Assessing quality of nursing care

Sally J Redfern, Ian J Norman, Deborah A Tomalin, Sarah Oliver

In parallel with developments in medicine much has been written in the nursing press, particularly over the past few years, on quality assurance and audit. In nursing early approaches focused on broad, proscriptive statements about the nature of nursing, why quality was often not achieved, and the changes required. Later, nursing practice was compared with pre-formulated standards and criteria, an approach that continues with greater acceptance of practitioner led and patient-centred initiatives.

The aims of this paper are to describe some of the quality assessment instruments used by nurses that are derived from externally generated criteria and claim to provide a valid index of the quality of nursing care delivered in a hospital ward. Both positive and negative features of these instruments are presented, and the authors' experience of some of these instruments together with findings on inter-rater reliability are summarised. This work forms part of a study of the validity of quality assessment instruments in nursing.

**Generic quality assessment instruments**

Composite measures were developed in the United States by incorporating criteria of nursing care quality, derived from professional knowledge, into checklists. These checklists, or generic quality assessment instruments, were used to obtain a measure of the quality of nursing care as a whole. The instruments focus on the performance of the nurse (for example, the Slater nursing competencies rating scale) or on the care received by patients (for example, the quality of patient care scale, the Phaneuf audit, and the Rush-Medicus nursing process methodology); many continue to be in widespread use.

The use of pre-formulated generic quality assessment instruments has increased appreciably in the United Kingdom over recent years, being used by some of the districts in all regional health authorities. Of these, Monitor is the most popular in its general, midwifery, junior, senior, psychiatry, health visiting, or district nursing forms. The quality of patient care scale, known as Qualpac, has its proponents, mainly in acute general and elderly care units.

**Monitor**

Adapted for use in the United Kingdom from the Rush-Medicus nursing process methodology developed in Chicago, Monitor was developed by Goldstone et al as part of the criteria for care system designed to establish nurse staffing levels and skill mix from analysis of nursing activities. The instrument consists of Patient Monitor, which assesses quality of nursing in relation to four main areas of care (planning and assessment, physical care, non-physical care, and evaluation of care), and Ward Monitor, with 43 items that describe and assess the procedures and management of the ward including structural factors such as staffing levels, grade mix, workload, support services, and environmental safety.

Patient Monitor contains four separate schedules for different categories of patient dependence. The schedule for patients in dependency group I (least dependent) contains 81 items, that for those in dependency group II contains 107 items, for dependency group III 148 items, and for dependency group IV (most dependent) 118 items. Each section of Patient Monitor contains several parts: planning of nursing care contains parts relating to the patient's admission, assessment and coordination of nursing care with medical care; meeting the patient's physical needs involves patient safety, providing physical comfort and rest, and hygiene, nutrition, and fluid and elimination needs; meeting the patient's non-physical needs contains parts relating to staff courtesy, patient privacy, rights, wellbeing, health promotion and prevention of ill health, and involvement of the patient's family; and evaluation of nursing care objectives refers to documentation of observations, treatment and care, and the patient's response to treatment.

Information to rate the items is obtained by examining nursing records, by talking to nurses and patients, and by observation. The items are rated with yes/yes, sometimes/no answers (see box). Unqualified "yes" answers score 1, qualified "yes" answers score half, and "no" answers score 0. Total scores are based on the percentage of "yes" answers, the closer the score to 100% the better the care.

**Senior Monitor**

Senior Monitor is the augmented and modified version of Monitor for use in assessing quality of care in elderly care wards. Its structure and content differ slightly from Monitor but it uses the same scoring system. Unlike Monitor, all 232 items within Senior Monitor are asked of each patient; patients are not classified into dependency groups with different schedules. Senior Monitor contains three additional sections – the patient's need for rehabilitation; care of the severely ill, terminally ill or dying patient; and the last offices (care of the
Assessing quality of nursing care

Reliability and validity of instruments

Generic instruments differ from each other in their focus and in the dimensions of nursing that they address, but they share some common characteristics, as follows:

- They evaluate care as a whole rather than focusing on nursing care related to specific patient problems.
- They are based on externally generated components of nursing relevant to quality.
- They claim to provide an “objective” assessment of the quality of nursing by translating key criteria into items and scores.
- They are usually administered by assessors external to the nursing team.

