Statistical Analysis Plan

1. ADMINISTRATIV INFORMATION

Title:

The Effect of Gluten in Adolescents with Minor Gastrointestinal Symptoms: A Double-Blind, Randomized Crossover Trial with Repeated Measurements

Trial registration number: Registered at clinicaltrial.gov NCT04639921

SAP: Version 1.2021-02-03

Based on Protocol: Version 6, 2020.08.20

SAP Revisions: None

Roles and responsibility: CC[1,2] and SH[1,2] were investigators of the study and took part of the funding and all required permissions. RC[2,3] supervised the statistical analysis plan, coming analyses and interpretation.

[1]Hans Christian Andersen Children's Hospital, Odense University Hospital, Denmark. [2]Department of Clinical Research, University of Southern Denmark. [3]Section for Biostatistics and Evidence-Based Research, the Parker Institute, Bispebjerg and Frederiksberg Hospital, Copenhagen, Denmark.

2. INTRODUCTION

Non-celiac gluten sensitivity (NCGS) is a newly described disease entity, where the individual shows signs of sensitivity to gluten, but with no evidence of IgE-mediated wheat allergy or CD (1). The extent and severity of NCGS is so far an enigma (2), but symptoms of NCGS may be commonly occurring, with a reported prevalence of up to 6% of the adult population (3). So far, only ten studies with a double-blind placebo-controlled food challenge have been performed showing divergent study outcomes probably due to different study designs and methods and a significant nocebo/placebo effect. In general, the studies are characterized by patient recruitment from a gastroenterological clinic limited to adults (4). Our study stands out in this respect, as we will investigate an unselected group of patients as adolescents (not recruited from an outpatient clinic), because patients with NCGS often report that their symptoms started at that time. Thus, our aim was to explore the effect of gluten on gastrointestinal symptoms in adolescents with minor gastrointestinal symptoms from a cohort based on the Danish National Birth Cohort.

3. STUDY METHODS

Trial design: A within-person, blinded, placebo-controlled food challenge (BPCFC) trial.

Randomization: Participants were randomized in blocks of four in the randomization module in RedCap administrated by the data-manager of the study. There was no stratification.

Sample size*:* Primary sample size calculation was based on McNemars test for comparison of two related proportions, specifying discordant proportion. The sample size was calculated to 29 participants assuming that 30% would be positive (a 30% increase in mean VAS score compared to placebo) to the challenge with gluten and 1% to the challenge with placebo. The level of significance was set to 0.05, the power to 80% and an expected drop-out of 5%. However, while preparing the statistical analysis plan (before looking at the actual data) it was decided to use the difference in the average VAS as the primary endpoint. A sample size of 30 participants (i.e. 30 paired samples) would correspond to a statistical power of 88.8% to detect (P<0.05) a difference of 15 VAS-units assuming a standard deviation of 25, with a correlation between measures of 0.5.

Framework: We will compare groups, periods, and the interaction between them, based on Repeated-Measures Linear Mixed Effects Model with participants modeled as a random effects variable.

Statistical interim analyses and stopping guidelines: Not relevant.

Timing of final analysis: All outcomes will be analyzed collectively.

Timing of outcome assessments: Period 1 is defined from day 1-7, wash-out period from day 8-14, and period 2 from day 15-21.

4. STATISTICLES PRINCIPLES

Confidence intervals and P-values: All 95% confidence intervals and *P* values will be two sided. We will not apply explicit adjustments for multiplicity, rather we will analyze and interpret the primary and secondary outcomes in a prioritized order: The analyses of the secondary outcomes will be performed in sequence until one of the analyses fails to show the statistically significant difference, or until all analyses have been completed at a statistical significance level of 0.05.

Adherence and protocol deviations: The participants will be instructed by a clinical dietician how to follow a gluten free diet. During the whole study period, they have access by e-mail and phone to the dietician. To ensure that the participants eat their granola bar they receive a reminder text message every day.

