Rating scale model for healthcare profiling of public providers in Naples

Nicola Lama[1], Giuseppe Signoriello[1], Daniela Nappo[2], Ciro Gallo[1]

[1] Department of Medicine and Public Health, Second University of Naples, Italy
[2] Department of Mathematics and Statistics, University of Naples "Federico II", Italy

Corresponding Author:
Nicola Lama
Department of Medicine and Public Health, Second University of Naples
Via Luciano Armanni 5, Napoli, 80138 - Italy
E-mail: nicola.lama@unina2.it

Summary

The aim of this study was to analyse the potential of the rating scale model (RSM) to evaluate a scale developed for measuring satisfaction with public healthcare providers across general practitioners (GPs). In a cross-sectional survey, a 4-item self-report questionnaire was administered to a sample of 108 consenting GPs working in the city of Naples. All the participants returned their questionnaire filled in. Statistically significant differences between RSM parameter estimates emerged, reflecting distinct profiles of satisfaction with the healthcare providers. The likelihood ratio test statistic and diagnostic statistics did not reveal critical violations of model assumptions. The survey method used in this study gave an excellent response rate. The psychometric properties of the 4-item scale analysed by the RSM provided preliminary evidence for profiling satisfaction with public healthcare.

KEY WORDS: *public health, evaluation studies, questionnaires/methods, quality of health care, Italy.*

Introduction

Evaluation of the effectiveness and efficiency of organisations and of the quality of the services they deliver is becoming a focus of rapidly increasing attention in several public sectors. This is particularly true of the Italian public healthcare system, now under growing pressure to justify its substantial tax-funded budget requests and expenditures. Customer satisfaction with publicly managed healthcare providers (PMHPs) operating within the Italian National Health Service (NHS) is an interesting topic and one of the different aspects that might be explored in this context. This can be done through surveys of PMHP customers, measuring specific items that contribute to the definition of overall satisfaction. Three broad categories of customers may be defined as possible targets of such surveys: citizen-patients, general practitioners (GPs) and PMHP staff. The members of the first category are the real customers (and, to an extent, the stakeholders) of public institutions. However, while surveys of citizen-patients would be suitable for evaluating tangible aspects like accommodation or waiting list management, these individuals are less equipped to convey competent judgements about the quality of the medical care delivered.

Therefore, in this study, we chose to focus on GPs because, unlike PMHP staff, GPs have access to the complete clinical history of their patients (citizen-patients) and also more direct dealings with, and therefore greater experience of, the different PMHPs. It was decided that overall satisfaction, because of its latent nature, would be assessed by summarising the surveyed GPs' responses to an ad hoc questionnaire. From this perspective, most of the appropriate methods proposed in the literature appear to be derived from the item response theory (IRT) (1, 2), originally developed in the field of psychometrics as a means of measuring people's abilities and attitudes by modelling data collected using tests and questionnaires.

The aim of this study was to analyse the potential of the rating scale model (RSM) (3) to explain the response pattern variability associated with the latent trait "global satisfaction", as observed in a cross-sectional survey of GPs in the city of Naples, Southern Italy.

Methods

Data were obtained through a self-report questionnaire administered to the members of the biggest GP cooperative group in Naples. These GPs are distributed over all the 10 districts of the city and provide care for more than 130,000 inhabitants (about 13% of the residents of Naples).

