among the patients with a suspicion of CD but normal villi and high γδ+ IELs (>4.3), 28% were EMA positive compared with only 8% with normal γδ+ IELs (<4.3). Unfortunately, the outcomes of these patients are not reported, so one cannot comment further based on this study about the usefulness of IELs in Marsh I or II patients.

Mino et al.[446] assessed the density of IELs in routinely stained specimens compared with specimens stained with the readily available CD3 antibody. Twenty-eight subjects with architecturally normal duodenal biopsies, which were well-oriented and demonstrated greater than 20 IELs/100 ECs were included in the study. AGA, EMA and tTG antibodies were measured. Subjects were divided in the groups listed in Table 50. Controls consisted of seven normal individuals, two patients with reflux, and two patients with irritable bowel syndrome.

Table 50: Results of study assessing IEL density in routinely stained specimens compared with specimens stained with the CD3 antibody[446]

	CD (n=8)	Treated CD (n=4)	Non-CD (n=16)	Controls (n=11)
Mean age	33.5	46.3	46.4	39.1
IELs/100 ECs by H&E staining	42.1	29.2	36.8	Not increased
IEL/100 ECs in villous tip by CD 3 staining	47.5	29.4	33.2	8.2

There were no statistically significant differences between any of the groups when IELs were measured with H&E staining. However, all pair-wise comparisons were statistically different, except between the treated CD group and the non-CD group, when villous-tip IELs were counted with CD3 staining. The authors conclude that villous tip IELs are more specific indicators of CD, particularly with CD3 staining (which is more readily available than staining for γδ+ IELs),

and suggest that the specificity of low grade Marsh lesions could be improved by these techniques.

In a similar study, Goldstein et al.[447] compared IEL density and villous distribution among patients suspected of CD. Twelve patients were diagnosed with CD based on histologic features and response to a GFD, whereas in 66 patients the diagnosis of CD was excluded based on biopsy, and supported by negative serology (and in some cases a lack of response to a GFD). Control cases consisted of patients with dyspepsia who underwent endoscopy and biopsy. The main results are summarized in Table 51.

Table 51: Results of study comparing IEL density and villous distribution among patients suspected of CD[447]

	CD (n=12)	Non-CD (n=66)	Controls (n=24)
Mean age	35.2	36.1	34.5
Iga EMA	8	3 (no response to GFD)	n/a
IgA AGA	5	13 (all EMA neg.)	n/a
Villous tip IELs	11.6	4.3	2.2
IELs distributed evenly along the villi	9/12 (75%)	3/68 (4%)	0
n/a = not applicable			

The authors found that the mean villous tip IEL density was significantly greater in the CD group than in the non-CD and control group. A more even distribution of IEL along the villi was also found to be significantly more common in the CD group compared with the other groups. However, this last point is controversial. Unfortunately, given that this is a small study, the authors did not look at differences in these characteristics among CD patients with different histologic grades.

Kuitumen et al.[448] compared the histologic features of children with untreated CD, treated CD, other GI disorders (cow's milk allergy, DH, congenital lactase deficiency, acrodermatitis enteropathica, and giardiasis) and a group of control subjects without GI pathology. Of the 52 children with CD in this group, all had severe villous atrophy. CD patients had the lowest enterocyte height, and the most intense IEL infiltration of the studied groups. The authors found no overlap between CD patients and controls for the density of IELs, villous height, crypt depth, and villous height to crypt depth; all these parameters were statistically different between the CD patients and controls.

Kaukinen et al.[449] studied 96 consecutive adults found to be ARA or AGA positive and compared them with 27 ARA- and AGA-negative patients with dyspepsia. All patients underwent duodenal biopsy and CD was diagnosed on the basis of a villous height to crypt depth of less than two and crypt hyperplasia. Twenty-nine patients met their biospsy criteria of CD (18 ARA- and AGA-positive patient, nine ARA-positive patients, and two AGA-positive patients). The 29 CD patients were placed on a GFD and of the 21 who were rebiopsied at 6 to 12 months, all showed unequivocal histologic improvement. The mean density of IELs in CD, serology positive, biopsy negative, and control patients were 87, 38, and 25 cells/mm, respectively. These numbers were statistically different. The mean density of $\gamma\delta$+ IELs among the CD patients was 16.6. Eleven serology-positive patients with normal villous structure (presumably Marsh I and II) expressed HLA DR and had higher levels of $\gamma\delta$+ IELs (mean of 13.4 cells/mm) than the non-CD controls. A repeat biopsy (time unspecified) was performed in 12 serology-positive patients

with normal villous structure at the time of the first biopsy. Ten of these had raised γδ+ IELs density on biopsy (Marsh I or greater). Five of these 12 were found to have villous atrophy (Marsh IIIa or greater). This study further illustrates the later development of CD in subjects with mild histologic changes, and suggests that although the specificity of villous atrophy may be high (all patients responded to a GFD), the sensitivity of villous atrophy (Marsh IIIa or higher) is lower than that of the serological test used in this study. This suggests that using a lower biopsy cut-off grade could improve sensitivity, albeit at the cost of specificity.

Using another approach, Wahab[450,451] identified 38 patients with symptoms of malabsorption who only demonstrated raised epithelial lymphocytes on duodenal biopsy (Marsh I). These patients were given a gluten challenge of 30g/day for 2 months, while maintaining their normal GFC. Twelve of 38 patients developed worsening mucosal lesions of crypt hyperplasia and partial or subtotal villous atrophy. After institution of a GFD all 12 patients showed improvement of their malabsorption, and improvement of their histology, suggesting that they truly had CD.

The same authors,[451] similarly studied 27 patients referred for malabsorption who were found to have a Marsh II lesion. HLA DQ2 or DQ8 was found in 21 of 27 patients (78%). The authors motivated 25 patients to follow a GFD, and all showed symptomatic improvement. The two patients who refused the GFD progressed to a Marsh IIIa lesion at follow-up. Although these data provide evidence of the true existence of CD in patients with Marsh II lesions, the frequency is unlikely to be as high as reported here. The high NPV of HLA DQ2/DQ8 suggests that at least some of the six testing negative likely don't have CD. In any case, this study adds further evidence to the notion that a Marsh III cut-off will miss some patients with CD.

In a very interesting study, Mahadeva et al.[452] identified all duodenal biopsies performed over a 1-year period with increased levels of IELs, yet normal villous structure. Biopsies were formalin fixed and stained with H&E. Other biopsies showing at least subtotal villous atrophy and increased IELs were considered as "suggestive of CD." Two normal control duodenal biopsies for every case of increased IELs with normal villous structure were also obtained. The upper limit of normal for IEL levels in this study was 22 IELs/100 ECs. Out of 626 biopsies assessed, 14 (2.2%) were found to have increased IEL and normal villous structure, whereas 15 (2.4%) cases of CD were identified. Normal histology was found in 502 (80.2%) of the biopsies. The biopsies with raised IELs had a mean of 38 IELs/100 ECs (range of 27-46). Control biopsies on the other hand had a mean of 12.4 IELs/100 ECs (range of 2-20). The presence of GI symptoms did not differentiate those with raised IELs from controls or CD patients in this cohort. Six of the 14 patients with raised IELs had positive EMA and/or unexplained anemia and were suggested as having "latent" CD by the authors. Unfortunately, follow-up in this group was incomplete with only three of these patients undergoing repeat biopsy. As with the previously described studies, the presence of patients evaluated for possible CD who have isolated increased IELs may contain a subset of true CD patients. In fact, if one assumes that the six EMA positive subjects with raised IELs do in fact have CD, then one can estimate that using a lower histologic grade to define CD in this population would have resulted in a sensitivity of biopsy of 100%, and a specificity of 98%—since only eight patients out of the studied sample of 531 would have been misclassified as having CD when in fact they did not. Of course, the expected specificity would not be as high as the one produced in this exercise since the authors do not tell us the histologic features or the diagnoses of the remaining 95 patients (626 biopsied, minus 502 normal, minus 15 CD, minus 14 raised IEL and normal villous structure = 95). However, taking this exercise further, if we assume that all of the other 95 patients were

misclassified as having CD, then the specificity would drop to a still respectable 83%. Clearly, this type of study is the starting point in assessing the diagnostic parameters of the biopsy itself as a test. However, what is needed to fully assess biopsy as a test is a clearer measure of the false positive and negative rates. This can only be accomplished by using a battery of tests (biopsy, serology, HLA) to act as a gold standard to initially identify all potential cases, and then a follow-up period (response to GFD or gluten challenge) to assess the permanence of the diagnosis and the utility of biopsy at various cut-offs when used alone.