Analysis populations: The primary analyses will be based on the Intention to Treat (ITT) population. The ITT principle asserts the effect of a treatment policy (that is, the planned treatment regimen), rather than the actual treatment given (i.e., it is independent of treatment adherence). Accordingly, participants allocated to a treatment group (ACB and BCA, respectively) were followed up, assessed and analyzed as members of that group, irrespective of their adherence to the planned course of treatment (i.e., independent of withdrawals and cross-over phenomena). We will analyze data from all participants who were randomized with the baseline measurement collected and thus available (i.e. the intention-to-treat population).

5. TRIAL POPULATION

Screening data: None

Eligibility: Inclusion criteria were participation in the GlutenFunen Cohort and more than three gastrointestinal symptom based on a questionnaire in the GlutenFunen cohort, ten was the maximum number of symptoms. Exclusion criteria were inflammatory bowel disease, transglutaminase Ig-A higher than reference range as indication of celiac disease and current antibiotic treatment.

Recruitment: 273 participants out 1266 participants had a symptom score higher than three. They were contacted by phone.

Withdrawal/follow-up: From the pilot study in 2019, we expect that 5% will drop-out from the BPCFC. We do not expect to lose any participant to follow-up. We expect that 10% of the VAS scores will be missing.

Baseline characteristics: We will report baseline demographic and clinical characteristics in Table 1.

6. ANALYSIS

Outcome definitions: The primary outcome is the mean VAS score when eating gluten compared to placebo. The VAS-score is from 0-100mm. There are 10 items and from these the average VAS symptoms score will be calculated and applied as the primary outcome measure.

Key secondary outcomes are the SF36 Mental component score and SF36 Physical component score (both scored from 0 to 100) and the Warwick Mental Health score (scored from 14 to70). Other secondary outcomes are gluten intake, gluten adverse events and every individual VAS score (scored from 0 to 100mm).The statistical estimates from the model with interaction terms etc. will be presented in Table 2, whereas the primary comparison between groups will be presented in Table 3.

Analysis methods: Our primary analyses will be based on the ITT population, including all randomized participants with available data at baseline. Missing data will be handled indirectly and statistically modeled using repeated-measures linear mixed models; the trajectory over time for the primary outcome measure will be presented in Figure 2. These models are considered valid assuming that data are 'Missing at Random' (MAR): i.e. "*Any systematic difference between the missing values and the observed values can be explained by differences in observed data*"(5). Contrasts between groups will be estimated based on repeated-measures analysis of covariance applied in mixed linear models (i.e., the main effect of group).

The primary statistical model will consist of fixed and random effects factors (i.e. covariates). Fixed effects define the expected values of the observations, and random effects define the variance and covariances of the observations. In this study, participants were randomly assigned to two treatment sequences (ACB vs BCA), and observations were made for each period over 7 day time points (and baseline) with the primary outcome measured at day 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21. Basically, there are three fixed-effect factors: group (three levels), period (three periods), and time (21 levels). Random effects result from variation between and within participants. We anticipate that measures on the same patient at different times are correlated, with measures taken closely together in time being more highly correlated than measures taken more apart in time; observations on different participants will be assumed as being independent (6).

The objectives of a repeated measures design are to make inferences about the expected values of the observations, that is, about the means of the populations from which participants are sampled. This objective is achieved by taking into account treatment and time effects in the model. Data will be analyzed using SAS and STATA, with the particular outcome variable (Yi) at baseline level (Y0,i) as a covariate, using a multilevel repeated measures mixed effects model with participants as the random effects factor based on a restricted maximum likelihood (REML) model (6).

For continuous outcomes (e.g., Yi score) the response (dependent) variable, and the baseline value (one for each participant), treatment group (three levels), period (three levels) and time (21 levels) will be included as covariates, as well as the interaction between treatment group and period; Patient ID will be handled as a random effect. This statistical model holds all between-group comparisons at all assessment points up to 21 days from baseline and allows for evaluation of the average effect (as visualized in the Figure 2 template below).

Missing data: Robustness is a concept that refers to the sensitivity of the overall conclusions to various limitations of the data, assumptions, and analytic approaches to data analysis. Robustness implies that the treatment effect and primary conclusions of the trial are not substantially affected when analyses are carried out based on alternative assumptions or analytic approaches.

