The reliability of patients’ judgements of care in general practice: how many questions and patients are needed?

Michel Wensing, Cees van de Vleuten, Richard Grol, Albert Felling

Abstract

Objectives—To estimate the number of questions and patients that are needed to achieve reliable measurements of patients’ judgements of care in general practice.

Design—Sensitivity study, using generalisability theory and real data from surveys of patients.

Subjects—739 patients with chronic illness from 23 general practitioners in the Netherlands.

Main measures—The reliability coefficients of scores per patient and scores per general practitioner for patients’ judgements of nine dimensions of care in general practice.

Results—For most dimensions the reliability per patient was 0.80 or higher if three questions were used, but for the evaluation of the “organisation of appointments” and “premises” five questions had to be used. To reach a reliability coefficient of 0.80 per general practitioner three questions and 90 patients, or five questions and 60 patients, were needed for most dimensions. Even more patients or questions were needed for the dimensions “availability for emergencies” and “continuity”. A reliability of 0.70 per general practitioner could be achieved if three questions and 60 patients were used, except for availability for emergencies and premises, for which more patients or questions were required.

Conclusions—Surveys of patients can only provide reliable information if the samples of questions and patients are large enough. It is important to distinguish between the reliability of scores per patient and the reliability per care provider, as well as between different dimensions of care. The reliability per patient is good for most dimensions if three questions are used, but a good reliability per care provider requires more questions or patients.

Keywords: patient satisfaction; questionnaires; reliability

Introduction

Research into patients’ satisfaction has been stimulated by the rise of medical audit and quality assurance in health care in the past decade. Surveys of patients are regarded as an important means for involving patients’ perspectives in the assessment and improvement of care delivery. In this context the results of the surveys are used to assess the quality of care delivered by a specific care provider or practice and to provide feedback to care providers who are supposed to change practice routines in response to this feedback. In such cases it is very important that patients’ judgements of care have been shown to be valid and reliable.

In medical audit the opinions of individual patients, expressed as responses to questions, are aggregated per care provider. Aggregated scores—such as the percentage of patients that are satisfied—are treated as characteristics of the care provider. This has important consequences. The care provider is now the object of measurement or unit of analysis, not the individual patient: the conclusions refer to care providers, not to patients. This implies that the reliability of scores per patient is not crucial here, but the reliability of the aggregated scores. The problem of the level of analysis has hardly been considered in research into patients’ satisfaction. An analysis of 40 studies on patients’ satisfaction in primary care showed that most reported the reliability of the answers of individual patients. Only in three studies was the accuracy of aggregated scores given, whereas in 10 studies reliability coefficients related to individual scores were given. In the remaining 27 studies no information at all was given on the reliability or accuracy.

The reliability of the aggregated score depends on the number of questions in the questionnaire and on the number of patients that responded in the survey. The more questions and patients, the more reliable the measurement will be. For reasons of efficiency no more questions and patients should be used than are needed to achieve reliable measurements. On the other hand, too few questions or patients could lead to unreliable information. Unfortunately, insight into the number of questions and patients that are needed to achieve a reliable aggregated score per care provider is limited. In a study on patients’ ratings of residents’ humanistic qualities it was found that for a reliability coefficient of 0.80 per resident 50 patients are needed who should each answer 50 questions. A limitation of this study was that these predictions were based on two patients per physician.
The reliability of scores from a measurement can be expressed as the consistency of the attained scores when the measurement is repeated. As it is often difficult to repeat measurements, in many situations one measurement is used to estimate the reliability. Tests were split into multiple parallel parts which were used to assess the reliability from repeated measurements. In the classic test theory and generalisability theory the reliability is the proportion of variation in scores that can be contributed to true differences between the units of analysis and not to irrelevant, random effects errors or to other fixed factors. Measurements should have at least interval level, that is invariant for linear transformations.

Two types of reliability coefficients are distinguished. Relative coefficients indicate the extent to which units of analysis can be distinguished from each other. Absolute coefficients indicate the extent to which the position of the units of analysis can be estimated within the range.

