Outcome measures in coeliac disease trials: the Tampere recommendations

Jonas F Ludvigsson,1,2 Carolina Ciacci,3 Peter HR Green,4 Katri Kaukinen,5,6 Ilma R Korponay-Szabo,7,8 Kalle Kurppa,9,10 Joseph A Murray,11 Knut Erik Aslaksen Lundin,12,13 Markku J Maki,14,15 Alina Popp,16,17 Norelle R Reilly,18,19 Alfonso Rodriguez-Herrera,20 David S Sanders,21 Detlef Schuppan,22,23 Sarah Sleet,24 Juha Taavola,25 Kristin Voorhees,26 Marjorie M Walker,27 Daniel A Leffler28

ABSTRACT

Objective A gluten-free diet is the only treatment option of coeliac disease, but recently an increasing number of trials have begun to explore alternative treatment strategies. We aimed to review the literature on coeliac disease therapeutic trials and issue recommendations for outcome measures.

Design Based on a literature review of 10,062 references, we (17 researchers and 2 patient representatives from 10 countries) reviewed the use and suitability of both clinical and non-clinical outcome measures. We then made expert-based recommendations for use of these outcomes in coeliac disease trials and identified areas where research is needed.

Results We comment on the use of histology, serology, clinical outcome assessment (including patient-reported outcomes), quality of life and immunological tools including gluten immunogenic peptides for trials in coeliac disease.

Conclusion Careful evaluation and reporting of outcome measures will increase transparency and comparability of coeliac disease therapeutic trials, and will benefit patients, healthcare and the pharmaceutical industry.

INTRODUCTION

Coeliac disease (CD) is an immune-mediated disease triggered by gluten exposure.1 Although characterised by small intestinal inflammation, consequences are widespread and linked to diverse manifestations that include osteoporosis,2 lymphoma,3,4 pneumonia3 and increased mortality.4 Symptoms vary, with some patients having diarrhoea and malabsorption (often termed ‘classical CD’), others suffering from constipation, fatigue and depression (non-classical CD) and some are asymptomatic (subclinical CD).5 The global prevalence of CD is about 1%–2%,8,9 but seems to be increasing.10,11

Lifelong adherence to a gluten-free diet (GFD) is the only available treatment for CD.1 For several reasons, patients find the GFD to be exceedingly burdensome.12 That is, it is socially restrictive13 and more expensive than ordinary food.14–16 Patients differ in their ability to adapt psychologically to CD. Some people have little difficulty in adopting the GFD, whereas for others, living with CD is a daily struggle.13 In addition to the burden of treatment, patients with CD frequently have ongoing symptoms and mucosal healing is slow and often incomplete. For these reasons, there is a need for alternative treatments of CD, as suggested by the intensive research efforts undertaken in different laboratories.17 Potential targets for treatment include glutenases, modified or pretreated gluten, gluten sequestrants, neutralising antibodies, inhibitors of intestinal permeability, lymphocyte blockers, including anti-interleukin-15, tissue transglutaminase (TG2) inhibitors, immune tolerance induction, exposure to hookworms and...
DQ2-blocking peptide analogues.\textsuperscript{17} \textsuperscript{18} Several of these drugs are now being tested in phase I or phase II trials, and a recent study found that novel therapies attract the interest of patients with CD more than any other disease-related topic.\textsuperscript{19}

Of importance is that treatment effects are measured against robust standards. A recent systematic review\textsuperscript{20} identified six histological, CD activity indices,\textsuperscript{21}–\textsuperscript{26} five patient-reported outcomes (PROs)\textsuperscript{27}–\textsuperscript{31} and four indices for endoscopic CD activity\textsuperscript{32}–\textsuperscript{35} that have been used for coeliac trials.

In the present paper, we have explored clinical, serological, histological and immunological outcome measures for performing clinical trials in CD, and importantly, have included the patient perspectives concerning recommendations for their use.

METHODS

Task force

Coauthors were invited by JFL and DAL with the aim to obtain a group with knowledge, experiences and interests that reflect the heterogeneity of outcome measures used in trials of CD. Most of the participating researchers were adult gastroenterologists (PHRG, CC, DSS, KKu, JM, JT, KEAL, DAL) but our group also included six paediatricians (JFL, NRR, KKu, MJM, AP, IRK-S), one pathologist (MMW), one basic scientist (AR-H) and two representatives of patient organisations (SS and KV). Members of this diverse collaboration originated from 10 countries. Most of the coauthors participated in the CD meeting organised by MJM in Tampere, Finland on 24–25 November 2016 (which provides the motivation for the subtitle of this paper).

Literature review

Coauthors were divided into seven teams of three to four individuals who jointly reviewed five domains: serology, histology, immunology, PROs and other outcome measures. The Karolinska Institutet library carried out literature searches for relevant papers up until 1 December 2016 (see online supplementary appendix). This search yielded 10 062 references. After reviewing titles and abstracts for these 10 062 references (online supplementary appendix), there remained 941 publications that were considered potentially relevant for this review and subsequently read in detail.

In this paper, we issued a number of recommendations. Where appropriate these were graded according to the method suggested by the Oxford Centre or Evidence-based Medicine,\textsuperscript{36} where grade A evidence represents the highest level of evidence and grade D the lower available evidence (generally based on the opinion of experts but with no preceding randomised trials, cohort or case-control studies). The appendix contains a detailed description of grade A–D. All recommendations were subject to a subsequent voting on a five-level scale (strongly disagree, disagree, not agree or disagree, agree and strongly agree).

Manuscript draft

JFL wrote the first draft of the paper. The text was then extensively revised by the coauthors. JFL and DAL supervised these revisions, but all authors contributed and agreed on the conclusions and the final wording of the paper.

This is a series of expert-based recommendations and did not meet the requirements of a formal systematic review (eg, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement\textsuperscript{11}). We aimed to highlight the state-of-the-art of designing intervention trials in CD.

RESULTS

Histology

Clinicopathological correlation is key to the diagnosis of CD in adults and children. In adults, confirmation of the diagnosis by duodenal biopsy is the gold standard.\textsuperscript{7} \textsuperscript{38} \textsuperscript{39} In Europe, a ‘biopsy-sparing’ protocol (with defined limitations for use) has been adopted for symptomatic children defined by an anti-TG2 titre ≥10 times the upper limit of normal, positive endomysial antibody (EMA) on a second blood draw and positivity for human leucocyte antigen haplotypes HLA-DQ2 and/or HLA-DQ8.\textsuperscript{40} To evaluate effective treatment for CD, quantitative histological assessment (morphometry) outperforms qualitative histology (eg, the Marsh score) in a trial setting.\textsuperscript{20} In clinical trials, optimised biopsy protocols should be followed for assessment of mucosal damage or healing.