If confidence is to be placed in evaluations of quality using generic instruments it is important that the instruments are reliable and valid. Yet most instruments require a judgement of quality. Inclusion of judgement, even by professionals, means that the reliability and validity of the instruments, as measured by conventional statistical techniques, is likely to be low. This conclusion is supported by the few studies that have compared scores obtained from different generic instruments. A correlation coefficient between Qualpacs and the Phaneuf Audit was only 0.01,18 similarly, correlations between Qualpacs and the Rush-Medicus audit were low, being close to zero between the physical and psychosocial subscales of both instruments.19

Other research comparing instruments has shown little agreement20 suggesting that quality is multidimensional and a combination of measures is necessary to achieve comprehensive assessment. It is important that instruments address all aspects of nursing care relevant to quality. Generic instruments face a difficulty because their preformulated criteria inevitably become outdated by the changing nature of nursing. For example, important topics such as health promotion and prevention are not adequately addressed by most instruments because these instruments emphasise disease rather than health.20-22

Inter-rater agreement has been achieved with quality assessment instruments. For example, during the development of Monitor inter-rater agreement was high for both Patient Monitor (91%-99%) and for Ward Monitor (80%-98%).15 In the original work on Qualpacs, correlation coefficients for inter-rater reliability ranged from 0.64 to 0.91 in three studies.3 Later research in the United States did not, however, achieve the predetermined standard of 0.75 for any of the subscales.19

The evidence for other aspects of reliability and for validity of generic instruments is scant. Monitor was found to discriminate consistently between wards, with scores ranging from around 50% to 80% or more in several studies.15 For Qualpacs internal consistency tests generated a Kuder-Richardson reliability coefficient of 0.96. Validity was tested by comparing average Qualpacs scores with independent judgements of the quality of care in the same wards. The correlation coefficient

<table>
<thead>
<tr>
<th>Examples of items and scoring format in Monitor*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source of information</td>
</tr>
<tr>
<td>Ask patient f Is the patient’s call light/buzzer/bell answered promptly?</td>
</tr>
<tr>
<td>To patient: In the past two days did someone come to you promptly when you pressed the bell (buzzer/light)?</td>
</tr>
<tr>
<td>Score</td>
</tr>
<tr>
<td>(Item 48, dependency group III)</td>
</tr>
<tr>
<td>Ask patient b Do nurses introduce themselves to the patient?</td>
</tr>
<tr>
<td>To patient: When a nurse first meets you, does she usually tell you who she is?</td>
</tr>
<tr>
<td>Score</td>
</tr>
<tr>
<td>(Item 111, dependency group III)</td>
</tr>
</tbody>
</table>

*From Goldstone et al11

- deceased – making seven sections in all. The number of parts in the first two sections has been increased.

Qualpacs

Qualpacs (the quality patient care scale), based on Slater’s nursing competencies rating scale, was designed in the United States by Wandelt and Ager.4 It provides a measurement of the quality of care in a ward based on observations of experienced nurse assessors. The scale contains 68 items classified into six sections; psychosocial individual, psychosocial group, physical, general, communication, and professional implications. Each item is accompanied by a set of cues that guide the observer with illustrative examples. The items observed are scored by reference to the standard expected of a first level (newly qualified) staff nurse, irrespective of the grade of nurse being observed. Thus the observer makes a judgement and rates the care given on a scale from 1 (poorest care) to 5 (best care). Scores are summed across items to generate section and total scale scores for a ward. The figure provides examples of items and response scales.

Psychosocial: individual

Actions directed toward meeting psychosocial needs of individual patients

- Patient is approached in a kind, gentle, and friendly manner D

Physical

Actions directed toward meeting physical needs of patients

- Nursing procedures are adapted to meet needs of individual patient for treatment D

D = direct observation

Examples of items and scoring format in Qualpacs (from Wandelt and Ager4)
was 0.52, and it was concluded that Qualpac discriminates successfully between wards.

Our study seeks to test the reliability and validity of Monitor, Senior Monitor, and Qualpac with a multiple method research design that includes observing nurses, interviewing them and patients as well as administering the instruments. Papers on the research design of the study and on Monitor and Senior Monitor are available elsewhere. As the research is still in progress this paper focuses on describing the problems encountered using these instruments, together with information on inter-rater reliability.

