Comprehensive assessment of general practitioners : a study on validity, reliability and feasibility

Summary

Quality of care in general practice has evolved into a new important field of work and research including research on quality assessment, i.e. the process of evaluating the current level of performance of General Practitioners (GPs), and quality improvement. The assurance of good quality of care in general practice requires a continuous repetition of the process of quality assessment and improvement. Within this frame, the idea of formal (re)assessment of practising GPs is becoming more and more an accepted fact, although a formal assessment procedure enabling a systematic valid, reliable and feasible assessment of practising GPs is still lacking. This educational study focuses on the development of such an assessment procedure that can be used as part of quality improvement in general practice.

The introduction of this thesis contains a description of the research problem at hand, the area of assessment of GPs and the high-stake issues in this field. This is followed by the research questions and the study design. It is obvious that no single assessment method can cover the extensive and complex domain of general practice care. Therefore, a comprehensive assessment approach is needed. It is also made clear that no assessment model exists which is able to cover all aspects of general practice care such as communication with patients, medical performance and practice management including GPs' knowledge and skills. It is concluded that it is desirable to develop such an assessment model, which can be used as a blueprint in the selection of assessment methods. Further, appropriate methods for the assessment of each separate aspect of general practice care are needed. It should be established under which conditions the assessment is best to be conducted. Should it take place in clinical practice, in a simulated situation, or by means of direct observation or by indirect tests such as written papers?

The relationship between the scores of different methods is unclear as is the predictive value of scores of written tests or tests in a simulated situation for actual performance. The relationship between process, i.e. communication with patients and medical performance, and structure, i.e. practice management, and the relationship between competence (knowledge and demonstrated skills) and actual performance are questionable as well.

Three research questions were therefore formulated on the development and testing of an appropriate model for comprehensive assessment of practising GPs, including different methods, which should cover the domain of general practice care as completely as possible. More specifically the following research questions are addressed:

1. What would be an appropriate model for comprehensive assessment of practising GPs, which covers the domain of general practice care as completely as possible?
2. To what extent do competence-based tests predict actual clinical performance?
3. What is the relationship between the GP's practice management (structure) and actual clinical performance of GPs in that practice (process)?

The study design showed the following steps (see figure 1 page 16). First, a literature search was carried out into models and blueprints for assessment of family practice care, leading to a new pyramid assessment model. Second, existing assessment instruments and methods were screened on their psychometric qualities, i.e. validity, reliability and feasibility. The methods with the best psychometric characteristics were selected for the main study by using the pyramid model. Third,

in a cross sectional study, GPs were assessed comprehensively. Knowledge tests were completed at home first, followed by a randomization of participants in order to control for order effects. Three groups of 50 GPs each, two groups of participants and one control group, comparable on personal and professional characteristics and on results on the knowledge tests, were formed. Group 1 and 2 were observed directly in both a simulated situation and daily surgeries, but in reverse order. Consultations were scored by peer-observers (N=35) on both medical performance and communication with standardized patients in the simulated situation and with real patients in daily practice. These patients participated as evaluators of the communication component as well. After observation of actual performance, GPs' practice management was assessed by non-physician observers visiting the practices. Finally, participants were sent a questionnaire about their acceptance of each method and their educational activities and plans as a result of their participation in the assessment and the feedback received.

Chapter 1 describes the development of a theoretical and practical model to be used as a blueprint for comprehensive assessment of practising GPs, taking into account recent views on medical expertise, the distinction between competence and performance, the state of the art in Continuing Medical Education (CME) and quality improvement, and recent developments in assessment methodologies. Computerised and hands-on literature searches were performed, focused on collecting models. Three existing frameworks were transformed into a new three-dimensional pyramid: Fabb's model, a systematic approach of the domain of general practice care; Miller's pyramid, reflecting the processes of clinical reasoning and the distinction between competence and actual performance; and Donabedian's framework, including process, structure and outcome of general practice care. For logistical and complexity reasons, patient outcomes were not included in the new Pyramid Assessment Model, except for the evaluation of quality of GPs' communication with patients by standardized patients and real patients.
Three subdomains in general practice were represented on one axis: medical performance, communication with patients, and practice management. Diseases of patients, classified in chapters of the International Classification of Primary Care (ICPC) and contextual factors were represented on the second axis. The third axis represented the different assessment levels: competence, i.e. knowledge and demonstration of skills, at the basis and actual performance at the top. Using this model, tasks related to the three subdomains could be connected with different diseases and complaints of patients and their contexts, whereas a proper assessment level, i.e. a level of competence or performance, could also be selected.
This model allows test developers to define precisely their objective for testing, i.e. what should be assessed, and where and how. For each of the domains and for each level (different) assessment instruments can be chosen (see figure 1, page 16).
A practical approach in using this blueprint is presented in this first chapter. Concerning the competence level, knowledge levels were assessed by written tests and the "shows how" level, i.e. the demonstration of skills, were assessed by a standardized observation test in a simulated situa-