Kaukinen et al.[453] performed a study partially fulfilling the above requirements. Ten patients with suspected CD but only Marsh I or II lesions were compared with 27 biopsy-normal controls. The suspected cases were assessed before and after a GFD. The main results are presented in Table 52.

Table 52: Results of study assessing patients with suspected CD and Marsh I or II, before and after a GFD[453]

	Histology	EMA+	TTG+	HLA DQ2	γδ+ IELs
Initially	Marsh III – 2 (patchy) Marsh II – 7 Marsh I – 1	8/10	9/10	9/9	Marsh III – 25 cells/mm Marsh I-II – 13 Controls – 1.4
After GFD	All Marsh II re-biopsied Marsh I – 2 Marsh 0 – 5	0/10	1/10 (Slightly elevated)	Same	Reported as decreased values not reported.

Although this is a small study with possible selection bias, the authors demonstrate that in a subset of patients suspected of having CD but without villous abnormalities, CD was diagnosed in all on the basis of a response to a GFD. Raised levels γδ+ IELs, positive serology, and HLA DQ2 positivity, supported the diagnosis of CD. Patients with CD and Marsh I-II lesions had significantly higher levels of IELs than controls. Unfortunately, this study did not include a larger sample of patients with Marsh I-II histology that included serology-negative subjects. Although it is clear based on this study that CD can exist in patients with Marsh I-II lesions with raised γδ+ IELs, it is difficult to generalize these results to an unselected sample of suspected CD patients.

In a somewhat complicated but important study, Kuakinen et al.[98] assessed 271 patients with suspected CD by biopsy. Forty-five patients were classified as having definite CD on the basis of a Marsh III lesion. While in 136 patients, CD was excluded on the basis of a Marsh 0 lesion and normal levels of γδ+ IELs. The remaining 76 patients had an uncertain diagnosis of CD based on biopsy (absence of villous atrophy) and underwent HLA DQ2 and DQ8 testing. In 59 of these patients, there were minor mucosal lesions or positive serological markers, while 17 were already on a GFD prior to biopsy. CD was excluded in 11 of these 17 patients on a GFD. Of the remaining 59 patients, CD was excluded in 22 because of a negative HLA DQ2/8 given the high NPV of this test, whereas 37 were DQ2/8 positive and remained with the suspicion of CD. Overall, CD was excluded in 33 of 76 patients. Among patients suspected of CD, but without villous atrophy, Marsh I-II lesions were found in 20 DQ2/8-positive patients versus in five DQ2/8-negative patients. Elevated levels of γδ+ IELs were found in 20 patients who were DQ2/8 positive compared with seven patients who were DQ2/8 negative, and IgA-EMA was found in 16 patients who were DQ2/8 positive compared with 0 patients who were DQ2/8 negative. Although data is not provided for some patients, one can estimate the sensitivity of

using a Marsh III cut-off. We know that CD was diagnosed outright in 45 out of 271 patients, but with subsequent testing a further 37 patients were found to be positive for HLA DQ2 or DQ8. At least 16 (EMA positive) and likely 20 (increased IEL counts) of these patients likely have CD. Based on these assumptions, the sensitivity of a Marsh III cut-off is between 69% (20 DQ2/8 patients with increased IELs have CD) and 74% (16 EMA and DQ2/8-positive patients have CD). The sensitivity would be lower if more of the DQ2/8 positive patients turned out to have CD. The specificity of that cuff-off would appear to be 100%, although we are not told if the Marsh III patients all improved on a GFD. Clearly using a biopsy cut-off lower than Marsh III would have increased the sensitivity, but unfortunately we are not given enough information to estimate this reliably.

This study with its battery of tests comes closer to the ideal design to estimate the diagnostic characteristics of biopsy, but unfortunately, it has significant short comings. To be fair the intent of the study was not to determine the sensitivity of a Marsh III cut-off. However, for the sake of future studies in this area, several design changes could have allowed this estimation. This study had two important positive aspects: it used a relevant clinically important population of patients suspected of having CD, and all the subjects underwent biopsy. However, it would have been ideal, if all the subjects also underwent HLA testing and serology. Furthermore, a follow-up of positive and negative patients, and or the assessment of the response to a GFD or the use a gluten-challenge in difficult to diagnose patients, would have allowed for the estimation of false positive and negative cases.

Relationship of serology to histology. As the data from the previous discussion suggests, CD clearly exists in patients with histological grades milder than Marsh IIIa. The fact that the sensitivity of biopsy is improved by using a lower grade as a cut-off brings up an important question. If the preceding statement is true, then what test is most sensitive for detecting CD with mild histologic changes—biopsy or serology? The issues surrounding this discussion have been addressed in the later portion of the serology discussion section, and a detailed narrative summary of the studies of the relationship of serology to histology can be found in Appendix H. However, to summarize, data from these studies as well as some data from Celiac 5 suggest that the sensitivity of serology drops with milder histologic grades, and suggests that serology alone would miss CD patients with mild histology grades.

In summary, CD exists in patients with histology grades less than Marsh IIIa. The sensitivity of biopsy at a Marsh IIIa or higher cut-off is likely less than that of serology with EMA or tTG. If lower Marsh grades are used, the sensitivity of biopsy increases, and it is possible that if morphometeric techniques including assessing IEL densities are used, the specificity may not suffer greatly. Ultimately, the question of the true sensitivity of biopsy can only be answered with a well-conducted study that attempts to identify all possible CD patients in a given clinically relevant population using multiple simultaneous tests (e.g., serology, HLA) in addition to biopsy. All patients, those who clearly have CD, those in whom CD seems excluded, as well as equivocal cases, need to be followed for the assessment of the permanence of their "diagnoses." Equivocal cases could also be considered for further testing, either with assessing response to a GFD or gluten challenge, to help in the clarification of their diagnosis. Although there are other potential variables to consider, with these measures, assessment of the false positive and false negative rates of biopsy, and hence a clearer estimate of the sensitivity and specificity, can be determined.

Celiac 2: Incidence and Prevalence of CD

Incidence in the General Population—Different Geographic and Racial/Ethnic Populations

The crude incidence of CD among western European and North American countries over the past 25 years has varied between 1 and 51 per 100,000, and the cumulative incidence by age 5 between 0.118 and 9 per 1,000 livebirths. Notable variations in CD incidence have not only been striking between neighbouring countries, such as is the case for Sweden and Denmark, but also between time periods for the same region, such was noted in the UK between the 70's and 80's as well as in Sweden over the 90's.

It is important to note that there were important methodological differences among the studies, from using patient registers[200] to actively screening at-risk patients.[128] Clinical practice also varied between time periods and regions. The advent of serological testing in the early 90's changed attitudes towards screening and identifying populations at risk with resulting higher detected incidences of CD. In some studies, active efforts were made to detect CD among asymptomatic subjects, such as the case in Finland where all subjects referred for endoscopy underwent small intestinal biopsy, independent of the cause for referral.[199] The incidence of CD is also expected to vary according to the genetic make-up of the studied population, although the prevalence of at-risk HLA haplotypes was only noted in one study.[128] These observations also highlighted the importance of dietary factors in triggering so-called CD epidemics among genetically predisposed populations. It would appear that breastfeeding bears a protective role, while early introduction of gluten, as well as the amount of gluten content in the diet may promote the early serological and pathological manifestations of CD. It is unknown whether these factors trigger an earlier expression of a disease which would become manifest anyway, or whether they trigger the appearance of a disease which may not otherwise occur, even later on in life.

In conclusion, caution should be exercised when extrapolating the noted incidence for one given region to a whole country, in particular in countries such as the US where there are differing population ethnicities among regions, between rural and urban areas, as well as between small and large cities. However, it remains that the true incidence and prevalence of CD are if anything greater than reported in clinical settings, since observations derived from screening and case-finding efforts were consistently greater than those relying on the diagnosis of clinically suspected cases. Lastly, it is important to bear in mind that, considering the large proportion of subjects with silent CD (the so-called celiac iceberg), observed incidences will depend upon the efforts spent screening cases, as is well illustrated by the difference in the relatively low incidence observed over 30 years in Olmstead county, where the majority of cases had clinically overt disease, as opposed to the very high incidence noted in Denver Colorado that resulted from a systematic and prospective screening of newborns and children at risk.

