JNM
J Neurogastroenterol Motil, Vol. 27 No. 4 October, 2021
pISSN: 2093-0879 eISSN: 2093-0887
https://doi.org/10.5056/jnm20179
Journal of Neurogastroenterology and Motility

Original Article

Can We Rely on the Rome IV Questionnaire to Diagnose Children With Functional Gastrointestinal Disorders?

Desiree F Baaleman,[1,2]* Carlos A Velasco-Benítez,[3,4] Laura M Méndez-Guzmán,[4] Marc A Benninga,[1] and Miguel Saps[5]

[1]Emma Children's Hospital, Amsterdam UMC, University of Amsterdam, Pediatric Gastroenterology, Amsterdam, the Netherlands; [2]Amsterdam UMC, University of Amsterdam, Gastroenterology and Hepatology, Amsterdam Gastroenterology Endocrinology Metabolism, Meibergdreef 9, Amsterdam, the Netherlands; [3]Program in Clinical Medicine and Public Health, University of Granada, Spain; [4]Universidad del Valle, Cali, Colombia; and [5]University of Miami, Miami, FL, USA

Background/Aims
To investigate the intra-rater (test-retest) reliability of the diagnosis of functional gastrointestinal disorders (FGIDs) as measured by the Questionnaire on Pediatric Functional Gastrointestinal Disorders, Rome IV version (QPGS-IV) in children.

Methods
A prospective cohort study was conducted in a public school in Cali, Colombia. Children and adolescents between 11 and 18 years of age were given the self-report Spanish version of the QPGS-IV at day 0 (baseline) and at day 2 (48 hours later).

Results
The study protocol was completed by 215 children, of which 97 (45%) were excluded from analysis due to the inability to follow the questionnaire's instructions. The final analysis included data of 118 children (mean age 15.0 ± SD 1.8 years old, 58.5% boys). The most common diagnoses were functional dyspepsia, functional constipation, and irritable bowel syndrome. We found a moderate intra-rater reliability (κ = 0.61-0.65) for diagnosing an FGID in general, a functional abdominal pain disorder, and the diagnosis of functional dyspepsia. We found a weak intra-rater reliability (κ = 0.46-0.54) for diagnosing a functional defecation disorder, functional constipation, irritable bowel syndrome, and the postprandial distress syndrome subtype of functional dyspepsia.

Conclusions
Our study shows that a large proportion of children cannot adequately complete the QPGS-IV and that the intra-rater reliability among those who did adequately follow the instructions is moderate. We advise to test the children's understanding of the instructions prior to completion of questionnaires and recommend to not rely exclusively on a self-reported questionnaire to select, recruit, or evaluate pediatric patients for FGIDs for research purposes.
(J Neurogastroenterol Motil 2021;27:626-631)

Key Words
Child; Functional gastrointestinal disorders; Prevalence; Test-retest reliability

Received: August 12, 2020 Revised: January 14, 2021 Accepted: March 8, 2021

*Correspondence: Desiree F Baaleman, MD
Emma Children's Hospital, Amsterdam UMC, University of Amsterdam, Pediatric Gastroenterology, Amsterdam, the Netherlands
Tel: +31-205662906, Fax: +31-205669181, E-mail: d.f.baaleman@amsterdamumc.nl

Introduction

Functional gastrointestinal disorders (FGIDs) are common clinical entities in children and adolescents.[1,2] There are no biological markers for FGIDs, and diagnosis relies on symptom-based criteria. In children older than 10 years of age, the Rome IV committee recommends self-report of symptoms to establish diagnosis and to assess the outcomes in clinical trials (patient-reported-outcome).[3] The Questionnaire on Pediatric Functional Gastrointestinal Disorders, Rome IV version (QPGS-IV) was adapted from the Questionnaire on Pediatric Gastrointestinal Symptoms by Walker et al[4] to facilitate diagnoses. The QPGS-IV has been used to estimate the prevalence of FGIDs in multiple studies in children.[1,5] Studies comparing the prevalence of FGIDs in children using the Rome III and Rome IV criteria have shown differences in prevalence predominantly in abdominal migraine, irritable bowel syndrome (IBS), and functional dyspepsia.[1,2] It is important to understand if those differences resulted from modifications of the Rome criteria or innate errors in measures. The results of all previous epidemiological studies were based on a single time survey per each respondent.[1,2] Relying on a one-time self-report measure does not account for the reliability or validity of the instrument nor the ability of the respondent to complete the survey. Therefore, we cannot definitively establish if a repeated measure could have resulted in a different response. This is particularly important in children, who may have poor recall of symptoms, limited understanding of questions or lose interest and get distracted during the completion of the survey. A common measure of reliability is the intra-rater reliability (consistency), in which the same subject is given the same measure on more than one occasion (test-retest-reliability). The objective of our study is to examine the intra-rater reliability for the diagnosis of an FGID in children and adolescents. In addition, we examined the intra-rater reliability for 2 major diagnostic groups and explored the intra-rater reliability for individual FGIDs.