tion. Actual performance and practice management ("does") were assessed by video observation of regular consultations (performance) and by a practice visit to assess practice management.

Chapter 2. In this chapter issues of validity, reliability and feasibility of video assessment of actual performance of GPs are described. These focused on the following questions: how can consultations be identified and selected for a valid assessment; how reliable is video assessment considering the number of consultations and observers required; how feasible is this method concerning technical aspects, logistics, costs and acceptance by GPs and patients?

Regular consultations of 93 GPs were video recorded during one week using two cameras, one in the consulting room and the other in the examination room. The receptionist informed patients about the video recording and asked permission. The GPs registered consultation and patient data in a logbook, from which 16 consultations per GP were selected. Preset criteria based on prevalence of complaints and diseases in general practice and on a nationally accepted job description were used in the selection procedure. The quality of communicative and medical performance of these consultations was assessed by trained peer-observers with a validated scoring instrument (MAAS-global).

The validity of the procedure was evaluated by checking the content of GPs' samples of consultations using specific sample criteria. Selection bias was estimated by multiple regression analysis, with sample characteristics as independent variables and scores on communication and medical performance as dependent variables. In this way the influence of GPs' personal characteristics on scores was estimated as well. The influence of observation on GPs and patients was assessed by a questionnaire. Generalizability theory was used to estimate reliability. Feasibility was assessed by a questionnaire, by keeping accounts, and by checking the technical quality of the video taped consultations.

The domain of general practice proved to be covered well in the samples; content validity was satisfactory. As regards the sample characteristics, only the total duration of consultations appeared to correlate significantly with the scores on both communication and medical performance: the longer the total duration of the 16 consultations the higher the performance scores. A majority (71%) of GPs reported not having been influenced by the observation, except in the first cases, and recognizing their usual daily performance in the videotaped consultations. An acceptable level of reliability was reached after 2.5 hours of observation, i.e. 12 cases observed by a single observer. The method was well accepted by both GPs and patients. The costs (£ 250 per GP) were acceptable. Video assessment of GPs in daily practice, performed as described, proved both to be valid by approximating real professional life as closely as possible, and appeared reliable and feasible for use in educational and quality improvement activities.

Chapter 3 reports on the comparison between observation of GPs' performance in a multiple station examination using standardized patients and observation of real surgeries in daily practice. Consultations of 90 GPs, divided into two groups, were videotaped both in a multiple station examination at the medical school using standardized patients and in their daily practice surgery with regular patients. Peer-observers assessed GPs' communication with patients and medical

performance with a validated instrument (MAAS-Global). Both groups passed through both assessments, but in reverse order. Content validity, criterion validity, reliability, i.e. generalizability, and feasibility of both methods were compared, taking time-order effects into account.

Content validity of practice video assessment was superior to the multiple station examination, since the domain of general family practice care was better covered. Regular consultations in practice were more authentic than the simulated cases, since initial and follow-up consultations with children and older patients in the natural daily context were included. Moreover, participants judged the videotaped practice consultations as "natural", whereas hardly any GP recognized his usual working style in the multiple station examination. Concerning concurrent validity between the two test methods, only the communication component of both methods correlated. In addition, real practice performance proved to be less influenced, thus more stable, than behaviour during the station examination: scores on the multiple station examination increased significantly when the practice video assessment had been done previously (group 2), whereas practice scores appeared to be consistent, irrespective of whether GPs had passed through the station examination before or not. Reliability of both methods, expected to be better in the more controlled multiple station examination, was comparable. The organization of practice video assessment was more flexible, less costly and better accepted by GPs than the multiple station examination.