Well-known classifications in histological assessment are described by Marsh and modified by Oberhuber\textsuperscript{38} and Corazza and Villanacci.\textsuperscript{41} Although grouped classifications are practical in clinical work, recent studies have shown imperfect reproducibility and reliability.\textsuperscript{26} \textsuperscript{42} Recently published recommendations for biopsy diagnosis of CD in adult patients for the number of biopsies and sites are available,\textsuperscript{38}\textsuperscript{39} with optimal laboratory processing alongside structured reports to include validated morphometric analysis.\textsuperscript{26} \textsuperscript{38}

It is recommended to take at least five duodenal biopsies, one or two from the duodenal bulb (D1) and four from the second part of the duodenum (D2).\textsuperscript{38} \textsuperscript{39} These biopsies should be taken across circular folds to avoid a crushing artefact.\textsuperscript{43} Endoscopists should take one biopsy specimen per pass of the forceps in that a single-biopsy technique improves the yield of well-oriented duodenal biopsy specimens.\textsuperscript{44} Biopsies from D1 and D2 should be reviewed separately by a pathologist.\textsuperscript{45}

When processed in the laboratory, biopsies should be orientated correctly and sectioned at three levels. In trial settings it is appropriate to always count intraepithelial lymphocytes (IELs) and state the number present/100 enterocytes (normal counts are ≤25/100 enterocytes).\textsuperscript{26} \textsuperscript{39} \textsuperscript{45} \textsuperscript{46} IELs can be counted in H&E stained sections; however, immunohistochemistry with CD3 is preferred by some pathologists.\textsuperscript{26} \textsuperscript{47} Frozen tissue specimens have been used to evaluate T-cell receptor gamma delta positive (γδ+) T cells,\textsuperscript{48} but new antibodies for use on paraffin-embedded specimens are now available.\textsuperscript{49} \textsuperscript{50} Identification of a high density of γδ+  T cells is relatively specific for CD and can be useful when histological diagnosis remains equivocal.\textsuperscript{48} \textsuperscript{49}

Immunohistochemistry to show deposited immunoglobulin A targeting TG2 in the small bowel mucosa is accurate in the detection of CD if patients are on a gluten-containing diet.\textsuperscript{51} These deposits have had 100% sensitivity\textsuperscript{52} and a mean specificity of 94%\textsuperscript{53} for CD. The deposits have been used in several gluten challenge studies to measure gluten reactivity.\textsuperscript{52} This technique requires the use of frozen tissue. The use of frozen tissue in clinical trials has shown variable results, which mirrored serum TG2 antibodies with increased deposits.\textsuperscript{54}

Morphometry, in which continuous variables such as the villous height-to-crypt depth ratio and IEL density are measured separately, overcomes certain problems encountered when using grouped classifications.\textsuperscript{26} \textsuperscript{55} Of note, a threshold change of 0.4 represents a measurable and likely clinically relevant difference between villous height:crypt depth ratio measurements. A villous height:crypt depth ratio of <2 is indicative of atrophy, active disease. Patients with treated CD have values above 3. Similarly, ≥30% change in T-cell IEL densities is considered clinically significant.\textsuperscript{26}
Importantly, morphometry, which has produced excellent reproducibility and reliability, has a significant role in clinical trials in which reliable and accurate measurements are a requirement. Whichever classification is used, two blinded observers should read the histology to ensure reliability in clinical trials. It should be noted that CD can be patchy and there is some intra- and even intrabiopsy variability in villous architecture and lymphocyte numbers contributing to sampling error and difficulties in interpretation. Given this, and that only a small proportion of the proximal small intestinal mucosa is evaluated by conventional biopsy review, new tools are needed for assessment of mucosal health. The optimal timing of biopsy to evaluate healing should be a clinico-pathological decision dependent on treatment offered and taking into account possible sampling errors by following protocols for biopsy sites.

In CD, there may be concurrent upper GI pathologies (e.g., *Helicobacter pylori* infection,57–58 lymphocytic gastritis59 and eosinophilic oesophagitis/oesophageal eosinophilia60) that should be assessed at initial endoscopy if clinical history is suggestive, and if present, reassessed post-treatment because these may contribute to ongoing symptoms not related to small intestinal damage.

Patients included in studies for CD therapy must have had an initial robust diagnosis. Occasionally, patients will have been diagnosed without histological confirmation.61 These patients should not be included in gluten challenge studies but instead included in trials of active CD treatment. Patients with a study entry biopsy confirming villous atrophy (VA) and a record of positive serology and permissive HLA status should be eligible for treatment studies in CD.

As per the current European Society for Paediatric Gastroenterology Hepatology and Nutrition guidelines,60 we are reluctant to suggest timelines for control biopsies for children, although a recent paper found that mucosal healing may not be as complete as previously assumed.62 For now, follow-up biopsy in children should be dictated by clinical needs.

**Table 1** summarises changes in mucosal histopathology, serology and symptoms in clinical trials with gluten challenge.

**Recommendations:** Histology is an essential outcome measure in any trials of CD treatment (grade B). Histology should be performed both before and at the end of the trial when healing or histological relapse is the primary outcome measure (grade B). In a gluten challenge study, successful treatment may be characterised as no change, or, in a treatment study, as histological improvement by a significant increase in the villous height:crypt ratio (>0.4 being considered relevant) and/or a ≥30% change in IEL densities (grade D). Additionally, histology may be useful as a criterion for study inclusion in which participants in gluten challenge studies should have a high villous height: crypt depth of >2–3 and participants in treatment trials should have a decreased villous height: crypt depth ratio of <2–3 (grade D). Histological evaluation should follow a priori histology protocols using quantitative measures.

Vote: agree: 7; strongly agree: 12.

### Serology

Serology is a cornerstone in the diagnostic workup for CD.18–21 IgA auto-antibodies to TG2 and IgG antibodies to deamidated gliadin peptides (DGPs) are central diagnostic tests for active CD. In IgA-sufficient patients, IgA-anti-TG2 is the most predictive and reproducible single test, although IgA EMA performs similarly well in some expert laboratories and is often used as a confirmatory test. IgG anti-DGP displays similar sensitivity as IgA anti-TG2 but has lower specificity. Selection of optimal serological tests is mandatory because not all commercial assays perform equally well.18–21 Importantly, calculation of results and thus numerical values for the same samples also may differ and only tests with a multipoint calibration curve give values proportional to serum antibody concentration.20 Differences in assays can make interpretation in and comparisons between clinical trials difficult. This difficulty is in part due to different epitopes that are detected by different tests, different calibration or to antibodies with lower avidity and specificity.20–23 Thus, an important shortcoming when using serology to evaluate the outcome of gluten challenge is the wide range of response. When the UK National External Quality Assessment consortium tested the same positive samples in 14 commercial anti-TG2 assays, large differences in antibody levels were found, consistent with substantial variability for antibodies used in the diagnosis of CD.63

Overall, a correlation exists between IgA-anti-TG2 antibody titres and the severity of mucosal damage by histology, as well as the histological outcome on a GFD.70–73 Yet, a recent meta-analysis found that serum TG2 and EMA often underestimate the degree of VA.71 Antibody titres below diagnostic cut-offs, thus, do not predict a normal or near-normal (Marsh I) histology. The one caveat is that biopsies usually sample a short segment of the (descending) duodenum, whereas active CD can affect large portions of the small intestine that occasionally extend down to the ileum.75 Therefore, a patient may be in clinical and serological remission with residual inflammation in the proximal duodenum but which is quantitatively much less extensive than before.74 Thus, in one of the largest studies to date,76 IgA anti-TG2 failed to detect 44% of persistent VA (Marsh III) in patients with CD on a GFD for >1 year.