**Model 1  Patients as units of analysis**
We assumed that both patients and questions were random samples from larger populations (random effects). No attention was paid to the fact that patients belong to a particular general practitioner so they were not nested within the general practitioner (see note). This model is used for the calculation of a generalisability coefficient which is comparable with coefficient a.

**SOURCES OF VARIATION**
(1) P patients
(2) I questions
(3) P*I3e interaction of patients and questions, error variation.

**RELATIVE RELIABILITY COEFFICIENT**
\[
P / \left( P + (P*I3e/n) \right)
\]
where n indicates the number of items.

This coefficient can be seen as the expected mean correlation between the questions in a particular dimension, that would be found if measurements were infinitely repeated with different random samples of questions in that dimension. By replacing n with other values than the actual sample size, projections of reliability coefficients are obtained from different measurement conditions.

**Model 2  General practitioners as units of analysis**
In this model patients are nested within general practitioners. Patients and questions are regarded as random samples for larger populations (random effects).

**SOURCES OF VARIATION**
(1) H general practitioners
(2) P:H patients nested within general practitioners
(3) I questions
(4) I*H interaction of questions and general practitioners
(5) I*P:H3e interaction of questions and patients nested within general practitioners, other interactions and error.

**RELATIVE RELIABILITY COEFFICIENT**
\[
H / \left( H + (P:H/n) + (I*P:H3e/n) \right)
\]
where n indicates the number of items and n the number of patients.

This coefficient is the expected mean correlation between the scores per general practitioner, if measurements were infinitely repeated with different random samples of patients and questions for that dimension.

**Note**
It can be argued that a nested model should be specified for the reliability per patient for a valid comparison with the reliability per general practitioner. However, estimations based on such a nested model gave similar results to model 1 (at maximum 0.01 difference in the estimation of reliability coefficients).

Questions were considered as random effects, as for each aspect of care many alternative questions can be formulated.

Furthermore, it is possible that reliability varies across different aspects of care that are evaluated by patients. For example, the evaluations of organisational aspects may have different reliability coefficients than the evaluations of interpersonal aspects of care. It is not known to what extent the reliability of scores per patient and the reliability of the aggregated score will differ in reality. Therefore we performed this study with real data from surveys of patients to estimate and predict the reliability of both individual and aggregated scores for different numbers of questions and patients and for different aspects of care.
The questionnaire focuses on several aspects of care, organised in nine dimensions that were found in statistical clustering (unpublished data). Most questions contained a six point answering scale ranging from poor, doubtful, fair, moderately good, good, to very good. For some aspects questions with a different answering scale were used as well, but these questions are not used in this study.

ANALYSIS

For the estimation of reliability coefficients we used the generalisability theory, which is an extension of the classic test theory. In a survey of patients this approach implies an analysis of the variation in patients' answers across different patients and different questions. The total variation is divided into different components. One such component is true variation: the variation which can be contributed to differences between the objects of measurements (units of analysis). Another component is error variation, which is the variation that cannot be explained by any of the factors studied. The proportion of true variation relative to the total variation, a figure between 0 and 1, is the reliability coefficient. In this study we consider a reliability coefficient of 0.95 as very good, 0.80 as good, and 0.70 as acceptable. As the number of relevant measurements increases, the reliability coefficient is higher. Reliability coefficients can be estimated for different hypothetical numbers of patients and questions, with real data used for the projections. In these decision studies different sources of variation and their relations can be specified.

PATIENT SAMPLE

For the analysis we used the computer program GENOVA, which requires a balanced data set (equal number of patients per general practitioner). The number of patients was very small for some general practitioners. Therefore the 23 general practitioners were selected for whom judgements of at least 15 patients were available. If more than 15 patients were available, a random sample of 15 was drawn. In this way a data set of 345 patients from 23 general practitioners was extracted. For the dimensions availability for emergencies and support less than 15 patients were available for some general practitioners. For these dimensions the estimations had to be based on eight and 14 patients, respectively, per general practitioner. For all analyses the original scores, expressed on the six point answering scale, were used.

