Reliability of pressure ulcer classification and diagnosis

Jane Nixon BSc MA PhD RGN
Deputy Director, Clinical Trials Research Unit, University of Leeds, Leeds, UK

Helen Thorpe BSc MSc
Senior Medical Statistician, Clinical Trials Research Unit, University of Leeds, Leeds, UK

Helen Barrow RGN
Pressure Trial CRN Co-ordinator, Seacroft Hospital, The Leeds Teaching Hospitals NHS Trust, Leeds, UK

Angela Phillips
Senior Trial Co-ordinator, Clinical Trials Research Unit, University of Leeds, Leeds, UK

E. Andrea Nelson BSc PhD RN
Senior Research Fellow, Centre for Evidence Based Nursing, Department of Health Sciences, University of York, York, UK

Susan A. Mason PhD RGN DN HV
Principal Research Fellow, Clinical Trials Research Unit, University of Leeds, Leeds, UK

Nicky Cullum PhD RGN
Professor, Centre for Evidence Based Nursing, Department of Health Sciences, University of York, York, UK

Accepted for publication 30 September 2004

Correspondence:
Jane Nixon,
Clinical Trials Research Unit,
University of Leeds,
17 Springfield Mount,
Leeds,
LS2 9NG,
UK.
E-mail: j.e.nixon@leeds.ac.uk

NIXON J., THORPE H., BARROW H., PHILLIPS A., NELSON E.A., MASON S.A. & CULLUM N. (2005) *Journal of Advanced Nursing* **50**(6), 613–623

Reliability of pressure ulcer classification and diagnosis

Aim. To assess the inter-rater reliability of the PRESSURE Trial pressure ulcer diagnosis (≥Grade 2) and skin classification for all grades between the clinical research nurse (CRN) team leader and CRNs working across different hospital sites; and CRNs and ward-based nurses.

Background. The United Kingdom National Health Service Health Technology Assessment Programme has funded a multi-centre, randomized controlled trial to compare the clinical and cost-effectiveness of alternating pressure mattress overlays and mattress replacements – PRESSURE Trial. Outcome skin assessments were recorded by qualified ward-based nurses daily, and expert CRNs twice weekly.

Method. Paired assessments were undertaken and skin assessed on seven body sites. The per cent agreement between nurses in the diagnosis of a pressure ulcer was determined and the Kappa statistic and confidence intervals calculated. Per cent agreement between nurses in classifying skin for all grades was also determined.

Results. Assessments were undertaken by 378 pairs: 16 paired patient assessments (107 site comparisons) by the CRN team leader and CRNs, and 362 paired patient assessments (2396 site comparisons) between CRNs and ward-based nurses. There was 100% agreement between the CRN team leader and CRNs in the diagnosis of a pressure ulcer, and the Kappa statistics indicated 'very good' agreement. There were only two (1·9%) disagreements in classifying skin for all grades between these

nurses. The agreement in the diagnosis of a pressure ulcer between CRNs and ward-based nurses varied by skin site, ranging from 93·6% to 100%, with the Kappa statistics indicating 'good' and 'very good' agreement. However, there were 508 (21·2%) disagreements in classifying skin for all grades.

Conclusions. Overall agreement and Kappa statistics indicated 'very good' and 'good' agreement between expert nurses, and between expert nurses and qualified ward-based staff, respectively. However, the high prevalence of normal skin concealed clinically important disagreements in both the diagnosis of pressure ulcers and skin classification for all grades.

Keywords: classification, inter-observer variation, nursing, pressure sore, pressure ulcer, reliability

Background

Pressure ulcers have been described as 'an area of localized damage to the skin and underlying tissue caused by pressure, shear and friction or a combination of these' (EPUAP 1999). They are complex lesions of the skin and underlying structures, and vary considerably in size and severity. The majority of pressure ulcers occur below the waist, with particularly vulnerable areas being the sacrum, buttocks and heels. The principal causative factor is localised pressure on an area of skin not adapted to the magnitude of such external forces.

Pressure ulcers have both cost and quality implications for health services and, whilst they are seen as largely preventable, there is no reliable body of evidence from high quality randomized controlled trials as to the best strategies for preventing them (Cullum *et al.* 1995, 2000).

The United Kingdom (UK) National Health Service (NHS) Health Technology Assessment Programme has funded a multi-centre, randomized controlled trial to compare the clinical and cost-effectiveness of alternating pressure mattress overlays with alternating pressure mattress replacements – the PRESSURE Trial. A number of methodological problems arise in the design and conduct of such a trial, not least the inability to blind the treatment allocations. Lack of blinding has the potential to bias outcome assessment. Options including daily assessments away from the bed, ordinary photography and high-resolution digital photography were considered, but these raised unacceptable inconvenience to and burden on patients. To minimize the potential for bias it was decided that qualified ward-based nursing staff (WN) would record daily skin assessments and Clinical Research Nurses (CRNs) would undertake assessments twice weekly in order to validate ward staff records, ward staff remaining blind to the CRN record. We investigated the inter-rater reliability of the skin classification scale as skin condition was the main endpoint for the PRESSURE Trial.