At their monthly administrative meeting in June 2006, all the GPs were informed of the study objective by the head of the cooperative group, who also illustrated the questionnaire. One of the authors (D.N.) handed out the questionnaires and then collected them in again, filled in, before the meeting ended (4). The GPs were asked to rate each item on the basis of their perceptions and expectations of the performances of all the 14 PMHPs in Naples (see Appendix, Table 1A): eight local hospitals, three larger hospitals considered to be of national importance, two medical schools and the National Cancer Institute (5). Since the aim of this study was not to publish a league table of the 14 PMHPs, the data collected are reported anonymously by randomly assigning each of them a number ranging from 1 to 14. In creating the questionnaire items, reference was made to the SERVQUAL model, developed by Parasuraman et al. (6) to measure customer satisfaction with a service. The questionnaire was then adapted in the light of the results of a study by Bertrand et al. (7) in which a focus group of GPs in Grenoble was interviewed to evaluate the quality of hospital healthcare. The following were identified as crucial areas: the organisation of patient care, the availability of hospital doctors, the flow of information, the speed of emergency admissions and patient-hospital doctor relationships. In this study, four conceptualised domains were considered: (i) Reliability, (ii) Responsiveness/Empathy, (iii) Tangibility, and (iv) Confidence. One specific item was then generated under each domain, the statement encapsulating the differ-

ent aspects considered. In particular, the first domain, (i) Reliability, covered organisational efficiency, the PMHP's management of appointments and the presence of an emergency department. (ii) Responsiveness/Empathy investigated the level of GP involvement in the decision-making process and the quality and formality/informality of the GP's relationships with the staff of the PMHP. (iii) Tangibility evaluated objective aspects like physical environment and facilities, equipment, personnel and waiting times. Finally, (iv) Confidence identified the level of the GPs' faith in the accuracy of diagnostic procedures undertaken and the efficacy of administered treatments, as perceived on the basis of their experience. The questionnaire items are set out in the Appendix, Table 2A.

Responses were graded on a Likert scale from zero to ten and coded for analysis in three categories: low (0-3), medium (4-6) and high (7-10) level of satisfaction.

Statistical analysis

The model used was Andrich's RSM (3), which extends the dichotomous Rasch model (8) to accommodate Likert scale data. This IRT model assumes unidimensionality, i.e., the items measure only one construct. Consequently, local independence of items is obtained such that the items are uncorrelated with each other, conditional on the latent trait. In this study the construct measured was global satisfaction. Assuming that N_V is the number of GPs, N_I is the number of items in the questionnaire and all items have an equal number of categories N_K, the RSM models the probability of the v-th GP responding with x category to the i-th item of the questionnaire as follows:

$$\log\left(\frac{P_{vi}(x)}{1-P_{vi}(x)}\right) = \sum_{k=0}^{x}\left[\theta_v - \left(\omega_k + \beta_i\right)\right] = \sum_{k=0}^{x}\left[\theta_v - \beta_{ik}\right]$$

$$v = 1\ldots N_V$$
$$i = 1\ldots N_I$$
$$x = 0\ldots N_K - 1$$

The model assumed that $P_{vi}(x)$ probability depends on the baseline level θ_v of the v-th GP's proclivity for

satisfaction and on the threshold ω_k at which two adjacent categories k and $k+1$ are equally probable adjusted by the i-th item critical satisfaction value, β_i, located on the latent variable scale. The β_{ik} parameters act as marks on the ruler of the latent variable, global satisfaction, against which the GP's proclivity level θ_v can be compared: the higher the value of a particular β_{ik} parameter, the higher will be the the v-th GP's baseline proclivity θ_v to opt for the $k+1$ rather than the k-th category.

Parameters were estimated by means of a conditional maximum likelihood approach (CML), assuming

that the GP's total raw score $r_v = \sum_{i=1}^{N_I} x_{vi}$ is a suffi-

cient statistic. Thus, by conditioning the likelihood onto r_v, the GP parameters θ_v can be neglected as nuisance parameters with the purpose of obtaining consistent estimates of item parameters β_{ik}. The estimates of θ_v can subsequently be obtained by conditioning their likelihood to the rv values and estimated β_{ik} values, as shown by Andersen (9, 10). Leaving aside the desirable properties of the estimators, CML stays close to the concept of specific objectivity proposed by Rasch (8, 9) which, in this study, would recommend that the measurement process should not be influenced by the specific characteristics of the surveyed GPs or by the peculiarities of the specific questionnaire adopted. This brings us to to the concept of the separability of item β_{ik} and the GP θ_v parameters, which do not need to be estimated simultaneously, as assumed by the CML approach. The goodness-of-fit was evaluated by means of the likelihood ratio test statistic proposed by Andersen (10). Model fit diagnostics based on standardised residuals were assessed by grouping data by GP (person-fit statistics) and by item (item-fit statistics) to isolate possible misfits.

