

ORIGINAL ARTICLE

Data quality and confirmatory factor analysis of the Danish EUROPEP questionnaire on patient evaluation of general practice

PETER VEDSTED, INETA SOKOLOWSKI & HANNE N. HEJE

Research Unit for General Practice, Institute of Public Health, University of Aarhus, Denmark

Abstract

Objective. The Danish version of the 23-item EUROPEP questionnaire measuring patient evaluation of general practice has not been evaluated with regard to psychometric properties. This study aimed to assess data quality and internal consistency and to validate the proposed factorial structure. **Setting.** General practice in Denmark. **Subjects.** A total of 703 general practitioners (GPs). Some 83 480 questionnaires were distributed to consecutive patients aged 18 or more attending practice during the daytime. A total of 56 594 eligible patients responded (67.8%). **Main outcome measures.** Data quality (mean, median, item response, missing, floor and ceiling effects), internal consistency (Cronbach's alpha and average inter-item correlation), item-rest correlations. Model fit from confirmatory factor analysis (CFA). **Results.** The distribution was skewed to the left for almost all items with a small floor effect (0.1–9.3%) and a ceiling effect larger than 15% (18.6–56.3%). Item response was high. For seven items “not applicable/relevant” represented more than 10% of the answers. Internal consistency was good. Item-rest correlations were below 0.60 for three items, and four items had lower correlations with their own domain than with other domains. CFA showed that four domains were highly correlated and that model fit was good for two indices (TLI and SRMR), acceptable for one index (CFI), and poor for three indices (chi-squared, RMSEA and WRMR). **Conclusions.** This study revealed high ceiling effects, a few items with low item-rest correlation and low item discriminant validity, and an uncertain model fit. There seems to be a need for developing response categories to bring down the ceiling effect and it is also unclear how to use the proposed domains. Future research should focus on evaluating the factorial structure when ceiling effect has been lowered, on whether items should be deleted, and on assessing the unidimensionality of each domain.

Key Words: Denmark, factor analysis, family practice, patient evaluation, questionnaires

Patient evaluation of healthcare services contributes to the basis of quality development in the healthcare sector [1,2], which, in turn, presupposes the availability of scientifically sound instruments facilitating valid and relevant data collection [3,4].

In 1993, the EU BIOMED study EUROPEP was launched to develop an international patient evaluation questionnaire for general practice [5,6]. As in other European countries, a national project based on this questionnaire, entitled “Danish Patients Evaluate General Practice” (DanPEP), was launched in 2002. This Danish version of the questionnaire was introduced as a tool for assessing patients' evaluation of general practice and consequently psychometric issues arose. So far, the development had focused on aspects of especially

content and construct validity and some on criterion validity [5–10]. The questionnaire was developed so that each item provided information and not as sum-scales. However, the items were categorized into five qualitatively developed domains (doctor–patient relationship, medical care, information and support, organization of services, and accessibility).

Thus, the question remains whether these domains can be used to categorize the items and as sum-scales. Further, data quality and internal consistency must also be assessed [11–13].

The aim of this study was to assess data quality and internal consistency and to validate the factorial structure with five domains of the EUROPEP questionnaire.

The EUROPEP evaluation questionnaire was developed using a comprehensive approach comprising literature studies, patient interviews, and priority and validation studies. However, we lack knowledge about its psychometric performance.

- Data from more than 50 000 patient evaluations indicated good quality data, good internal consistency, and low floor effect.
- The ceiling effect was consistently very high.
- Two of six goodness-of-fit indices showed poor, two showed good, and one index showed acceptable fit of the questionnaire domains.
- There is a need to develop the response categories and the factorial structure of the questionnaire.

Material and methods

The questionnaire

The EUROPEP questionnaire was developed on the basis of a systematic literature review, patient interviews, and an empirical study designed to determine patient priorities [6–8]. A cross-national validation study of the 23-item questionnaire was performed before its introduction [9,10].

The patients were asked to evaluate their GP on a five-point scale (1–5) ranging from “poor” (1), through “acceptable” (3) to “excellent” (5) (2 and 4 had no text). Alternatively, patients could use the response option “Not applicable/relevant” [5]. All items were scored in the same direction. The EUROPEP questionnaire was not developed as a rating scale, but as a collection of several equal items (“indicator” variables [11]) measuring a construct.

