

Starreveld scoring method in diagnosing childhood constipation

Fredericus T. Kokke · Judith S. Sittig · Annemiek de Bruijn · Tjeerd Wiersma · Rick R. Van Rijn · Jan L. Limpen · Roderick H. Houwen · Kathelijnn Fischer · Marc A. Benninga

Received: 21 December 2009 / Revised: 18 March 2010 / Accepted: 25 April 2010 / Published online: 1 July 2010
© The Author(s) 2010. This article is published with open access at Springerlink.com

Abstract

Background Four scoring methods exist to assess severity of fecal loading on plain abdominal radiographs in constipated patients (Barr-, Starreveld-, Blethyn- and Leech). So far, the Starreveld score was used only in adult patients.

Objective To determine accuracy and intra- and inter-observer agreement of the Starreveld scoring method in the diagnosis of functional constipation among pediatric patients. In

addition, we compared the Starreveld with the Barr scoring method.

Materials and methods Thirty-four constipated and 34 non-constipated children were included. Abdominal radiographs, obtained before treatment, were rated (Starreveld- and Barr) by 4 observers. A second observation after 4 weeks was done by 3 observers. Cut-off level for the Starreveld score, accuracy as measured by the area under the receiver operator characteristics curve, and inter- and intra-observer agreement were calculated.

Results Cut-off value for the Starreveld score was 10. AUC for Starreveld score was 0.54 and for Barr score 0.38, indicating poor discriminating power. Inter-observer agreement was 0.49–0.52 (Starreveld) and 0.44 (Barr), which is considered moderate. Intra-observer agreement was 0.52–0.71 (Starreveld) and 0.62–0.76 (Barr).

Conclusion The Starreveld scoring method to assess fecal loading on a plain abdominal radiograph is of limited value in the diagnosis of childhood constipation.

F. T. Kokke (✉)

Pediatric Gastroenterology, Wilhelmina Childrens Hospital, KE 1.144.3, PO Box 85090, Utrecht 3508 AB, The Netherlands
e-mail: f.t.m.kokke@umcutrecht.nl

J. S. Sittig · J. L. Limpen

Pediatrics Wilhelmina Childrens Hospital, Utrecht, The Netherlands

A. de Bruijn · T. Wiersma

Radiology, Rijnstate Hospital, Amhem, The Netherlands

R. H. Houwen

Pediatric Gastroenterology, Wilhelmina Childrens Hospital, Utrecht, The Netherlands

R. R. Van Rijn

Radiology, Academic Medical Center Amsterdam, Amsterdam, The Netherlands

K. Fischer

Epidemiology, Center for Health Sciences and Primary Care, Utrecht, The Netherlands

M. A. Benninga

Pediatric Gastroenterology, Emma Children's Hospital/Academic Medical Center, Amsterdam, The Netherlands

Keywords Constipation · Scoring · Abdominal radiograph · Child

Introduction

Constipation is a common gastrointestinal complaint in children with a prevalence ranging from 0.77% to 29.6% both in Western and non-Western countries [1]. The symptoms may vary from mild and short-lived to severe chronic constipation with faecal impaction and the involuntary loss of faeces. Medical history together with a thorough physical examination is generally sufficient for diagnosis and treatment of most children with constipation. However, many clinicians additionally order a plain abdominal radiograph to

assess the presence of retained stool or enlargement of the distal gastrointestinal tract to confirm the diagnosis. Others use this test to evaluate severity of constipation, to evaluate treatment or to convince parents that constipation is the cause of their child's complaints.

To date three scoring systems have been described to assess the severity of faecal loading using an abdominal radiograph in constipated children [2–4]. These papers described a good diagnostic accuracy, with more than 80% of the constipated and non-constipated patients identified correctly. When evaluated by others however, accuracy was lower with an area under the curve (AUC) in the receiver operator characteristics of 0.68 for the Leech method [5] 0.84 and 0.74, respectively, for the Barr and Blethyn scoring methods [6]. Another important parameter for the usefulness of these methods, intraobserver and interobserver agreement, was good to excellent in the original description of these methods [2–4]. Although some investigators could reproduce this for the Leech [7] and Barr scores [8], others could not, finding a much lower intra- and interobserver agreement [5, 6, 9].