Reliability of pressure ulcer classification

The severity of pressure ulcers varies from erythema of intact skin to tissue destruction involving skin, subcutaneous fat, muscle and bone, and a number of classification systems have been developed (Witkowski & Parish 1981). The purpose of a pressure ulcer classification system is to standardize record-keeping and provide a common descriptor of ulcer severity for the purposes of clinical practice, audit and research.

At an international level, attempts to standardize classification have resulted in consensus between the American Agency for Health Policy and Research [Agency for Health Care Policy and Research (AHCPR) 1992] and the European Pressure Ulcer Advisory Panel (EPUAP 1999) in their pressure ulcer classifications. However, pressure ulcer classification is based on the clinical manifestations and tissue layer affected rather than underlying histopathology, and there is an ongoing debate about the description, inclusion and clinical assessment of erythema as a Grade/Stage 1 pressure ulcer (Bethell 2003). In addition, previous research has demonstrated problems in the inter-rater reliability of pressure ulcer classification scales when used by a large number of clinical staff (Nixon *et al.* 1998).

Early prevalence studies reported problems with the reliability of pressure ulcer classification. Barbenel *et al.* (1977) reported 'there were major disagreements in the reporting of Grade 1 sores, and this grade was discarded from the survey analysis'. In a subsequent study by David *et al.* (1983) ward nurses identified patients with pressure ulcers and graded skin areas from memory. Researchers then assessed 1589 skin sites on 821 patients to verify the site and grade of pressure ulcers. Of the 1589 skin sites assessed, there was agreement for only 778 (49·0%) sites.

More recent research shows wide variability in both simple percentage and chance-corrected rater agreement. Chance-corrected agreement can be calculated using the Kappa

Table 1 The Kappa statistic (adapted from Altman 1991)

Value of Kappa (κ)	Strength of agreement
< 0·20	Weak
0·21–0·40	Fair
0·41–0·60	Average
0·61–0·80	Good
0·81–1·00	Very good

statistic, which has a value range from 0 to 1 such that 1 is perfect agreement and 0 indicates only chance agreement (Altman 1991). Various guidelines are used to interpret the Kappa statistic and classify findings, for example as good, average, fair and weak (see Table 1). Buntinx *et al.* (1995) examined agreement between six observers who clinically assessed and graded 27 ulcers using one classification scale. They reported inter-observer percentage agreement for all pairs of observers ranging from 40% to 80% and an overall Kappa statistic of 0·42 or 'fair' agreement. There were, however, methodological weaknesses with this study, including use of a small number of assessors and patients for repeated clinical assessments.

Healey (1995) asked 109 clinical nurses to grade 10 photographs using four classification scales. The agreement between raters for the four classification scales ranged from 39% to 67%. Severe ulcers were found to have the highest reliability, and worst agreement was observed in the classification of skin redness. Chance-corrected agreement for the four classification scales ranged from 0·37 to 0·22 or 'fair' agreement. In a similar study by Russell and Reynolds (2001), 97 nurses graded 12 digital photographs using two classification scales and there was 30·2% and 61·9% agreement between raters for the two scales. Chance-corrected agreement was not reported. Both these studies have methodological weaknesses, including the use of photographs and a small number of pressure ulcers for repeated assessment.

Two further studies report inter-rater reliability, undertaken to prepare staff for participation in research. In a two-centre randomized controlled trial, a pre-trial assessment of inter-rater reliability was conducted. Of 133 paired assessments undertaken by two CRNs and 92 hospital ward-based nurses, there was 97·8% agreement between the CRN and ward nurse assessments. Chance-corrected agreement was not reported. The level of agreement was reduced for co-assessments undertaken during the trial and, of 171 paired assessments all undertaken by ward staff, there was 91·5% agreement (Nixon *et al.* 1998). It was suggested that, whilst planned inter-rater reliability assessments involving self-selected staff and immediately following training give good inter-rater reliability, data collected during the normal course of daily practice may have reduced reliability.

A pilot study conducted during the preparation for a large multi-centre prevalence survey assessed inter-rater reliability in nursing home, hospital and home care settings (Bours *et al.* 1999). In the nursing home and hospital a total of 344 (on 23 patients) and 674 (on 45 patients) observations were made, respectively, by two nurses on the same occasion. In the home care setting a wound care nurse made second assessments for a total of 1348 (on 90 patients) observations. Nurses were in agreement about the staging of pressure ulcers in 94% (nursing home), 99·7% (hospital) and 98% (home care) of patients, and chance-corrected agreement using Kappa was 0·97 (nursing home), 0·81 (hospital) and 0·49 (home care). Bours *et al.* (1999) highlight the difficulties in the interpretation of agreement using Kappa when a large proportion of patients have no pressure ulcers. Disagreements were largely related to the classification of normal skin and Stage 1 pressure ulcers, which were defined as 'discolouration of intact skin – light finger pressure applied to the site does not alter the discolouration'.