Materials and Methods

We conducted a prospective school-based cohort study evaluating the intra-rater reliability of an FGID diagnoses as measured by the QPGS-IV at a public school in Cali, Colombia. A Spanish version of the QPGS-IV was already available to us, as we translated the questionnaire for a previous study according to the guidelines of the Rome foundation for translation and localization.[1,6] In short, 2 bilingual physicians provided reverse translation of the questionnaire. After initial translation, the questionnaire was adapted by a randomly selected focus group of 20 clinic patients from Cali with an equal division between sexes and with ages ranging from 8 to 18 years of age. The adapted version was then translated back into English and a member of the research team who was not involved in the translation reviewed the final version to assure fidelity with the original English version of the QPGS-IV. Informative material, a questionnaire covering the child's medical history, and consent and assent forms were sent to the homes of schoolchildren/adolescents between 11 and 18 years of age. The need for the second questionnaire was explained to the families in accordance with the aim of our study; we explained that we wanted to investigate if the children answered similarly 2 days apart. Children with reported organic gastrointestinal disorders (eg, gastritis, inflammatory bowel disease, and Hirschsprung disease), gastrointestinal complaints that could mimic FGIDs by causing abdominal pain, or comorbid painful conditions frequently associated with FGIDs were excluded. Children of families that were eligible and consented, completed the self-report Spanish version of the QPGS-IV in class on day 0 (baseline) and on day 2 (48 hours later). After completion of both questionnaires, children who did not follow instructions on the questionnaire due to misunderstanding of instructions, inappropriate reading comprehension, or not paying attention to the instructions, were excluded from analyses. Sections of the questionnaire instructed in bold letters that a specific answer to a previous question should prompt to skip a specific section. Children who failed to follow those instructions were considered unable to complete the questionnaire accurately.

Survey Administration

In each classroom and for each administration, 2 members of our research team distributed self-report paper surveys. Parents were not present in the classroom. The researchers provided instructions on completion of the survey without disclosing the objective of the study and repeated this again after 48 hours. As the accuracy of recall of symptoms of children has been questioned,[7,8] we chose a short interval between survey administrations to facilitate that the content of the responses would not change, as questions had a 30-day reference period. On the other hand, the interval was considered long enough to decrease the likelihood that a child would automatically repeat the answers on the second administration based on their recall of recent answers.