Prevalence in the General Population—Different Geographic and Racial/Ethnic Populations

The included prevalence studies demonstrated important differences in execution, tests for prevalence assessment, and in patient sampling, making pooled estimates of prevalence unreliable. Furthermore, the discussions regarding the operational characteristics of the serological tests themselves, the influence of disease prevalence on the PPVs and NPVs of these tests, and the criteria by which clinical and histological CD is defined, have to be kept in mind when considering the results of this section. The last point regarding the histologic definition of CD is particularly important in this setting, since one-third of the included studies did not seek histologic confirmation of serology diagnosed CD, and in another four studies, a large proportion of the serology-diagnosed patients did not undergo histologic confirmation. Finally, because of the previously discussed concerns regarding the sensitivity of serological tests in lower grade histological lesions, and the potential for missing true CD patients based on histologic criteria that require villous atrophy, the true prevalence of CD in the general population may still have been underestimated in these studies.

With these points in mind, the results of this report suggest that the prevalence of CD in the general unselected populations of North America and Western Europe is quite high and likely falls within the range of 0.5% to 1.26% (1:200 to 1:79). Smaller sample-size studies tended to give wider estimates ranging from 0.17% to 2.67%. Among the studies from the US, the range of prevalence was 0.4% to 0.95% in adults, and 0.31% in children. In Italy, the range of prevalence was between 0.2% and 0.8%, whereas the Scandinavian countries, Ireland and the UK, tended to show a higher prevalence of CD of approximately 1.0% to 1.5%, although there were also studies from those same countries that showed a lower prevalence.

In summary, the prevalence of CD in Western populations is likely close to 1% (1:100) and may be higher in Northern European countries. A firm estimate of the prevalence is impeded by between-study differences, and uncertainties regarding the performance of serological tests at these relatively "low" prevalences, compared with the 40% to 60% prevalences in the studies of the diagnostic characteristics of these same tests (Celiac 1).

Prevalence of CD in Patients with Suspected CD

The prevalence of CD is greatly affected by the study population. In populations where the diagnosis of CD is clinically suspected, either because of the presenting symptoms or the presence of associated conditions, its prevalence varied between 1.1%[307] and 50%.[301] This illustrates well how the patient selection process will influence the prevalence of the condition—studies reporting very high prevalence had populations that originated from tertiary, referral centers, while studies reporting low prevalence had populations that tended to originate from general practice. Although the report of the large American study of CD prevalence in at-risk and not-at-risk individuals did not specify how their subjects had been gathered,[206] we can assume that these were derived from community practices, considering their large number.

Altogether the variations between the study populations, the diagnostic criteria and the study design were such that it was inappropriate to statistically combine the observed prevalence to obtain a summary measure. Nonetheless, considering studies with subjects who were not originating from a specialized referral centre, the observed prevalence of CD in subjects with symptoms or conditions associated with CD ranged between 1% and 4%.

Prevalence of CD in Patients with Type I Diabetes

The findings of this report suggest that the prevalence of CD in patients with type I diabetes is higher than the prevalence in the general not-at-risk population. These findings appear to be consistent across the studied age groups, and by the screening method. Although the magnitude of the risk of CD among patients with diabetes varied to some degree from study to study, many of these differences can be explained by issues of study design. An overall pooled estimate of the prevalence of CD in diabetes could was not calculated due to these study differences.

Almost uniformly, the prevalence of CD by biopsy was to some degree lower than the prevalence by serology. This may reflect the fact that there were some false-positive serology results in the prevalence of CD seen in these studies. Additionally, all these studies used some degree of villous atrophy to make a diagnosis of CD, which may underestimate the true biopsy prevalence of CD, since CD patients with Marsh I or II lesions were not considered. The prevalence by biopsy seemed to be lower still in studies that require subtotal or greater villous atrophy to make a diagnosis of CD. Furthermore, the prevalence by biopsy was uniformly low, as would be expected, in studies in which a large proportion of the screen-positive patients did not undergo biopsy. In these studies, the prevalence by biopsy was typically less than two percent, which likely represents an underestimation of the true prevalence of CD in this population.

The prevalence of CD by serology varied greatly with lows near 1% and highs close to 12%. However, the majority of studies, and particularly those using EMA or tTG, demonstrated prevalences in the range of 4% to 6%. Although the prevalence by biopsy also varied, the typical study with complete biopsy confirmation of serology-positive patients demonstrated prevalences in the range of 3% to 6%.

This evidence report has gathered the reported studies examining the relationship between diabetes and CD. Baring in mind the limitations noted above, we believe there is sufficient evidence to show individuals with type I diabetes are at higher risk of CD. The prevalence of CD in this population is likely between 3% and 6%.

Prevalence of CD in Relatives of Patients with CD

The prevalence is CD in relatives of patients with CD is elevated, both in first-degree and second-degree relatives. That prevalence varied between 2.8%[246] and 17.2%[235] in first-degree relatives and between 2.6%[206] and 19.5%[235] in second-degree relatives. The prevalence remains elevated among first cousins, and was 17% in the only study of these subjects.[235]

We have identified several factors that can be responsible for the variation in the observed prevalence. In particular, the selection of the families, of the relation to the index case, the diagnostic criteria, and the choice of study design.

The prevalence of CD appears to be generally higher in families with multiple known cases, such as reported by Book et al.[235] and Mustalahti et al.[241] Most other studies referred to their subjects as originating from a “CD family,” without systematically documenting the proportion of families with multiple known cases of either CD or DH.

As expected, in studies that looked at various degrees of relation, the risk was greatest in the first-degree relatives.[206,235,239] However, Book et al.[235] found no difference in prevalence

between second-degree relatives and first cousins, i.e., 19.5% (95% CI: 15.1-23.9) and 17.0% (95% CI: 6.4-27.7), respectively.

Also, the age of the screened population might be a factor even beyond infancy, since it has been observed by prospective serological[248] and histological[237] follow-up studies that the serological and histological markers of CD can develop after an initial negative screen in a genetically predisposed individual. Therefore, a one-time assessment or screen in these individuals may be insufficient.

The serological diagnosis of CD will be affected by the diagnostic accuracy of the test. Fortunately, 11 out of 12 studies that used serological screening were EMA-based, a test with good diagnostic accuracy in populations with relatively high prevalence, such as relatives of CD patients. The single non-EMA study[236] used AGA, a test with a lower sensitivity and specificity than EMA, but all seropositive subjects underwent a confirmatory intestinal biopsy.

The histologic diagnostic criteria also affect the reported prevalence, as was well illustrated by the study by Tursi et al.,[249] where Marsh grades of I and II were also considered diagnostic, resulting in a prevalence of 44.1%.

The study design, especially whether all at-risk individuals are biopsied as opposed to solely those that satisfy a non-invasive criteria, is also to be considered. The EMA-based serological tests can miss milder forms of enteropathy as has been discussed, and this may explain why the prevalence of CD was generally higher in studies where all identified relatives were biopsied.

Prevalence of CD in Patients with Anemia

The results of this report demonstrate an increased prevalence of CD in patients with IDA. The prevalence is highest (between 10% and 30%) in studies of patients with GI symptoms, or in patients who have no gross lesions seen at initial investigation. CD appears to also be common in premenopausal women, both with (4.5%) and without (33%) heavy periods. Overall, in asymptomatic IDA patients assessed by serology or biopsy, the prevalence of CD was between 2.3% and 6%. Therefore, patients with IDA, particularly those without a clearly identifiable cause, should be evaluated for CD as part of their investigation.

Prevalence of CD in Patients with Low BMD

The studies of the prevalence of CD in patients with low BMD suggest that between 0.9% and 3% of patients with osteoporosis have CD. As a comparison, Fasano et al.[15] found that in the United States 0.75% of the general not-at-risk population, and 4.55% of first degree relatives of CD patients were found to have CD.

The results from these studies should be interpreted within the context of some methodological limitations. Three of them used AGA as the initial screening test to prompt further investigation, and we have shown that the sensitivity of this test is not high. Furthermore, the biopsy criteria used to define CD was either not reported, or required the presence of subtotal, or greater villous atrophy (Marsh IIIb or greater). We have also shown that CD exists in patients with lower grade histological lesions. Furthermore, the study results are contradictory. Two showed a risk of CD higher than the general population,[296,298] while the other two did not. In particular, the study by Mather et al.[297] found that seven out of the 96 screened patients were positive for EMA-ME, but none of these were positive on biopsy. From what we have seen regarding the specificity of this test being close to 100% (and therefore the

PPV would be expected to be high as well), it is unlikely that there are so many false positives even if the prevalence of CD was low, and raises the question of whether early grade CD patients remained undiagnosed. As such, it is difficult to draw any firm conclusions about the true prevalence of CD in this population, given the contradictory results, the fact that lower grade lesions were not considered, and that no follow-up data was provided on the patients who screened positive for serology but did not meet the biopsy criteria. Taking into account these limitations, it is likely that the prevalence of CD in patients with osteoporosis is higher than that in the general population.