Normal serology is generally required for entry into a gluten challenge study to ensure that participants do not have severe VA prior to gluten challenge. Conversely, serology above diagnostic cut-offs has not been used as an inclusion criterion for treatment studies in non-responsive CD because many people with a normal serology will still have VA and ongoing symptoms.74 Participants with elevated serologic tests may respond better to some therapies. While this remains to be confirmed, stratification by CD antibody levels at study entry should be considered. In clinical trials, serology may be used in assessing change during gluten challenge or to monitor longer treatment studies. The antibody response to gluten challenge depends on four factors: duration of the previous GFD, daily amount of ingested gluten, duration of gluten intake and individual factors.

Patients with CD may tolerate different levels of gluten exposure. When low (1–3 g/day) or moderate (3–5 g/day) amounts of gluten were administered to 25 Finnish patients with CD in remission for 12 weeks, only 67% of the patients with CD showed signs of mucosal inflammation and 43% developed positive IgA-anti-TG2 antibodies.65 However, in a US study of 20 adult patients with CD in remission challenged with 3 or 7.5 g gluten/day for 2 weeks, Marsh III histology developed in 68% of the patients with CD, whereas anti-TG2 and anti-DGP antibodies increased in only 25% for anti-TG2 and 30% for anti-DGP. Remarkably, positivity for both antibodies increased to 55% and 45% 2 weeks after the end of gluten challenge. A recent study from Norway showed even lower proportion of patients responding serologically after 2 weeks.78 Therefore, the histology at week 2 and serology at week 4 combined showed a gluten response in nearly 90%, with no difference between both doses.77 Notably, some rare patients who had been on a GFD for years may develop a tolerance to gluten ingestion that may last
# Table 1 Coeliac disease: results of clinical trials with gluten challenge, changes in mucosal histopathology, serology and symptoms

<table>
<thead>
<tr>
<th>Patient, n</th>
<th>Amount of gluten</th>
<th>Challenge duration</th>
<th>Mean reduction in VhCRD*</th>
<th>Mean increase in CD3+ IELs</th>
<th>Marsh-Oberhuber classification</th>
<th>Serology</th>
<th>Symptoms</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lähdeaho et al 2014</td>
<td>16 1.5 g</td>
<td>6 wk</td>
<td>0.6†</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>Trend towards more symptomatic in GSRS, CDQ and SF-36 questionnaires</td>
</tr>
<tr>
<td></td>
<td>21 2.0 g</td>
<td>6 wk</td>
<td>0.8†</td>
<td>NA</td>
<td>Statistically significant increase†</td>
<td>NA</td>
<td>1/21 seroconverted to positive EMA</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>13 3.0 g</td>
<td>6 wk</td>
<td>1.1†</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>13 6.0 g</td>
<td>6 wk</td>
<td>1.7†</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>–</td>
</tr>
<tr>
<td>Leffler et al 2013</td>
<td>10+10§ 3 and 7.5 g§</td>
<td>Day 3</td>
<td>0.52¶</td>
<td>3.1 per 100 ECs¶</td>
<td>NA</td>
<td>Mean increase in IgA-TG2 titre was 1.03¶</td>
<td>Significant change in both CSh and GSR51</td>
<td>Significant change in both CSh and GSR51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Day 14</td>
<td>1.09†</td>
<td>19.22 per 100 ECs¶</td>
<td>10/20 showed histological response</td>
<td>Mean increase in IgA-TG2 titre was 2.86†</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Tack et al 013</td>
<td>7 7 g</td>
<td>2 wk</td>
<td>NA</td>
<td>NA</td>
<td>3/7 deteriorated¶</td>
<td>1/7 seroconverted to positive IgA-TG2</td>
<td>No significant change in CDQ</td>
<td>5/11 had increased deposits of TG2 IgA at end.</td>
</tr>
<tr>
<td>Brottveit et al 2013</td>
<td>11 Four slices of bread</td>
<td>Day 4</td>
<td>Data not presented¶</td>
<td>6/11 deteriorated</td>
<td>NA</td>
<td>4/11 had increased abdominal symptoms</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Lähdeaho et al 2011</td>
<td>10 3–5 g</td>
<td>12 wk</td>
<td>1.3†</td>
<td>39 per mm of epithelium†</td>
<td>NA</td>
<td>5/10 had increase in IgA-TG2 titre</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>11 1–2 g</td>
<td>12 wk</td>
<td>1.0†</td>
<td>20¶</td>
<td>NA</td>
<td>4/10 had increase in IgA-TG2 titre</td>
<td>8 developed mild to moderate symptoms*</td>
<td>5/10 had increased deposits of TG2 IgA.</td>
</tr>
<tr>
<td>Catassi et al 2007</td>
<td>10 10 mg</td>
<td>3 months</td>
<td>–1%¶</td>
<td>NA</td>
<td>1/5 deteriorated</td>
<td>4/10 had increased deposits of TG2 IgA.</td>
<td>3/10 showed clinical symptoms</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>13 50 mg</td>
<td>3 months</td>
<td>–20%††</td>
<td>0%</td>
<td>None</td>
<td>No change in IgA-TG2</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Holm et al 2006</td>
<td>10 14 g (range 7–19 g)</td>
<td>3–24 months</td>
<td>Statistically significant increase†</td>
<td>Statistically significant increase†</td>
<td>NA</td>
<td>EMA and IgA-TG2 positive in 5/10 at 1 month</td>
<td>4/10 developed symptoms at histological relapse</td>
<td>–</td>
</tr>
<tr>
<td>Catassi et al 1993</td>
<td>10 100 mg</td>
<td>4 wk</td>
<td>0.2†</td>
<td>8 per 100 ECs†</td>
<td>NA</td>
<td>4/10 seroconverted to positive IgA AGA</td>
<td>None</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>10 500 mg</td>
<td>4 wk</td>
<td>0.5†</td>
<td>15 per 100 ECs†</td>
<td>NA</td>
<td>6/8 seroconverted to positive IgA AGA</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

*Over 0.4 is considered a statistically significant change.26
†Significant change compared with baseline (P≤0.05).
‡Significant change compared with baseline (P<0.001).
§The study consisted of a low and high gluten group with no significant differences between the groups in histology or serology. In the study, the data were presented after combining the two groups.
¶Non-significant change (P>0.05).
**Data not presented.
††Significant change compared with placebo (P<0.05).
CDQ, Celiac Disease Questionnaire; CSh, Celiac Symptom Index; EC, epithelial cells; EMA, endomysial antibody; GSRS, Gastrointestinal Symptom Rating Scale; IFN, interferon; IL, interleukin; NA, not available; SF-36, 36-Item Short-Form; TG2, tissue transglutaminase; TNF, tumour necrosis factor; VhCRD, villous height crypt depth ratio; wk, week.
Coeliac disease

for several years. Overall, high serologic titres, or significant increases in serologic titres, are predictive of VA, but substantial mucosal changes may occur without a significant change in serology.

