Diagnostic Value of the Reflux Disease Questionnaire in General Practice

M.C. Aanen\textsuperscript{a} M.E. Numans\textsuperscript{b} B.L.A.M. Weusten\textsuperscript{c} A.J.P.M. Smout\textsuperscript{a}

\textsuperscript{a}Gastrointestinal Research Unit, \textsuperscript{b}Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, and \textsuperscript{c}Department of Gastroenterology, Sint Antonius Hospital, Nieuwegein, The Netherlands

Abstract

Objectives: This study determined the diagnostic and therapeutic response of the Reflux Disease Questionnaire (RDQ) using the symptom association probability (SAP) as reference. In addition, the RDQ's construct validity and its relationship to quality of life (QOL) were ascertained. Methods: Seventy-four patients with GORD symptoms (age 51 years (22–78); △ 62%) derived from primary care completed the RDQ, GRSRS and QOLRAD before and after a 2 weeks’ course of esomeprazole 40 mg daily. The SAP was determined by a 24-hour pH recording before PPI treatment. The diagnostic abilities of the RDQ (total and 4 dimensions scores) were assessed with the area under the curve (AUC) of a receiver operating curve. RDQ scores before and after PPI treatment were compared with Wilcoxon tests. Multiple linear regressions assessed the RDQ’s construct validity (GSRS) and relationship to QOL (QOLRAD). Results: The AUCs were low for all RDQ dimensions (AUC 0.6). In the SAP-positive patients all RDQ dimensions improved (p < 0.0001) while the scores of the SAP negatives did not (heartburn p < 0.01; GORD and total score p < 0.05; regurgitation and dyspepsia n.s.). The RDQ was related to the total and reflux GRSRS dimensions while the food and drink QOL dimension was linearly associated with the RDQ. Conclusions: The RDQ is a valid and reliable questionnaire with excellent construct validity and a good relationship to QOL. The diagnostic value of the RDQ in primary care is limited, but combination with an additional PPI treatment course might improve the RDQ's ability to discriminate GORD patients according to their SAP outcome.

Introduction

Gastro-esophageal reflux disease (GORD) is common in the western population [1]. Most GORD patients are diagnosed and treated in the primary care setting. Despite the fact that the disease itself is benign since the prevalence of complications and severe GORD-related morbidity is low, GORD can severely reduce a patient’s quality of life [2, 3]. This is why GORD treatment is usually started empirically in an early stage. Following treatment start, diagnostic procedures are postponed and a favorable response to proton pump inhibitors (PPI) is widely accepted to validate the diagnosis [4]. However, the increase in long-term use of PPIs without a proper diagnosis has urged researchers to look for other means to improve GORD diagnostics. One of the tools that might help the general practitioner (GP) is a validated GORD questionnaire to support diagnostic accuracy. It is likely that the use of a questionnaire would reduce the...
inter-observer variability in comparison with clinical history taking. Besides aiding in the diagnosis of GORD, a questionnaire would also enable the physician to quantify therapeutic response.

The Reflux Disease Questionnaire (RDQ) is a promising new questionnaire that was specially designed to be used in the primary care setting [5]. Extensive research has found this questionnaire to be reliable, valid, responsive and above all practical [5–7]. Furthermore, the RDQ outcome seems to correlate well with quality of life [6, 8]. However, data on its diagnostic validity is still lacking [7].

This validation should be performed in a population that represents patients in whom primary care physicians consider the diagnosis of GORD and in this population the most relevant diagnostic reference test for GP patients should be used, which, in our opinion, is a measure of the symptom-reflux association, such as the symptom association probability (SAP) [9]. The SAP objectively determines with a Fisher exact test whether symptoms are due to reflux events taking all symptom episodes and reflux events into account.

The aim of this study was therefore to assess the diagnostic and therapeutic response of the RDQ questionnaire using the SAP outcome as determined by 24-hour pH recording as reference standard in a primary care population. The secondary aim was to ascertain the construct validity of the questionnaire and to specify the RDQ’s relationship with quality of life.