PRESSURE Trial skin classification and endpoints

The classification scale used in the PRESSURE Trial was adapted from international classification scales (AHCPR 1992, EPUAP 1999) in order to meet practical data collection requirements for the purpose of research (Table 2). Specifically, Grade 0 (no skin changes) was included to clearly distinguish skin assessment of normal skin from missing data. Grade 5 (black eschar) was included as a separate grade until wound debridement enabled classification by tissue layer. In addition, blanching and non-blanching erythema were recorded and classified as Grade 1a and 1b respectively and were referred to as 'skin changes' (Nixon *et al.* 1999).

A pressure ulcer and the primary endpoint for the PRESSURE Trial is defined as the development of a new pressure ulcer (Grade 2 or above) after randomisation and before discharge or trial completion. In addition, secondary endpoints include the time to development of new ulcers and

Table 2 Pressure ulcer and skin classification scale (adapted from Nixon *et al.* 1999)

Grade	Description
0	No skin changes
1a	Redness to skin (blanching)
1b	Redness to skin (non-blanching)
2	Partial thickness wound involving epidermis/dermis only (i.e. skin break or blister)
3	Full thickness wound involving subcutaneous tissue
4	Full thickness wound through subcutaneous tissue to muscle or bone
5	Black eschar

the maximum grade of new pressure ulcers. Healing endpoints include time to healing. The definition of a pressure ulcer as a Grade 2 or above is used because of concerns about the reliability and validity of non-blanching erythema, and the need to minimize the potential for bias in the trial, as assessors were not blind to the mattress interventions.

Whilst not included in the primary endpoint, Grade 1 skin changes were classified and recorded. Non-blanching erythema is an important independent predictor of Grade 2 pressure ulcer development, increasing the odds approximately six-fold (Allman *et al.* 1995, Nixon 2001). Therefore secondary analysis for the PRESSURE Trial will include adjustment for Grade 1b at baseline.

There is evidence that there are pathological differences between normal skin and blanching erythema (Witkowski & Parish 1981, Nixon 2001) and for this reason blanching erythema is classified as Grade 1a, that is, distinct from normal skin. There is very limited prospective evidence and it is unclear whether blanching erythema is predictive of subsequent pressure ulcer development. Use of this data in the PRESSURE Trial will be exploratory.

It was important, therefore, to assess the reliability of the diagnosis of a pressure ulcer (Grade 2 or above) and skin classification for all grades (secondary endpoints and secondary analysis).

The study

Aim

The aim of this study was to assess the inter-rater reliability of the PRESSURE Trial pressure ulcer diagnosis (Grade 2 or above) and skin classification for all grades:

- between the CRN team leader and CRNs, working across different hospital sites and
- between CRNs and WNs.

Design

A multi-centre inter-rater reliability study was designed. Patients and WNs from medical, elder care, orthopaedic and vascular surgical wards across eight hospitals sites (four NHS Trusts) were invited to participate in the study by the PRESSURE Trial CRNs.

Participants

Patients who were aged over 18, bedfast or chairfast on the day of the CRN ward visit and able to provide consent were invited to participate. Paired patient assessments were undertaken and skin assessed on seven body sites including the sacrum, left and right buttocks, left and right hips and left and right heels. Assessed skin was graded using the classification scale detailed in Table 2.

Preparation of nurses

The CRNs were all experienced clinical nurses with at least 3 years post-registration experience in care of older people, medical, vascular surgery or orthopaedic nursing and an interest in tissue viability. They were given additional preparation in skin assessment using the skin classification scale (Table 2). This included provision of the study protocol, published articles detailing the skin classification scale and clinical assessment methods, participation in a 2-day training programme during study set-up (including discussion of skin assessment and issues of reliability) and one-to-one discussion with the CRN team leader.

The WN preparation included one-to-one or small group explanations of the study's skin classification scale by the CRNs, emphasizing differences from any scale in clinical use. Information about the study was provided for each ward, including a study protocol, a poster giving details of the study, and a poster detailing the skin classification scale including a description and photographs for each grade.

Data collection

CRN team leader and CRN agreement

The CRN team leader made a planned site visit and, together with the CRN, recruited patients from the research wards including older, medical, orthopaedic and vascular inpatients. Skin inspection was performed simultaneously by both assessors, but recorded separately. Up to four patients were assessed by both nurses, and where possible this included at least one patient with a pressure ulcer. The CRN team leader returned all documentation for analysis.

CRN and WN agreement

The CRNs made planned ward visits to assess four patients with each WN who had received an explanation of both the study and skin assessment scale and had agreed to participate in the pre-trial inter-rater reliability study. Patient recruitment and assessments were undertaken as detailed above and the CRNs returned all documentation for analysis.