Three scoring systems were specifically designed for and evaluated in children [2–4]. A fourth was only used in adults [10]. As this Starreveld scoring system might be applicable in children as well, we assessed the accuracy of this method in the diagnosis of functional constipation in children, as well as its intra- and interobserver agreement. Furthermore, we compared the performance of the Starreveld score with the Barr score, the oldest and most widely used method for evaluating constipation on a plain abdominal radiograph.

Materials and methods

Study population

Between September 2001 and April 2004 all children with functional constipation ages 7–12 years and referred by general practitioners and public health physicians to the outpatient clinic of a large teaching hospital (Hospital Rijnstate Arnhem the Netherlands) were eligible for this study. All children had to fulfil at least 2 out of 4 criteria of constipation: stool frequency <3 per week, ≥ 2 episodes of faecal incontinence per week, periodic passage of very large amounts of stool at least once every 7–30 days, or a palpable abdominal or rectal mass at physical examination [11]. Medical history, defecation frequency, faecal incontinence frequency, faecal consistency using the Bristol stool form scale and passage of a large amount of faeces were recorded in a standardized bowel diary [12]. Children with organic causes of constipation, including Hirschsprung disease, spina bifida, hypothyroidism, metabolic or renal abnormalities, mental retardation and children using drugs influencing

gastrointestinal function (laxatives or other medications), pre- or probiotics, or antibiotics in the previous 4 weeks before the first visit were excluded from the study.

Controls consisted of a group of children fulfilling the Rome II criteria for functional non-retentive faecal incontinence (FNRFI) and functional abdominal pain (FAP) [13].

Participation in the study was voluntary and written informed consent was obtained before the start of the study. The medical ethics committee of the hospital approved the protocol.

Abdominal radiography and scoring methods

Starreveld score

The Starreveld score quantifies the amount of faeces in four different bowel segments (ascending colon, transverse colon, descending colon and recto-sigmoid). For each bowel segment faecal stasis is scored as follows: no faeces (1), small amount of faeces (2), moderate faecal stasis (3), or severe faecal stasis (4). Therefore, the minimum score is 4 and maximum score is 16. A cut-off point at which the score is considered positive for constipation was not provided by Starreveld in his original paper [10].

Barr score

The Barr score quantifies the amount of faeces in four different bowel segments (ascending colon, transverse colon, descending colon and rectum) and also the quality of faeces, i. e. granular and rock-like faeces. Minimum score is 0 and maximum score is 22. A radiograph is considered positive for constipation when the score is ≥ 10 points [2].

Observers

Four observers, a medical student (JS), a resident radiologist in an academic medical centre (AdB), a senior radiologist in a large teaching hospital (TW) and a senior paediatric radiologist in an academic centre (RvR) independently scored the same abdominal radiographs in random order. The student was trained to apply the two scoring systems by a senior radiologist on two occasions. All observers were blinded to the patient characteristics. To assess intraobserver agreement, all abdominal radiographs were rated a second time by 3 of the 4 observers (JS, AdB, TW) after an interval of 4 weeks.

Statistical analysis

Nonparametric tests were used to compare general characteristics between children diagnosed with functional constipation

and the control group with FNRFI and FAP. Absence or presence of constipation was compared for different scores in both methods. For the Starreveld score the optimum cut-off value was determined by the lowest Youden index: i.e. sum of false-positives and false-negatives [14]. A cut-off of 10 was used for the Barr scoring method [2].

A receiver operator characteristic (ROC) plot was constructed for both the Starreveld and Barr scoring method. The area under the ROC curve (AUC) was used as a single indicator of diagnostic accuracy. The AUC can be interpreted as the probability that a randomly chosen case with functional constipation has a higher Starreveld or Barr score than a randomly chosen control (FNRFI or FAP). The perfect test has an AUC of 1.