It is important to note that the serological tests were developed for the diagnosis of CD. Currently, the Food and Drug Administration (FDA) have only cleared use of coeliac serologies as an aid for diagnosis of patients with suspected CD. This restriction limits how serologies can be used in regulatory trials, although they are routinely used for monitoring in clinical practice. To date, no manufacturer has submitted a claim for use of serological tests for disease monitoring and the FDA is only able ‘to review submitted claims’. Nonetheless, well validated IgA anti-TG2 and IgG anti-DGP tests will be important assets to clinical studies by helping to monitor CD activity during short and especially longer duration gluten challenge trials. Their further validation in ongoing trials may lead to FDA approval as secondary or combined (with histology or symptoms) primary end points in phase II and III clinical studies (table 1).

**Recommendations:** Although serology is not approved for use as a primary pivotal clinical trial outcome by the FDA, IgA TG2 and IgG DGP should, at a minimum, be measured at study entry and at completion in trials of CD (grade B). For entry into a gluten challenge study, participants should have near-normal titres, whereas for treatment studies, titres may be either elevated or normal, with stratification by serologic titre at study entry possibly as an a priori analysis (grade B). Choice of assay should be made with care and attention to be paid to dynamic range, especially around or below the cut-off for normal ranges. Preferably, an assay with a calibration curve should be applied. Although cut-off ranges for diagnosis may not be optimal for monitoring response or predicting VA, any significant increase during a trial suggests increasing CD activity and may be used as a key outcome in some studies.

Vote: agree: 2; strongly agree: 17.

**Immunological tools to measure treatment outcomes**

Known innate, and particularly adaptive, immune mechanisms in CD, are prime candidates to measure treatment outcomes with and without gluten challenge. These may be non-invasive blood tests, duodenal biopsies with histological assessment by immunocytochemistry, gene expression signatures or in vitro culture. With in vivo gluten challenge, there is rapid immune activation in the duodenum. One study found that interferon (IFN)-γ was increased both at baseline and with gluten challenge and for this reason does not appear to be a useful measure of disease activity. A whole-blood IFN-γ release assay is a much more promising measure for identifying immune responsiveness to gluten.

Treated patients with CD and healthy individuals show highly variable differences in serum cytokines and chemokine levels. Gluten challenge, however, induces a wave of cytokine release. In the lamina propria of the duodenal mucosa gluten peptides are taken up by dendritic (antigen-presenting) cells with surface HLA-DQ2 or DQ8 MHC molecules to stimulate gluten peptide reactive CD4+ T cells. On day 6 of postgluten challenge an increase of active disease gluten responsive T cells was seen in peripheral blood. These T cells can be demonstrated in Enzyme-Linked ImmunoSpot (ELISPOT) assays of cytokines when rechallenged with gluten ex vivo. Overall, 80%–90% of treated patients with CD in remission will show a positive response on challenge. T cells can also be demonstrated by binding to HLA-DQ-gliadin peptide tetramers, a construct consisting of multimers of HLA-DQ2 or HLA-DQ8 molecules bound to a gluten peptide and a reporter molecule giving a signal in flow cytometry. Here, also, 80%–90% of challenged patients with CD in remission will have a positive test. However, both the ELISPOT and the HLA-DQ2-gluten tetramer tests are confounded by large interindividual differences and small number of studied patients. Recently, the HLA-DQ2-gluten tetramer technology has demonstrated disease-specific T cells in the peripheral blood even without a gluten challenge. This, together with the demonstration of restricted T-cell receptors, may lead to new outcome measurements.

Serum IgA anti-TG2 antibody levels as biomarkers of disease activity seem to be a useful tool (see the ‘Serology’ section). Peripheral blood B cells may also prove to be a potential source of future biomarkers. The local mucosal production of antibodies targeting extracellular TG2 in vivo has shown potential in diagnosis, but was not informative beyond standard histology in a clinical drug trial with gluten challenge.

**Recommendations:** Although several immunological markers are under development as potential outcome measures, they have not been validated for therapeutic trials. At this point, they should only be used as exploratory outcomes in phase II and III clinical trials (grade D).

Vote: agree: 2; strongly agree: 17.

**Gluten immunogenic peptides as a compliance measure**

Symptom monitoring, serology and histology are at best indirect measures of GFD adherence with imperfect overall accuracy. Similarly, diet questionnaires are poor predictors of gluten exposure. Gluten immunogenic peptides (GIPs), including the 33-mer peptide from [2]-gliadin, are resistant to GI digestion. Because of this resistance, GIP can be detected in faeces or urine and thus provides direct evidence and likely quantitation of intake. A clinical trial examined correlations between faecal GIPs and traditional methods to monitoring the GFD. The majority (83.7%) of children with CD under 3 years of age had faeces negative for GIPs. Among those aged ≥13 years, faecal positivity for GIPs rose to 39.2%. More males were positive for GIPs in faeces compared with females (60% vs 31.5%, P=0.034). Serum IgA anti-TG2 was found negative in 40 of the 56 patients with GIP-positive stools. Today’s data suggest that GIP testing may be superior to questionnaires or anti-TG2 antibodies. Furthermore, a strong correlation has been demonstrated between the absence of GIPs in urine and healing of the intestinal epithelium. The first therapeutic clinical trials using the technology are ongoing (NCT02637141, NCT02633020). Whereas coeliac trial investigators and sponsors had to previously guess whether patients were consuming gluten or adhering to the GFD, this technology decreases the guesswork with data to accurately interpret results and outliers. The assay could possibly be developed for quantitative detection of the rate of glutamine residue deamidation in trials aiming at interfering with transglutaminate activity.

**Recommendations:** GIP testing is a promising tool for evaluating and selecting patients for clinical trials in CD aimed at reducing toxicity related to gluten exposure (grade D). Hence, it should be considered in future trials, especially for trials in non-responsive CD for therapies that are designed to prevent symptoms because of inadvertent gluten exposure (grade D).