Celiac 3: Risk of Lymphoma in CD

The association between malabsorption and lymphoma is a concept that has evolved over the past century. The observation that a significant proportion of patients with intestinal lymphoma also had villous atrophy at a distance from the malignancy, or had previously been diagnosed with CD, led to the publication of several series on the topic.

Although the objective of the task order was not to determine the risk of CD in lymphoma per se, the broad coverage of our search strategy also allowed us to systematically appraise the literature on this question, and were able to identify only two controlled studies on this association, which we describe here.[454,455]

Johnson et al.[455] performed a retrospective search of the five main pathology laboratories serving Northern Ireland to identify all the incident cases of small bowel lymphomas (SBL) and small bowel adenocarcinoma from 1987 to 1996. The clinical presentation of the cases, as well as the presence or absence of villous atrophy at a distance, were noted. The prevalence of CD in this group of SBLs was compared with that of the general population in Northern Ireland, as observed from serological screening of the population at large.[188] There were 13 cases of CD (gender not reported) out of 69 cases of SBL, all of which were ETCLs. Only one out the 13 CD cases was known to have CD prior to the diagnosis of SBL. The OR of CD in SBL was 27.98 (95% CI: 11.88-65.81) compared with the general population. The OR of unrecognized CD in SBL was 15.72 (95% CI: 9.71-25.45) compared with the general population.

In a prospective multicenter Italian study conducted between 1996 and 1999, Catassi et al.[454] screened newly diagnosed adult patients with NHL for CD using EMA and AGA testing; EMA-positive or IgA-deficient patients underwent small bowel biopsy. There were six cases of CD out of 653 patients with NHL (prevalence 0.92%). Three had B-cell and three had T-cell lymphomas. Four out of six cases had lymphoma primarily located in the gut. Two patients were known to have CD for more than 1 year, one of whom was poorly adhering to a GFD. Two cases had been diagnosed with CD within 1 year of the diagnosis of NHL, whereas two other cases had no prior CD diagnosis. The prevalence of CD among these NHL patients was compared with that observed in two Italian studies which performed large scale screening for CD.[126,222] The OR of CD in NHL was 3.1 (95% CI: 1.3-7.6) compared with an age-and sex-matched population.

These observations point to a clear association between CD and lymphoma. To determine the degree of association, or to quantify the risk of lymphoma in CD, we searched the literature for controlled studies of the incidence of lymphoma in CD. Unfortunately, the majority of publications on lymphoma in CD were uncontrolled. Typically, patients diagnosed with CD in a single institution were followed over time and the incident cases of lymphoma were described,

along with characteristics of the affected patients, the course of their CD and the histological type of lymphoma. Unfortunately, such studies provide little confidence to estimate the true risk of lymphoma in CD, since lymphoma per se will occur in the general population. The incidence of lymphoma has to be compared with "controls," matched on various characteristics such as age, sex, period and population. Any study that did not adjust the observed incidence to the expected incidence for age- and sex-matched individuals of the same population was deemed uncontrolled and excluded.

Cohort studies, either prospective or retrospective, constituted the majority of controlled studies. The incidence of lymphoma in a cohort of biopsy-proven CD patients, calculated as the number of lymphomas divided by the number of patient-years of follow up, was compared with that of an age- and sex-matched population from the same geographic area and time-period.

The SIR therefore represents the likelihood of lymphoma in CD patients relative to those who do not have CD in the same population. The value of the denominator reflects the incidence of lymphoma in a given population, so that it is not possible to pool SIR's from different populations.

The AR, however, is a measure of association that provides information about the absolute excess risk of disease in CD patients compared with "non-afflicted" individuals. This measure is defined as the difference between the incidence rates in the CD patients and normal population and, in a cohort study, can be calculated as the difference of cumulative incidence (risk difference) or incidence densities (rate difference) depending on the study design. The AR is a measure of risk which can be pooled; however, since incidence rates were reported in only two studies, we had insufficient data to generate a representative summary statistic.

Furthermore, studies varied greatly at several levels, in particular with respect to the definition of an incident case of lymphoma, the reported outcome measure, and the CD population selection.

Studies differed in their definition of observed cases of lymphoma, in the following manners:

1. Inclusion of malignancies that antedated the diagnosis of CD. In one American study, the number of at-risk years was calculated both from the time of CD diagnosis and from the time of onset of symptoms that could be attributed to CD.[340] In a prior national survey to patients with CD,[456] these authors had collected evidence to support that there is usually a long duration of symptoms before a diagnosis of CD is made in the United States, so that they considered this account justifiable. However, authors from other countries would specifically exclude the malignancies that were diagnosed prior to CD, assuming that it was unknown whether these were truly "at-risk" periods and that this account could falsely inflate the incidence of lymphoma in CD.[333] Considering that publications uniformly calculated and reported the incidence ratio based on the time period from the CD diagnosis, this is the measure of risk that we selected.
2. Inclusion of malignancies that were recognized simultaneously to the diagnosis of CD (i.e., within 1 to 12 months of diagnosis). In some cases, the diagnosis of CD can be unknown until the presentation of lymphoma. This fact highlights the possibility that lymphoma can occur in asymptomatic patients with CD. Although the importance of such cases is undeniable, the account of such cases can introduce bias and inflate the incidence of lymphoma in CD. In other words, the simultaneous diagnosis of CD and lymphoma is similar to an incident case in a patient with a "zero" duration of follow-up, i.e., is closer to a measure of prevalence than incidence. The inclusion of cases of

lymphoma occurring in patients with previously undiagnosed CD should theoretically be related to all cases of CD, diagnosed and undiagnosed, in order to give an accurate estimate of incidence, which is obviously impossible. However, some studies chose to include such cases, while others excluded them from the incidence calculation. This distinction was noted in the results presentation.

3. Exclusion of malignancies that were diagnosed incidentally at autopsy. In their large Swedish cohort of individuals hospitalized with CD, Askling et al.[337] also excluded unsuspected autopsy diagnoses of lymphoma, assuming that such entities would have been silent during life, and that they therefore could not be controlled for in the comparator group.
4. Case definition of lymphoma. Lymphomas are broadly categorized as Hodgkin's lymphomas and NHLs. The lymphomas that have been associated with CD have typically been of the NHL type, and so the majority of studies sought cases of NHL, with the exception of the Scottish study from Logan,[336] where both Hodgkin's and NHLs were reported.

The reported outcome measures also varied and impaired our ability to combine observations. Some studies reported the incidence of lymphoma, while others, relying on death certificates for ascertainment of outcomes, reported on the mortality from lymphoma.

Finally, the patient selection also varied, along with the reporting of the circumstances that led to the diagnosis of CD. These factors limited our ability to draw conclusions on the risk of lymphoma in symptomatic versus asymptomatic patienst with CD.

We were also unable to find controlled data on the risk of lymphoma in refractory CD, an objective which had been suggested by the TEP. We did find, however, two prospective studies and one retrospective study that could lend support to the notion that the risk of lymphoma in refractory CD is greater than that of responsive CD.[457-459]

In the Netherlands, Wahab et al.[457] prospectively followed 158 biopsy-proven CD patients to assess the recovery of histological changes with a GFD over time. There were 11 incident cases of refractory CD with more than 5-years of follow-up, five of whom developed ETCL, in contrast to none of the remaining GFD-responding CD patients.

Goerres[458] reported on 18 patients diagnosed with refractory CD between 1998 and 2000, gathered from all over the Netherlands, whom they treated with azathioprine and prednisone. There were three men and 15 women, with a mean age of 58 years (range 39-82). Subtypes of IEL populations were analyzed by flow cytometry, allowing for the classification of refractory CD patients into two types: type I refractory CD (n=10), in which a normal IEL population is seen, and type II refractory CD (n=8), in which an aberrant IEL population is present. All of the patients with type I refractory CD responded to combined azathioprine-prednisone therapy, whereas none of the patients with type II refractory CD showed a response. In fact, six of the eight patients with type II refractory CD developed EATL within a 3-year period, and a seventh patient died with blastic T-cell-like cells in the small bowel and the liver, and myeloproloferative changes in the bone marrow. The authors concluded that type II refractory CD is a premalignant condition with a very poor prognosis.