Interobserver agreement was assessed by two-way intra-class correlation coefficient (ICC) for ordinal data on the first observation session for total Starreveld and Barr scores comparing the data from the four observers. Kappa and ICC were classified according to arbitrary cut-off values as poor (<0.20), fair (0.21–0.40), moderate (0.41–0.60), good (0.61–0.80) or very good (0.81–1.00) agreement.

Intraobserver agreement was calculated using Cohen’s K statistics for ordinal data comparing the data of the first and second observation from three observers.

Two statistical software packages were used: R statistics [(R Development Core Team 2006); R: A language and environment for statistical computing; R Foundation for Statistical Computing, Vienna, Austria; ISBN 3-900051-07-0, URL <http://www.R-project.org>] and specifically the separately downloaded *irr* Package (Version 0.62), and SPSS-PC v.17.0 (SPSS Inc., Chicago, IL, USA).

Results

Baseline characteristics

A total of 34 children fulfilling the criteria of childhood constipation were included and compared to 34 non-constipated children. Baseline characteristics of the two groups are summarized in Table 1. Significant differences between constipated children and controls were found with respect to defecation frequency, incontinence and abdominal and rectal scybala.

Performance of Starreveld and Barr scores

When applying the Youden index, an optimal cut-off level of ≥ 10 for the Starreveld score was found, using the mean score of the first observation of four observers. Youden index, positive and negative predictive value (PPV, NPV) according to the different Starreveld cut-off levels and the recommended Barr cut-off level of ≥ 10 are shown in Table 2. Using the optimal cut-off level for the Starreveld score, 23/34 constipated children were correctly labelled, while 14/34 non-constipated children were mislabelled. Using the cut-off level for the Barr score only 14/34 constipated children were correctly labelled, while 20/34 non-constipated children were mislabelled as constipated.

Similar results were found when computing a ROC curve. The AUC (using the mean of the scores from the first observation of all four observers) for Starreveld scoring method was 0.54 [95% confidence interval (CI): 0.40–0.68], only slightly above a result expected by chance, while the AUC for the Barr score was even worse (0.38,

Table 1 Clinical characteristics of the study and control groups. Values are mean (range) or numbers (%). *BSFS* Bristol stool form scale [12]

	Controls (n=34)	Constipation (n=34)	Significance
Characteristics			
Boys/girls	21/13	18/16	0.47
Median age in years (range)	9.0 (7–12)	9.0 (7–12)	0.08
Median duration of complaints in months (range)	18 (1–84)	12 (1–104)	0.27
Median defecation frequency / week (range)	7 (2–14)	3 (1–7)	<0.01
Median score stools (BSFS 1 or 2)	5 (15)	4 (12)	0.66
Painful defecation (%)	8/34 (24)	6/34 (18)	0.44
Abdominal pain (%)	20/34 (59)	27/34 (79)	0.13
Loening-Baucke criteria			
Stool frequency <3/week	5/34 (15)	25/34 (74)	<0.01
Incontinence ≥ 2 times/week (%)	19/34 (56)	9/34 (26)	0.02
Large amounts of stool (%)	3/34 (9)	8/34 (24)	0.28
Abdominal and / or rectal scybala (%)	0/34 (0)	26/34 (76)	<0.01

Table 2 Sensitivity, specificity, Youden index, PPV and NPV according to different cut-off values of the Starreveld score and the standard cut-off for the Barr score. *Youden index* [14] sum of false positives in the control group and false negatives in the constipated

Method (n=68)	Positive if \geq	Constipated radiological /clinical	Youden index	PPV	NPV
Starreveld	7	65/32	35	0.49	0.33
Starreveld	8	61/31	33	0.51	0.57
Starreveld	9	52/29	28	0.56	0.69
Starreveld	10	37/23	25	0.62	0.65
Starreveld	11	31/16	33	0.52	0.51
Starreveld	12	20/10	34	0.50	0.50
Barr	10	34/14	40	0.41	0.41

95% CI: 0.25–0.52; Fig. 1). Interestingly, the AUC obtained was significantly different between observers, with the highest AUC obtained of 0.72 and lowest of 0.28 for the Starreveld score and a highest vs. lowest AUC for the Barr score of 0.63 vs. 0.09. There was no correlation between experience in evaluating abdominal radiographs and the AUC obtained.