Ethical considerations

Ethical approval for the PRESSURE Trial protocol, which included the inter-rater reliability assessments, was obtained

from a Multi-centre Research Ethics Committee and the Local Research Ethics Committee of each participating centre.

Permission to approach WNs for study participation was given by nurse managers and ward managers. Ward nurses were approached by the CRNs or CRN team leader, shown ward-based information about the trial and inter-rater reliability study (protocol) and given verbal information about the inter-rater reliability study. Participation of WNs was voluntary and the right to refuse participation without giving a reason was respected.

Permission to approach patients for study participation was given by the ward nurse-in-charge. Information about the study was given to patients by the CRN or CRN team leader and patient consent obtained prior to participation. The right of patients to refuse without giving reasons was respected. Further, patients remained free to withdraw at any time, without giving reasons and without prejudicing any further nursing care or treatment.

Data analysis

To assess the inter-rater reliability of the pressure ulcer diagnosis, the per cent agreement between nurses in grading a skin site with either a pressure ulcer (Grade 2 and above) or no pressure ulcer (Grade 0, 1a and 1b) was determined and the Kappa statistic calculated. To assess the inter-rater reliability of skin classification for all grades, assessments for all skin sites were pooled and percent agreement between nurses in classifying skin for all grades (Grades 0, 1a, 1b, 2, 3, 4 and 5) was determined.

The Kappa statistic can verify that agreement exceeds the level of agreement that is likely to happen by chance. It requires independency of patients and is influenced by the prevalence within categories. Chance agreement is more likely to happen if there is a small number of assessors classifying a small number of skin areas. Therefore, patients were only included in the study on one occasion, and the inclusion criteria aimed to obtain a patient sample which included at least one in four patients with an existing pressure ulcer of Grade 2 or above. In addition, nurses from all pressure trial wards were included.

Reporting per cent agreement is also important because Kappa is dependent on the prevalence of the categories, and values of Kappa generated from different studies are not easily comparable (Altman 1991). Assessments for the seven skin sites were analysed separately, assuming each skin site to be independent, and strength of agreement was categorized using established guidelines (Table 1). Assessments were also pooled and analysed overall.

Results

A total of 378 paired assessments was undertaken by 116 nurses for the inter-rater reliability assessments during the period from December 2000 to February 2001. These included 16 paired assessments between the CRN team leader and four CRNs, and 362 paired assessments between six CRNs and 109 WNs. This generated data for 2646 skin sites: 112 site comparisons between the CRN team leader and CRNs, and 2534 site comparisons between the CRNs and WNs.

Excluding site comparisons with missing data (because of the presence of dressings or limb amputation, for example) resulted in a final sample of 2503 skin site comparisons. This included 107 site comparisons on 16 patients between the CRN team leader and the CRNs, and 2396 site comparisons on 362 patients between the CRNs and WNs.

CRN team leader and CRN agreement

Pressure ulcer diagnosis

The per cent agreement in the diagnosis of a pressure ulcer between nurses and corresponding Kappa statistics for the seven skin sites and overall are given in Table 3. There was 100% agreement for all skin sites between the CRN team leader and the four CRNs, and the Kappa statistics indicate 'very good' agreement for all sites in relation to the assessment of pressure ulcer/no pressure ulcer. Confidence intervals for the Kappa statistics are not reported due to the 100% agreement between the CRNs, resulting in standard errors of zero for each Kappa statistic, and hence the upper and lower 95% confidence limits for each statistic are equal to 1·0.

Skin classification – all grades

Agreement between the CRN team leader and CRNs for the 107 paired site assessments are shown in Table 4. There was a total of two (1·9%) disagreements between the CRN team

Table 3 Pressure ulcer diagnosis: Clinical Research Nurse (CRN) team leader and CRN agreement

Skin site	Percentage of agreement	Kappa statistic
Sacrum	100 (16/16)	–*
Left buttock	100 (16/16)	1·0
Right buttock	100 (16/16)	1·0
Left heel	100 (14/14)	1·0
Right heel	100 (16/16)	1·0
Left hip	100 (14/14)	–*
Right hip	100 (15/15)	–*
All areas	100 (107/107)	1·0

*A Kappa statistic is not given for these particular skin sites as all nurses graded patients as having no pressure ulcer; hence there is only one non-zero level in the 2 × 2 table.

Table 4 Skin classification – all grades: Clinical Research Nurse (CRN) team leader and CRN agreement – all sites

	CRN assessment							
Grades	0	1a	1b	2	3	4	5	Total
CRN team leader assessment								
0	47	0	0	0	0	0	0	47
1a	1	30	0	0	0	0	0	31
1b	0	1	18	0	0	0	0	19
2	0	0	0	6	0	0	0	6
3	0	0	0	0	4	0	0	4
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
Total	48	31	18	6	4	0	0	107

leader and CRNs. Both were only 1 grade different: Grades 0 and 1a (1), 1a and 1b (1). It is noteworthy that the areas of disagreement between the CRN team leader and CRNs were in relation to the assessment of normal skin, blanching and non-blanching erythema.