Loss to follow-up and missing data for various reasons is difficult to avoid in randomized trials and in particular in pragmatic trials. We will apply the analysis framework suggested by White et al (2011) in which missing data related to the ITT approach depend on making plausible assumptions about the missingness of the data and including all participants in subsequent sensitivity analyses:(7)

#1. Attempt to follow up all randomized participants, even if they withdraw from allocated treatment (i.e., contact all individuals unless they explicitly stated that they had withdrawn their consent)

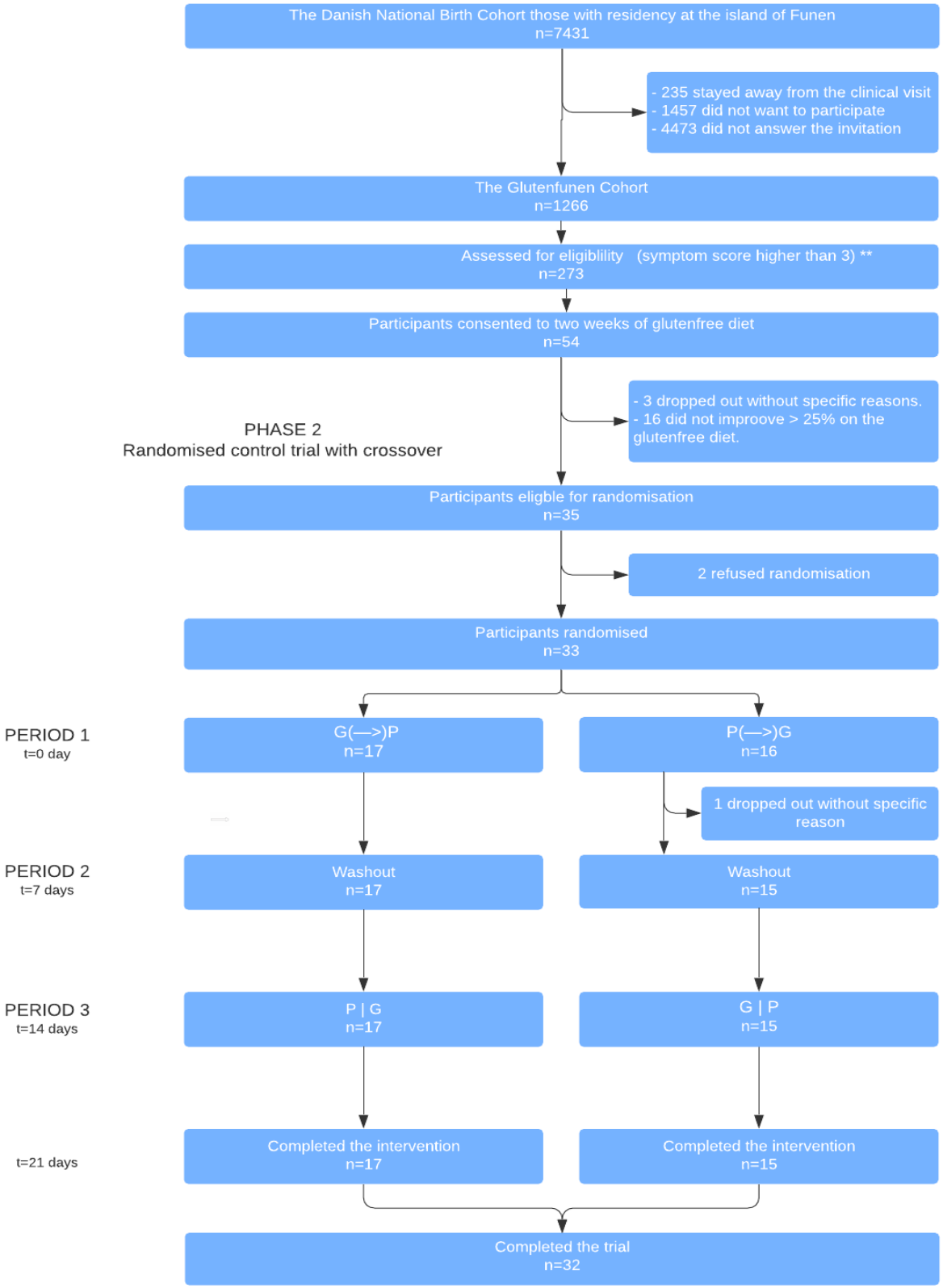
#2. Perform a main analysis of all observed data that are valid under a plausible assumption about the missingness of the data (i.e., Model-based: data as observed; using linear mixed models, assuming that data are '*Missing at Random*' [MAR])