Measurements

The questionnaires were reviewed to assess prevalence of

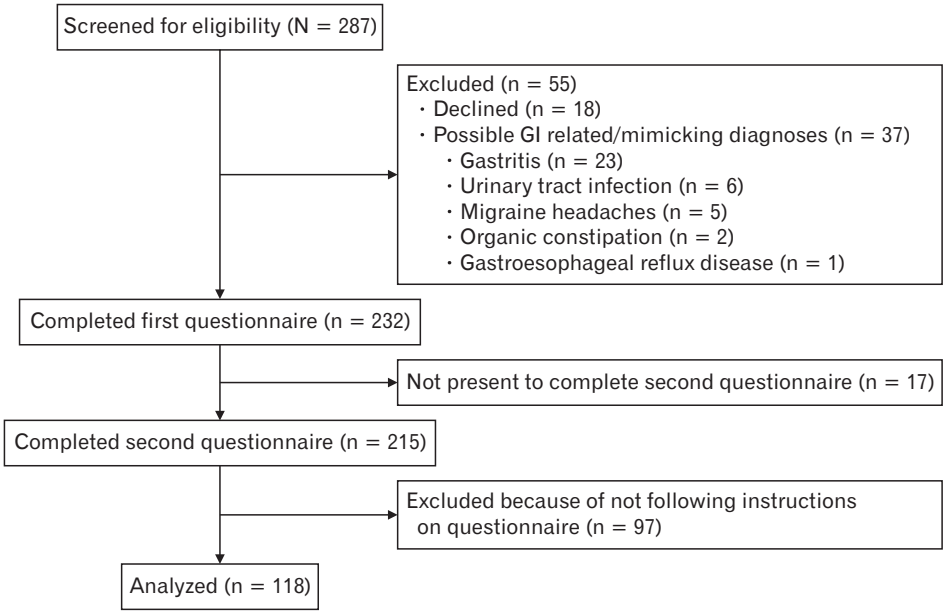

Figure. Patient flow chart. GI, gastrointestinal.

FGIDs and the intra-rater reliability of the diagnoses. The intra-rater reliability was analyzed for the presence of an FGID in general, for the presence of an FGID in one of the major diagnostic groups (functional abdominal pain disorders and functional defecation disorder), and for individual FGID diagnoses. The intra-rater reliability was not analyzed when less than 5.0% of the children were diagnosed with one particular disorder due to difficulties in interpreting the reliability measurements. With a high percentage of responses falling into a single nominal category (in this case "no") the percentage of agreement is by definition high and so is the expected/chance agreement (resulting in possible low kappa values which may not necessarily reflect low rates of overall agreement).[9]

Statistical Methods

The statistical analysis included measurements of central tendency (average, standard deviation, and percentage) and measurements reflecting the intra-rater reliability. The percentages of agreement and the coefficient of Cohen's kappa (κ) including 95% confidence intervals were calculated. Kappa values for agreement were interpreted according to the following magnitude guidelines: 0.00-0.20 (none), 0.21-0.39 (minimal), 0.40-0.59 (weak), 0.60-0.79 (moderate), 0.80-0.90 (strong), and > 0.90 (almost perfect).[10]

Sample size was calculated by using the previously reported prevalence rate for an FGID diagnosis of 21.2%.[1] Preliminary outcomes of intra-rater reliability within a 2-week interval by van Tilburg et al[11] showed chance-corrected agreements mostly in the range of 0.64-0.78. Because of our shorter time-interval we expected an average agreement rate of around 75.0%. The minimum acceptable level of agreement was set at 40.0%. In according with guidelines on the minimum sample size requirements for kappa agreement test we used a two-tailed test with 80.0% power at α = 0.05 for kappa statistics and calculated that a sample size of 53 subjects was needed to establish agreement between 2 ratings for the presence of an FGID diagnoses in general.[12] The sample size needed to measure the intra-rater reliability of the 2 largest diagnostic subgroups based on their reported prevalence (functional abdominal pain disorders 8.2% and functional defecation disorders 10.8%) was 117 subjects. As we did not know how many children were to be excluded, we oversampled to obtain enough data. Sample sizes calculations for the subgroup of nausea and vomiting disorders, and individual FGIDs exceeded our possible study sample.

The study was approved by school's teachers and principal and by the local Institutional Review Board (No. 024-2019).

Results

We invited 287 children to participate in the study (Figure). Parents of 269 (93.7%) children gave consent. Out of those children, 37 children (13.8%) were excluded from participation. After exclusions, 232 children completed the first questionnaire, 17 of them (7.3%) were not present to complete the second questionnaire. Both questionnaires were completed by 215 children, of which 97 (45.1%) were excluded from analysis due to not following instructions of the questionnaire. In conclusion, data of 118 children were

analyzed, mean age of 15.0 years (± SD 1.8 years), 58.5% were boys. Children who were excluded because of not following the instructions of the questionnaire had a mean age of 14.6 years (± 1.7 years) and 45.8% were boys, this was not different compared to the included children. Mean response time of the questionnaires was 12.0 ± 6.0 minutes.

