PRESSURE ULCERS are a considerable burden on patients, carers and the NHS (Royal College of Nursing (RCN) 2000) and, consequently, prevention of pressure ulcers is of major concern to most nurses. To help them to achieve this, many risk assessment tools have been developed (Bergstrom et al 1987, Chaloner and Franks 1999, Flanagan 1993, Pritchard 1986, Williams 1992), so that resources can be targeted at patients most in need. Many trusts have developed algorithms that allocate resources to patients, in particular support surfaces such as mattresses and cushions, based on their risk assessment scores (Shipperley 1998).

To be of practical use, risk assessment tools must be reliable. Inter-rater reliability is ‘the degree to which two raters or observers, operating independently, assign the same ratings or values for an attribute being measured or observed’ (Polit and Beck 2004). It is an important element of risk assessment tools as ‘they must consistently identify the “at risk” population regardless of the user’ (Bridel 1994).

Literature review

Despite the importance of inter-rater reliability, little research has been done in this area (Bridel 1994, Flanagan 1997). In an attempt to discover if this was true of the Waterlow risk assessment tool (Waterlow 1996), which is the most frequently used system in UK acute hospitals (Scott and Newens 1999), a literature search was conducted using CINAHL (1982-2004) and Medline (1996-2004), using the keywords pressure ulcer, risk assessment, Waterlow and inter-rater reliability. Relevant references in the articles found were also accessed. Only those articles written in English were retrieved.

Only four relevant studies were found. Dealey (1989) reported on 20 patients who were each assessed using the Waterlow pressure ulcer risk assessment scale by either two or three nursing students. The Wilcoxon Test rejected the null hypothesis that there was no difference in the risk scores arrived at by individual nurses and the patient’s actual score, that is, there is a significant difference between the scores obtained by the nurses in the study and the gold standard score.

Conclusion

The results show poor inter-rater reliability when using the Waterlow pressure ulcer risk assessment scale. Part of the problem is that nurses are not using the tool in the way it was intended.

Jennifer Kelly is lecturer, University of East Anglia, Queen Elizabeth Hospital, King’s Lynn, Norfolk. Email: j.kelly@uea.ac.uk

Keywords

Pressure ulcers; Risk assessment scales; Waterlow score

These keywords are based on the subject headings from the British Nursing Index. This article has been subject to double-blind review. For related articles and author guidelines visit the online archive at www.nursing-standard.co.uk and search using the keywords p86-92w32 13/4/05 4:24 pm Page 86
In Watkinson’s (1996) study a total of 12 different assessors assessed nine older patients using three different assessment scales (Braden, Douglas and Waterlow). Only on one occasion was consensus reached among assessors, with identical scores being awarded to a patient – this occurred when the Douglas score was used. Only when the scores were allowed to differ by four points did the percentage agreement reach more than 50 per cent. This was true for all three scales. In Cook et al.’s (1999) study, an adapted Waterlow scale was used on 15 patients in two wards. Each patient was assessed daily by two different nurses over a period of seven days. A total of 28 clinical nurses were involved and a total of 210 assessments obtained. Statistical analysis demonstrated weak or a moderate degree of inter-rater reliability.

In all four of the studies, real patients were used with all their complexities. It is, therefore, difficult to identify if lack of reliability is due to different perceptions of the patients by the assessors, or due to different interpretations on how to use the Waterlow tool. The Braden scale with its six categories: sensory perception; moisture; activity; mobility; nutrition; and friction and shear, and clear descriptors for each score, has been shown to have a high inter-rater reliability (Bergstrom et al 1987, Pang and Wong 1998, Ramundo 1995). In her comparison of the Norton, Waterlow and Braden scales, Bridel (1993) considered Braden to be the most reliable of those described in the literature, and this is probably due to its operational definitions. Although the Waterlow scale does not include assessment guidelines on the tool itself, like the Braden scale, Waterlow (1996) has provided guidelines on how the tool is supposed to be used and these have been updated with the recent revision of her tool (Waterlow 2005).