#3. Perform sensitivity analyses to explore the effect of departures from the assumption made in the main (#2) analysis (i.e., a non-responder-imputation: using the value at baseline to replace missing data will correspond to a non-responder imputation; these models will potentially be informative even if data are 'Missing Not At Random' [MNAR])

#4. Account for all randomized participants, at least in the sensitivity analyses (covered by #2 and #3 above, plus the corresponding analyses based on the per protocol population).

The interpretation of the corresponding statistical measures of uncertainty of the treatment effect and treatment comparisons will involve consideration of the potential contribution of bias to the statistical inference such as P-values and 95% confidence intervals. Ad#1+2: Our primary analysis population will be all participants with available data at baseline, statistically modelled using repeated-measures linear mixed models (see above). These models will be valid if data are 'MAR'. Ad#3+4 Sensitivity: We will analyze all variables, with missing data being handled by a simplistic single-step non-responder imputation (baseline observation carried forward). If the different sensitivity analyses are in agreement, and the analyses on the sensitivity analyses and the main analysis lead to essentially the same conclusions, confidence in the trial results is increased.

Figure 1 Flowchart



** symptom score is based on a questionnaire from the GlutenFunen cohort. It is the number of gastrointestinals symptoms pr week, with a maximum score of ten.

Table 1. Baseline demographic and clinical characteristics

	Period 1 Gluten	**Period 1** Placebo	**Baseline total** Combined
Basic characteristics			
Age, years[a]			
Sex, females (n%)[b]			
Height, SD-score[c]			
Weight, SD-score[c]			
Body-Mass-Index (BMI) kg/m^2			
BMI, SD-score[c]			
Number of gastrointestinal symptoms (at least 4 out of 10)[d]			
IgE Wheat positive (n%)[e]			
On diet to obtain weight loss (n%)[f]			
Medication, fasting or vomiting to obtain weight loss (n%)[g]			
Wanting to be thin or afraid of weight gain (n%)[h]			
Exercising to obtain weight loss (n%)[i]			
Autoimmune disease among 1. Degree relatives (n%)[j]			
Primary outcome:			
VAS Symptom Score, 0-100			
Key secondary outcomes:			
SF36 Mental Component Score, score: 0-100[g]			
SF36 Physical Component Score, score: 0-100[g]			
WMHs, score: 14-70[g]			
Other secondary outcomes			
Gluten intake, (n%)			
VAS_1, 0-100mm			
VAS_2, 0-100mm			
VAS_3, 0-100mm			
VAS_4, 0-100mm			
VAS_5, 0-100mm			
VAS_6, 0-100mm			
VAS_7, 0-100mm			
VAS_8, 0-100mm			
VAS_9, 0-100mm			
VAS_{10}, 0-100mm			

* Values are means ± SDs, or median and interquartile range (depending on the empirical data distribution); unless otherwise stated.

[a] the age was calculated at the age for the participants September 1st 2020.

[b] Based on the Danish national cpr number

[c] Height, weight and BMI were reported as SD scores to allow for age and sex using the Danish references for growth(8). The values are obtained from the Glutenfunen Cohort including participants from 2018-2020.

[d] Based on the questionnaire to the participants in the Glutenfunen Cohort. The score reflects the number of gastrointestinal symptoms where the maximum was 10. To be included in this study a symptom score higher than three was necessary.

[e] IgE Wheat was measured at day 0. Higher than 0.35 kU/l was considered positive

[f] Was considered positive if the participants answered “every day” or “often” or “several times”, it was considered negative if “ a couple of times” or “never” or “do not know”. The time scale was the last year.