RELIABILITY PER PATIENT

Next, for each dimension two approaches were taken in the estimation of reliability coefficients. In the first approach three components were distinguished in the variation of patients' judgements of care: (a) differences between patients, (b) differences between questions, (c) differences between a combination of patients and questions, and (d) error variation (which cannot be distinguished from the patient by question interaction term

Table 1 Patients' judgements of care in general practice (n=739)

<table>
<thead>
<tr>
<th>Dimensions (questions)</th>
<th>Mean (SD)</th>
<th>Poor</th>
<th>Fair</th>
<th>Good</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organisation of appointments (9)</td>
<td>4.7 (0.83)</td>
<td>2.3</td>
<td>25.0</td>
<td>72.7</td>
</tr>
<tr>
<td>Availability for emergencies (3)</td>
<td>4.4 (1.25)</td>
<td>12.9</td>
<td>25.7</td>
<td>61.4</td>
</tr>
<tr>
<td>Premises (3)</td>
<td>4.3 (1.11)</td>
<td>10.9</td>
<td>30.5</td>
<td>58.6</td>
</tr>
<tr>
<td>Continuity (4)</td>
<td>4.7 (1.00)</td>
<td>3.6</td>
<td>22.6</td>
<td>73.8</td>
</tr>
<tr>
<td>Cooperation (4)</td>
<td>4.4 (1.19)</td>
<td>9.7</td>
<td>29.7</td>
<td>60.6</td>
</tr>
<tr>
<td>Medical care (6)</td>
<td>4.6 (0.99)</td>
<td>3.6</td>
<td>28.2</td>
<td>68.2</td>
</tr>
<tr>
<td>Relation and communication (10)</td>
<td>4.7 (0.98)</td>
<td>3.4</td>
<td>26.8</td>
<td>69.8</td>
</tr>
<tr>
<td>Information and advice (6)</td>
<td>4.6 (1.01)</td>
<td>3.2</td>
<td>28.9</td>
<td>67.9</td>
</tr>
<tr>
<td>Support (6)</td>
<td>4.3 (1.26)</td>
<td>11.8</td>
<td>30.0</td>
<td>58.2</td>
</tr>
</tbody>
</table>

* Scores for the dimensions, expressed on the six point scale, were collapsed into three categories: 1-2.49 = poor; 2.5-4.5 = fair; 4.51-6 = good.
for technical reasons). The reliability coefficient is the ratio between \((a)\) and \((\alpha)\) and \((c)\) in which \((\alpha)\) is divided by the sample size of the questions of interest (box 1 for formulae). As the differences between questions \((b)\) was not included, only relative differences between patients are referred to. The resulting coefficients are therefore comparable with the coefficient \(\alpha\).

### RELIABILITY PER GENERAL PRACTITIONER

In the second approach the reliability of aggregated scores is estimated—that is, scores that are aggregated per general practitioner. In this approach it is acknowledged that patients belong to a specific general practitioner: they are nested within a general practitioner. The following components were distinguished in the variation: \((a)\) differences between general practitioners, \((b)\) differences between patients

<table>
<thead>
<tr>
<th>Questions (n)</th>
<th>1</th>
<th>3</th>
<th>5</th>
<th>7</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organisation of appointments</td>
<td>0.48</td>
<td>0.73</td>
<td>0.82</td>
<td>0.87</td>
<td>0.89</td>
</tr>
<tr>
<td>Availability for emergencies*</td>
<td>0.74</td>
<td>0.90</td>
<td>0.94</td>
<td>0.95</td>
<td>0.96</td>
</tr>
<tr>
<td>Premises</td>
<td>0.43</td>
<td>0.69</td>
<td>0.79</td>
<td>0.84</td>
<td>0.87</td>
</tr>
<tr>
<td>Continuity</td>
<td>0.61</td>
<td>0.82</td>
<td>0.89</td>
<td>0.92</td>
<td>0.96</td>
</tr>
<tr>
<td>Cooperation†</td>
<td>0.70</td>
<td>0.87</td>
<td>0.92</td>
<td>0.94</td>
<td>0.95</td>
</tr>
<tr>
<td>Medical care</td>
<td>0.74</td>
<td>0.89</td>
<td>0.93</td>
<td>0.95</td>
<td>0.96</td>
</tr>
<tr>
<td>Relation and communication</td>
<td>0.64</td>
<td>0.84</td>
<td>0.90</td>
<td>0.93</td>
<td>0.94</td>
</tr>
<tr>
<td>Information and advice</td>
<td>0.71</td>
<td>0.88</td>
<td>0.92</td>
<td>0.94</td>
<td>0.96</td>
</tr>
<tr>
<td>Support‡</td>
<td>0.83</td>
<td>0.94</td>
<td>0.96</td>
<td>0.97</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Table 2 Estimated reliability coefficients for scores per patient for different numbers of questions (simulation study)