This analysis was performed using the eRm package (11) for R software (12).

Results

All of the 108 GPs gave their consent to participate in the research (see Table 1): most of them were males (78.7%), aged 51-60 (82.4%) and had more than a thousand patients (81.5%) in their care. On av-

Table 1. Characteristics of the sample of 108 general practitioners.

Gender	
Female	23 (21.3%)
Male	85 (78.7%)
Age in years	
30-40	0 (0.0%)
41-50	16 (14.8%)
51-60	89 (82.4%)
61-70	3 (2.8%)
Number of patients	
501-1000	20 (18.5%)
>= 1000	88 (81.5%)
Years in practice	
Mean(SD)	24.24 (3.64)
Satisfaction with Reliability	
Low	233 (15.4%)
Medium	718 (47.5%)
High	455 (30.1%)
Missing	106 (7.0%)
Satisfaction with Responsiveness/Empathy	
Low	196 (13.0%)
Medium	724 (47.9%)
High	453 (30.0%)
Missing	139 (9.1%)
Satisfaction with Tangibility	
Low	232 (15.3%)
Medium	764 (50.5%)
High	389 (25.7%)
Missing	127 (8.5%)
Satisfaction with Confidence	
Low	212 (14.0%)
Medium	709 (46.9%)
High	465 (30.8%)
Missing	126 (8.3%)

erage, they had been practising as GPs for 24.2 years (SD = 3.6).

The distributions of the GPs' responses for each item are reported in Table 1. Overall responses in the low, medium and high satisfaction categories totalled 873 (14.5%), 2915 (48.2%) and 1762 (29.1%), respectively, while 498 (8.2%) of responses were missing. The missingness mechanism was assumed to be independent of response, since it was mostly attributable to GPs' lack of familiarity with some of the PMHPs (e.g. local hospitals distant from the GP's surgery). Thus, whenever a GP failed to provide more than two responses on a given PMHP, all his/her responses relating to that PMHP were exclud-

ed from the analysis. The remaining 167 (2.8%) missing values were imputed using the modal value of the relevant item within each subset of PMHP response values.

The estimates of the β_{ik} parameters were obtained by fitting the RSM model to the dataset of responses related to each PMHP. Figure 1, in which the Reliability item response category probability curves (13) are plotted for the model of PMHP no. 1, provides an example of the results given by the model. These curves describe the probability of responding in the low, medium or high satisfaction category as a function of the level of trait (global) satisfaction. The intersection points between two probability curves reflect the β_{ik} thresholds at which two adjacent categories are equally probable. The extension of the tails of the probability curves inform on the discrimination capacity of the item categories. Very similar plots were observed for the other items and PMHPs. The analysis for the four items and the fourteen PMHPs is reported in Figure 2, where both point and 95% confidence interval estimates of β_{ik} parameters are plotted. These estimates represent the thresholds between the low/medium and the medium/high satisfaction categories. The lower the β_{ik} threshold estimate, the greater the probability of recording a higher level of satisfaction, given the v-th GP's proclivity for satisfaction θ_v. Usually the two thresholds are consistent (i.e. both lower or higher than the average), but there are two situations in which they are not. In the first, they are close to each other (i.e. the low/medium
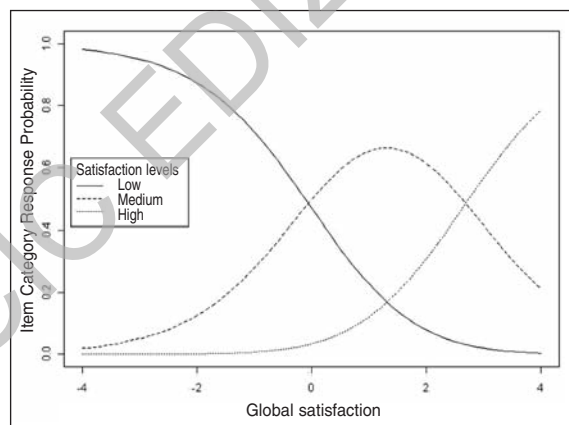