**Aim**

To identify whether lack of inter-rater reliability with the Waterlow scale is due to different perceptions of patients by nurses or different interpretations of the Waterlow tool.

**Method**

The sample consisted of 110 qualified nurses who used the Waterlow pressure ulcer risk assessment tool in their daily work. They were delegates at five one-day refresher in-house study days on pressure ulcer prevention and management and they were given a case study (Box 1) and an uncompleted copy of the Waterlow risk assessment tool. The participants were asked to complete a risk assessment for the patient in the case study. They were encouraged not to discuss the case study with other delegates, and were given the opportunity to ask the speaker for any further information. After each delegate had obtained a total risk score, the scores were collected. The results were presented to the delegates, and an informal discussion ensued as to how the nurses had arrived at the scores they had. This methodology allowed the researcher to collect the information presented in this study, at the same time as making the delegates aware of how Waterlow intended the tool to be used.

The scores obtained by the delegates were compared with the ‘gold standard’ of 18. This was identified by an expert panel of three tissue viability clinical nurse specialists who, guided by Waterlow’s (1996) guidelines on how the tool should be used, achieved a 100 per cent consensus after brief discussions.

The data were analysed using the Wilcoxon Signed Rank Test. This is a non-parametric test, which was used to test whether the median of a distribution was different from a specified value. It was thus being used to test the null hypothesis that there was no difference between the median of the nurses’ score and the ‘gold standard’. This test is more powerful than the sign test because, in addition to the type of difference between two scores (either under or over-rating), the Wilcoxon Signed Rank Test also measures the magnitude of these differences (Bland 2000). Although the Wilcoxon Signed Rank Test is usually used to test matched pairs, it is legitimate to compare a single sample with a constant to determine if the median of the distribution is equal to the specified constant value, as has been done here.

**Results**

In total 110 nurses were involved over five study days, and the scores they obtained are shown in Table 1. The results show that the nurses tended to over-rate (n=72, 65 per cent) rather than to under-rate (n=25, 23 per cent) the patient’s risk of developing a pressure ulcer, with only 13 of the 110 nurses accurately rating the patient’s score as 18. The Wilcoxon Signed Rank Test rejected the null hypothesis that there was no difference in the risk scores arrived at by individual nurses and the patient’s actual score (T=827, P<0.001 highly significant). T is the test statistic and is a measure of whether the median of the nurses’ score is different from the ‘gold standard’, and P is the probability of the result being obtained by chance where <0.001 means that probability is less than one in a thousand.

**Discussion**

Clearly part of the problem in this exercise is that the nurses did not have the actual patient in front of them and the case study does not provide the
**Box 1**

**Case history**

**Name:** Bertha  
**Next of kin:** Husband  
**Date of birth:** March 24 1926  
**Reason for admission:** For open reduction of comminuted Colles’ fracture of right wrist under general anaesthetic on June 6 2004  
**Past medical history:**  
1954 – Tonsillectomy  
1971 – Abdominal hysterectomy and bilateral removal of ovaries  
1991 – Right-sided stroke from which she made a full recovery  
**Vital signs on admission:** Pulse 89; blood pressure 175/95mmHg; temperature 36.4°C; urine no abnormalities detected; weight 63kg (138lb)  
**Operation notes recorded in care plan:** Bertha returned to the ward after open reduction of the Colles’ fracture. The procedure took two and half hours due to a carpal tunnel defect. Sutures to skin. External fixator  
**Mental state:** Easy-going and cheerful  
**Vital signs for June 9:** Pulse 95; blood pressure 180/95mmHg; temperature 38.5°C and having rigors  
**Diet for June 8:**  
Breakfast – two cups of tea  
10am – one cup of tea and one rich tea biscuit  
Lunch – two sausages, mashed potato, mushy peas, redcurrant jelly  
3pm – one cup of tea and one rich tea biscuit  
Supper – consommé, one slice of bread and butter and two cups of tea  
8pm – one cup of tea  
**Elimination:** Incontinent of urine; constipated  
**Medication:**  
Diclofenac 50mg – orally – as necessary  
Ferrous sulphate 200mg – orally – three times daily  
Vancomycin 500mg – intravenously – six hourly