In a French national cooperative study, the clinical information and tissue specimen necessary for IEL subpopulation analysis were gathered from 21 patients diagnosed with refractory CD between 1974 and 1998.[459] There were five men and 16 women, with a mean age of 51 years (range 29-73 years). Nine of the 21 patients (43%) died from severe malnutrition

and/or lymphoma (three patients) after a mean of 6.7 (range 1-14) years after the onset of symptoms of refractory CD. A phenotypically abnormal IEL population associated with evidence of clonality was found in eight of the nine patients that could be tested. The authors suggested that refractory CD may be the missing link between CD and ETCL.

This systematic review identified nine controlled studies that met inclusion criteria. The major observation of our review is that the risk of lymphoma in CD was significantly increased compared to an age-matched population from the same region and period in 8 out of 9 studies. The SIR (NHL) varied from 2.66[338] to 42.7,[333] whereas, the SMR from NHL or lymphoma in CD varied from 11.4[337] to 69.3.[339] This increased risk persists even when the cases that are diagnosed with lymphoma simultaneously or within 1 year of the diagnosis of CD are excluded from the calculation.

Some observational studies suggest that the risk of lymphoma, relative to patients of the same age without CD, may be highest in individuals who were diagnosed during adulthood,[336,337] and appears to decrease with adherence to a GFD, as shown by several authors.[333,336-339] It is also interesting to note that the only study that did not report a significant increased risk of lymphoma was one where 75% of patients were on a strict GFD.[338]

The differential risk of lymphoma among patients diagnosed with CD in adulthood versus childhood may indicate that early diagnosis and treatment with a GFD is protective. The possibility that a GFD may be protective is also supported by Askling et al.[337] who found that the risk of lymphoma dropped to unity after 15 years of follow up. Limitations in the designs of these studies, however, prevents firm conclusions. These studies have followed relatively few patients diagnosed as children through middle age when the risk of lymphoma rises, and they may not have accounted for other factors (severity of symptoms, or other marker of disease activity) which might affect risk. The distinction between childhood and adult diagnosis of CD in the published cohorts relies on the presence or absence of CD-related symptoms during childhood, which has historically been a key factor in CD diagnosis. Based on the observations from these groups of patients, it would seem that continuous gluten exposure and ongoing mucosal damage sets the stage for malignancy later on in life. It remains unclear, however, why some individuals would have persistent mucosal damage in the absence of symptoms. Would these individuals also carry other characteristics that modulate their risk of malignancy? As we tap into the base of the “celiac iceberg” through systematic screening, we will hopefully in the future be able to observe the incidence of lymphoma in child and adult CD populations who were identified through population screening, and placed on a GFD despite them being asymptomatic during that period of their lives. The notion that lymphoma arises from prolonged antigenic stimulation should be confirmed if the risk of lymphoma is, as expected, lower than historical CD cohorts in those individuals.

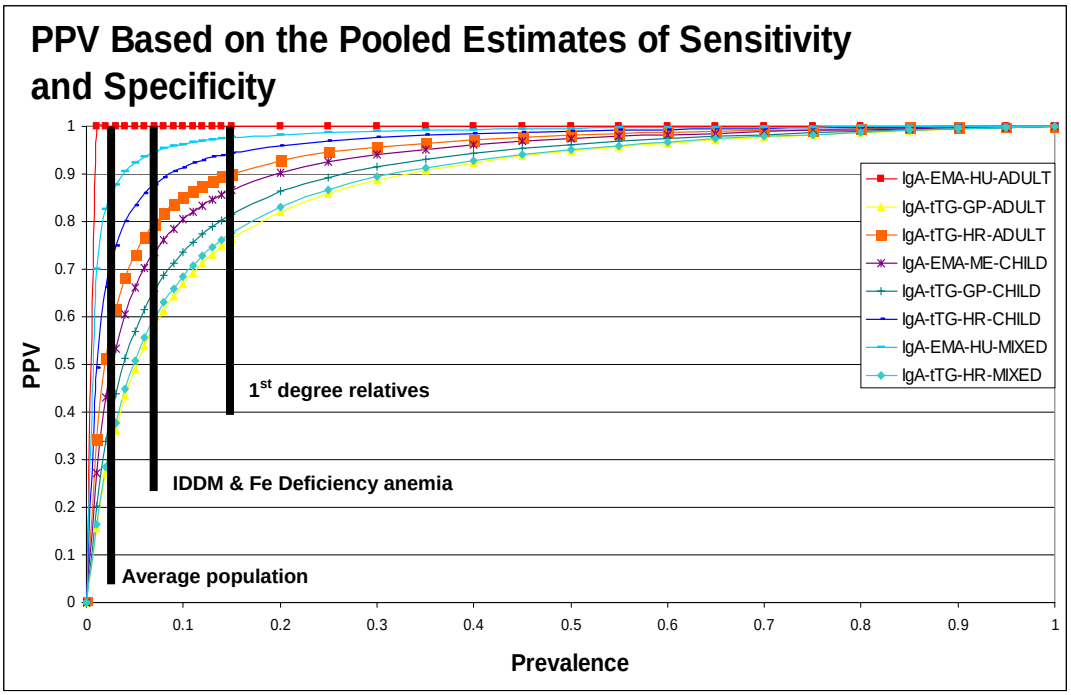
Celiac 4: Consequences of Testing for CD

The search strategy did not identify any studies that would allow us to address the specific benefits and harms of testing with different strategies for CD. At present, there is inadequate information from the published literature on the benefits and harms of screening and the potential risks of undetected CD. Prospective trials of screening would be helpful to provide the data necessary to construct the tables that depict the consequences of screening specific populations.

Information on the consequences of screening will come from the currently ongoing large population based prevalence studies.

The consequences of such issues as false-positive results were dealt with in the Celiac 1 Discussion. As discussed in that section, the definition of CD used and the prevalence of CD in the test populations, have a great impact on the diagnostic parameters of the available tests. We have presented data that show that the sensitivity of the available tests declines considerably when applied to patients with low-grade histological lesions. Unfortunately, there is insufficient data to address the question of what is the consequence of missing patients with low-grade histological lesions if serological screening alone is used. As described in Celiac 1, all the diagnositic test studies of the various serological markers were undertaken in study populations in which the prevalence of CD exceeded the that observed in most clinical situations. We have shown that the positive predictive value, which is predominately influenced by the test specificity and the prevalence of CD in the test population, drops from the reported values to much lower values when the test is applied in typical clinical populations. To illustrate this point, Figure 31 highlights the expected PPV when applied to different test populations.

Figure 31: PPV based on pooled estimates of sensitivity and specificity



As can be seen from Figure 31, the PPV—the probability that a positive test result actually represents true CD—drops with the prevelance of the population in which the test is applied. This relationship holds true for all the summary curves, but differ in degree. It is important to note that the PPV is predominantly influenced by the specificity of the test and prevalence. Since we have identified that the specificity of EMA and tTG is quite high, the major influence on the PPV in these analyses is the prevelance of CD in the population being tested. The practical importance of this discussion, is that despite having very high specificity, the use of these serological markers in low-prevalence populations would be expected to result in high false-positive rates. Below a prevalence of 5%, the false-positive rates may be as high as 30% to

50% based on our estimates. This may seem counterintuitive, given that the specificity is greater than 95% and close to 100% in some cases. One must keep in mind that unless the specifity actually equals 100%, the prevalence of CD will influence the PPV. As the specificity approaches 100%, the influence of the prevalence decreases. The same interplay occurs between the negative predictive value (the probability that a person with a negative test does not have CD), and the sensitivity of the test. However, in this case, the NPV rises as the prevalence of the disease falls (see Celiac 1 Figures). Given that we have identified that EMA and tTG have a sensitivity in the range of 95%, the NPV would be expected to be very high (>96%), particularly in low-prevalence populations. This would mean that the false-negative rates with these tests are less than 1% to 4%. These data would then suggest that a negative test result would have a high probability of being a true negative result, but that a positive test would have to be considered in light of the expected prevalence of CD in the tested population. If the expected prevalence is in the range of 10% or lower, then the possibility that the result represents a false-positive should be considered. Lastly, one must not forget the discussion regarding the true sensitivity of these serological markers when lower grade CD lesions are considered. The studies by Rostami et al.[16] and others, suggest that the sensitivity can be lower than 80%. In fact, both Rostami et al.[16] and Tursi et al.[424] suggest that the sensitivity for grades less than Marsh IIIa, is in the range of 30% to 40%. If this is the case, then the nearly perfect NPV discussed above would be expected to fall, particularly in groups with a higher prevalence of CD. For example, if the sensitivity was really 75%, then the NPV would drop to 88% (12% false negatives) if a population of patients with suspected CD was tested. However, because of the strong influence of a low prevalence (<15%) on the NPV, the NPV will remain higher than 90%, as long as the sensitivity of the test is greater than 50%.