Vote: agree: 4; strongly agree: 15.
Clinical trial end points

The ideal clinical trial end point should be clearly linked to an outcome important to patients, reliable, responsive to treatment, clinically or physiologically proximal to the outcome of interest and efficient and scalable for use in diverse clinical trial settings. Although in some areas biomarkers may be used as primary outcomes in clinical trials, these rarely have sufficient data for regulatory acceptance. In contrast, clinical outcome assessments can be more easily linked to patient well-being: clinical outcome assessments are grouped into PROs, clinician-reported outcomes, observer-related outcomes and performance outcomes. Clinician-reported outcomes, such as physician global assessment, are of limited value as they do not directly assess patient status and generally do not correlate well with PROs. Observer-related outcomes can be vital in specific populations, such as young children, where direct response is not possible. Symptom-focused PROs are the main clinical outcome assessment in use in CD and in gastroenterology overall and are the focus of this section, although other clinical outcomes will also be discussed.

From a clinical/practical perspective, PROs can be helpful to monitor patient status and target quality improvement initiatives. A growing number of digital devices allow patients to track and transmit symptom data to their physician; however, for these to be useful for practising clinicians, PROs must be easy to administer and interpret, as well as to allow feasible clinical interventions. In research and clinical trials, the key features of PROs that should be considered are high responsiveness to change and low participant burden.

For patients, the ideal PRO must focus on the symptoms or disease attributes most meaningful to them while again minimising time and complexity of use. In CD, this is particularly important given the significant impact on emotional, mental and social well-being due to the constant vigilance required by the GFD. Finally, from a regulatory standpoint, the ability to use a PRO in a pivotal clinical trial to support a labelling claim depends on whether its characteristics (eg, concept being measured, content validity, conceptual framework, intended population, format, scoring) are satisfactory and clearly documented in a regulatory dossier, which is now available for only a few PROs.

Presently available PROs frequently used in CD or developed for CD include the Gastrointestinal Symptom Rating Scale (GSRS), the Celiac Symptom Index (CSI), the Celiac Disease PRO (CeD-PRO), the Celiac Disease Symptom Diary (CDS Diary) and the Celiac Disease Assessment Questionnaire (CDAQ). Of these, the GSRS has been used most frequently, ranging from natural history to the effects of the GFD to clinical trials of novel therapeutics. The GSRS was developed originally for peptic disease and irritable bowel syndrome, but because the symptoms of many GI disorders overlap, it has proven useful for a variety of GI disorders, including CD. However, it is not optimised for CD and would not be appropriate for use in pivotal trials. Conversely, the CSI was developed specifically for CD and has been used in many cross-sectional and interventional studies. However, its development was completed before the 2009 FDA guidance and thus the CSI lacks much of the documentation necessary for regulatory clearance. Conversely, both the CDS Diary and the CeD-PRO were developed specifically for regulatory approval of CD therapeutics and are preferred instruments for this purpose.

The CDAQ was recently developed and assesses a variety of domains: symptoms, dietary burden, worry, stigma, and social isolation. As such, it is a hybrid of the health-related quality of life (HRQoL) tools discussed below and a symptom PRO. It is unclear if this instrument were developed and documented in line with regulatory guidance: in the CDAQ both constipation and diarrhoea are evaluated in a single combined question that may make changes in these important symptoms difficult to assess. However, the incorporation of these or similar HRQoL domains related to disease burden is critical in ensuring that outcomes are relevant and meaningful to patients.

Across these instruments, there is significant overlap, which is expected given the limited number of GI symptoms in general (table 2): common to all the PROs are diarrhoea, abdominal pain, bloating, and nausea. It should be acknowledged that while PROs in general may be developed and tested in specific diseases, they will not discriminate between diseases and therefore scores

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Symptoms assessed across PROs in CD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symptom</td>
<td>CDS Diary</td>
</tr>
<tr>
<td>Abdominal pain</td>
<td>X</td>
</tr>
<tr>
<td>Diarrhoea</td>
<td>X</td>
</tr>
<tr>
<td>Bloating</td>
<td>X</td>
</tr>
<tr>
<td>Nausea</td>
<td>X</td>
</tr>
<tr>
<td>Constipation</td>
<td>X</td>
</tr>
<tr>
<td>Fatigue*</td>
<td>X</td>
</tr>
<tr>
<td>Headache*</td>
<td>X</td>
</tr>
<tr>
<td>Skin rash</td>
<td>X</td>
</tr>
<tr>
<td>Flatulence</td>
<td>X</td>
</tr>
<tr>
<td>Burping</td>
<td>X</td>
</tr>
<tr>
<td>Problems thinking*</td>
<td>X</td>
</tr>
<tr>
<td>Incomplete emptying</td>
<td></td>
</tr>
<tr>
<td>Loss of appetite</td>
<td></td>
</tr>
<tr>
<td>Hunger pains</td>
<td></td>
</tr>
<tr>
<td>Stomach rumbling</td>
<td></td>
</tr>
<tr>
<td>Reflux</td>
<td></td>
</tr>
</tbody>
</table>

*Removed from final PRO as per the Food and Drug Administration recommendations. †Assessed in a single domain.

CD, coeliac disease; CDAQ, Celiac Disease Assessment Questionnaire; CDS Diary, Celiac Disease Symptom Diary; CSI, Celiac Symptom Index; GSRS, Gastrointestinal Symptom Rating Scale; PRO, patient-reported outcome.
Coeliac disease

for patients with different GI disorders will overlap. There is also a poor correlation between symptoms, histology, and serology that is due in part to a different time to response after gluten exposure or coexisting symptoms due to irritable bowel syndrome or undetected food allergy. Thus, it is also clear from recent clinical trials that many symptomatic patients have no histological or serological evidence of active CD and many patients with significant enteropathy have few or no symptoms. Meanwhile complete recovery of small intestinal lesions is very rare in adult CD patients despite symptom disappearance.

In addition, PRO use in CD can be challenging because of symptom heterogeneity (eg, asymptomatic or paucisymptomatic patients) and variable extraintestinal manifestations, for which no PROs are available. In contrast to disorders such as chronic constipation or headache in which one symptom defines the condition, symptoms in CD can vary substantially between individuals and hence careful attention to PRO use is mandatory. Moreover, responsiveness to change of non-symptom-based dimensions in PROs can vary significantly and must be carefully assessed in relation to the intervention under investigation. For example, measures of quality of life (QoL) may be less amenable to change if the intervention does not reduce dietary burden, social isolation and stigma, which are closely linked to managing the GFD. Comparing overall PRO mean scores at baseline and postintervention may dilute the treatment effect if not all domains change or if changes cancel out each other (eg, diarrhoea improves but results in constipation by disrupting bowel movements). Another option is to compare the means only in prespecified domains (eg, include diarrhoea but not constipation). However, this approach may result in a highly selected population that is not representative. A more sophisticated approach is to limit primary assessment of the effect of intervention to the symptoms most bothersome to a specific patient and then to include all individuals with bothersome symptoms in the final PRO assessment. Even with this approach, for treatment trials, only patients with symptoms measured by the primary PRO outcome can be enrolled. Given the heterogeneity of CD, it is likely that sequential trials will need to be performed with different outcomes in order to understand the utility of a particular therapy. For example, a therapy found to be useful in improving GI symptoms in CD could be assessed in a later trial assessing itch in patients with dermatitis herpetiformis.