GP and patient populations

From 2002 to 2005, 703 GPs participated in the DanPEP survey. Each GP personally distributed questionnaires to a number of consecutive patients aged 18 or more attending the surgery and able to read and understand Danish. The GPs were divided into three groups because other aims of the project were to study the effect of reminders and of using postal questionnaires to patients. Hence, 121 GPs each distributed 130 questionnaires to consecutive patients [14], 391 GPs each distributed 100 questionnaires to consecutive patients, whereas the questionnaires from 191 GPs were sent directly to 150 patients by the secretary. A total of 83 480 questionnaires were distributed. Non-responders from the two latter groups received a reminder from the DanPEP secretariat 3–5 weeks after the

consultation [9]. Patients completed the questionnaire at home and returned it to the DanPEP secretariat. A total of 56 594 eligible patients responded (67.8%). Questionnaires were produced and optically scanned using Cardiff TELEform.

Analyses

The analyses consisted of two parts: First, we assessed the data quality, internal consistency and correlations between items and domains and between domains. Second, we explored the five-domain structure using confirmatory factor analysis.

Analyses were performed with Stata 9 and factor analysis with Mplus 4.1 [15]. The analyses included questionnaires in which at least 50% (12+) of the items had been answered.

Data quality was assessed in terms of mean with standard deviation, median, percentage of missing data, number of “not applicable/relevant” answers and extent of ceiling and floor effects. Floor and ceiling effects between 1% and 15% were defined as optimal [16]. Internal consistency was assessed using Cronbach’s alpha and average inter-item correlation. Only answers on the 1–5-point scale were included. We defined an alpha of 0.80 as the lowest acceptable value [3,4]. In contrast to alpha, the average inter-item correlation is independent of the number of items and sample size when measuring internal consistency. We regarded an average inter-item correlation of at least 0.50 as good [12].

The correlation analyses assessed whether each item had a high correlation with the sum score of the rest of the scale (internal item convergence) and a higher correlation with the items in its own scale than with those of other scales (item discriminant validity) [3,17,18]. Correlations were fixed at a minimum of 0.60 to reflect a high level of internal convergence [11]. We defined a sufficient item discriminant validity as a correlation with the items in its own scale two standard deviations above that obtained with other scales (calculated as the 95% confidence intervals for the coefficients based on Fisher’s transformation [Stata, ci2-option]) [12].

The factorial structure was evaluated by confirmatory factor analysis (CFA) where items were analysed as categorical measures with a variance-adjusted weighted least-squares method (WLSMV) estimator. The objective of the CFA was to explore to what degree the correlations between the variables could be explained by the five domains (factors). Thus, we defined a basic model where an item was linked to its own domain (see Table I), with unspecified correlation between domains. A number of indices are available to assess the fit of a model

Table I. Data quality (mean value (mean), standard deviation (SD), median, item response, answers in the “Not applicable/relevant” category, missing answers, and answers in lowest (floor) and highest (ceiling)) and internal consistency (Cronbach’s alpha and average inter-item correlation for the domain).

	n	Mean ¹	SD	Median ¹	Item response	“Not relevant”	Missing	Floor	Ceiling
Doctor-patient relationship (average inter-item correlation =0.64, alpha =0.92)									
1	55877	4.16	0.91	4	99.7	0.3	0.4	1.3	43.5
2	55630	4.23	0.88	4	99.6	0.7	0.4	1.0	47.0
3	55390	4.22	0.88	4	99.6	1.1	0.4	1.0	45.9
4	52936	4.15	0.89	4	99.2	5.1	0.8	0.9	39.8
5	55841	4.31	0.85	5	99.6	0.3	0.4	0.8	51.6
6	43838	4.68	0.61	5	99.0	21.0	1.0	0.1	57.9
Medical care (average inter-item correlation =0.68, alpha =0.92)									
7	49908	4.03	0.94	4	99.2	10.5	0.8	1.5	33.0
8	47444	4.04	0.93	4	99.0	1.0	1.0	1.3	31.0
9	55260	4.27	0.89	5	99.4	1.1	0.6	1.0	49.9
10	54567	4.23	0.88	4	99.4	2.3	0.6	0.9	46.0
11	45339	4.05	1.09	4	99.1	18.5	0.9	2.8	36.1
Information and support (average inter-item correlation =0.72, alpha =0.91)									
12	52692	4.20	0.86	4	99.6	5.9	0.4	0.7	41.4
13	53574	4.10	0.94	4	99.7	4.4	0.3	1.2	39.7
14	43825	4.04	1.01	4	99.3	21.4	0.7	1.8	32.0
15	51870	4.16	0.84	4	99.4	7.2	0.6	0.6	37.5
Organization of services (average inter-item correlation =0.67, alpha =0.80)									
16	52404	4.05	0.96	4	99.6	6.4	0.4	1.5	37.0
17	38895	3.97	0.98	4	99.2	30.0	0.8	1.3	25.1
Accessibility (average inter-item correlation =0.52, alpha =0.86)									
18	53071	4.06	0.93	4	98.4	4.0	1.6	1.3	36.9
19	55449	3.92	1.10	4	99.1	0.5	0.9	3.7	38.2
20	54522	3.40	1.23	3	98.9	1.9	1.1	9.3	22.1
21	49547	3.56	1.20	4	98.9	10.8	1.1	6.3	24.1
22	55407	3.49	1.02	3	99.1	0.6	0.9	4.4	17.7
23	38168	4.22	0.94	4	98.6	30.7	1.4	1.1	33.8