[g] Was considered positive if the participants answered “every day” or “several times per week” or “once at week” or “1-3 times per month” and negative if “never” or “do not know” or “less than once per month”

[h] Was considered positive id the participants answered “every day” or “often” or “sometimes”, it was considered negative if “ rarely” or “never” or “do not know”.

[j] Based on the questionnaire to the participants in the Glutenfunen Cohort.

[g] Measured at day 0

Figure 2. A visualization of the expected trajectories according to group, period, time, and interaction between time and group

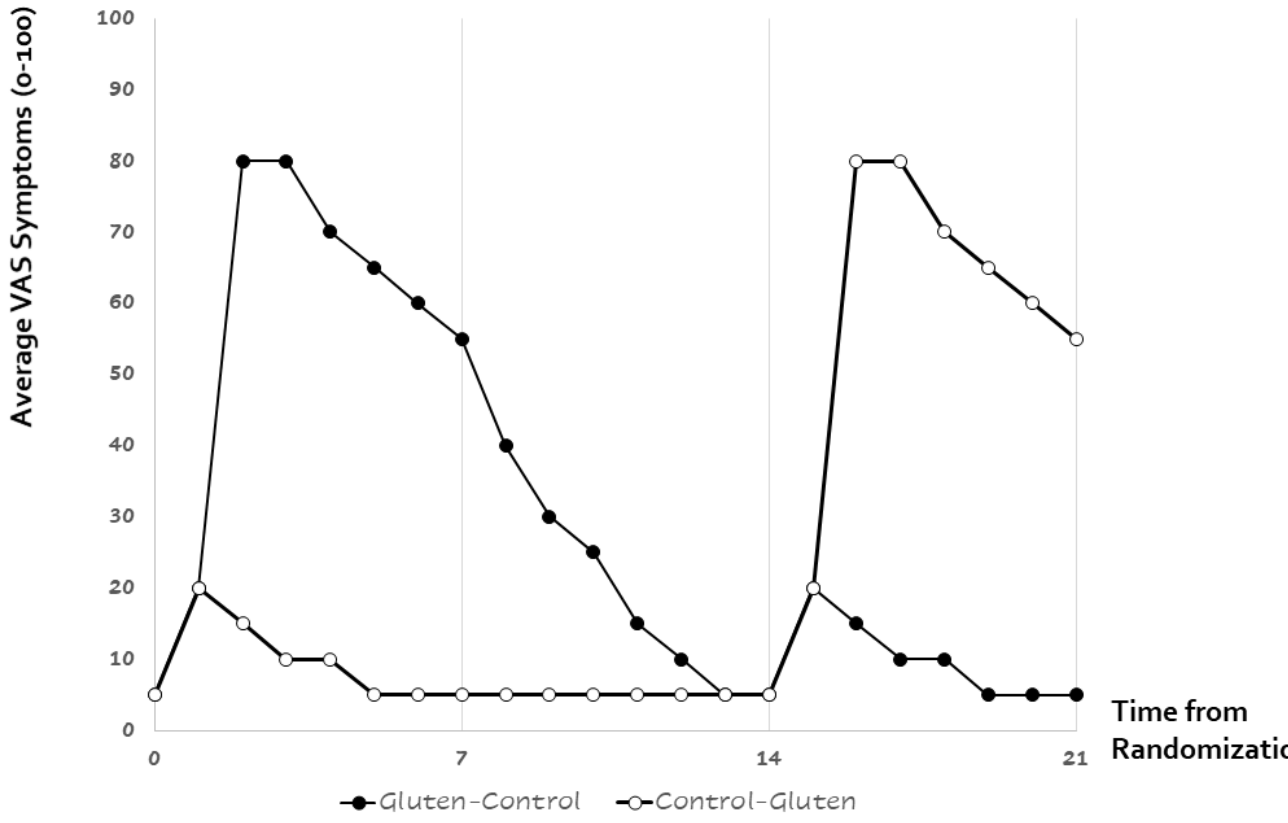


Table 2. Summary results for each study group and period

	Period 1		Period 2	Period 3		Statistical Tests: Fixed effects		
	Gluten	Control	Wash-out	Gluten	Control	Period*Group	Period	Group
Primary Outcome								
Average VAS, 0-100mm								
Key Secondary Outcomes								
SF36 Mental Component Score, score: 0-100								
SF36 Physical Component Score, score: 0-100[g]								
WMHs, score: 14-70								
Other Secondary Outcomes								
Gluten intake, (n%)								
VAS_1, 0-100mm								
VAS_2, 0-100mm								
VAS_3, 0-100mm								
VAS_4, 0-100mm								
VAS_5, 0-100mm								
VAS_6, 0-100mm								
VAS_7, 0-100mm								
VAS_8, 0-100mm								
VAS_9, 0-100mm								
VAS_{10}, 0-100mm								