Level of information that could have been obtained from a real patient. However, the nurses were all supplied with exactly the same information, thereby controlling for possible variance due to differing levels of knowledge possessed by staff about the patient being assessed (Watkinson 1996). This suggests that the lack of inter-rater reliability and the extreme range of scores might be due to the nurses not using the tool as it was intended (Waterlow 1995). This hypothesis is supported by looking at how the nurses arrived at their scores, and admission that they were unaware of written guidance on how the Waterlow assessment tool should be used.

To give Bertha a score, the nurse needs to calculate the patient’s body mass index (BMI), which is not easy to do without a calculator. The nurses needed to ask for the patient’s height, as this was intentionally not provided in the case study, however few did. The patient’s height is 1.83m (BMI of 18.8), making her BMI below average. Many nurses scored her as average. Once the nurse knows that this score is based on BMI, reliability should be 100 per cent, especially if BMI tables are used and the nurses are not required to do the arithmetic. It is, therefore, interesting that in Edwards’s (1995) study there was 22.5 per cent disagreement (n=9 out of 40 patients) between the two researchers when assessing this risk factor. Waterlow (2005) in her revised tool has clarified this aspect of the tool by giving the formula for calculating BMI and the BMI ranges for the different build scores.

The issue of continence was problematic. The patient does not fit into any of Waterlow’s categories as she is described as incontinent. This problem was highlighted in Cook et al’s (1999) study. Some nurses correctly argued that being incontinent was worse than being occasionally incontinent as far as risk of pressure ulcer development was concerned and so scored the patient as two. Two means her risk of developing pressure ulcers from continence problems is not as low as when she is fully continent (score 0) but not as high as if she is incontinent of both urine and faeces (score 3).

For skin type some nurses assumed Bertha was healthy and gave her zero, some argued she was an older person so she must have paper thin skin, while others correctly identified that she was pyrexial and having rigors, so she would have clammy skin. Of concern were those nurses who gave her a score of three because she had a wound, even though it was a surgical wound, as surgical wounds do not on their own increase risk of pressure ulcer development, but do so by affecting mobility (Morison 2001). In Edwards’s (1995) study it was this category of skin type that caused the most disagreement (55 per cent or n=22 out of 40 patients) between the two researchers. Waterlow (2005) in her revised tool has again clarified this aspect by identifying that ‘discoloured’ refers to grade 1 ulcers, while broken spot refers to grade 2-4 on the European Pressure Ulcer Advisory Panel Scale (EPUAP 1998).

Mobility was difficult to assess because the nurses did not see the patient, and the whole range of scores from fully mobile to chairbound were allocated. However, it would appear that even when a patient is present agreement is still problematic (Edwards 1995). Age and sex gave the nurses no problems, except for three who did not do the arithmetic correctly to work out that Bertha was 78 years old. However, it has been suggested that gender should be removed from the tool as it does not significantly predict risk (Anthony et al 2003, Papanikolaou et al 2002).
For appetite most nurses awarded Bertha with a score of zero, as they believed that her appetite was average for an older person. However, Waterlow (1996) indicates that an average appetite constitutes eating three meals a day, and Bertha did not fulfil this criterion and thus is classified as having a poor appetite. This section has been considerably revised in the new Waterlow (2005) tool with the incorporation of the malnutrition screening tool (Ferguson et al 1999).

In the special risk category Bertha has anaemia as demonstrated by her prescription for iron tablets. Twenty three (21 per cent) of the nurses missed this. Many (n=32, 30 per cent) did not know what terminal cachexia was, despite the fact that they were using the assessment tool on a daily basis. This demonstrates that some nurses are prepared to ignore things they do not understand rather than ask for clarification. Some nurses gave the patient a score of five for peripheral vascular disease (PVD) as they assumed that because she was old and had hypertension she also had PVD. Also, many nurses gave Bertha a score of between four and six because she had had a stroke in the past, not recognising that she had made a full recovery and so it did not increase her present risk of developing pressure ulcers. One criticism of Waterlow’s tool is that neurological deficit is scored between four and six which allows for ‘individual interpretation by each assessor’ and, hence, possibly decreased inter-rater reliability (Cook et al 1999).