Methods

Patients

Seventy-four patients (mean age 51 years (22–78); 62% male) who completed a 24-hour pH recording and exhibited symptoms during this recording were analyzed in this study. These patients were recruited directly during a GP consultation (34%) or indirectly by advertisement in a local newspaper (66%). In case of recruitment by advertisement the patient’s GP was consulted before inclusion. All patients had symptoms suggestive of reflux disease (i.e. heartburn, regurgitation, acid taste, burning sensation in the epigastric region, epigastric pain and chest pain) for at least twice a week for the past 3 months. The subjects had not used an acid-suppressant drug for at least 4 weeks before entry. Furthermore, none of the subjects had undergone gastrointestinal surgery.

Informed written consent was obtained before the start of the study and the protocol was approved by the Medical Ethics Committee of the University Medical Center, Utrecht.

Study Protocol

All patients with symptoms suggestive of GORD were asked to fill in the RDQ. In order to assess the RDQ’s construct validity and the RDQ’s assessment of quality of life, the Gastrointestinal Rating Scale (GERS) and Quality of Life in Reflux and Dyspepsia (QOLRAD) questionnaires were simultaneously completed with the RDQ 1 week before a 24-hour pH recording [5, 10, 11]. After the pH recording patients used 40 mg esomeprazole daily during 2 weeks. After these 2 weeks the patients completed the three questionnaires again. The SAP was determined after the pH recording but the patients and their physician were kept oblivious of the results until the protocol was completed.

24-Hour pH Monitoring

The 24-hour pH recording was performed after identification of the lower oesophageal sphincter (LOS) with manometry. The manometric recording was performed with a 10-channel silicone rubber catheter with a reversed sleeve sensor (DentSleeve International Ltd, Mississauga, Ont, Canada) which was perfused at a rate of 0.2 ml/min with degassed water, using hydraulic flow restrictors (DentSleeve International Ltd, Mississauga, Ont, Canada). The pressures were recorded with external pressure transducers (Abbott, Sligo, Ireland). The 24-hour pH recording was performed with a glass pH catheter with in-built reference electrode (Ingold, Urdorf, Switzerland) that was transnasally placed 5 cm above the LOS. The pH catheter was calibrated with 3.2 and 7.4 pH buffers solutions. The pH catheter was then attached to a digital datalogger (MMS, Enschede, the Netherlands) which used a sampling frequency of 2 Hz. All patients were instructed to record their symptoms by pressing the event marker button on the datalogger and at the same time specifying the symptom in a diary card. In the diary card also the times of consumption of meals and beverages and the recumbent time were noted. Patients were instructed to restrict their intake to 3 meals and 3 drinks during the 24 h at standardized times. Meals and drinks had to be consumed within 30 and 15 min, respectively. Patients were encouraged to maintain their normal daily activities during the 24-hour pH study.

After the 24-hour recording period the data from the datalogger was transferred to a personal computer.

Analysis of 24-Hour pH Data

The 24-hour pH data were analyzed automatically (MMS, Enschede, the Netherlands), excluding all eating and drinking periods. The SAP is calculated by dividing 24-hour pH data into consecutive 2-minute segments. For each of these 2-min segments, it is determined whether reflux occurred, providing the total number of two-minute segments with (total R+) and without (total R–) reflux. Then, for each symptom episode, it is determined whether reflux did (S+R+) or did not (S–R–) occur in the preceding two-minute period. Subtraction of S+R+ from total R+ results in S–R+ and subtraction of S+R– from total R– results in S–R–. A 2 × 2 contingency table is then constructed in which the number of 2-min segments with and without symptoms and with and without reflux are tabulated. Fisher’s exact test is used to calculate the probability (p) that the observed distribution could have been brought about by chance. The SAP is calculated as (1 – p) × 100%. By statistical convention, the SAP is considered positive if it exceeds 95% [9, 12].

Analysis of Questionnaires

The English version of the RDQ comprises 12 questions assessing the frequency and severity of heartburn, acid regurgitation and dyspeptic complaints which are scored on a 5-point Lik-
The Dutch version of RDQ which has been translated from English to Dutch and re-translated back for validity. The 12 items are combined into 3 dimensions: heartburn, regurgitation, dyspepsia. The mean of all three dimensions gives a total score ranging from 0 to 5. The specific GORD dimension is determined by the mean of the dimensions heartburn and regurgitation.