Projections were based on 15 patients per general practitioner except for * based on sample of 8 per general practitioner; † based on sample of 12 per general practitioner; ‡ based on sample of 14 per general practitioner.

### RELIABILITY PER PATIENT

Table 2 shows the reliability coefficients of the scores per patient that were estimated. If only one question was used, the reliability was acceptable for five out of the nine dimensions: availability for emergencies, cooperation, medical care, information and advice, and support. The reliability was good for many dimensions if three questions were used. The exceptions were organisation of appointments and premises, where five questions were needed. There was a very good reliability in five out of the nine dimensions if nine questions were used.

### RELIABILITY PER GENERAL PRACTITIONER

Table 3 shows the reliability coefficients of the scores per general practitioner. In general the reliability of general practitioner was lower, especially for smaller samples of patients. For example, for availability for emergencies the reliability was 0.90 per patient and 0.45 per general practitioner if three questions and 30 patients were used. If one question and 30 patients were used, the reliability was only acceptable for the dimensions medical care and information and advice. If three questions and 30 patients were used, the reliability per general practitioner was also acceptable for cooperation as well as relation and communication. The remaining five dimensions had coefficients below 0.70 (organisation of appointments had a coefficient of 0.69 (box 3)). The general trend was that increases in the number of questions did not affect the reliability per general practitioner very much, whereas...
increases in the number of patients had stronger effects. If three questions and 60 patients were used, the reliability per general practitioner was acceptable for all dimensions except availability for emergencies and premises. For a reliability of 0.70 both dimensions required five questions and availability for emergencies required 90 respondents.

The number of questions and patients needed for good reliability differs between the dimensions. For medical care and information and advice only three questions and 60 patients were needed. A choice can be made between three questions and 90 patients, or five questions and 60 patients for the dimensions organisations of appointments, cooperation, relation and communication, and support (for some of these combinations the coefficients were 0.78 or 0.79). For the remaining dimensions even higher numbers were required for good reliability. For premises nine questions and 90 patients were needed, whereas for continuity seven questions and 90 patients were required. The reliability of the evaluations of the availability for emergencies remained below 0.80, even if nine questions and 90 patients were used.

So, three questions and 90 patients or five questions and 60 patients would provide good reliability for most dimensions. Finally, a very good reliability (a coefficient of at least 0.95) was not achieved for any dimension within the range of number of questions and patients that were analysed.

Discussion

The results of surveys of patients’ judgements of care should be reliable, but to save resources no more measurements should be made than necessary. This study provided insight into the number of questions and patients needed to achieve reliable scores. Increasing the number of questions may result in a good reliability of scores per patient, but not necessarily in a good reliability of scores per care provider. In medical audit the reliability of the scores per care provider is usually more relevant than the reliability of the scores per patient. It was estimated that a good reliability (0.80) of the aggregated score per general practitioner or practice can be achieved for most dimensions if three questions and 90 patients, or five questions and 60 patients, are used. For some dimensions, especially those concerning the organisation of care, more questions and patients are needed. Box 3 shows how the number of questions affects the reliability coefficients for the dimension organisation of care.