CRN and WN agreement

Pressure ulcer diagnosis

The per cent agreement in the diagnosis of a pressure ulcer between nurses and corresponding Kappa statistics for the seven skin sites and overall are given in Table 5. There was 93·6% to 100% agreement between CRNs and WNs. The Kappa statistics calculated indicate 'good' and 'very good' agreement; 95% confidence intervals for the Kappa statistics are reported and, in general, they confirm 'average' to 'very good' agreement. Due to the large sample some confidence intervals are narrow, but conversely due to the prevalence of the categories some are extremely wide (e.g. right hip) and therefore interpretation of the Kappa statistic is difficult.

Of the 2396 paired site assessments, there were 77 (3·2%) disagreements between CRNs and WNs in relation to the diagnosis of pressure ulcer (Table 6). The 77 disagreements were observed on 50 patients, 13·8% of patients assessed by CRNs and WNs. Disagreements included: both nurses recording a pressure ulcer but at different sites such as buttock and sacrum, left hip and right hip (seven patients); the CRN recording a pressure ulcer when the WN did not (24 patients); the WN recording a pressure ulcer when the CRN did not (14 patients); and both recording a pressure ulcer but one recording more than one ulcer (five patients). Disagreements were observed for all skin sites, apart from left hip, and there were fewer disagreements for hip and heel areas compared with buttocks and sacrum (Table 5).

The 77 disagreements were associated with 38 different WNs and, of these, 16 staff recorded one disagreement, 8 recorded two disagreements, 11 recorded three disagreements and 3 recorded four disagreements.

Table 5 Pressure ulcer diagnosis: Clinical Research Nurse and Ward Nurse agreement

Skin site	Percentage of agreement	Kappa statistic	95% confidence interval	
Sacrum	95·3 (322/338)	0·80	0·70	0·89
Left buttock	93·6 (334/357)	0·67	0·55	0·80
Right buttock	93·8 (334/356)	0·62	0·48	0·77
Left heel	96·5 (333/345)	0·78	0·66	0·90
Right heel	99·1 (342/345)	0·95	0·89	0·99
Left hip	100 (330/330)	1·00	1·00	1·00
Right hip	99·7 (324/325)	0·67	0·05	0·99
All areas	96·8 (2319/2396)	0·77	0·72	0·82

Skin classification – all grades

Agreement between CRNs and WNs for the 2396 paired site assessments for all grades is detailed for each site and pooled overall (all sites) in Tables 7–14. There were a total of 508 (21·2%) disagreements between CRNs and WNs: 419 were one grade different (such as 0/1a, 1a/1b and so on), 68 were two grades different [including 0 and 1b (21), 1a and 2 (46) and 3 and 5 (1)], and 21 were more than two grades different [including 0 and 2 (13), 0 and 3 (1), 1a and 3 (3), 2 and 5 (4)].

Table 6 Pressure ulcer diagnosis: Clinical Research Nurse (CRN) and Ward Nurse (WN) agreement

	WN		
	No pressure ulcer	Pressure ulcer	Total
CRN			
No pressure ulcer	2175 (90·8%)	35 (1·5%)	2210
Pressure ulcer	42 (1·8%)	144 (6·0%)	186
Total	2217	179	2396

Table 7 Skin classification – all grades: Clinical Research Nurse (CRN) and Ward Nurse (WN) agreement sacrum

	WN assessment							
Grades	0	1a	1b	2	3	4	5	Total
CRN assessment								
0	153	14	4	3	0	0	0	174
1a	27	58	9	3	0	0	0	97
1b	3	3	13	1	0	0	0	20
2	1	5	2	28	3	0	0	39
3	1	0	0	0	4	1	0	6
4	0	0	0	0	0	0	0	0
5	0	0	0	1	0	0	1	2
Total	185	80	28	36	7	1	1	338

Table 8 Skin classification – all grades: Clinical Research Nurse (CRN) and Ward Nurse (WN) agreement left buttock

	WN assessment							
Grades	0	1a	1b	2	3	4	5	Total
CRN assessment								
0	166	15	1	2	0	0	0	184
1a	29	70	12	7	0	0	0	118
1b	2	3	8	0	0	0	0	13
2	1	8	4	19	1	0	0	33
3	0	1	0	2	5	0	0	8
4	0	0	0	0	0	1	0	1
5	0	0	0	0	0	0	0	0
Total	198	97	25	30	6	1	0	357

Table 9 Skin classification – all grades: Clinical Research Nurse (CRN) and Ward Nurse (WN) agreement right buttock