Notes: Only questionnaires with at least 12 items answered were included. Each item began as follows: What is your opinion of the GP and/or general practice over the last 12 months with respect to . . .

¹Based on answers on answers to the categories 1–5 and excluding “Not applicable/not relevant”.

based on categorical data and we present the six indices which have been shown to be useful in assessing model fit [19–22].

1. Chi-squared goodness-of-fit statistic assesses the discrepancy between the sample and fitted covariance matrix (the null hypothesis is that the model fits the data. An insignificant test indicates good fit (p-value above 0.1). The chi-squared is extremely sensitive to sample size (about 200 cases) and in large samples it tends to result in a rejection of the model. For this reason, additional fit indices that are less sensitive are recommended (using the non-centrality parameter and taking into account sample size and degrees of freedom).
2. Comparative fit index (CFI) assesses fit relative to a null model and ranges from 0 to 1 with values of 0.90–0.95 indicating acceptable and over 0.95 good fit.
3. Tucker Lewis index (TLI) adjusts for the number of parameters of the model and is interpreted as CFI.

4. Root mean square error of approximation (RMSEA) expresses the lack of fit per degree of freedom of the model. Values are interpreted as follows: ≤ 0.05 indicates very good, > 0.05 – 0.08 good and ≥ 0.10 poor fit.
5. Standardized root mean square residual (SRMR) is the average of the differences between the observed and predicted correlations and has a range from 0 to 1. Values of less than 0.08 indicate good fit.
6. Weighted root mean square residual (WRMR) is also a residual based measure where values above 0.90 indicate poor fit.

Results

Data quality analysis

For all items except 20, 21, and 22 the distribution was skewed to the left (Table I). The item response was high with a small number of missing answers (0.3–1.7%). The response category “not applicable/relevant” was used by 0.3–30.7% of responders. For

Table II. Correlations between domains and correlations between items and (1) the rest of the items in its own scale (item-rest correlation) and (2) the other domains (Dim 1 to Dim 5).