An important feature of this study is the assumption that patients are not independent with respect of the care that they have received from their general practitioner. For this reason a model was used in which patients were nested within general practitioners. This is a phenomenon that is often found in medical audit and quality assessment, in which there are often fewer measurements (patients, charts, consultations) than conclusions (care providers, institutions). Ignoring this phenomenon can result in serious bias.

This study focused on the technical aspects of reliability calculations. The question can be raised whether it is always possible to increase the number of patients and questions to achieve more reliable scores. Sixty or 90 patients for each general practitioner or practice is quite a lot, particularly if the expected non-response (20%–40%) is considered as well. In medical audit this sample size may be difficult to achieve, as the methods and procedures need to be feasible and acceptable for the care providers involved. Nevertheless, it is important to realise that the consequence of non-achievement is a lower reliability, so there is a strong need to be cautious in the interpretation of the results.

Finding more questions for a particular dimension of care (such as continuity) may be a challenge. On the one hand questions used for a particular aspect of care were seen as a random sample from all possible questions on that dimension. Differences in the content of the questions were ignored, as a factor analysis showed that all these questions refer to one particular dimension (unpublished data). On the other hand, we thought that questions should have a specific content which is different from the other questions. If questions are too similar, patients may feel patronised and care providers may find the information

In the original questionnaire nine questions were used for this dimension. The principal factor analysis provides factor loadings for each question, which indicate the importance of the question for the dimension. One way of selecting questions would be to choose the questions with the highest factor loadings. Other criteria, such as item response, skewness of the distribution, or the sensitivity to differences between general practitioners, are ignored here. The questions, their factor loadings, and corresponding reliability coefficients per patient and per general practitioner are shown if 60 patients responded. It can be seen, for example, that the estimated reliability coefficients per general practitioner were 0.80 if five questions (questions a to u and including e) were used.

<table>
<thead>
<tr>
<th>Factor loading</th>
<th>Reliability per patient</th>
<th>Coefficients per general practitioner (n=60 patients per general practitioner)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a Helpfulness of practice assistant in making convenient appointment</td>
<td>0.74</td>
<td>0.48</td>
</tr>
<tr>
<td>b Correctness of practice assistant</td>
<td>0.72</td>
<td>0.65</td>
</tr>
<tr>
<td>c Possibilities for getting an appointment quickly</td>
<td>0.71</td>
<td>0.73</td>
</tr>
<tr>
<td>d Telephone accessibility for making an appointment</td>
<td>0.68</td>
<td>0.79</td>
</tr>
<tr>
<td>e Convenience of consultation hours</td>
<td>0.68</td>
<td>0.82</td>
</tr>
<tr>
<td>f Convenience of hours for telephone consultation</td>
<td>0.68</td>
<td>0.85</td>
</tr>
<tr>
<td>g Clarity of information about practice organisation</td>
<td>0.66</td>
<td>0.87</td>
</tr>
<tr>
<td>h Accessibility of the general practitioner for telephone consultations</td>
<td>0.62</td>
<td>0.88</td>
</tr>
<tr>
<td>i Acceptability of waiting time at the latest visit</td>
<td>0.58</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Box 3 Selecting questions for the dimension organisation of appointments.
The reliability of patients' judgements of care in general practice

not specific enough. So the challenge was to find questions for each dimension that are different from each other, but still refer to this dimension.

This study had several limitations. The general practitioners who used the handing out procedure may have been responsible for some non-response, which could have introduced bias in the reliability coefficients per general practitioner. A specific problem is the low response rate for specific dimensions. Regression analyses (not reported) suggested that the response to each item is randomly distributed over different categories of patients, although chronically ill patients who visit the practice often tended to fill in more questions in the questionnaire. Furthermore, it is unclear whether the results can be generalised to other questionnaires for patients. For example, increasing the number of answering categories usually results in higher reliability coefficients.12

Preferably, the reliability is analysed for each questionnaire, with different approaches as shown in this article. It is important to use appropriate methodology in these situations instead of standard approaches that are common in other disciplines.

8 Brennan RL. Elements of generalizability theory. Iowa City: ACT, 1983.