Table 1. Prevalence of Functional Gastrointestinal Disorders Based on Rome IV Criteria on Day 0 and Day 2

FGIDs	FGIDs (N = 118)	
	Day 0	Day 2
Subjects with FGIDs	43 (36.4)	31 (26.3)
Subjects with 1 FGID	34 (28.8)	23 (19.5)
Subjects with 2 FGIDs	9 (7.6)	8 (6.8)
Functional nausea and vomiting disorders	5 (4.2)	5 (4.2)
Cyclic vomiting syndrome	0 (0.0)	0 (0.0)
Functional vomiting	2 (1.7)	1 (0.9)
Functional nausea	0 (0.0)	0 (0.0)
Rumination syndrome	0 (0.0)	1 (0.9)
Aerophagia	3 (2.5)	3 (2.5)
Functional abdominal pain disorders	30 (25.4)	22 (18.6)
Functional dyspepsia	19 (16.1)	14 (11.9)
Postprandial distress syndrome	17 (14.4)	9 (7.6)
Epigastric pain syndrome	2 (1.7)	5 (4.2)
IBS	7 (5.9)	7 (5.9)
IBS-constipation	1 (0.9)	3 (2.5)
IBS-diarrhea	0 (0.0)	0 (0.0)
BS-mixed	2 (1.7)	3 (2.5)
IBS-unclassified	4 (3.4)	1 (0.9)
Abdominal migraine	1 (0.9)	0 (0.0)
Functional abdominal pain-not otherwise specified	3 (2.5)	1 (0.9)
Functional defecation disorders	17 (14.4)	12 (10.1)
Functional constipation	16 (13.6)	11 (9.3)
Non retentive fecal incontinence	1 (0.9)	1 (0.9)

FGIDs, functional gastrointestinal disorders; IBS, irritable bowel syndrome.
Data are presented as n (%).

Prevalence of Functional Gastrointestinal Disorders

On the first day of testing, 43 children (36.4%) met the diagnostic criteria for at least one FGID (Table 1). The most common diagnosis was functional dyspepsia (19 children, 16.1%). Nine children (7.6%) met diagnostic criteria for 2 FGIDs; the most common combination was functional dyspepsia and functional constipation (5 children, 4.2%). On the second day of testing, 31 children (26.3%) met diagnostic criteria for at least 1 FGID. The most common diagnosis was again functional dyspepsia (14 children, 11.9%). Eight children (6.8%) met criteria for 2 FGIDs. The most common combination was functional dyspepsia and functional constipation (4 children, 3.4%). When comparing the results of the 2 testing days, the number of children meeting criteria for an FGID was higher on the first testing day (36.4% vs 26.3%), this difference was not statistically significant. On both testing days, functional dyspepsia was the most common FGID, followed by functional constipation and IBS. There were no significant differences between the prevalence of a diagnosis on testing day 1 and day 2.

Intra-rater Reliability

The reliability of the presence of an FGID diagnosis was considered moderate with an agreement of 83.1% and a kappa value of 0.61 (Table 2). The agreement between both testing days for the diagnostic group of functional abdominal pain disorders was moderate ($\kappa = 0.65$) and the agreement for the diagnostic group of functional defecation disorders was weak ($\kappa = 0.49$). We found

Table 2. Intra-rater Reliability for All Functional Gastrointestinal Disorders, Functional Gastrointestinal Disorder Subgroups and Individual Functional Gastrointestinal Disorders

FGIDs	Percentage agreement	κ	95% CI	Agreement strength
Subjects with FGIDs	83.1	0.61	0.43-0.79	Moderate
Functional nausea and vomiting disorders	-	-		Not interpretable[a]
Functional abdominal pain disorders	88.1	0.65	0.47-0.82	Moderate
Functional dyspepsia	90.7	0.61	0.44-0.79	Moderate
Postprandial distress syndrome	89.9	0.49	0.31-0.67	Weak
Irritable bowel syndrome	94.9	0.54	0.36-0.72	Weak
Functional defecation disorders	90.0	0.49	0.32-0.67	Weak
Functional constipation	90.0	0.46	0.28-0.63	Weak

[a]Due to low prevalence.
FGIDs, functional gastrointestinal disorders.

a moderate intra-rater reliability for the diagnosis of functional dyspepsia ($\kappa = 0.61$), and a weak intra-rater reliability for IBS ($\kappa = 0.54$), functional constipation ($\kappa = 0.46$), and the postprandial distress syndrome subtype of functional dyspepsia ($\kappa = 0.49$).