		Item-rest ¹	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
Dim 1	Doctor–patient relationship		–	0.89	0.92	0.86	0.64
1	Making you feel you had time during consultations?	0.73	–	0.67	0.69	0.62	0.51
2	Interest in your personal situation?	0.82	–	0.72	0.75	0.67	0.47
3	Making it easy for you to tell him or her about your problems?	0.82	–	0.71	0.75	0.65	0.46
4	Involving you in decisions about your medical care?	0.77	–	0.72	0.75	0.66	0.48
5	Listening to you?	0.84	–	0.73	0.75	0.66	0.47
6	Keeping your records and data confidential?	0.55	–	0.53	0.55	0.48	0.37
Dim 2	Medical care		–	–	0.92	0.87	0.65
7	Quick relief of your symptoms?	0.78	0.66	–	0.70	0.63	0.50
8	Helping you feel well so that you can perform your normal daily activities?	0.79	0.66	–	0.70	0.62	0.49
9	Thoroughness?	0.81	0.77	–	0.77	0.68	0.49
10	Physical examination of you?	0.80	0.76	–	0.77	0.68	0.48
11	Offering you services for preventing diseases?	0.69	0.65	–	0.70	0.63	0.46
Dim 3	Information and support		–	–	–	0.92	0.65
12	Explaining the purpose of tests and treatments?	0.79	0.74	0.74	–	0.71	0.50
13	Telling you what you wanted to know about your symptoms and/or illness?	0.83	0.76	0.76	–	0.72	0.51
14	Help in dealing with emotional problems related to your health status?	0.80	0.76	0.75	–	0.71	0.49
15	Helping you understand the importance of following his or her advice?	0.76	0.74	0.75	–	0.70	0.52
Dim 4	Organization of services		–	–	–	–	0.66
16	Knowing what s/he had done or told you during previous contacts?	0.67	0.67	0.66	0.70	–	0.49
17	Preparing you for what to expect from specialist or hospital care?	0.67	0.68	0.71	0.76	–	0.52
Dim 5	Accessibility		–	–	–	–	–
18	The helpfulness of the staff (other than the doctor)?	0.58	0.44	0.44	0.46	0.44	–
19	Getting an appointment to suit you?	0.71	0.42	0.43	0.43	0.42	–
20	Getting through to the practice on the phone?	0.70	0.36	0.37	0.38	0.37	–
21	Being able to speak to the GP on the telephone?	0.71	0.41	0.42	0.43	0.42	–
22	Waiting time in the waiting room?	0.58	0.39	0.40	0.40	0.42	–
23	Providing quick services for urgent health problems?	0.61	0.56	0.58	0.58	0.54	–

Only questionnaires with at least 12 items answered were included.

¹Item-rest = item-rest correlation = correlation between the item and the sum of the other items in its own domain.

7 items (6, 7, 11, 14, 17, 21, and 23) this category represented more than 10% of the answers. Floor effect was small (range 0.1–9.3%) and all items had a ceiling effect larger than 15% (range 18.6–56.3%).

Cronbach's alpha ranged from 0.80 to 0.92, and average inter-item correlation ranged from 0.64 to 0.72.

Item-rest correlation (Table II) (internal item convergence) (range 0.55–0.84) was below 0.60 for three items (6, 18, and 22). For four items (6, 11, 16, and 17) the correlation with own domain (item discriminant validity) was lower than with other domains. Four domains (dim 1 to dim 4) were highly correlated (ranged 0.86–0.92). These domains had lower correlations with the "Accessibility" domain.

Confirmatory factor analysis

The CFA showed high factor loading for all items (Table III). The indices for model fit (Table IV) show that the model fits the data well for two indices (TLI and SRMR), acceptably fit for one index (CFI), and poorer for three indices (chi-squared, RMSEA, and WRMR).

Table III. Results of the confirmatory factor analysis showing the standardized factor loadings and standardized residuals for each item when modelled with its own domain.

		Standardized factor loading	Standardized residual variance
Doctor-patient relationship			
1	Making you feel you had time during consultations?	0.867	0.249
2	Interest in your personal situation?	0.930	0.135
3	Making it easy for you to tell him or her about your problems?	0.923	0.148
4	Involving you in decisions about your medical care?	0.904	0.182
5	Listening to you?	0.943	0.110
6	Keeping your records and data confidential?	0.782	0.389
Medical care			
7	Quick relief of your symptoms?	0.883	0.221
8	Helping you feel well so that you can perform your normal daily activities?	0.879	0.227
9	Thoroughness?	0.946	0.105
10	Physical examination of you?	0.935	0.126
11	Offering you services for preventing diseases?	0.844	0.288
Information and support			
12	Explaining the purpose of tests and treatments?	0.909	0.173
13	Telling you what you wanted to know about your symptoms and/or illness?	0.930	0.135
14	Help in dealing with emotional problems related to your health status?	0.919	0.155
15	Helping you understand the importance of following his or her advice?	0.908	0.175
Organization of services			
16	Knowing what s/he had done or told you during previous contacts?	0.871	0.241
17	Preparing you for what to expect from specialist or hospital care?	0.914	0.164
Accessibility			
18	The helpfulness of the staff (other than the doctor)?	0.788	0.379
19	Getting an appointment to suit you?	0.816	0.334
20	Getting through to the practice on the phone?	0.836	0.301
21	Being able to speak to the GP on the telephone?	0.868	0.246
22	Waiting time in the waiting room?	0.734	0.461
23	Providing quick services for urgent health problems?	0.934	0.128