The QOLRAD, a disease-specific quality-of-life questionnaire, covers five dimensions: emotional distress, sleep disturbance, problems with eating and drinking (food and drink problems), limitations in physical and social functioning and lack of vitality. The Dutch QOLRAD is similar to the original English version. Responses were rated on a 7-grade Likert scale. The lower the score, the more severe the impact was on daily functioning during the past week. The QOLRAD has been shown to be reliable and valid [11, 13].

The GSRS includes 15 items combined into five symptom clusters addressing to what extent different gastrointestinal symptoms were bothersome in the past week. The five symptom clusters depict reflux, abdominal pain, indigestion, diarrhea and constipation. The GSRS has a seven-graded Likert type scale where 1 represents absence of bothersome symptoms and 7 very bothersome symptoms. All questions were translated into Dutch. The GSRS is documented to be reliable and valid [13, 14].

**Statistics**

In order to determine significant differences a Mann-Whitney test was used for unpaired data and a Wilcoxon test for paired data. A p < 0.05 was considered statistically significant. Data is presented in median and interquartiles (25–75th).

The ability of the RDQ to discriminate SAP-positive from SAP-negative patients was quantified by using the receiver operating curve (ROC) [15]. The area under the curve (AUC) denotes the discriminative power of a diagnostic model and can range from 0.5 (no discrimination, like flipping a coin) to 1.0 (perfect discrimination). A value of 0.7–0.8 is considered to represent a reasonable diagnostic test and a value of >0.8 represents a good discriminative diagnostic test [16]. In case of a good discriminative AUC, an optimal cut-off point can be determined at the largest angle of the curve, most closely related to the left upper corner of the figure.

To compare the RDQ dimensions with the pretreatment quality of life outcome of the QOLRAD and to assess the construct validity with the GSRS a stepwise multiple regression was performed. The 5 RDQ dimensions were subsequently chosen as the dependent factor. In a multiple linear regression model, adjusted R square (adjR²) measure the proportion of the variation in the dependent variable accounted for by the explanatory variables thereby making adjustments for the number of explanatory variables inserted into the model. The adjusted R squares can take on any value between 0 and 1, with a value closer to 1 indicating that a greater proportion of variance is accounted for by the model. In this study all adjusted R squares are shown in percentages.

**Results**

Similar pretreatment scores were found for the five RDQ dimensions that were studied. Also posttreatment the symptom scores of the five RDQ dimensions were also similar. However, PPI treatment significantly reduced all RDQ symptom scores (fig. 1).

Seventy percent of the studied subjects had a positive SAP. Patients with and without a positive SAP had similar pretreatment RDQ scores (table 1). On PPI treatment, patients with a positive SAP appeared to have a greater improvement of their RDQ scores: only in subjects with a positive SAP all RDQ dimensions scores were significantly reduced (table 1). Treatment decreased the total score significantly more in SAP-positive than in SAP-negative patients (p < 0.05).

The ROC analysis showed that all RDQ dimension scores had an AUC <0.6 with the SAP as reference stan-
Fig. 2. ROC curves of the total RDQ scores and the 4 RDQ dimensions versus the SAP outcome. The sensitivity represents the true-positive fraction and 1 – specificity the false-positive fraction at different cut-off points. All dimensions scores followed the broken reference line, indicating that every improvement in false-positive rate is matched by a corresponding decline in the false-negative rate corresponding with areas under the curves below 0.6 (table 2). Therefore, no diagnostic cut-off point could be identified.

Table 1. Pre- and posttreatment scores for all RDQ dimensions for SAP-positive and SAP-negative patients