**Monitor and Senior Monitor**

**SUMMARY OF PROBLEMS**

Several practical difficulties were encountered in the process of administering Monitor and Senior Monitor; these have been described in detail elsewhere and are summarised here.

**Structure of instruments** – As the items in the manual are not grouped by information source (that is, consult patient’s records, ask patient, ask nurse, observe) assessors have to leaf through the whole manual when covering items whose answers are to be found from the same source, with the result that items may be missed. This problem was overcome by regrouping the information and providing assessors with a checklist of all items listed under each source.

**Accessing information** – During the study patient’s records were not always available to the assessors when needed; also, talking to or observing the patient was sometimes impossible or unacceptably intrusive.

**Interpreting items and their responses** – Agreement between assessors in interpretation is essential for achieving reliability, and many of the items required detailed discussion before agreement was reached. Another problem occurred with cues accompanying certain items when, for example, the question in the cue was different from that in the item. The assessors also found it difficult to accept that confirmation of an item did not necessarily convey appropriate care. For example, a “yes” score was awarded if there was “… a written statement of care given to pressure areas on the skin,” even when that statement conveyed inadequate care. A high score therefore did not necessarily represent high quality care.

**Disruption to ward staff** – A major problem was finding the appropriate time, or sometimes any time, to talk with nurses and patients without appreciably interfering with nursing care.

**Length of time to reach consensus** – It took much longer than suggested in the manual for our assessors to reach agreement on every ambiguous item, which is an essential prerequisite for reliability.

Most of these difficulties encountered, although irritating, can be overcome during the planning stages of the exercise. It is also essential to devote enough time to discussion and to training assessors in order to achieve acceptable inter-rater reliability.

**INTER-RATER RELIABILITY**

Inter-rater reliability was assessed during pilot work and again towards the end of the main fieldwork. For Monitor the mean percentage agreement across patients in two surgical wards was 79% (range 58% to 89%) in the first ward and 89% (86% to 93%) in the second. For Senior Monitor in the first ward (care of the elderly) a mean percentage agreement of 89% was reached (84% to 96%) and in the second ward (surgical) 80% (73% to 84%). The mean percentage agreement for section scores varied much more widely but was never below 65% and usually over 75%. These figures are not far removed from Goldstone’s criterion of 80% for high agreement.

To account for possible overestimation, an accepted limitation of percentage agreement, inter-rater reliability was also assessed using the intra-class correlation coefficient. For each instrument the coefficients for both wards were well above the criterion of 0.75, being 0.89 or higher. However, section scores varied rather more, ranging from 0.60 to 0.98 for Monitor and from 0.40 to 0.95 for Senior Monitor. All were significant at p < 0.001 except for two that achieved p < 0.01. Because the intra-class correlation coefficient emphasises correlation between ranks rather than agreement between absolute scores low coefficients will emerge when the range of scores is narrow; this was the case in our study and accounts for the lower coefficients that emerged when absolute scores were very close.

Therefore, apart from the practical difficulties encountered with these two instruments inter-rater reliability of both Monitor and Senior Monitor reached acceptable levels.

**Qualpac**

**SUMMARY OF PROBLEMS**

Qualpac presented the following difficulties.

**Unwieldiness** – The main schedule, with its 68 items, is a manageable length, but if the observer carries around the 20 pages of cues to the items, as is advised, the whole makes a bulky document to handle in the ward.

**Verbal obscurity** – Items in the schedule are described in the passive voice – for example, “patient receives explanation and reassurance when needed” – and this could be simplified to “nurse explains and reassures.” Such simplification would promote speed and certainty in scoring, both of which are essential to accuracy when rating many observations that occur in rapid succession.

**Unequivocal identification of items** – Difficulties were encountered with the comprehensiveness of the items, mutual exclusivity, and coherence. For example, some items were so general in scope that they encompassed broad areas of nursing care (“patient receives nurse’s full attention”) whereas others were more specific but not independent of that area (“patient is given an opportunity to explain his feelings”). Many items overlapped and some were multiple items, making a single score impossible to award.
Assessing quality of nursing care

Cues – The cues provided examples to aid the observer in interpreting the items, but it proved difficult to avoid rating the cue rather than the item; our solution was to abandon the cues after the training period.