Finally, the interobserver agreement using the ICC was only moderate (Table 3) for all observers. Intraobserver agreement was moderate to good for the Starreveld score (Kappa range 0.52–0.71) and good for Barr score (Kappa range 0.62–0.76).

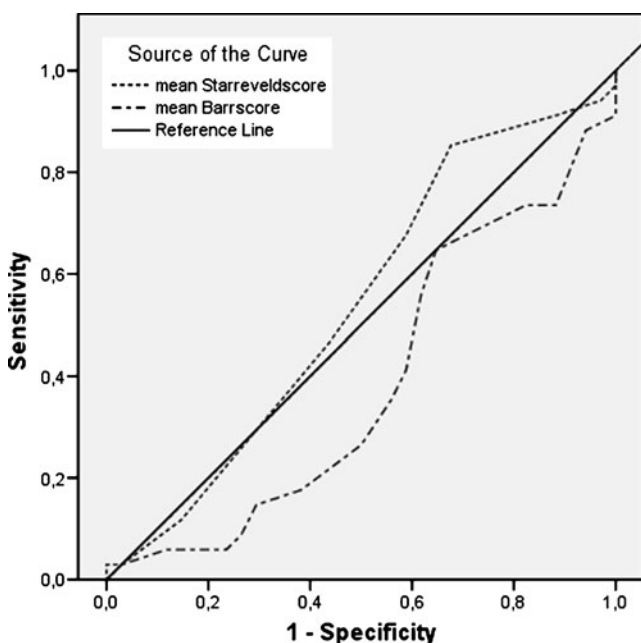


Fig. 1 ROC-curve for both mean Starreveld and mean Barr scores generated by 4 observers

group, the optimum being the lowest index; *PPV* proportion of children with positive test results who are correctly diagnosed with constipation; *NPV* proportion of children with negative test results who are correctly diagnosed as not being constipated

Discussion

In this study we show that both the Starreveld and the Barr scoring method for assessing faecal loading on a plain abdominal radiograph are of limited value in the diagnosis of paediatric constipation. Although the Starreveld score performed better than the Barr score, diagnostic discrimination of both methods was poor.

This study was conducted using strict criteria for constipation as described by Loening-Baucke [11]. For FAP and FNRFI the Rome II criteria were applied [13]. Similar control groups have been used by others [5]. However it cannot be excluded that in patients with functional abdominal pain and non-retentive faecal incontinence an overfilled colon is found more frequently than in the general population. A control group as used by Jackson et al. [6], consisting of patients with trauma, ureteric colic, insertion of a ventriculo-peritoneal drain or nonspecific abdominal pain might have given a better representation of the “normal” population.

Our results in children differ from those obtained by Starreveld in adults. While in the original study scores given by the four individual observers were highly significantly correlated, we obtained only a moderate interobserver agreement [10]. In addition, Starreveld

Table 3 Interobserver agreement according to observer and scoring method after a 4 week interval using ICC in a two-way model. Interpretation of agreement: poor (<0.20), fair (0.20–0.39), moderate (0.40–0.59), good (0.60–0.79) or very good (0.80–1.0)

	Observers	ICC	95% CI
Starreveld 1st	4	0.52	0.40–0.62
Starreveld 2nd	3	0.49	0.36–0.60
Barr 1st	4	0.44	0.24–0.59
Barr 2nd	3	0.44	0.31–0.56

described a significant correlation between the actual image as seen on the abdominal radiograph and defecation frequency. However, no controls were included, so the actual performance using a ROC curve could not be assessed. Our analysis actually showed a diagnostic accuracy which, with an AUC of 0.54, was only marginally above results that can be obtained by chance.