Particularly noteworthy is that while great progress has been made in PROs in CD, there is limited experience using the more recent PROs in languages other than English and none has been validated for use in paediatric populations. Although it is expected that the PROs developed for CD in adults will be applicable for paediatric populations, this still requires validation. For example, young children and teens may define improved QoL differently because of unique challenges in school, social settings and peer relationships. Furthermore, improvement in extraintestinal symptoms, including behavioural changes, may be more relevant among paediatric patients. Development of formats suitable for caregivers of children unable to independently complete the questionnaires must also be advanced. Development of responder definitions and minimal clinically important change are additional areas requiring development to realise the potential of CD PROs in clinical research.

Given that the relative SD of histology is substantially smaller than that of symptoms, for a study where several hundred patients are needed to adequately detect differences in symptoms, fewer patients may be needed for the histology end point. Under such circumstances, researchers may choose to perform histology on a random sample of study participants provided that they have sufficient study power for the histology end point.

Recommendations: Clinical end points must be included in trials of CD and generally PROs should be a primary outcome in studies of treatment of active CD, generally late phase II and III (grade D). Thus far, there is insufficient evidence to recommend one specific scale, although the CDSSD and the CeD-PRO appear most likely to meet regulatory requirements (grade D). Given the heterogeneity of symptoms in patients with CD, it is adequate to limit analyses to certain domains, either in the study overall or to allow for participant-specific symptom assessments. However, such decisions should be made at study entry and rigorously documented.

Vote: agree: 3; strongly agree: 16.

Health-related quality of life assessment

Clinical trials must also consider the ongoing psychological burden of CD to better understand the outcomes. Many studies in different settings suggest that CD has a considerable impact on HRQoL. In untreated CD, GI symptoms and extraintestinal issues, such as isolated anaemia, fatigue and malaise, may be responsible for reduced HRQoL. In general, the treatment of CD results in significant improvement in the HRQoL of symptomatic patients. Even in patients with silent or asymptomatic screening-detected CD, improvement in both symptoms and HRQoL has been shown in numerous studies.

Impairment in HRQoL may also contribute to and be impacted by psychological disorders (eg, anxiety and depression) seen in patients with CD. Whereas anxiety generally improves after diagnosis and treatment CD, depression may exist before and persist after diagnosis. Moreover, anxiety and depression alone or through their impact on HRQoL can impair dietary compliance. The interaction between mood disorders, GFD adherence and HRQoL is incompletely understood and should be addressed in future trials. Additionally, the burden of a GFD may lead to ongoing HRQoL impairment despite symptom relief and improved physical well-being after dietary intervention.

Because of small numbers of studies and variations in study designs, populations and HRQoL measures, we lack a complete understanding of degrees and drivers of HRQoL in individuals with untreated and treated CD. There are also few studies about the effect of CD treatment in the outcome of depression and anxiety. Therefore, further studies are required if we wish to know more about this specific aspect of CD.

Although HRQoL is generally not accepted by regulatory agencies as a primary outcome in pivotal trials, this is a key outcome for both patients and clinicians and may determine whether a therapy succeeds or fails. Arguably, a main goal of therapeutics in CD, in addition to histological and symptom improvement, is to improve HRQoL. Indeed, it is probable that some patients in histological and symptomatic remission could have a significant HRQoL benefit from pharmacological therapy owing to a reduction in burden of the GFD and anxiety regarding potential exposures. There are several HRQoL scales that have been specifically developed for or used in CD (table 3).

Recommendations: HRQoL in CD is complex and multidimensional and may be a more relevant concern to patients than specific symptoms. In any trial aiming to improve CD control (as opposed to gluten challenge studies that aim to prevent worsening), measurement of HRQoL should be considered a critical end point that may help to determine the overall value of a therapy or intervention to both patients and payors (grade D).

Vote: agree: 4; strongly agree: 15.
<table>
<thead>
<tr>
<th>Instrument Description</th>
<th>Instrument Author Participants</th>
<th>Instrument description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Generic</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zung Self Rating Depression Scale</td>
<td>Ciacci 2003149</td>
<td>581 CD</td>
</tr>
<tr>
<td>Gastrointestinal Quality of Life Index</td>
<td>Casellas 2008150</td>
<td>340 CD</td>
</tr>
<tr>
<td>COPE (brief)</td>
<td>Smith 2011151, Sainsbury 2013152</td>
<td>156 CD, 189 CD</td>
</tr>
<tr>
<td><strong>Adult instruments</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EuroQol-5D</td>
<td>Casellas 2008153, Grey 2010154, Norström 2011155, Casellas 2015156, Ramirez-Cervantes 2015157</td>
<td>340 CD, 783 CD, 1031 CD, 366 CD on GFD&gt;1 year, 80 CD on GFDp6months</td>
</tr>
<tr>
<td>Beck Depression Inventory</td>
<td>Borghini 2016158, Nachman 2009159, Nachman 2010160</td>
<td>210 CD (70 untreated vs 70 on GFD 6–12 months vs 70 on GFD&gt;12 months)</td>
</tr>
<tr>
<td>Psychological General Well-Being Index</td>
<td>Ford 2012161, Smith 2011162, Mustalahti 2002163, Mahadev 2016164, Borghini 2016165, Paukola 2011166, Peräaho 2004167, Paavola 2011168, Barratt 2011169, Ukkola 2011170, Johnston 2004171, Häuser 2006172, Häuser 2007173, Barratt 2011174, Barratt 2011175, Ramirez-Cervantes 2015176</td>
<td>288 CD, 211 (71 screen-detected CD vs 140 symptom-detected CD), 210 CD (70 untreated vs 70 on GFD 6–12 months vs 70 on GFD&gt;12 months), 466 CD (96 screen detected CD vs 370 symptom-detected CD vs 110 non-CD controls), 39 CD, 98 CD (54 screen-detected CD vs 44 symptom-detected CD), 698 CD, 156 CD, 596 CD</td>
</tr>
<tr>
<td>Hospital Anxiety and Depression Scale</td>
<td>Häuser 2007177, Häuser 2008178, Barratt 2011179, Barratt 2011180, Ramirez-Cervantes 2015181</td>
<td>516 CD, 446 CD, 225 CD, 225 CD, 80 CD on GFDp6months</td>
</tr>
<tr>
<td>The Short-Form 36-Item QoL measure</td>
<td>Bakker 2013182, Usal 2002183, Usal 2007184, Barratt 2011185, Hallert 1998186, Viljamaa 2005187, Johnston 2004188, Häuser 2005189, Häuser 2006190, Nachman 2009191, Nachman 2010192, Hopman 2009193, Tontini 2010194, Aksan 2015195</td>
<td>57 CD + TID, 58 CD, 129 CD, 225 CD, 225 CD, 89 CD, 98 CD, 14 CD, 516 CD, 446 CD, 132 Newly diagnosed CD, 53 Newly diagnosed CD, 53 CD, 43 CD, 205 CD</td>
</tr>
</tbody>
</table>