Discussion

In agreement with other authors we found a skewed distribution for almost all items [23]. Consequently, the ceiling effect was high indicating that the full evaluation range was not captured [16], which may lower the responsiveness of the questionnaire [24]. This obviously calls for a change in the response categories as the ceiling effect was seen for all items. We saw a small number of missing responses, but for seven items more than 10% of respondents found the questions irrelevant or not applicable. These items represented situations where respondents were supposed to have experienced health problems and thus they were not relevant to all respondents in general practice. Answers to these questions may consequently be divided depending on which patients respond.

We found a few items with low item-rest correlation and higher correlations to other domains than their own. Thus, it is possible that some items should be assigned to another domain. We also saw high correlations between four of the five domains, which may be a result of cross-correlation. Still, we found a high internal consistency of the domains which

Table IV. Model statistics of the CFA of the Danish version of the EUROPEP questionnaire (n=20 072).

Statistics	Chi-squared	df	p	CFI ¹	TLI ¹	RMSEA	SRMR	WRMR
Model fit for basic model with five domains	19748.3	79	<0.0001	0.928	0.991	0.111	0.041	8.871

¹Number of free parameters = 33. Chi = squared = chi-squared goodness-of-fit with df = degrees of freedom and p = p-value (reference: $p \geq 0.1$). CFI = Comparative fit index (reference: 0.90–0.95 = acceptable, >0.95 = good). TLI = Tucker Lewis index (reference: 0.90–0.95 = acceptable, >0.95 = good). RMSEA = root mean square error of approximation (reference: ≤ 0.05 = good, ≥ 0.10 = poor fit). SRMR = standardized root mean square residual (reference: <0.08 = good fit). WRMR = weighted root mean square residual (reference: >0.90 = poor fit).

indicates that the items behave reasonably similarly within the domains. However, this would also be seen if there was a high cross-correlation between items from different domains.

The high correlations between domains seem to be confirmed by the CFA. The proposed model of the EUROPEP questionnaire did not fit the data for each goodness-of-fit measure. The chi-squared was expected to reject the model because of the high sample size. Even if we ignored this test, two indices showed good, two poor, and one acceptable fit. Thus, we are not able to determine the fit of the proposed factorial structure. We made additional analyses (not shown) with different values for correlations between the domains, but these models did not improve the results compared with the basic model. We also considered removing one or more items, but as seen in Table III, no item had a particularly low loading. The goodness-of-fit indices may, however, have been strongly affected by the highly skewed response distributions.

Strengths and weaknesses

We obtained a high response rate, minimizing the risk of selection bias due to dropout. The questionnaires were handed out personally by the GPs, which may have given the GPs the possibility of excluding some patients (e.g. those with the most negative attitudes) from the study. The patients were included when attending the surgery, and frequent attenders were thus more likely to be included. However, this selection method ensured that the patients were able to evaluate the GP, which in turn enabled them to better answer the questions. These issues of selection bias may have affected the actual scores, but are unlikely to have changed the factorial structure of the questionnaire. We have previously shown that although non-responders may evaluate their GPs more negatively, this would only result in a small change in the complete evaluation [14].

However, in addition to providing very high statistical precision the large sample size also led to narrow confidence intervals for correlation coefficients and affected Cronbach's alpha, which is

sample-size sensitive (there is no simple rule here, but the results should be treated with caution if alpha is below 0.85 in a large sample [11]).

Implications and future research

This study revealed problems with the Danish version of the EUROPEP questionnaire: high ceiling effects, some items with low item-rest correlation and low item discriminant validity, and an uncertain model fit of the proposed factorial structure. Consequently, there seems to be a need to improve the Danish version of the questionnaire, and the study especially emphasizes the need to develop new response categories to lower the ceiling effect. Subsequently, there also is a need to assess the factorial structure of the questionnaire again due to changed variances and to explore possible composite subscale scores and how the categorical response categories behave based on item response theory. Further, future research should focus on trying to provide an external anchor which would make it possible to analyse whether the indices (multi-item model) can be used without loss of information compared with the use of a single-item model, evaluate whether items should be deleted and finally assess the unidimensionality of each domain.