*Values are Least Squares Means (SE) unless otherwise stated.
Statistical tests are based on Repeated-Measures Linear Mixed Effects Model (participants modeled as a random effects variable).

Table 3. Summary results for each study group and period

	Gluten	Control	Difference between groups (95%Confidence Interval)	P-value
Primary Outcome				
Average VAS, 0-100mm				
Key Secondary Outcomes				
SF36 Mental Component Score, score: 0-100				
SF36 Physical Component Score, score: 0-100[g]				
WMHs, score: 14-70				
Other Secondary Outcomes				
Gluten intake, (n%)				
VAS_1, 0-100mm				
VAS_2, 0-100mm				
VAS_3, 0-100mm				
VAS_4, 0-100mm				
VAS_5, 0-100mm				
VAS_6, 0-100mm				
VAS_7, 0-100mm				
VAS_8, 0-100mm				
VAS_9, 0-100mm				
VAS_{10}, 0-100mm				

*Values are Least Squares Means (SE) unless otherwise stated. Statistical test are based on Repeated-Measures Linear Mixed Effects Model (participants modeled as a random effects variable)

Appendix table 1. Sensitivity analysis using a single-step non-responder imputation

	Gluten	Control	Difference between groups (95%Confidence Interval)	P-value
Primary Outcome				
Average VAS, 0-100mm				
Key Secondary Outcomes				
SF36 Mental Component Score, score: 0-100				
SF36 Physical Component Score, score: 0-100[g]				
WMHs, score: 14-70				

*Values are Least Squares Means (SE) unless otherwise stated. Statistical test are based on Repeated-Measures Linear Mixed Effects Model (participants modeled as a random effects variable)

REFERENCES

1. Catassi C, Elli L, Bonaz B, Bouma G, Carroccio A, Castillejo G, et al. Diagnosis of Non-Celiac Gluten Sensitivity (NCGS): The Salerno Experts' Criteria. Nutrients. 2015;7(6):4966-77.
2. Husby S, Murray J. Non-celiac gluten hypersensitivity: What is all the fuss about? F1000Prime Rep. 2015;7:54.
3. Carroccio A, Mansueto P, Iacono G, Soresi M, D'Alcamo A, Cavataio F, et al. Non-celiac wheat sensitivity diagnosed by double-blind placebo-controlled challenge: exploring a new clinical entity. Am J Gastroenterol. 2012;107(12):1898-906; quiz 907.
4. Francavilla R, Cristofori F, Castellaneta S, Polloni C, Albano V, Dellatte S, et al. Clinical, serologic, and histologic features of gluten sensitivity in children. J Pediatr. 2014;164(3):463-7 e1.
5. Detry MA, Lewis RJ. The intention-to-treat principle: how to assess the true effect of choosing a medical treatment. JAMA. 2014;312(1):85-6.
6. Garcia-Dominguez JM, Martinez-Gines ML, Carmona O, Caminero AB, Prefasi D, Maurino J, et al. Measuring burden in caregivers of people with multiple sclerosis: psychometric properties of the CSI questionnaire. Patient Prefer Adherence. 2019;13:101-6.
7. White IR, Horton NJ, Carpenter J, Pocock SJ. Strategy for intention to treat analysis in randomised trials with missing outcome data. BMJ. 2011;342:d40.
8. Tinggaard J, Aksglaede L, Sorensen K, Mouritsen A, Wohlfahrt-Veje C, Hagen CP, et al. The 2014 Danish references from birth to 20 years for height, weight and body mass index. Acta Paediatrica. 2014;103(2):214-24.