Bertha’s operation took more than two hours, and so many nurses gave her a score of five. However, Waterlow (1996) maintains that this score ‘is only to be awarded for the period of 48 hours following surgery, unless the patient remains severely ill’ and she has made this clear on her revised tool (Waterlow 2005). Similarly, awarding a score of four for being on anti-inflammatory drugs (diclofenac) is not appropriate as these only increase pressure ulcer risk if they are taken over a long period of time, and the ‘as necessary’ suggests that Bertha is only taking them for post-operative pain control.

**Use of categories**

This small study indicates poor inter-rater reliability, although as Waterlow’s assessment tool categorises patients into three groups this is not quite as bad as it initially appears. A total of 30 (27 per cent) nurses did classify Bertha correctly as being of ‘high risk’, 12 (11 per cent) put her in the lower category of ‘at risk’ and 68 (62 per cent) thought she was ‘very high risk’. If the aim of using a risk assessment score is to prevent pressure ulcer development, and Waterlow’s assessment tool was designed to go with a prevention and treatment policy, then this over scoring should result in an improved prevention and treatment policy being implemented. However, as resources in NHS trusts are limited, this arguably unwarranted demand on resources will not be acceptable to budget holders, especially if the Waterlow scale over predicts risk (Hamilton 1992).

However, further studies are required to identify if equipment allocation reflects risk score allocation. Schoonhoven et al (2002) questioned this belief, and the draft National Institute for Clinical Excellence (NICE 2003) guidelines state that the choice of pressure-relieving equipment ‘should not be based solely on the scores from risk assessment scales’ but on a holistic patient assessment. The practice of allocating resources based on a patient’s risk assessment score might be fundamentally flawed. There is no evidence to link a particular score to particular equipment.

### TABLE 1

<table>
<thead>
<tr>
<th>Waterlow risk score</th>
<th>Number of nurses obtaining score</th>
<th>Number of nurses in each risk category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 11</td>
<td>0</td>
<td>At risk – 12</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>5</td>
<td>High risk – 30</td>
</tr>
<tr>
<td>17</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>18*</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>4</td>
<td>Very high risk – 68</td>
</tr>
<tr>
<td>20</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>110</td>
<td></td>
</tr>
</tbody>
</table>

* Gold standard score
and it could be argued that complicated algorithms might be an unnecessary distraction that delays implementation of care, and might actually contribute to an increased risk of pressure ulcer development. If risk assessment tools are used only as ‘an aide-memoire and are not intended to replace clinical judgement’ (NICE 2003), then, perhaps, all they need to do is help to identify whether or not patients are ‘at risk’ and not the degree of that risk. If this is the case the results of this study are not quite so concerning. If a score of 15 or more equals ‘at risk’, then only 11 participants (10 per cent) wrongly classified Bertha as not being ‘at risk’.

**Conclusion**

This study identified poor inter-rater reliability when using the Waterlow risk assessment tool.

**References**


Part of the problem is that nurses are not using the tool in the way it was intended. It was hoped that the nurses involved in this study could be followed up after six months to see if inter-rater reliability had improved as a result of taking part in this study and having discussed how the Waterlow tool was supposed to be used. However, an administration error resulted in the loss of the nurses’ contact details, and so a follow-up post test was not possible. However, this could form the focus for future study to see if active teaching on how to use Waterlow’s tool improves inter-rater reliability. It will be interesting to see if the additional guidelines on how to score the patient, written on the revised tool (Waterlow 2005), improve inter-rater reliability NS.

**Acknowledgement**

The author would like to thank Gibson D’Cruz, University of East Anglia, for his generous help with the statistics.