**CD-specific scales**

**Continued**
## Table 3  Continued

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Author</th>
<th>Participants</th>
<th>Instrument description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Celiac Disease Questionnaire (child)</td>
<td>Häuser 2006</td>
<td>516 CD</td>
<td>28-item self-report, 7-point Likert scale. Measures four domains: emotional and social problems, disease-related worries and GI symptoms.</td>
</tr>
<tr>
<td></td>
<td>Zampieron 2011</td>
<td>187 CD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ford 2012</td>
<td>288 CD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Marchese 2013</td>
<td>171 CD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sainsbury 2013</td>
<td>189 CD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pouchoz 2014</td>
<td>211 CD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Aksan 2015</td>
<td>205 CD</td>
<td></td>
</tr>
<tr>
<td>Celiac Disease Quality of Life Survey</td>
<td>Mahadev 2016</td>
<td>211 CD</td>
<td>20-item self-report, 5-point Likert scale, four subscales: limitations, dysphoria, health concerns and inadequate treatment.</td>
</tr>
<tr>
<td></td>
<td>Zingone 2011</td>
<td>230 CD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Casellas 2013</td>
<td>298 CD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Casellas 2015</td>
<td>366 CD (CD&gt;1 year)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lee 2016</td>
<td>2138 CD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rodriguez-Almagro 2016</td>
<td>1230 CD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Russo 2017</td>
<td>19 CD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Castilhes 2015</td>
<td>103 CD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dowd 2017</td>
<td>220 CD</td>
<td></td>
</tr>
<tr>
<td>Fatigue scales</td>
<td>Siniscalchi 2005</td>
<td>130 CD</td>
<td>8-item self-report for physical fatigue and five items for mental fatigue. The score ranges from 0 to 26.</td>
</tr>
<tr>
<td>The Fatigue Severity Scale (child)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daily Fatigue Impact Scale Questionnaire</td>
<td>Jordà 2010</td>
<td>51 CD</td>
<td>8-item self-report questionnaire, five alternative responses per item (from 0=no problem, to 4=extreme problem).</td>
</tr>
<tr>
<td>Eating disorder scales</td>
<td>Passamani 2013</td>
<td>100 CD</td>
<td>16-item self-report. Measures the behavioural aspects of binge eating, as well as feelings and thoughts associated with binge eating.</td>
</tr>
<tr>
<td>Eating Disorder Inventory (child)</td>
<td>Passamani 2013</td>
<td>100 CD</td>
<td>91-item self-report (range 0–78); 11 scales, eg, inadequacy, interceptive awareness, fear of maturity, asceticism, impulsivity and social insecurity.</td>
</tr>
<tr>
<td>Eating Disorder Examination (EDE)</td>
<td>Wagner 2015</td>
<td>259 CD adolescents</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Karwautz 2008</td>
<td>283 CD adolescents</td>
<td></td>
</tr>
<tr>
<td>Eating Disorder Examination (EDE)</td>
<td>Wagner 2015</td>
<td>259 CD adolescents</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Karwautz 2008</td>
<td>283 CD adolescents</td>
<td></td>
</tr>
<tr>
<td>Eating Disorder Examination Questionnaire (EDE-Q)</td>
<td>Karwautz 2008</td>
<td>283 CD adolescents</td>
<td>28-item, self-report. Uses same subscales as the EDE interview.</td>
</tr>
<tr>
<td>Sleep disorders scale</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pittsburgh Sleep Quality Index (child)</td>
<td>Zingone 2010</td>
<td>60 CD</td>
<td>19 self-rated items (15 multiple choice and 4 write in)+5 questions rated by a bed partner or roommate (range 0–57).</td>
</tr>
<tr>
<td>Pediatric scales</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TNO-AZL Children’s QOL questionnaire (TACQOL)</td>
<td>Kolsteren 2001 van Kopp 2009</td>
<td>133 CD children</td>
<td>Generic–two forms for children and parents: The TACQOL comprises seven scales (range0–16); Generic: Dutch Children TNO-AZL QOL Questionnaire 25 items with four.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>32 CD children</td>
<td>Domains: emotional, social, familial, and physical.</td>
</tr>
<tr>
<td>DUX-25 (child)</td>
<td>van Kopp 2009</td>
<td>32 CD children</td>
<td>Ages 5–16; 25-item self-report with four domains: physical, emotional, social and home functioning.</td>
</tr>
<tr>
<td>Inventory of Life Quality in Children and Adolescents</td>
<td>Wagner 2008</td>
<td>365 CD adolescents</td>
<td>9-item self-rating questionnaire</td>
</tr>
<tr>
<td>Berner Subjective Well-being Inventory</td>
<td>Wagner 2008</td>
<td>365 CD adolescents</td>
<td>Nine areas: school, family, social peer contacts, interests and leisure activities, physical health, psychological health, overall QOL, judgement and disease and therapy-associated burden.</td>
</tr>
<tr>
<td>PenDQOL core: Pediatric Quality of Life Measurement Inventory</td>
<td>Sud 2012 Biagetti 2015 Sevin 2016</td>
<td>28 CD+ type 1 diabetes children</td>
<td>Four areas: physical functioning (eight items), emotional functioning (five items), social functioning (five items) and school functioning (five items).</td>
</tr>
<tr>
<td>WHOQOL-BREF (child)</td>
<td>de Lorenzo 2012</td>
<td>33 CD child/parent dyads</td>
<td>26-item self-report (parent), 5-point Likert scale. Six domains: loneliness, general health perception, physical and psychological functioning, social and environmental.</td>
</tr>
<tr>
<td>EuroQol-5D (child)</td>
<td>Nordyke 2011</td>
<td>153 screen-detected CD children and 66 known CD vs non-CD controls</td>
<td>5-item self-report, 5-point Likert scale, plus a general health rating (scored on a 20 cm visual analogue scale) (child friendly pilot and proxy used).</td>
</tr>
<tr>
<td>DISABKIDS Chronic (short version)</td>
<td>Bystrom 2012</td>
<td>160 CD child/parent dyads</td>
<td>12-item self-report Likert scale for ages 8–18 (additional available version for ages 4–7) mental, social and physical domains.</td>
</tr>
</tbody>
</table>

Continued
DISCUSSION

Strengths and limitations

In this review, a large number of authors reviewed the literature on treatment outcomes in coeliac trials to issue recommendations for future trials. Our research team was multidisciplinary and the treatment outcomes we have evaluated in the paper reflect the expert views of the authors.