	WN assessment							
Grades	0	1a	1b	2	3	4	5	Total
CRN assessment								
0	162	18	2	2	0	0	0	184
1a	31	77	14	9	0	0	0	131
1b	0	2	7	2	0	0	0	11
2	3	4	1	16	0	0	0	24
3	0	1	0	2	1	1	0	5
4	0	0	0	0	1	0	0	1
5	0	0	0	0	0	0	0	0
Total	196	102	24	31	2	1	0	356

Table 10 Skin classification – all grades: Clinical Research Nurse (CRN) and Ward Nurse (WN) agreement left hip

	WN assessment							
Grades	0	1a	1b	2	3	4	5	Total
CRN assessment								
0	303	8	0	0	0	0	0	311
1a	6	7	3	0	0	0	0	16
1b	0	0	0	0	0	0	0	0
2	0	0	0	1	0	0	0	1
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	2	2
Total	309	15	3	1	0	0	2	330

Table 11 Skin classification – all grades: Clinical Research Nurse (CRN) and Ward Nurse (WN) agreement right hip

	WN assessment							
Grades	0	1a	1b	2	3	4	5	Total
CRN assessment								
0	295	7	3	0	0	0	0	305
1a	5	8	3	0	0	0	0	16
1b	1	0	1	0	0	0	0	2
2	0	1	0	1	0	0	0	2
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
Total	301	16	7	1	0	0	0	325

Table 12 Skin classification – all grades: Clinical Research Nurse (CRN) and Ward Nurse (WN) agreement left heel

	WN assessment							
Grades	0	1a	1b	2	3	4	5	Total
CRN assessment								
0	81	14	0	0	0	0	0	95
1a	45	106	12	2	1	0	0	166
1b	3	18	30	2	0	0	0	53
2	0	7	0	15	1	0	2	25
3	0	0	0	0	1	0	0	1
4	0	0	0	0	0	0	0	0
5	0	0	0	1	0	0	4	5
Total	129	145	42	20	3	0	6	345

Table 13 Skin classification – all grades: Clinical Research Nurse (CRN) and Ward Nurse (WN) agreement right heel

	WN assessment							
Grades	0	1a	1b	2	3	4	5	Total
CRN assessment								
0	79	16	0	0	0	0	0	95
1a	44	116	12	0	0	0	0	172
1b	2	21	23	1	0	0	0	47
2	1	0	1	15	0	0	0	17
3	0	0	0	2	3	0	0	5
4	0	0	0	0	0	0	0	0
5	0	0	0	0	1	0	8	9
Total	126	153	36	18	4	0	8	345

Limitations

The levels of agreement between CRNs and WNs raise important issues in relation to the limitations of summary measures for inter-rater agreement and problems associated with the diagnosis of early pressure ulcers (both Grade 1b and Grade 2) which affect trial design, monitoring clinical performance and professional accountability.

As a test statistic, Kappa can verify that agreement exceeds chance levels; however, there has been controversy over its use to quantify the level of agreement among two or more

Table 14 Skin classification – all grades: Clinical Research Nurse (CRN) and Ward Nurse (WN) agreement all sites

	WN assessment							
Grades	0	1a	1b	2	3	4	5	Total
CRN assessment								
0	1239	92	10	7	0	0	0	1348
1a	187	442	65	21	1	0	0	716
1b	11	47	82	6	0	0	0	146
2	6	25	8	95	5	0	2	141
3	1	2	0	6	14	2	0	25
4	0	0	0	0	1	1	0	2
5	0	0	0	2	1	0	15	18
Total WN	1444	608	165	137	22	3	17	2396

raters (Byrt *et al.* 1993, Guggenmoos-Holzmann 1993, Lantz & Nebenzahl 1996, Nelson & Pepe 2000). One of the difficulties is that Kappa can be affected in complex ways by the presence of bias between raters (Byrt *et al.* 1993). In this study, however, there is approximate symmetry between the two discordant proportions (Table 6). Hence there appears to be no systematic difference in the way the nurses use the skin classification scale. That is, WNs do not appear to under- or over-estimate any more than CRNs, and there is no obvious bias.

Another difficulty associated with the use and interpretation of Kappa is that its value depends on the proportion of subjects (prevalence) in each category (Altman 1991). This is clearly a limitation in the present study, where the majority of skin sites have no pressure ulcer identified by either the CRN or WN (Table 6). This dependency of the Kappa statistic is particularly illustrated in the Kappa calculated for the right hip site (Table 5). Only one disagreement was observed (per cent agreement 99·7%), yet the 95% confidence interval of the Kappa statistic indicates that the true value of Kappa lies between 'weak' and 'very good' agreement.

Translated overall, the Kappa statistic for CRN and WN agreement for all skin sites pooled is 'good' (Table 5). If CRNs are taken as the 'gold standard', the proportion of pressure ulcers that are correctly identified by WNs is 144 of 186 (77·4%) (Table 6). Alternatively, the proportion of no pressure ulcers correctly identified by WNs is 2175 out of 2210 (98·4%). However, these percentages are influenced by the high prevalence of 'no pressure ulcer'.