Therefore it has to be concluded that assessment for quality improvement of GPs by video observation in daily practice is superior to video assessment in a simulated situation.

Chapter 4 compares the predictive values of two written medical knowledge tests and a standardized multiple station examination for GPs' actual medical performance in daily practice.

Two groups of GPs (N=46 and N=44) were assessed using a general medical knowledge test and by a knowledge test on technical skills, followed by the multiple station examination using standardized patients and the practice video assessment of real surgeries according the study design described. In both groups the predictive value of medical knowledge tests, ranging from 0.43-0.56 (disattenuated Pearson correlation), proved to be comparable with the predictive value of the multiple station examination for actual medical performance (0.33-0.59). The overall explained variance of scores of the practice video assessment by scores on the knowledge tests was moderate (35%). GPs' professional characteristics did not contribute to the explanation of variation in performance scores.

In conclusion, medical knowledge tests can predict actual clinical performance to the same extent as a multiple station examination. Compared with a station examination, a knowledge test may be a good alternative method for assessment procedures of a large number of practising GPs, since knowledge tests can be used on a broad scale with relatively low resource investments. In addition, these tests have a far better predictive value for actual performance than GPs' personal characteristics, such as age, gender, College membership or working single-handedly.

Chapter 5 describes the relationship between practice management (structure) and actual clinical performance (process) in general practice. The precise relationship between these dimensions is tenuous. Analysis of their mutual relationship may yield insight into the way they contribute to

outcome. A study is described in which the practice management of 93 GPs was assessed by a practice visit performed by a non-physician observer using a validated instrument (VIP), directly after the GPs had been videotaped in their daily surgeries,.

Pearson correlations (observed and disattenuated for unreliability of the instruments) between scores on 22 practice management dimensions and scores of 16 selected cases on medical performance and communication were calculated. The predictive value of specific practice management aspects for actual performance was determined by multiple regression analysis.

Nine practice management dimensions proved to correlate significantly with medical performance and so did five dimensions with actual communication. Overall, most associations were weak.

Combined with demographic variables (age for medical performance and working single-handedly for communication), 26 percent of variance in medical performance scores and 11 percent of variance in scores of communication with patients could be explained by only five practice management dimensions. Organization of quality assessment activities, i.e. assessment with the help of data from the medical insurance, prescriptions, referrals and diagnostics, explained most of the variation in medical performance (ten percent). Two practice dimensions, i.e. delegation of medical tasks to the practice assistant and working single-handedly, explained six respectively five percent of variance in communication with patients scores.

In conclusion, practice management (structure) and actual performance (process) seem largely independent constructs. However, some practice management dimensions might be linked up with GPs' performance. In all, quality improvement and assessment activities should emphasize that practice management is different from actual performance. Structure and process may contribute to patient outcome independently from each other.

Chapter 6 contains an explorative study concerning observation of GPs' communication with patients assessed by peers, real patients and standardized patients (SPs), focused on the question whether such assessments of GPs provide relevant additional information about GPs' communicative performance, from both the professional perspective and patients' perspective.

Two groups of 43 GPs each went through the "Multiple Station Examination" using SPs and through the "Practice Assessment" in regular surgeries, following the study design described before (see page 16). GPs' communicative performance was evaluated by peers with the MAAS-Global, and by SPs and real patients with an instrument derived from the MAAS-Global.

The correlation between the scores given by standardized patients, scores given by real patients and scores given by peer-observers was assessed. Results showed that SPs were to some extent consistent with peers in the simulated situation. Second, both real patients and SPs valued GPs' quality of communicative performance more positively than peers. Third, real patients probably differ from both SPs and peers in evaluating the quality of GPs' communication substantially, since no correlation was found between patients' scores and SPs' scores or peers' scores.

Finally, the quality of GPs' communication with SPs in a simulated situation may differ from the quality of GPs' actual communication with real patients in daily practice. This aspect has been analysed in detail in Chapter 3.

It has been hypothesized that SPs' evaluation of GPs' communicative performance was comparable with the observation "at distance" by peers, since they evaluated as uninvolved healthy persons about hundred GPs in a short and independent relationship with the observed GP. Real patients, on the other hand, are more involved in the lasting relationship with their GP. Moreover, real patients may be focused on the GP's final conclusion, i.e. bad or good news, concerning their serious complaints and therefore they may be less interested in the communication process. Being consumers, they seem to have an other perspective on GPs' communication with patients than professionals.