Acknowledgements

The authors would like to thank Ms Gitte Hove, MLISc, for data management. This part of the DanPEP project was funded by the Central Committee on Quality and Informatics in General Practice, the Ministry of the Interior and Health, and the regional committees on quality development in the counties of Aarhus, Frederiksborg, Funen, Southern Jutland, and Vejle and the municipalities of Copenhagen and Frederiksberg. They are grateful to the many patients and their general practitioners for their contributions to this study.

References

- [1] Wensing M, Baker R. Patient involvement in general practice care: a pragmatic framework. *Eur J Gen Pract* 2003;9:62–5.

- [2] Anden A, Andersson SO, Rudebeck CE. Concepts underlying outcome measures in studies of consultations in general practice. *Scand J Prim Health Care* 2006;24:218–23.
- [3] Sitzia J. How valid and reliable are patient satisfaction data? An analysis of 195 studies. *Int J Qual Health Care* 1999;11:319–28.
- [4] McDowell I, Newell C. *Measuring health: A guide to rating scales and questionnaires*. Oxford: Oxford University Press; 1996.
- [5] Wensing M, Mainz J, Grol R. A standardised instrument for patient evaluations of general practice care in Europe. *Eur J Gen Pract* 2000;6:82–7.
- [6] Grol R, Wensing M. Patients evaluate general/family practice. The EUROPEP instrument. EQUIP, WONCA Region Europe, 2000.
- [7] Wensing M, Mainz J, Ferreira PL, Hearnshaw H, Hjortdahl P, Olesen F, et al. General practice care and patients' priorities in Europe: An international comparison. *Health Policy* 1998;45:175–86.
- [8] Grol R, Wensing M, Mainz J, Ferreira P, Hearnshaw H, Hjortdahl P, et al. Patients' priorities with respect to general practice care: An international comparison. European Task Force on Patient Evaluations of General Practice (EUROPEP). *Fam Pract* 1999;16:4–11.
- [9] Grol R, Wensing M, Mainz J, Jung HP, Ferreira P, Hearnshaw H, et al. Patients in Europe evaluate general practice care: An international comparison. *Br J Gen Pract* 2000;50:882–7.
- [10] Mainz J, Vedsted P, Olesen F. How do the patients evaluate general practitioners? (in Danish). *Ugeskr Laeger* 2000;162:654–8.
- [11] Fayers PM, Machin D. *Quality of life: Assessment, analysis and interpretation*. Chichester: Wiley; 2000.
- [12] Streiner DL, Norman GR. *Health measurement scales. A practical guide to their development and use*. New York: Oxford University Press; 2003.
- [13] Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: Theory and application. *Am J Med* 2006;119:166.
- [14] Heje HN, Vedsted P, Olesen F. A cluster-randomized trial of the significance of a reminder procedure in a patient evaluation survey in general practice. *Int J Qual Health Care* 2006;18:232–7.
- [15] Muthén LK, Muthén BO. *Mplus user's guide*. Los Angeles: Muthén & Muthén; 2007.
- [16] McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: Are available health status surveys adequate? *Qual Life Res* 1995;4:293–307.
- [17] Foster SL, Cone JD. Validity issues in clinical assessment. *Psychol Assess* 1995;7:248–60.
- [18] Ware-JE J, Gandek B. Methods for testing data quality, scaling assumptions, and reliability: The IQOLA Project approach. *International quality of life assessment. J Clin Epidemiol* 1998;51:945–52.
- [19] McDonald RP, Ho MHR. Principles and practice in reporting structural equation analyses. *Psychological Methods* 2002;7:64–82.
- [20] Muthen B, Kaplan D. A comparison of some methodologies for the factor-analysis of non-normal Likert variables. *Br J Math Stat Psychol* 1985;38:171–89.
- [21] Bentler PM. Comparative fit indexes in structural models. *Psychol Bull* 1990;107:238–46.
- [22] Hu L-T, Bentler PM. Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multi-disciplinary Journal* 1999;6:1–55.
- [23] Milano M, Mola E, Collecchia G, Carlo AD, Giancane R, Visentin G, et al. Validation of the Italian version of the EUROPEP instrument for patient evaluation of general practice care. *Eur J Gen Pract* 2007;13:92–4.
- [24] De Vet HC, Terwee CB, Bouter LM. Current challenges in clinimetrics. *J Clin Epidemiol* 2003;56:1137–41.