Indeed, the high prevalence of skin areas assessed as having no pressure ulcer conceals the level of disagreement between CRNs and WNs in identifying pressure ulcers. Of the 186 pressure ulcers reported by CRNs, 42 (22·6%) are not identified by WNs (under-reporting) (Table 6). Despite this relatively poor agreement of pressure ulcer diagnosis, the Kappa statistic and its 95% confidence interval for all skin sites suggest 'good' agreement between raters. These proportions, however, assume that CRN assessments are always 'correct' and that within- and between-CRN variability does not exist. Clearly we cannot assume this, and so these results should not be over-interpreted.

Discussion

The good levels of agreement between the CRN team leader and CRNs suggest that CRNs are able to interpret clinical observations of skin changes (such as Grade 1a and 1b) and pressure ulcers in a consistent and reliable way. The two disagreements observed between the CRN team leader and CRNs were only one grade different and were associated with the assessment of normal skin, blanching erythema and non-blanching erythema, illustrating the difficulties of skin assessment even when undertaken by expert nurses. These results justify the definition of the PRESSURE Trial primary endpoint of a Grade 2 pressure ulcer or above.

Overall, this study suggests that, even when a pressure ulcer is defined as a Grade 2 skin lesion, there are clinically important differences in reporting by qualified WN and expert nurses. This level of disagreement is further increased when pressure ulcers are defined using the European and American pressure ulcer classification systems, which include non-blanching erythema, equivalent to a Grade 1b. If CRNs are assumed to be the 'gold standard' and Grade 1b skin lesions are defined as pressure ulcers, then the proportion of grade 1b ulcers not identified by WNs, resulting in the incorrect classification of 'no pressure ulcer' (i.e. skin sites assessed by the WNs as Grade 0 and 1a lesions), is 58 out of 146 (39·7%) (Table 14).

From a trial design perspective, this poor level of agreement between CRNs and qualified WN in assessing Grade 1b skin changes justifies the PRESSURE Trial endpoint defined as a Grade 2 skin lesion. Whilst some of the disagreements are simply the result of site confusion (for example, between left and right), in relation to trial design, the lack of reliability in Grade allocation by body site has serious consequences in the determination of outcome. The results highlight the need to use co-assessments or expert assessors to validate endpoints, estimate the number of misclassifications and undertake sensitivity analyses to assess the effect of misclassifications on the treatment difference in studies using qualified ward nurse assessments to record endpoints (Nixon *et al.* 1998).

In terms of the wider debate on the definition of a pressure ulcer and the description, inclusion and clinical assessment of Grade/Stage 1 pressure ulcers (Bethell 2003), it has been clearly determined that non-blanching erythema is associated

with a sixfold increase in risk of subsequent skin loss (Allman *et al.* 1995, Nixon 2001) and therefore is clinically important. It should be classified and recorded for practice in order to identify patients at risk of subsequent skin loss and to prompt active interventions. Our results indicate the need for further investigation of the impact of poor reliability in nursing assessment of non-blanching erythema upon patient outcome.

The difficulties in monitoring clinical performance using prevalence and incidence have been highlighted in National Guidelines (AHCPR 1992) and by a number of authors (Dealey 1991, Clark & Cullum 1992, Clark & Watts 1994, Bridel *et al.* 1996, McGough 1998). Problems associated with interpretation of clinical performance measures include different inclusion and exclusion criteria and definition of the population 'at risk' (that is, the denominator population); difficulties in establishing the denominator population from patient information systems, particularly in areas with a high patient through-put and ward transfers; the use of various classification scales with inclusion and exclusion of Grade/Stage 1 pressure ulcers; reliability of data sources, ranging from direct observation of patients by trained researchers to retrieval of data from patient records; and insufficient knowledge of risk factors to allow case-mix adjustment. Our results further challenge the reliability and validity of clinical performance monitoring that uses large numbers of clinically-based staff to identify, record and report pressure ulcer prevalence and incidence. The limitations of such data cannot be overemphasized and their value is highly questionable.

Professional issues are raised by our findings, including nursing skill and competence in skin assessment, documentation and record-keeping. The UK Nursing and Midwifery Council (NMC) states that Registered Nurses 'have a duty of care to patients and clients, who are entitled to receive safe and competent care' (NMC 2002a, p. 3) and should be able to demonstrate 'full account of your assessment and the care you have planned and provided' (NMC 2002b). The International Council of Nurses (ICN 1997, p. 44) defines competence as 'a level of performance demonstrating the effective application of knowledge, skill and judgement'. Specific competencies include: accurately interpreting objective and subjective data and their significance for safe delivery of care; carrying out relevant and systematic health and nursing assessment; analysing, interpreting and documenting data accurately; and evaluating data to modify care planning (ICN 2003).