Discussion

This is the first study to measure the intra-rater reliability of a self-report questionnaire for the pediatric Rome IV criteria. We found that 45.1% of children were not able to follow the instructions on the questionnaire. Once those children were excluded, the intra-rater reliability of diagnosing a child with an FGID using the QPGS was moderate when corrected for chance ($\kappa = 0.61$). We found a moderate intra-rater reliability for the diagnostic group of functional abdominal pain disorders ($\kappa = 0.65$) and a weak intra-rater reliability for the diagnostic group of functional defecation disorders ($\kappa = 0.49$). Although we knew that we would not be able to obtain a sample size large enough to analyze each of the diagnoses with confidence, we thought the analysis was still worthy and could provide useful information for future studies. The found intra-rater reliability was moderate for diagnosing functional dyspepsia ($\kappa = 0.65$) and weak for diagnosing functional constipation ($\kappa = 0.46$), IBS ($\kappa = 0.54$), and postprandial distress syndrome ($\kappa = 0.49$). Functional dyspepsia, functional constipation, and IBS were consistently the most common diagnoses on both days of testing (day 0 and day 2).

Together, our findings speak of the utility of the use of questionnaires for diagnosing FGIDs while providing a cautionary note at the time of interpreting the results. Only 54.9% of the children in our population followed all the instructions of the questionnaire; that alone questions the reliability of the questionnaire. Adequate completion of the questionnaire seemed unrelated to the age of the participants. A thorough explanation of the questionnaire and coaching of the children throughout completion may help the understanding of the children, which could improve the reliability and the validity of the diagnoses. Indeed, higher intra-rater reliability rates of the previous version of the QPGS-III were found after face-to-face interviews with children.[13]

Alternatively, it is reasonable to consider completion of the questionnaires by the parents. A previous study by van Tilburg et al[11] looked into this alternative and studied the intra-rater reliability of 40 parents and 18 children. They compared the children's diagnoses according to the children and their parents, who completed the QPGS-III at baseline and after a 2-week follow up. They found that children had a moderate intra-rater reliability for most diagnoses and the intra-rater reliability of their parents were considerably lower for all diagnoses, except functional dyspepsia. These results are in line with our intra-rater reliability scores, and those of other studies which have shown that children do not have lower intra-rater reliability compared to their parents. This indicates that completion of the questionnaire by parents will likely not improve the reliability of the diagnoses.[14]

If the results of pediatric studies do not reach perfect agreement, even when parents are included, a question could be raised about whether the absence of perfect reliability is exclusively inherent to the pediatric Rome criteria or to the Rome criteria in general. A comparison with adult literature can help solve this conundrum. Palsson et al[15] assessed the intra-rater reliability of the adult Rome IV criteria as measured by the Rome IV Diagnostic Questionnaire at baseline and 20-40 days follow up in 140 adults. Compared to Palsson et al,[15] we found a higher reliability in diagnosing children with functional dyspepsia ($\kappa = 0.65$ vs $\kappa = 0.53$), and similar reliability in diagnosing children with IBS ($\kappa = 0.54$ vs $\kappa = 0.51$), and functional constipation ($\kappa = 0.46$ vs $\kappa = 0.44$). This implies that the limitations in reliability are not limited to the pediatric population.

Strengths of our study include the novelty of the study, the high participation rate (93.7%), the low dropout rate (12.6%), and the assessment of adequate completion of the questionnaire, an aspect that has not been previously studied in children completing questionnaires using the Rome criteria. However, multiple limitations should be considered. First, this study included children within a specific age range (11-18 years) located in 1 public school in Cali, Colombia, and the questionnaire was translated into a Spanish version. Therefore, the results cannot be generalized to all age groups, languages, or other geographic areas. Second, supplementary investigations for organic disease were not performed, and some of the diagnoses, therefore, may be inaccurate. Third, we used the exact same questionnaire to evaluate intra-rater reliability and a relatively short interval (48 hours), by this, some children may have remembered what they filled out the first time, which may have falsely increased our found levels of agreement. Fourth, we did not exclude children with 2 FGIDs from analysis (overlap) which may have caused an overestimation of the agreement in diagnosing an FGID in general. However, excluding those children would have underestimated the diagnosis of an FGID in general and could have resulted in inaccurate reliability outcomes for our other measures. Lastly, the current reported reliability outcomes of individual diagnosis have to be interpreted with caution and should be considered preliminary. However, the inclusion of these data may be valuable

for the conceptualization of the problem and to guide sample size calculations in future studies.