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Pretreatment</th>
<th>Posttreatment</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>median 25–75th</td>
<td>median 25–75th</td>
<td></td>
</tr>
<tr>
<td>SAP-positive</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heartburn</td>
<td>1.75 1.50–3.44</td>
<td>0.00 0.00–1.50</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Regurgitation</td>
<td>1.75 0.94–2.81</td>
<td>0.25 0.00–1.50</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Dyspepsia</td>
<td>1.88 0.94–3.06</td>
<td>0.00 0.00–1.25</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>GORD</td>
<td>2.13 1.00–2.63</td>
<td>0.63 0.00–1.50</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Total score</td>
<td>1.83 1.33–2.88</td>
<td>0.42 0.00–1.50</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>SAP-negative</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heartburn</td>
<td>2.13 0.38–3.13</td>
<td>0.50 0.00–2.31</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Regurgitation</td>
<td>1.38 0.00–3.25</td>
<td>0.75 0.00–2.50</td>
<td>n.s.</td>
</tr>
<tr>
<td>Dyspepsia</td>
<td>1.50 0.25–3.13</td>
<td>0.75 0.00–2.75</td>
<td>n.s.</td>
</tr>
<tr>
<td>GORD</td>
<td>1.56 1.09–3.16</td>
<td>1.25 0.00–1.78</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Total score</td>
<td>1.67 0.88–3.04</td>
<td>1.00 0.42–1.96</td>
<td>&lt;0.05</td>
</tr>
</tbody>
</table>

p values relate to differences between pre- and posttreatment scores.

Table 2. Area under the curve (AUC) for all RDQ dimensions found during ROC analysis against the SAP outcome as reference standards

<table>
<thead>
<tr>
<th>Indices</th>
<th>Dimensions</th>
<th>AUC</th>
<th>SEM</th>
<th>95% asymptotic confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>lower</td>
</tr>
<tr>
<td>SAP</td>
<td>heartburn</td>
<td>0.51</td>
<td>0.08</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>regurgitation</td>
<td>0.54</td>
<td>0.08</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>dyspepsia</td>
<td>0.52</td>
<td>0.08</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>GORD</td>
<td>0.55</td>
<td>0.08</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>total score</td>
<td>0.56</td>
<td>0.08</td>
<td>0.40</td>
</tr>
</tbody>
</table>

The RDQ dimension scores from patients recruited by advertisement or by their GP showed similar AUC values both below 0.6 (data not shown). No RDQ cut-off value could be determined that would separate SAP-positive from SAP-negative patients (fig. 2).

The RDQ dimensions GORD and heartburn were found to be linearly related with the GSRS reflux score and they could be explained by the GSRS for 57% and 29%, respectively. Also regurgitation and the total RDQ score were positively related to GSRS reflux score (regurgitation adjR² 40% p < 0.001; total RDQ adjR² 44% p < 0.001). By adding the total score of bothersome gastroin-
testinal symptoms the regression model improved (regurgitation: reflux and total adjR² 47% p < 0.05; total RDQ: reflux and total adjR² 60% p = 0.01). The dyspepsia dimension was solely related to the total GSRS score (adjR² 34% p < 0.001).

Assessment of the relationship between the QOLRAD scores and the RDQ scores showed that the total RDQ score and the RDQ dimension GORD could both be explained for 40% and the RDQ dimension heartburn for 33% by the QOLRAD dimension food and drink problems (all 3 regressions p < 0.001). Regurgitation was also associated with problematic food and drink intake and the variance of this dimension could be explained for by 19% (p < 0.01). The dimension physical and social dysfunction accounted for an extra 5% elucidation of regurgitation (p < 0.05), but this was due to three outliers. After exclusion of these outliers regurgitation was only associated with food and drink problems (adjR² 17%, p < 0.001). The dyspepsia dimension did not relate to a specific QOLRAD domain but was linearly related to the total QOLRAD score (adjR² 16%, p < 0.001).

Discussion

The main findings of our study are: (1) The RDQ cannot be used to identify GORD patients as defined by a positive SAP outcome. The areas under curves were lower than 0.6 for all dimensions and subsequently no optimal cut-off value could be determined. (2) After PPI treatment all RDQ scores were significantly reduced in SAP-positive patients but not in SAP-negative patients and the reduction in total RDQ score after treatment was greater in SAP-positive than in SAP-negative patients. (3) The RDQ questionnaire is specific for reflux symptoms but not for other bothersome gastrointestinal symptoms as assessed by GSRS. (4) The RDQ relates to QOL, i.e. to the dimension food and drink problems.