Expected Outcomes of Treatment of CD

The four studies of diabetes and CD in children/adolescents that evaluated the impact of a GFD found that body composition parameters improved on the GFD, but HbA1c levels did not improve. Some studies observed an increase in the insulin requirements after introduction of a GFD, which could be explained by improved absorption of nutrients.

The results of studies on anthropometrics and body composition in CD patients are variable due to differences in populations, and methods used to evaluate body composition. Overall, weight and BMI improves after starting a GFD. Individuals with CD may have a lower BMI when compared with controls because of lower daily energy intakes, particularly in those who strictly follow a GFD.

A few small studies have evaluated the impact of the diet on nutritional parameters in newly diagnosed symptomatic CD patients. These studies found that nutritional status does improve in the majority of subjects with CD on a GFD. Certain biochemical parameters such as ferritin may take longer to normalize. There is evidence that the recovery of nutritional status is linked to improvement of villous atrophy. Larger studies of nutritional status in those with classical and silent CD patients and the relationship of biochemical values to changes in histological grade on small bowel biopsy and compliance with the GFD would be helpful.

Compliance with the GFD was assessed in adolescent populations in three studies and the results varied. Compliance with a strict GFD was greater in those who were symptomatic, compared with those who were diagnosed via a screening program. Another study in adults by Ciacci et al.[460] looked at the correlation between intestinal biopsy and compliance (assessed by

dietary interview) and found that that intestinal damage was significantly associated with dietary compliance. Low or very low compliance with a GFD had a PPV of 92.8%, and good compliance had a negative PPV of 96.8%. This study also suggested that those with more severe symptoms at diagnosis were more likely to have better compliance. Given the poorer compliance in those without symptoms, different strategies to promote adherence with the GFD may need to be developed if screening for CD is promoted.

The justification for screening the general population for CD would be strengthened by well-conducted comprehensive cost-effective analyses. Only one study[360] appeared to include the majority of the components that have been recommended for the reporting of cost-effectiveness analyses (CCOHTA, Guidelines for Economic Evaluation of Pharmaceuticals: Canada, 1997). None of the analyses incorporated the use of health related quality of life or utility assessments.

Fractures/BMD/Osteoporosis/Osteopenia

There were a number of methodological limitations in the studies that examined bone-related consequences of CD. Limitations included: selection of representative cases and controls, ascertainment of the outcome and failure to identify and control for relevant co-interventions such as calcium and vitamin D.

The issue of whether fractures are increased with individuals with CD appears to be somewhat controversial based on results of the included studies. Both Thomason et al.[394] and Vestergaard et al.[388] did not find increased fracture rates for CD subjects, whereas, the recent population-based study by West et al.[385] did find an increased rate of fractures. This is an important issue to clarify since osteoporotic fractures are one of the key reasons for promoting strict adherence to the GFD and for making decisions about screening. In some studies, the sample sizes were small and may not have been large enough to detect an increased risk in fractures in subjects with CD relative to controls. In addition, methodologies and study populations varied, and not all studies controlled for duration of CD. Moreno et al.[392] found that the risk of fracture in subclinical and silent cases of CD was not significantly different from that of controls. Overall, the risk of fracture seemed to increase with age as one would anticipate and may be greater in those patients who were clinically symptomatic. Based on results of current studies, the risk of fracture appears to be highest prior to diagnosis of CD and diminishes once individuals are on GFD. This latter finding would be consistent with the increase in BMD that is seen after 1 year on a GFD. Additional population based fracture studies would be useful to clarify the relative and absolute risk of fracture in CD and to determine if it differs in asymptomatic cases.

Overall, the studies consistently documented an increased prevalence of osteoporosis/osteopenia in newly diagnosed patients relative to controls. There was a significant increase in BMD, especially within the first year of being on a GFD. Some of the variability in the results could be attributed to proportion that were compliant with the diet and use of co-interventions such as calcium and vitamin D. Moreno et al.[392] found that the lumbar spine BMD did not differ in groups according to clinical presentation, but they did find a significantly lower T score of the femoral neck BMD in classically symptomatic cases versus subclinical or silent cases. Mustalahti et al.,[378] however, found that BMD in the spine was lower in asymptomatic cases.

Based on the two studies in children,[352,377] BMD appears to normalize in children after treatment with a GFD. The normalization of BMD in children would support the need for early

diagnosis of CD and treatment. However, in children skeletal growth may affect BMD, with some of the change relating to changes in growth. Most studies of BMD in adults on a GFD have found that the BMD is still reduced at all sites when compared to normal controls. One study suggested that those without secondary hyperparathyroidism at time of diagnosis may normalize their BMD, but this finding was not replicated. A large BMD study with baseline and follow-up small bowel biopsy data, and documentation of clinical presentation, percent compliance with the GFD and adjustment of co-interventions is recommended to give us accurate information on bone-related consequences of CD.

Mortality

The majority of observational studies have demonstrated an increase in overall mortality rate (SMR of 2 or greater) in subjects with CD when compared with the general population. The increase in mortality can be attributed to deaths from malignant diseases, respiratory, and digestive diseases. The increase in mortality appears to be greatest within the first 3 years after diagnosis and declines over time. The mortality rate seems to increase with longer delays in diagnosis and poor adherence to the GFD. Perhaps one of the most important points from the Corraro study,[362] is that the mortality rate was not increased compared to the general population for those individuals who had mild symptoms or were asymptomatic. This latter result has potential implications for population screening for CD.

Celiac 5: Promoting or Monitoring Adherence to a GFD

Monitoring Adherence to a GFD

Some of the same concerns expressed in the other celiac objectives, regarding clinical definitions, histological criteria, and the performance of the serological tests, are repeated when the results of the studies on monitoring adherence to a GFD are considered. Foremost in facilitating the interpretation of these studies is the question of what to consider as the histological criteria to define recovery on a GFD. Certainly normalization to Marsh 0 would constitute recovery, but what about improvement to Marsh I or II, or even accepting Marsh IIIa? The distinction has important implications for assessing the strength of the correlation between histological and serological improvement, and in this regard, different studies have adopted different cut-offs.

It is clear from the presented studies that improvement of symptoms does not offer an accurate assessment of adherence to a GFD as judged by interview or by biopsy. This point is illustrated in the study by Kluge et al.[461]. In follow-up of 18 adult patients with CD, all patients felt well and appeared to be clinically in remission. Nonetheless, only 17% of the patients reported being on a strict GFD. Biopsy assessment of eight patients showed six with total villous atrophy including one patient who reported strict adherence to GFD. The remaining two patients did not have villous atrophy but the mucosa was not normal, including an excess of IELs. Thus, small amounts of gluten may provoke a histologic change without clinical symptoms which may be an important reason why adherence to GFD may be less than perfect. In other words, non-compliance does not necessarily translate into noticeable consequences for the patient.

Furthermore, it is increasingly recognized that most CD patients don't have symptoms, so reliance on symptomatic improvement is clearly not adequate.

There is good evidence that mucosal recovery following institution of GFD is slower and more incomplete than previously assumed, especially in adults.[405,411,414] Whether this slow recovery is due to dietary transgression, inadvertent gluten intake or whether this is simply the natural history of the disease is less clear. This has definite implications for the interpretation of both biopsy and serology results in monitoring adherence to GFD, particularly in the short run.

With the advent of the newer and more sensitive serologic tests for CD (EMA, tTG), the possibility of a reduction in the need for follow-up biopsies and a move towards non-invasive serological monitoring has been proposed. The question arises as to whether serology can detect dietary transgressions and reasonably mirror histological improvement on a GFD.

A number of studies show that values of serologic markers will fall with increasing duration of GFD, whether one looks at IgA-AGA, IgA-EMA, or IgA-tTG. As well, several studies suggest that in both adults and children, increasing degrees of non-compliance with a GFD, are more likely to be associated with positive serologic tests.[396,402,408]. The question, however, is *not* whether serology can pick-up major transgressions such as with a gluten challenge which it is clearly capable of assessing,[400,404] but rather if serology can pick-up milder degrees of dietary non-compliance and reasonably reflect histological status. A high rate of falsely-negative serology with lesser degrees of dietary transgression would diminish serology as a means of accurately monitoring adherence.