Figure 1. Reliability Item Category Response Probability curves under the Rating Scale Model of the publicly managed healthcare provider (PMHP) no. 1. Intersections between curves identify points of equi-probability between the related categories, measured by β_{ik} parameter estimates.

threshold is higher than the average, while the medium/high threshold is lower than the average), see e.g. PMHP no. 14 in the Confidence item (Fig. 2) which indicates a reduced probability of achieving a medium degree of satisfaction. In the second, when the two thresholds are distant, the reverse is true since, in this case, the medium level of satisfaction is the most probable response. The superimposed reference bars (dotted lines) indicate the inverse-variance-weighted average of β_{ik} parameter estimates.

In our study, PMHP no. 2 appears to be more satisfactory than average with respect to the Tangibility item since a lower GP's θ_v level is sufficient to observe a medium as opposed to low satisfaction response. On the contrary, PMHP no. 2 appears to be less satisfactory than average in the Empathy domain, since a far higher GP's θ_v level is required in order to observe a high (as opposed to a medium) or a medium (as opposed to a low) satisfaction response. PMHP no. 11 seems to be associated with a similarly low level of satisfaction on the Tangibility item. For example, the PMHP no. 14 seems to show a good medium/high satisfaction thresholds for all the investigated domains, with the exception of Confidence, where it exhibits a less satisfactory low/medium threshold. PMHP no.8 shows a good medium/high satisfaction threshold for Reliability and Confidence, while PMHP no. 6 reveals a poor medium/high satisfaction threshold for the Tangibility domain.

Table 2 reports the goodness-of-fit diagnostics of the RSM models. The LR test statistics (10) did not show evidence to reject the null hypothesis of goodness-of-fit for each investigated model. Person-fit and item-fit diagnostic statistics did not reveal critical violations of model assumptions. In particular, the number of statistical significant deviations (p < 0.05) shown by Person-fit diagnostics did not exceed the number of statistically significant results one might expect by chance. Similar results were observed on item-fit statistics with only one statistically significant p-value emerging i.e., that relating to the Empathy item for PMHP no. 14.