The RDQ, assessing frequency and severity of reflux symptoms to facilitate the diagnosis of GORD in primary care, presently is the best-designed GORD-specific questionnaire, due to the fact that both expert opinions and patient’s interview analyses were used in its development [5, 7]. According to our study results, however, RDQ does not deliver the solution for misclassification problems in primary care. This finding needs further discussion. It has been shown that clinical history taking alone has a low specificity for diagnosing GORD and is riddled with a high interobserver variability, despite a high sensitivity [17]. Nowadays, most GPs have adopted the PPI test for diagnosing GORD, although this has also shown to yield a low specificity, despite a high sensitivity [17–19]. Due to the rebound effect of PPIs even more patients are false-positively labeled [20]. This leads to the conclusion that currently no adequate diagnostic test for GORD can be performed in primary care and thus, the vast majority of patients with suspected GORD symptoms have to be treated empirically. Subsequently, in most patients a diagnosis is never made, which may lead to a life-long treatment with a PPI without knowing whether this is the best treatment option.

Performing a more invasive and costly investigation in every patient is not possible in a primary care setting. Furthermore, due to the fact that no absolute gold standard exists, it is unclear which diagnostic test would be most useful. Endoscopy is irrelevant in a primary care population in which the majority of GORD patients do not have acid-induced oesophageal lesions [3, 17]. We believe that the occurrences of GORD symptoms are caused by reflux events. Calculation of the SAP determines whether a true relationship exists between symptoms and reflux events and is therefore a more relevant procedure than only an assessment of the percentage of time with oesophageal pH below 4 [9, 12]. For the evaluation of the diagnostic properties as well as the responsiveness to treatment of the RDQ questionnaire we have used the SAP outcome as the reference standard for presence of GORD. A questionnaire validated against the SAP that would demonstrate whether reflux symptoms of primary care patients are due to acidic reflux events would greatly improve GORD diagnosis in primary care [12].

Disappointingly however, the RDQ was unable to discriminate SAP-positive from SAP-negative patients, despite the fact that our study population was comprised of patients in whom a GP indeed might use a questionnaire and/or short-term PPI treatment to support diagnostic decisions [18, 21, 22]. Although our study population was ‘GORD-enriched’, this does not influence the characteristic test properties, i.e. sensitivity and specificity [23]. The observation that patients recruited from two different sources did not show different outcomes in AUC suggests that the performance of the RDQ is not due to issues related to patient selection.

Several studies have shown that the response to PPI treatment, when simply evaluated with symptom improvement or patient satisfaction scores, results in a large group of false-positively labeled patients [18, 19]. It is possible that the additional use of the RDQ would enable
physicians to improve the diagnostic test characteristic of the PPI test. Our data suggest that patients with a positive SAP had significantly better posttreatment RDQ scores than patients with a negative SAP while the reduction in total RDQ score was significantly greater in SAP-positive patients. However, these results only show differences between groups and may not be applicable to individual patients. Nevertheless, our results suggest that the RDQ could be used as an additional help in differentiating GORD from non-GORD in the relevant population of patients suspected of GORD in primary care, when short-term empirical treatment with PPI is considered. Further research is needed to determine whether the response of the RDQ scores to PPI treatment can indeed be used diagnostically.

The RDQ has been shown to have excellent construct validity for GORD which we confirmed with the GSRS questionnaire [5, 7]. All reflux-specific RDQ domains were strongly related to the reflux dimension of the GSRS as the reflux dimension could explain 50–30% of the variation for most RDQ domains. Concerning the QOL, most RDQ dimensions were related to problems encountered during eating and drinking (food and drink problems) indicating that the QOL of patients was diminished due to the fact that they had less appetite, avoided eating due to their reflux symptoms and that certain food items increased their symptoms. In contrast, other studies suggested that the RDQ correlates with all QOL dimensions of the QOLRAD. However, the highest degree of correlation was found also with food and drink problems [6, 8]. It is known that GORD patients experience the most problems with the consumption of food and drink and little for social and physical functioning [6, 24]. The total RDQ scores and 3 of the 4 RDQ dimensions scores (GORD, heartburn and regurgitation) could be explained for 20–40% by the specific QOLRAD dimension food and drink problems.

In conclusion, the RDQ is a valid and reliable questionnaire with excellent construct validity and a good relationship to quality of life. However, in our primary care population the diagnostic value of the RDQ with a positive SAP as reference was limited. An empirical short term PPI treatment course with pre- and post-treatment evaluation might improve the RDQ’s ability to discriminate between individuals with and without GORD.

Acknowledgement

We acknowledge Astrazeneca, the Netherlands, for their support.

References