In both adults and children, the sensitivity of serology for picking-up dietary transgressions based on interview or self-reporting is disappointing.[401,402,410,415] One conflicting study[412] showed a good correlation between serology and adherence. This likely reflects the way patients were categorized, and it is likely that in this study, patients with lesser degrees of dietary transgression were categorized as compliant. In general, there is a significant rate of normal serology in patients identified as not adhering to a GFD. Furthermore, evidence from several studies suggests that serology, regardless of the actual test used, does not adequately reflect the mucosal state in adults.[398,403,407,409,409,413] Surprisingly, it seems that serology may be normal, not only in Marsh I or II lesions, but also when there is villous atrophy present.[398,407,409,413] Although the specificity of various serologic markers for villous atrophy seems better than sensitivity,[398] the NPV of serology would suggest that a negative test does not offer high assurance of the absence of villous atrophy.

As discussed earlier, mucosal recovery can be a slow process. It may be that serologic markers may better reflect histology in long-term follow-up. Certainly, in the range of follow-up of these studies (6-30 months), serology may be negative despite villous atrophy. There is evidence that even in longer follow-up, serology does not accurately reflect adherence.[398,402,410]

In younger patients, IgA-AGA and IgA-EMA-ME may better represent the mucosal state.[397,415] These studies are in keeping with the impression that in children and adolescents, mucosal recovery is faster and more complete. In children, serology seems to be a better marker of the absence of villous atrophy. Still, serology may be negative in the face of lesser degrees of histologic abnormality without villous atrophy.[397] The significance of such lower-grade biopsy abnormalities, although, is unclear.

It is possible that IgA-AGA may rise faster with non-compliance to GFD than other markers.[396,400] However, there is little direct evidence to show superiority of one serologic test over another in monitoring adherence.

Perhaps an important question that arises from this discussion, with particular relevance to symptomatic CD patients, is: "is it good enough for CD patients to show symptomatic improvement and a corresponding fall in, or normalization of, a sensitive serological marker without need for 'normalization' of the intestinal mucosa?" Unfortunately, this question is not an easy one to answer since many of the outcome studies in CD, particularly for lymphoma and mortality, did not specifically address differences in histologic grade. Furthermore, we identified no clear evidence suggesting that refractory sprue was the result of dietary indiscretion as opposed to a different spectrum of CD. Nonetheless, histological improvement appears to be important. For example, one study[356] demonstrated that osteoporotic patients with CD on a GFD who had Marsh III lesions had lower median Z-scores than those with grades less than Marsh III, while another study demonstrated a significant correlation of nutritional status measured by histomorphometric index, with the severity of the histological biopsy grade.[346] In the former study as well as one other study,[358] histologic grade correlated with degree of IDA, all suggesting that the goal of monitoring should be to assess degree of histological improvement.

It can be concluded that the return of serologic markers to normal is associated with duration of GFD and degree of patient compliance. Unfortunately, the correlation remains imperfect, especially in adults, and seems to reflect gross rather than minor degrees of dietary transgressions. Serological tests seem to have a higher specificity than sensitivity for dietary transgressions. It is recognized that this area is controvercial and that clinicians are moving away from routine follow-up biopsy as a means to assess dietary compliance. It seems reasonable to suggest that improvement in clinical parameters, and disappearance of serological markers would be an adequate measure of response to a gluten free diet. In children, because of their faster and more complete mucosal recovery, this strategy of using serology may be an appropriate means to monitor adherence. In adults, however, the situation is somewhat more complex. Therefore, while serology certainly can be an adjunct means to monitor adherence to a GFD, consideration should be given to assessing histological improvement since some evidence exists to suggest that mucosal improvement to at least below a Marsh III appears to be important from an outcomes perspective. If biopsy is to be utilized as a means of assessing adherence to a GFD in adults, the timing of the biopsy needs to take into consideration the slower mucosal healing in adults, and should therefore be performed after 1 year to 1.5 years of a GFD.

Interventions to Promote Adherence to a GFD

Changes in dietary habits are difficult to attain and maintain. The barriers to compliance are many. No interventions to promote compliance with GFD have been studied and found to be effective. Adding to the difficulty of assessing any proposed intervention is the lack of certainty as to how best to measure GFD compliance.

The existing evidence suggests a positive correlation between parental socioeconomic status, education, knowledge of CD, and the compliance of their children.[416,418] Compliant children may also have a better knowledge of CD[420] than those children who are non-compliant. Improved knowledge in adults also appears to correlate with compliance.[419] It is, therefore, not unreasonable to suggest that interventions designed to improve knowledge about CD in general, and about GFD, and specifically how to identify gluten-containing products, would likely improve compliance with a GFD. Improving knowledge regarding gluten-containing food products and additives would also likely improve self-confidence in choosing gluten-free foods as suggested by Lamontagne et al.[419] Improved knowledge of outcomes of untreated CD may

also improve compliance. Such information interventions, however, would need to be prospectively evaluated to ensure that they perform as expected.

Membership in a local celiac society appears to be an effective means of promoting compliance with a GFD. This is not surprising since such organizations provide CD patients with not only improved knowledge regarding their disease, and the intricacies of the GFD, but also provide emotional and social support.

It is interesting that one study[417] has demonstrated lower rates of compliance in children detected by screen as compared with those diagnosed on the basis of symptoms. It seems logical that if there are no obvious detrimental symptoms from a gluten-containing diet, that children and likely adults will be less likely to be compliant. The authors speculate that since screen-detected patients had a higher mean age of diagnosis, compliance might be promoted by earlier identification. They speculate that earlier detection would avoid the difficulty of changing formed eating habits.

Is early detection of CD an effective intervention to promote compliance? It appears rational that it would be easier to follow a GFD if it were introduced at an earlier age. There are some interesting observations [417]that suggest that diagnosis in early childhood is associated with improved compliance.[421] Unfortunately, the issue of compliance in asymptomatic screen-positive individuals casts doubt on the positive downstream effects of screening asymptomatic populations for CD, particularly if the low-compliance rates in asymptomatic individuals can be reproduced in other studies.

In summary, it is suggested by the results of this report that a multidisciplinary approach to patient and parent education and support by physicians, dieticians, and celiac societies, possibly employing formal knowledge and decision support interventions that involve the patient (and parent) directly, are likely to improve compliance in individuals diagnosed with CD. Formal testing of interventions and programs would be valuable.

Strength of the Body of Evidence

Celiac 1

Overall, the quality of the diagnostic studies assessed in the Celiac 1 objective was quite good, due largely to our stringent inclusion criteria. However, 59% of the included studies reported using a selected patient population that may not be representative of a clinically-relevant population. This is likely related to study design. In addition, only 11% of the studies reported on whether the reference test was reported without knowledge of the index test. However, we felt that this was not a major threat to the validity of the studies.

Two other factors that affect the interpretation of these results, yet were not captured in the quality assessments, are the threshold effects for determining the positivity of a serological test, and the high prevalence of CD in these studies (see above). With these considerations in mind, the overall strength of the evidence is quite good.

Celiac 2

The overall quality of reports of the included studies in the Celiac 2 objective was found to be marginal to fair. For example, most of the studies did not report on whether the patients were consecutively enrolled, a factor that could contribute to selection bias. However, setting aside the quality of individual studies, from a policy perspective, the strength of the evidence is fairly good in that the study populations were selected to reflect that of a North American/Western European descent, that should reflect the demographics of the US population.

Celiac 3

The studies included in the Celiac 3 objective were found overall, to be of good quality. Again, the overall strength of the evidence is due largely to the stringent inclusion criteria, such as the requirement for the reporting of standardised rates for the outcomes based on rates from the local general population, and the overall good quality of the included studies.

Celiac 4

The majority of studies included in this objective were single group "before–after" studies, although some had in addition a comparative healthy control group. We could not identify any quality instruments for this type of study design and in general, this type of study is considered weak, particularly in the absence of a control group. Overall, however, the strength of the evidence for this objective is fair to good and suggests that the results can be used for policy decisions with the understanding that this area of CD research is still relatively new and requires further high quality studies.

Celiac 5

The majority of studies in this objective were also of a "before–after" design. However, in this setting, this design may not pose a major limitation, since the purpose of the study is to assess the change in serology and histology after introduction of a GFD. In this regard, the strength of the evidence for monitoring adherence to a GFD is fairly good. However, there is almost a complete absence of studies of interventions for the promotion of adherence to a GFD.