Discussion

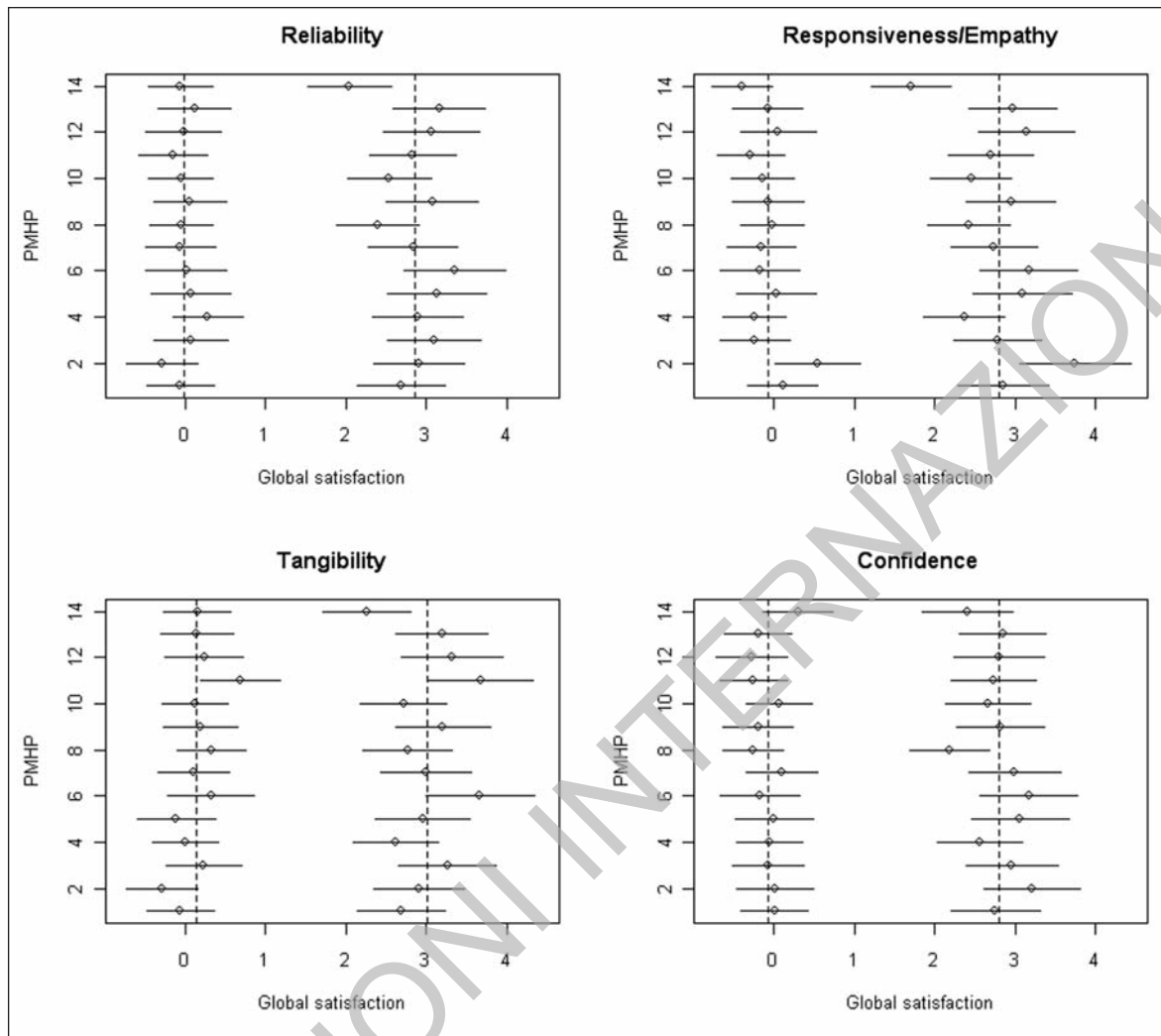The idea of considering GP perception of quality arose a response to the presence of contrasting results

Figure 2. Forest plot of point and interval estimates of β_{ik} for the 14 publicly managed healthcare providers (PMHP), located on the scale of the latent variable global satisfaction. Vertical dashed lines indicate the weighted averages of the estimates of β_{ik} using the inverse variance method.

and a lack of consensus on the definition of "the customer". Initial studies on this topic considered the level of citizens' satisfaction with their health care system (14, 15). In both these studies Italy showed the highest level of public discontent among the different countries surveyed. The EURISKO study surveyed 10,000 people and found that only 36% of the sample was at least fairly satisfied with the NHS (16, Chapter X, p. 212). Contrasting results were previously published in the WHO Report 2000 (17) where the interviewed healthcare experts ranked the Italian public health system second among the investigated 191 countries. The controversial issue of how to evaluate the performance of health systems was discussed

by Blendon et al. (18) who pointed out that the WHO Report relied on a survey of public health experts, many of whom did not reside in the countries they were rating and suggested that both public and expert views should be considered. Similar conclusions were reached by Bertrand et al. (7), who suggested including evaluations of GPs' satisfaction in the procedure of continued improvement of the quality of care within French establishments. GPs are, indeed, experts and can provide reliable judgements on quality of care, which also take into account the clinical history and disease complexity of their patients who are referred to the PMHPs. Moreover, their perception of satisfaction develops through their close dealings

Table 2. p-values of Andersen's LR and item-fit diagnostic statistics stratified by the 14 healthcare providers (PMHPs). The person-fit column reports the number of statistically significant person-fit diagnostics ($p < 0.05$).