We performed an extensive literature review of more than 10,000 papers, and based our deliberations on personal experiences and expertise from our treatment of patients with CD with clinical trials and our research in CD. There are already a number of guidelines for reporting treatment outcome in CD. Hence, our paper is not meant to give authoritative advice on the study design, which is not yet possible because of the developing nature of the field, but to complement available literature with our expertise with a focus on how to measure response to non-GFD treatment.

CD is a lifelong disease in which the GFD is the only treatment option. However, we suspect that soon other alternative treatments will become commonplace.

Regulatory agencies are responsible for evaluating new therapies based on risks and benefits to patients in how they feel, function or survive. These aims are of intuitive value to patients and clinicians and sufficiently broad that they should form the foundation of any interventional clinical trial. Despite this, the precedent in many fields, including gastroenterology, has been the use of poorly validated outcomes of limited applicability to clinical practice. Although CD adversely impacts survival and function, these outcomes are generally not feasible to assess in clinical trials because of low prevalence and long latency. This leaves the options of measuring how patients feel—mainly using PROs, histology and serology—to assess changes in risk of future adverse events. Ideally, a treatment should result in improvement in more than one outcome measures (PROs, histology and serology), and that is possible using a coprimary end point. However, coprimary end points should be used with caution since they decrease study power and the number of patients is often limited.

There are currently several well-designed and partially validated PROs developed for CD that should be considered standards for trials in the future. Assessment of extent (degree) of enteropathy (intestinal architecture and IEL assessment) should be considered a critical outcome in clinical trials of CD. However, it is recognised that technical limitations of duodenal biopsy as a reflection of overall small intestinal mucosal disease limit the potential value of histology as an end point. Development techniques that better evaluate enteropathy across the small intestine are applicable in clinical practice and relevant to patients.

Author affiliations

1Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
2Department of Pediatrics, Örebro University Hospital, Örebro, Sweden
3Coeliac Center at Department of Medicine and Surgery, Sccoli Medica Salernitana, University of Salerno, Salerno, Italy
4Celiac Disease Center at Columbia University, New York, USA
5Celiac Disease Research Center, Faculty of Medicine and Life Sciences, University of Tampere, Tampere, Finland
6Department of Internal Medicine, Tampere University Hospital, Tampere, Finland
7Coeliac Disease Centre, Heim Pál Children’s Hospital, Budapest, Hungary
8Department of Paediatrics, Faculty of Medicine, University of Debrecen, Debrecen, Hungary
9Celiac Disease Research Center, Faculty of Medicine and Life Sciences, University of Tampere, Tampere, Finland
10The Mayo Clinic, Rochester, Minnesota, USA
11Institute of Clinical Medicine and K.G. Jebsen Coeliac Disease Research Centre, Faculty of Medicine, University of Oslo, Oslo, Norway
12Department of Gastroenterology, Oslo University Hospital, Oslo, Norway
13Science Center, Tampere University Hospital, Tampere, Finland
14Tampere Centre for Child Health Research, Faculty of Medicine and Life Sciences, University of Tampere, Tampere, Finland
15Institute for Mother and Child Health Bucharest, University of Medicine and Pharmacy ‘Carol Davila’, Bucharest, Romania
16Tampere Centre for Child Health Research, University of Tampere, Tampere University Hospital, Tampere, Finland


1419
Coeliac disease

---

**Contributors** JFL and DL initiated the study. JFL coordinated the study and wrote the first draft, the text was then extensively revised by the coauthors. All authors contributed to the literature searches, contributed to the writing of the manuscript and approved the final version of the manuscript.

**Funding** JFL was supported by the Swedish Research Council (522-2409-195) and the Swedish Society of Medicine while writing the draft of this paper. DS received research support from the German Research Foundation (DFG) and the Ministry Research and Development (MBMF) and the Leibniz Foundation. DSS received an educational grant from Biocad and Simtomet to undertake an investigator-led research study on point-of-care tests and an educational grant from Dr Schär (a gluten-free food manufacturer) to undertake an investigator-led research study on gluten sensitivity. KKa was supported by the Academy of Finland, the Competitive Research Funding of the Tampere University Hospital and The Sigrid Juselius Foundation. IRK-S was supported by the Hungarian Research Fund (Grant NKFI 120392). MIM was supported by the Competitive State Research Financing of the Tampere University Hospital (grant 9U038).

**Competing interests** PHRG: scientific advisory board of Alvina Pharmaceuticals, ImmunogenX and Immunosant. JAM: serves on the advisory board of Celimmune, was a consultant to BioLineRx, GlaxoSmithKline (GSK), Genentech, UCB Biopharma and Genmark Pharmaceuticals Ltd and is a consultant to Immunosant, Institute for Protein Design (IP-Pharmaceuticals Biologic), Takeda Pharmaceutical Company, Ltd., Innovate Biomanufacturing, Inc., Intrexon, ZGPharma Inc., Boehringer-Ingehelm and Immunosant. KEAL: Immunosant, Regenon and Alvina Pharmaceuticals. DSS: holds a patent and receives royalties for the TGZ-antibody assay, has received an educational grant from Coeliac UK, Biocad, Simtomet and Dr Schär to undertake an investigator-led research study on CD and/or gluten sensitivity. NRR: clinical advisory board for Immunosant. DAL: Medical Director for Takeda Pharmaceuticals Asia Pacific, co-author detecting gluten peptides in human fluids (Patent No. US 20170160288 A1), consultant for Vircell. IRK-S: patent on rapid coeliac antibody detection licensed by the University of Tampere to Labsysy Om, Finland. MIM: serves on the Advisory Board of Celimmune, USA, Immunosant, USA and Innovate Pharmaceuticals Inc, USA; is consultant to FinnMedi Oy, Finland and Jibol Oy, Finland via his own company Maki HealthTech Oy, Finland; is an inventor in the patent Methods and Means for Detecting Gluten-Induced Diseases, USA (Patent No. 7,361,480—USA, European Patent No. 1390753). The patent resulted in a commercial product from FinnMedi at the Tampere University Hospital and the University of Tampere, a coeliac disease point-of-care test, Biocad Celiac Test, licensed to Labsysy Diagnostics Oy (former AnBiotech Oy), Finland.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

---

**REFERENCES**

Coeliac disease

an endoscopic approach that avoids the need for biopsy: a double-center study. 


Coeliac disease


Carver CS. You want to measure coping but your protocol’s too long: consider the brief COPE. *Int J Behav Med* 1997;4:92–100.


Coeliac disease