Future Research

This review has allowed us to identify several areas in need of future research. Perhaps the most important of these is a need for the development of a consensus on the definition of CD in the era of advanced serological testing. As discussed in the report, this distinction of what one calls CD has profound implications for each of the requested task order objectives. Do screen-positive patients without villous atrophy have CD. Certainly the preliminary evidence suggests that this is the situation in many cases. However, what is required is a new definition of a gold standard for the diagnosis of CD. This new gold standard may include a combination of serology, biopsy and HLA testing. Such a gold standard, when used in studies with a time dimension (e.g., response to a GFD or gluten challenge; extended follow-up), would help answer some of the uncertainties identified in this report including: the real performance of the serological tests when low-grade lesions are considered CD; the diagnostic performance of biopsy alone; the outcomes of patients with these low-grade lesions; and, those that would be "missed" using current screening strategies. Even in the absence of a new gold standard, we could not identify a well-conducted study of the diagnostic performance of the various serological markers when applied to an average population (i.e., one with a prevalence of CD in keeping with the range identified for average risk), with the entire cohort being investigated equally (i.e., all are biopsied). Such a study would at least be able to shed light on the performance of these tests in average-risk patients, and since all patients are biopsied, the relationship of histology to serology could be further assessed.

On a similar theme, we have identified multiple studies that suggest the importance of histological improvement on a GFD. This is a controversial area since in common clinical practice, clinicians are moving away from routine follow-up biopsy. It seems reasonable to believe that improvement in clinical parameters with loss of serological markers is adequate evidence of response to a GFD. In children, this issue may be less important since histological improvement is much more rapid and complete than in adults, and correlation with serology seems better. However, we have identified multiple studies in adults that suggest poor correlation between serology and improvement of histology on a GFD, and other studies that suggest that serology is useful for detecting gross dietary indiscretion, but not minor occurrences. Therefore, the question that arises is what constitutes adequate improvement on a GFD, and what are the criteria to define this improvement. Based on the lymphoma literature that suggests that this malignancy may arise from chronic antigenic stimulation and immune activation, what are the outcomes of adults with clinical improvement, yet persistent histological abnormalities? Are some histological features, such as reduction of mucosal lymphocytes, more important markers of improvement and possibly prognosis than other features such as villous height?

We feel that clarification of these fundamental questions is necessary for the conduct of future studies in all areas of CD, and in particular studies of the diagnostic tests and the outcomes in CD, since these are so dependent on the definitions discussed above.

Conclusion

This report has provided a systematic review on five broad areas of CD, with each of these areas including important sub-components. Perhaps one of the most important findings of this report is the understanding of the importance of how one chooses to define CD in the era of serological testing, and how this apparently clear-cut task has profound implications on all the results presented in this report. Specifically, can CD be diagnosed solely on the basis of serology? Is some degree of villous atrophy necessary for the diagnosis of CD? These questions have important implications downstream of the diagnosis as well. Do CD patients without symptoms or villous atrophy have the same risk of complications as those with villous atrophy? Is serological improvement on a GFD sufficient to reduce CD complications or must there be documented histological improvement, and what degree of histological improvement is necessary?

The results of the Celiac 1 objective suggest that in the era of EMA and tTG antibody testing, AGA testing in both children and adults has a limited role. The sensitivity and specificity of EMA and tTG are quite high (over 95% for sensitivity, and close to 100% for specificity), as are their PPVs and NPVs, but as previously discussed, one has to be aware that the reported diagnostic parameters are taken from studies in which the prevalence of CD was, for the most part, much higher than that seen in usual clinical practice and certainly the PPV of these tests may not be as high as reported when these tests are applied in general population screening. The bulk of the evidence on the diagnostic characteristics of these tests was derived from studies that defined CD as having at least some degree of villous atrophy. We have identified studies that suggest that the sensitivity of these tests drops, at times significantly, when applied to populations with CD with lower-grade histological lesions. This not only has implications regarding those patients with “mild” CD who were missed during screening efforts, but also puts into question the nearly perfect NPV of these tests.

HLA DQ2/DQ8 testing appears to be a useful adjunct in the diagnosis of CD. The test has high sensitivity, in excess of 90% to 95%, but because around 30% of the general population and an even higher proportion of “high-risk” subjects including diabetics and family members also carry these markers, the specificity of this test is not ideal. The greatest diagnostic utility of this test appears to be its NPV.

Biopsy itself, when used with a strict cut-off requiring villous atrophy, appears to have high specificity, but poor sensitivity. Using lower grade cut-offs clearly improves sensitivity, but because of the wide differential of causes of histological lesions similar to Marsh I to IIIa, the specificity suffers. The use of histomorphometric measures, such as quantification of $\gamma\delta$+ IELs, are likely to allow for the use of lower-grade cut-offs while maintaining reasonable specificity. Ultimately, a trial utilizing multiple diagnostic tests in an attempt to capture as many CD patients in a clinically-relevant population as possible, with a time dimension including a response to a GFD or gluten challenge, is required to fully assess the diagnostic characteristics of biopsy alone. This type of study would be able to characterize the false-positive and false-negative rates if all studied patients are followed forward in time.

The included prevalence studies demonstrated important differences in execution, tests for prevalence assessment, and in patient sampling, and their results also have to be interpreted in the light of some of the limitations that have been identified regarding the diagnostic performance of the tests for CD. Nonetheless, the results of this report suggest that CD is a very common disorder with a prevalence in the general population that is likely close to 1:100 (1%).

Several high-risk groups with a prevalence of CD greater than that of the general population have been identified including those suspected of having CD, family members of CD patients, type I diabetics, and those with IDA or low BMD. Additionally, the review identified multiple other high-risk groups such as those with Down Syndrome, short stature, and infertility, to name a few, though their inclusion was beyond the scope of this report. These results would suggest that at the very least, high-risk groups should be screened for CD. If the performance of the noninvasive serological tests can be verified in the relatively "low prevalence" situations in general unselected populations, then population screening may also be advisable, particularly if a greater understanding of the consequences of missing early low-grade CD can be obtained, and the issues of low-compliance with a GFD of asymptomatic screen identified patients can be addressed.

CD is known to be associated with GI lymphoma. The results of this report confirm this strong association, with the limitations indicated in the text. Nonetheless, the report identified SIR for lymphoma that ranged from 4 to 40, and SMR that ranged from 11 to 70. GI lymphoma is believed to arise as a result of chronic antigenic stimulation, which leads to the development of a clonal T-cell population with usually a refractory intermediate stage. We have identified epidemiologic data that supports this notion, and suggests that a diagnostic delay, and in particular diagnosis of CD in adulthood, as apposed to in childhood, is associated with poorer outcomes. Fortunately, several studies suggest that adherence to a GFD reduces the risk of lymphoma in CD patients. These findings underscore the importance of early diagnosis and treatment of CD.

The consequences of testing for, and identifying CD patients, is expected to have a positive impact on patient outcomes be it either from a reduced risk of lymphoma with early diagnosis and treatment of CD or from improvements in nutritional status, BMI, and BMD. The consequences of testing in at-risk and symptomatic patients appears to be more straightforward since these patients appear to be more compliant with a GFD and would be expected to benefit from this intervention. The data is less clear for asymptomatic screen-identified patients, particularly those who are truly silent and/or don't have fully developed villous atrophy since, on the one hand the outcome of such patients has not been extensively studied, and on the other hand, compliance with a GFD appears problematic, particularly for those diagnosed in adulthood.

Finally, no specific interventions have been identified that promote adherence to a GFD, but education of patients and family members about CD and about the intricacies of the GFD through multidisciplinary teams, and participation in local CD societies, has been show to improve compliance. Therefore, the development and evaluation of formal educational interventions in collaboration between healthcare professionals and CD societies would appear to be a means to build on the methods that appear to already improve patient compliance. Monitoring of adherence to a GFD appears to be important, since improvement in histologic grade has been associated with improved BMD, IDA, and nutritional status. The serological markers appear to be adequate for detecting gross dietary indiscretion, and responding to gluten challenge, but unfortunately, they have poor sensitivity for detecting lesser degrees of dietary indiscretion, and have inadequate correlation with histological improvement at least in the short-term. It is true that histological improvement tends to lag behind clinical and serological improvement, especially in adults in whom improvement may never be complete, but even considering this, a negative serological test has been shown to miss patients with persistent villous atrophy. The recognition of persistent villous atrophy appears to be important since

improvement beyond this level is associated with the improved outcomes listed above. It should be noted, however, that we could not identify a controlled study that objectively determined the level of histological improvement that would be associated with improved outcomes, and this is an area for future study. Although somewhat controversial, nonetheless, based on this report it would appear that follow-up biopsy, at least 1 year after GFD in adults to document improvement of the histological grade, would be valuable.