In conclusion, our study shows that the child-reported Spanish version of the QPGS-IV in a Colombian population has a moderate intra-rater reliability for an FGID diagnosis in general. We recommend not relying on this questionnaire exclusively to select, recruit, or evaluate children for research purposes. In addition, we advise to provide a thorough explanation of the questionnaire, and possibly use mock questions to assure the understanding of the questions. Larger studies are needed to investigate the accuracy, as well as the reliability and validity of the pediatric Rome IV criteria, to assess the reliability of low-prevalent diagnoses, and to compare with previous versions of the Rome criteria.

Financial support: None.

Conflicts of interest: None.

Author contributions: Carlos A Velasco-Benítez, Laura M Méndez-Guzmán, and Miguel Saps: design of the work; Carlos A Velasco-Benítez and Laura M Méndez-Guzmán: acquisition and analysis of data for the work; Desiree F Baaleman, Carlos A Velasco-Benítez, Marc A Benninga, and Miguel Saps: interpretation of data for the work; Desiree F Baaleman, Carlos A Velasco-Benítez, Laura M Méndez-Guzm, and Miguel Saps: drafted the initial manuscript; and Desiree F Baaleman, Marc A Benninga, and Miguel Saps: critically revised the manuscript for important intellectual content. All authors approve of the final version of the manuscript as submitted and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Saps M, Velasco-Benitez CA, Langshaw AH, Ramírez-Hernández CR. Prevalence of functional gastrointestinal disorders in children and adolescents: comparison between rome III and rome IV criteria. J Pediatr 2018;199:212-216.
2. Robin SG, Keller C, Zwiener R, et al. Prevalence of pediatric functional gastrointestinal disorders utilizing the rome IV criteria. J Pediatr 2018;195:134-139.
3. Hyams JS, Di Lorenzo C, Saps M, Shulman RJ, Staiano A, van Tilburg M. Childhood functional gastrointestinal disorders: child/adolescent. Gastroenterology 2016;150:1456-1468, e2.
4. Walker LS, Caplan-Dover A, Rasquin-Weber A. Manual for the questionnaire on pediatric gastrointestinal disorders. Nashville, TN: Department of Pediatrics, Vanderbilt University School of Medicine 2000.
5. Russo M, Strisciuglio C, Scarpato E, Bruzzese D, Casertano M, Staiano A. Functional chronic constipation: rome III criteria versus rome IV criteria. J Neurogastroenterol Motil 2019;25:123-128.
6. Rome Foundation, "Guidelines = rome translation project", https://theromefoundation.org/products/rome-translation-project/guidelines/, Published by the Rome Foundation (accessed 27 Sep 2021).
7. Chogle A, Sztainberg M, Bass L, et al. Accuracy of pain recall in children. J Pediatr Gastroenterol Nutr 2012;55:288-291.
8. van der Plas RN, Benninga MA, Redekop WK, Taminiau JA, Büller HA. How accurate is the recall of bowel habits in children with defaecation disorders? Eur J Pediatr 1997;156:178-181.
9. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005;37:360-363.
10. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012;22:276-282.
11. van Tilburg MA, Squires M, Blois-Martin N, Leiby A, Langseder A. Test of the child/adolescent rome III criteria: agreement with physician diagnosis and daily symptoms. Neurogastroenterol Motil 2013;25:302-e246.
12. Bujang MA, Baharum N. Guidelines of the minimum sample size requirements for Kappa agreement test. Epidemiol Biostat Public Health 2017;14:e12267.
13. Özgenç F, Akaslan Kara A, Demiral Yilmaz N, et al. Validity and reliability study of the pediatric rome III questionnaire for Turkish children and adolescents. Turk J Gastroenterol 2016;27:129-135.
14. Mellor D. Furthering the use of the strengths and difficulties questionnaire: reliability with younger child respondents. psychol Assess 2004;16:396-401.
15. Palsson OS, Whitehead WE, van Tilburg MA, et al. Development and validation of the rome IV diagnostic questionnaire for adults. Gastroenterol 2016;150:1481-1491.