From the educational perspective, peer observation or evaluation by standardized patients combined with the evaluation by real patients may provide relevant information about GPs' communication with patients. SPs represent patients' perspective by giving high absolute scores as well as the professional perspective by being consistent with peers. Evaluation by real patients is important, since they represent consumers' needs. On the other hand, the low scores given by peers may be an effective stimulus for GPs to improve their communicative performance.

Finally, within the scoringlists used, absolute high scores might be relatively low, which is useful for the selection of topics for quality improvement.

Therefore, both not-involved observers (peers and SPs) and involved real patients should be used in evaluating GPs' communication with patients.

Chapter 7 contains the general discussion and recommendations for research, and describes GPs' evaluation of the assessment procedure. The overall conclusion of the study is that a valid and reliable comprehensive assessment of individual practising GPs should include both an assessment method for direct (video) observation of GPs' communicative and medical performance in daily practice and a method for practice visitation to assess GP's practice management. Concerning the pyramid assessment model, it is argued that further study should be performed in order to include Gps' attitude and patient outcomes in a model for assessment. Clinical tasks to be assessed should be formulated before the implementation of a procedure for comprehensive assessment, since the model is global. In relation to video assessment in daily practice, the area of tension between standardization of samples of regular consultations, content validity and reliability needs further research. Strict standardization may decrease content validity, since each sample has to be representative for the working style of each GP. Assessment in a simulated situation may be useful in specific CME activities focused on specific skills. For screening purposes on a broad scale this method is costly, not well accepted by GPs and therefore less useful than video assessment in daily practice.

Moreover, less costly medical knowledge tests predict actual clinical performance to the same extent as a multiple station examination and are therefore a good alternative method. Since these tests have a far better predictive value for actual performance than GPs' personal characteristics, such as age, gender, College membership or working single-handedly, it is concluded that these tests should be developed for the assessment of GPs' communication with patients and practice management as well. Concerning practice management, it is argued that this aspect of general

practice care has a low predictive value for actual performance. Structure and process may contribute to patient outcomes independently from each other, which needs further research. Finally, with regard to patients' evaluation of GPs' communicative performance, real patients being consumers may have another perspective than peers or standardized patients. Real patients' satisfaction with GPs' communicative performance, which may be considered as patient outcome, showed no relationship with peers' scores of this communicative process. This finding requires further research as well.

The participation in the comprehensive assessment including the feedback given was evaluated by the GPs using a questionnaire. Peer-observers and GPs of the control group were sent a questionnaire as well. In this study feedback was given in a limited extent in order to minimize a possible order effects, which would bias the relationship between methods used. The questionnaire contained four main questions concerning the following aspects: the profit of participation in the assessment versus its efforts (1), the effects on GPs' insight in their way of functioning (2), GPs' plans for concrete changes in daily practice care (3) and GPs' activities on medical education as a result of participation (4). GPs and peer observers viewed a positive balance between profit and effort of participation. GPs of the control group, who participated on the knowledge tests only, viewed the balance as less positive than observers and assessed GPs. Observers and assessed participants viewed both the medical knowledge test, the assessment of practice management, the multiple station examination and the practice video assessment as methods which gave them insight in their way of functioning, the latter two for communication aspects particularly. However, a majority made no plans for concrete changes, whereas a minority performed individual educational activities mainly, i.e. self directed learning and immediate changes in performance or practice management.

GPs' preference for assessment methods to be included in a comprehensive procedure favoured the video-assessment in daily practice (medical and communicative performance) combined with the assessment of practice management by visitation and a written medical knowledge test.

A majority of GPs preferred a combination of written and oral feedback and showed a slight preference for an unknown colleague, provided that a GP would be the mediator.

In this study, the "quality circle" correctly started in practice observation and collecting data from practice. However, the next stages of this circle, i.e. the evaluation of information (performance versus targets) including setting priorities in implementing changes, the agreement of criteria including setting target standards and a subsequent observation of practice to evaluate changes, have not been performed, since this methodological study was focused on the relationship between methods used. In relation to these limitations, the educational effects of the study on the participants are encouraging. Further research on feedback as a part of continuous quality improvement activities is necessary, in order to make assessment activities including feedback as effective as possible.