The other three scoring systems for evaluating constipation using an abdominal radiograph also had good sensitivity and specificity results in the original publications [2–4]. However, when in a subsequent evaluation a ROC curve was obtained, the AUC of the Leech score did not exceed 0.68 [5]. For the Barr and Blethyn scores the AUC obtained was 0.84 and 0.74 respectively, when scoring was done by an experienced radiologist, but lower when performed by a student or trainee [6]. Interestingly, in our study more experience did not result in an improved AUC. The best AUC, 0.72 for the Starreveld and 0.63 for the Barr score, was obtained by the student. This AUC, which is still far from ideal, is similar to values obtained by others for the Leech, Blethyn and Barr scores [5, 6].

In our study interobserver variability for both the Starreveld and Barr score was not good. Similar results were obtained by others for both Barr and Blethyn scores, although the Leech score performed unexpectedly well in another evaluation [5–7]. However, we and others found a good agreement between the two evaluations of the same observer at different time points [5, 7]. Obviously each observer develops their own interpretation of the original guidelines, resulting in considerable interobserver variability. However, each observer remains consistent in time given the acceptable intraobserver agreement.

Conclusion

The four scores developed for evaluating constipation using an abdominal radiograph did well on initial evaluation [2–4, 10]. However, on subsequent independent evaluation, both in the current study and in others, these good initial results could not be repeated [5, 6]. Given both the suboptimal AUC and the large interobserver variability

the abdominal radiograph should not be part of the routine work-up of childhood constipation.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

1. van den Berg MM, Benninga MA, Di Lorenzo C (2006) Epidemiology of childhood constipation: a systematic review. *Am J Gastroenterol* 101:2401–2409
2. Barr RG, Levine MD, Wilkinson RH et al (1979) Chronic and occult stool retention: a clinical tool for its evaluation in school-aged children. *Clin Pediatr (Phila)* 18:674–679
3. Blethyn AJ, Verrier Jones K, Newcombe R et al (1995) Radiological assessment of constipation. *Arch Dis Child* 73:532–533
4. Leech SC, McHugh K, Sullivan PB (1999) Evaluation of a method of assessing faecal loading on plain abdominal radiographs in children. *Pediatr Radiol* 29:255–258
5. de Lorijn F, van Rijn RR, Heijmans J et al (2006) The Leech method for diagnosing constipation: intra- and interobserver variability and accuracy. *Pediatr Radiol* 36:43–49
6. Jackson CR, Lee RE, Wylie AB et al (2009) Diagnostic accuracy of the Barr and Blethyn radiological scoring systems for childhood constipation assessed using colonic transit time as the gold standard. *Pediatr Radiol* 39:664–667
7. van den Bosch M, Graafmans D, Nievelstein R et al (2006) Systematic assessment of constipation on plain abdominal radiographs in children. *Pediatr Radiol* 36:224–226
8. Rockney RM, McQuade WH, Days AL (1995) The plain abdominal roentgenogram in the management of encopresis. *Arch Pediatr Adolesc Med* 149:623–627
9. Benninga MA, Buller HA, Staalman CR et al (1995) Defaecation disorders in children, colonic transit time versus the Barr-score. *Eur J Pediatr* 154:277–284
10. Starreveld JS, Pols MA, Van Wijk HJ et al (1990) The plain abdominal radiograph in the assessment of constipation. *Z Gastroenterol* 28:335–338
11. Loening-Baucke V (1990) Modulation of abnormal defecation dynamics by biofeedback treatment in chronically constipated children with encopresis. *J Pediatr* 116:214–222
12. Lewis SJ, Heaton KW (1997) Stool form scale as a useful guide to intestinal transit time. *Scand J Gastroenterol* 32:920–924
13. Rasquin-Weber A, Hyman PE, Cucchiara S et al (1999) Childhood functional gastrointestinal disorders. *Gut* 45(Suppl 2):II60–II68
14. Youden WJ (1950) Index for rating diagnostic tests. *Cancer* 3:32–35