Delimiting an interaction – As is always a problem in observation, deciding when an interaction began and ended required frequent discussion until agreement was reached.

Identification of sections – Identifying the appropriate section for the interaction observed was not always easy. For example, the section “communication” referred to communications made by the givers of care on behalf of the patient rather than communication between them and the patient. Also the section entitled “general” could be deleted and its items allocated to other sections or omitted altogether when alternatives exist elsewhere.

Rating interaction – The standard of measurement – that is, the care expected of a first level staff nurse – required considerable discussion by the raters to reach an approximate consensus, and it was never easy to articulate the details of this consensus. We also had to become accustomed to rating unqualified nurses and other carers against the same standard.

Omitted care – The manual specified that omitted care should be rated the same as “poorest care,” which is an important distinction.

One of the notable strengths of Qualpacs is that it treats the patient as a whole rather than a bundle of conditions; another is its unequivocal reliance on professional judgement, which is essential when rating professional interactions. It has several limitations but, like Monitor, most can be dealt with during preparatory training.

INTER-RATER RELIABILITY

We calculated inter-rater reliability of Qualpacs with percentage agreement, expressed as a ratio between the scores given by the two observers, and the intra-class correlation coefficient. The mean ratio percentage agreement across patients in the first ward (care of the elderly) was 97 (range 84 to 106), in the second ward (surgical) 98 (93 to 110), and in the third ward (surgical) 104 (93 to 118). Qualpacs was administered in all three wards because it is applicable to any patient whereas Senior Monitor rather than Monitor is designed for elderly care wards. We accepted mean ratio percentages that were reasonably close to 100 (between 85 and 115) as indicating sufficient agreement for our purposes, and all three total scale values achieved this. Section scores also showed acceptable agreement except for one, with mean ratio percentages ranging from 83 to 111.

The intra-class correlation coefficients for Qualpacs total scale scores for the first two wards exceeded the criterion, at 0·98 and 0·86 respectively, but fell below it for the third ward (0·67). Section scores achieved coefficients that ranged from only 0·17 to 0·99, although all were significant (p < 0·001). As with Monitor and Senior Monitor low coefficients occurred when the ranges of the scores were narrow, the percentage agreements for items that received low intra-class correlation coefficients were high, indicating close agreement between raters in absolute scores.

Overall the level of inter-rater reliability reached for all three instruments was acceptable for our purposes, and this suggests that the limitations of each instrument were adequately overcome.

Discussion

Several issues emerge from the above description. Firstly, there is the question of how to improve the way in which assessors use the instruments. The research indicates that several key points in all of the tools could be clarified to facilitate better and more consistent use. This general point on utilisation has a direct impact on the second area under investigation – namely, inter-rater reliability. To achieve acceptable scores for inter-rater reliability clear preparation and training of assessors were required.

Though the question of validity of the instrument is not addressed here, it is important to reflect on the utility and appropriateness of off the shelf tools. Such instruments are relatively expensive audit or quality measures to use correctly, given the need for proper staff training and support, and the cost effectiveness of the use of these types of tools as repeat measures of nursing quality must be considered. Is it necessary to hold such detailed information for all wards across a hospital to obtain a baseline measure of quality? Would it be more cost effective to identify several more precise indicators of quality and use these as baseline measures, the assumption being that if particular targets are achieved then others will also have been reached? Such questions must be addressed in order to direct resources to the research and development necessary to produce reliable and valid measures of quality.

If the idea of a generic, off the shelf measure is assumed to be worth retaining, does it need to be as lengthy and cumbersome as the Monitors and Qualpacs? The findings reported in this paper suggest that these instruments may be too comprehensive and that it might be possible to produce an equally effective but much shorter tool for measuring quality of nursing care as a whole. Moreover, it is essential that users’ views are represented; instruments like the Monitors and Qualpacs are generated entirely from the views of nursing professionals and researchers.

Our research is currently addressing these issues with generic instruments. We have interviewed patients and their nurses to elicit their views on quality of nursing care. Comparison of these views with items in the Monitors and Qualpacs will establish the relevance of the instruments to quality and will assure the patient’s contribution to the measurement of quality.

The research on which this paper is based is funded by the Department of Health. We thank the other members of the research team, Keith Jacka and Romana Kebe, for their help.