PMHP no.	LR-test	Person-fit	Reliability Item-fit	Empathy Item-fit	Tangibility Item-fit	Confidence Item-fit
1	0.593	2	0.781	0.999	0.787	0.737
2	0.123	1	0.555	0.995	0.824	0.978
3	0.148	1	0.978	0.079	0.998	0.983
4	0.666	1	0.992	0.501	0.872	0.938
5	0.119	2	0.853	0.872	0.948	0.766
6	0.364	3	0.975	0.668	0.944	0.902
7	0.565	4	0.998	0.102	0.970	0.988
8	0.542	0	0.950	0.199	0.999	0.908
9	0.875	3	0.951	0.449	0.984	0.943
10	0.690	3	0.994	0.100	0.993	0.962
11	0.603	0	0.993	0.269	0.998	0.888
12	0.159	1	0.956	0.441	0.971	0.970
13	0.714	1	0.989	0.889	0.645	0.943
14	0.715	0	0.976	0.016	0.977	0.999

with and considerable experience of all the PMHPs in the area in which they reside.

The response rate of the GPs enrolled in this study was excellent, although this is not surprising given the survey method and sampling strategy used.

The model exhibited a reasonably good fit to the data. Item-fit diagnostics detected a possible violation of the fit assumption for the Empathy domain, a result that could indicate that this statistical technique is not able to model properly the β_{ik} for the item. In order to test this hypothesis, the Partial Credit Model (19) could be considered, thereby augmenting the flexibility at the expense of a greater number of parameters to be estimated. Another possible interpretation of this item-fit diagnostics result is that the Empathy might be a multi-factorial concept. If this is the case, it would be advisable to reformulate the item splitting it into its different sub-concepts. This technique should be further developed to accommodate surveys making provision for a repeated measurement design and multidimensional latent structure.

The RSM parameter estimates led to a preliminary exploratory comparison between the levels of satisfaction with the investigated PMHPs. With the aim of providing PMHPs with meaningful information on their healthcare performance, the approach proposed in this study deliberately avoids trying to summarise the results in a final composite score. In fact, this would imply confounding each domain effect by arbitrarily assigning weights to the components. The resulting overall performance ranking list would be difficult to interpret, especially when there is no clear heterogeneity among the institutions (20), or even used as a punitive mechanism or to "name and shame" individual PMHPs uncritically. The actual interest in this methodology should, in fact, be focused on its ability to develop a scale on the latent variable satisfaction suitable to evaluate the effect of new policies and intervention in healthcare.

Conclusion

The survey method used in this study gave an excellent response rate. The psychometric properties of the adopted scale analysed by RSM model provided preliminary evidence in profiling public healthcare satisfaction.

vey. The authors thank Dr Gaetano Piccinocchi, head of the COMEGEN cooperative for introducing and promoting the questionnaire among the GP members.

References

1. van der Linden W, Hambleton R. Handbook of Modern Item Response Theory. New York: Springer-Verlag 1997.

2. Baker F, Kim SH. Item Response Theory, 2nd edition. New York: Marcel Dekker 2004.

3. Andrich D. A rating formulation for ordered response categories. Psychometrika 1978; 43: 561-573

4. Nappo D. "La qualità nel Sistema Sanitario Nazionale: il caso dei medici di base" Thesis. "Federico II" University of Naples 2006.

5. France G, Taroni F, Donatini A. The Italian health-care system. Health Econ 2005; 14 (Suppl 1): S187-S202.

6. Parasuraman A, Zeithaml VA, Berry LL. Conceptual model of service quality and its implication for future research. Journal of Marketing 1994; 49: 41-50.

7. Bertrand D, François P, Romanet P. General physicians' viewpoint on the evaluation of health care quality in the hospital. Sante Publique 1999; 11: 453-463.

8. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danish Institute for Educational Research 1960.

9. Andersen EB. Polytomous Rasch models and their estimation. In: Fischer G, Molenaar I eds Rasch Models: Foundations, Recent Developments, and Applications. New York: Springer 1995: 271-292.

10. Andersen EB. A goodness of fit test for the Rasch model. Psychometrika 1973; 38:123-140.

11. Mair P, Hatzinger R. Extended Rasch modeling: The eRm package for the application of IRT models in R. Journal of Statistical Software 2007; 20(9): 1-20.

12. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing 2009.

13. Ostini R, Nering ML. Polytomous Item Response Theory Models. Series: Quantitative Applications in the Social Sciences. Beverley Hills; SAGE Publications 2005.

14. Mossialos E. Citizens' views on health care systems in the 15 Member States of the European Union. Health Econ 1997; 6: 109-116.

15. Blendon RJ, Leitman R, Morrison I, Donelan K. Satisfaction with health systems in 10 nations. Health Aff (Millwood)1990; 9: 185-192.

16. Ministero della Sanità, World Health Organization, European Center for Environment and Health. Health in Italy in the 21st century. Rome: Available at: http://www.euro.who.int/Document/Hms/itahfa21/omsindice.htm (Accessed May 8, 2008)

17. World Health Organization. Statistics World Health Report 2000. Available at: http://www.who.int/whr/2000/en/whr00_en.pdf (Accessed May 8, 2008)

18. Blendon RJ, Kim M, Benson JM. The public versus The World Health Organization on health system performance. Health Aff (Millwood) 2001; 20: 10-20.

19. Masters GN. A Rasch Model for Partial Credit Scoring. Psychometrika 1982; 47: 149-174.

20. Spiegelhalter D. Ranking institutions. J Thoracic Cardiovasc Surg 2003; 125: 1171-1173.

APPENDIX

Tab 1A: Complete list of the 14 assessed PMHPs

LOCAL HOSPITALS:
- Ospedale SS. Annunziata
- Ospedale Ascalesi
- Ospedale C.T.O.
- Ospedale Santa Maria del Popolo degli Incurabili
- Ospedale S. Maria di Loreto Mare
- Ospedale dei Pellegrini
- Ospedale San Paolo
- Ospedale San Gennaro

MEDICAL SCHOOLS
- Azienda Universitaria Policlinico della Seconda Università di Napoli
- Azienda Ospedaliera Universitaria "Federico II"

HOSPITALS OF NATIONAL IMPORTANCE
- Ospedale Cardarelli
- Ospedale Monaldi
- Ospedale Santobono-Pausilipon

NATIONAL CANCER INSTITUTE
- Istituto Nazionale per lo studio e la cura dei tumori "Fondazione Pascale"

Tab 2A: List of questionnaire items.

- Reliability: ORGANIZZAZIONE DELLE STRUTTURE OSPEDALIERE
  Considerando la sua esperienza lavorativa, attribuisca a ciascuna delle seguenti strutture ospedaliere di Napoli un giudizio in merito all'efficienza della struttura

- Responsiveness/Empathy: PARTECIPAZIONE AL PROCESSO DECISIONALE
  Considerando la sua esperienza lavorativa, associ alle seguenti strutture ospedaliere un punteggio in base alla disponibilità dei medici in esse operanti ad instaurare un contatto collaborativo con i medici di base.

- Tangibility: ASPETTI TANGIBILI DELLA STRUTTURA OSPEDALIERA
  Considerando la sua esperienza lavorativa, esprima un giudizio considerando l'insieme degli aspetti tangibili (aspetti alberghieri, tempi di attesa, presenza di un reparto specialistico, cortesia del personale)

- Confidence: FIDUCIA
  Considerando la sua esperienza lavorativa, associ alle seguenti strutture ospedaliere un punteggio in base al suo grado di fiducia che lei ripone nella struttura