The problem of poor standards of nursing documentation are highlighted by quality assurance and research reports (Gunningberg *et al.* 2000), but our results raise fundamental questions about nursing practice in this area of care. It is unclear whether poor nursing documentation is related to lack of time and attention to skin assessment, nursing competence and skill in assessing skin, difficulties in making an accurate diagnosis of pressure ulcer, or simply failing to document skin assessments clearly.

Our results suggest that the issues relate to both the nursing competence and skill in assessing skin, and difficulties in making an accurate diagnosis of a pressure ulcer. For example, over half [45 of 77 (58·4%)] of the disagreements were associated with only 14 of the 109 (12·8%) WNs and a small but clinically important number of skin areas were more than two grades different, including four pressure ulcers assessed as Grade 3 by one nurse and Grade 0 or 1a by another. This raises questions about the competency and skill of nurses in assessing skin. However, there were also a clinically important number of patients where the CRN recorded a Grade ≥2 pressure ulcer when the WN did not (24 patients, 6·6% of the 362 patients), and the WN recorded a pressure ulcer when the CRN did not (14 patients, 3·9% of the 362 patients). This suggests that the diagnosis of a pressure ulcer may be difficult to make.

The WNs involved in the study were self-selected and had received recent ward-level explanation and guidance on skin assessment and the classification scale used in the trial. They were able to refer to the scale during skin assessment and were aware that a concurrent skin assessment was being undertaken by a CRN. Further investigation exploring methods to improve the diagnosis of pressure ulcers, assessment of clinical competency, impact of clinical competency on decision-making and appropriateness of nursing care interventions.

Conclusions

Overall per cent agreement and Kappa statistics indicate 'very good' and 'good' agreement between expert nurses, and between expert nurses and qualified WN respectively, but clinically important disagreements in both the diagnosis of pressure ulcers and skin classification for all grades are concealed by the high prevalence of normal skin with no skin changes. The results raise important issues in relation to the limitations of summary measures for inter-rater agreement, problems associated with the diagnosis of early pressure ulcers (both Grade 1b and Grade 2) which affect trial design, monitoring clinical performance and professional accountability.

The study suggests that, even when a pressure ulcer is defined as a Grade 2 skin lesion, there are important differences in the reporting of ulcers by qualified WN and expert nurses. This level of disagreement is further increased

What is already known about this topic

- The inter-rater reliability of pressure ulcer classifications has been reported previously.
- Previous studies have methodological weaknesses, including patient numbers, use of photographs to grade skin, repeated assessments of the same patients, and reporting of simple agreement (that is, per cent agreement).

What this paper adds

- The limitations of summary measures used to quantify levels of agreement.
- There are clinically important differences in pressure ulcer reporting by qualified ward-based and expert nurses.
- Nurses have difficulty in making an accurate assessment and diagnosis of Grade 1 and Grade 2 pressure ulcers.

if the pressure ulcer definition includes non-blanching erythema. From a trial design perspective, the level of disagreement in assessing non-blanching erythema justifies the PRESSURE Trial endpoint defined as a Grade 2 skin lesion. In relation to monitoring clinical performance, we found important limitations of pressure ulcer reporting by clinical staff. The results suggest that further investigation exploring the most appropriate pressure ulcer definition for practical applicability and assessment of clinical competency in this field is required.

Acknowledgements

This study was supported by a grant from the National Health Service Research and Development Health Technology Assessment Programme. The views and opinions expressed in the paper do not necessarily reflect those of the NHS Executive. We are grateful to the PRESSURE Trial CRNs (Monica Clark, Richard Buckland, Ann Fotheringham, Fiona Corcoran, Dawn Parkes and Helen Marson) and staff of the four participating NHS trusts (Leeds Teaching Hospitals NHS Trust, North Tees and Hartlepool NHS Trust, York Health Services NHS Trust and Scarborough and North East Yorkshire Healthcare NHS Trust), including the nursing staff and patients who voluntarily gave their time.

Author contributions

JN, AP, EAW, SAM and NC conceived and designed the study. MB and AP collected the data. JN and HT analysed the data and drafted the manuscript. EAW, SAM and NC critically revised the paper. HT provided statistical expertise. JN, EAW, SAM and NC obtained funding. NC was the principal investigator.

References

Agency for Health Care Policy and Research (AHCPR) (1992) *Pressure Ulcers in Adults: Prediction and Prevention: Quick Reference Guide for Clinicians*. US Department of Health and Human Services, Rockville.

Allman R.M., Goode P.S., Patrick M.M., Burst N. & Bartolucci A.A. (1995) Pressure ulcer risk factors among hospitalized patients with activity limitation. *Journal of the American Medical Association* **273**, 865–870.

Altman D.G. (1991) *Practical Statistics for Medical Research*. Chapman and Hall, London.

Barbenel J.C., Jordan M.M., Nicol S.M. & Clark M.O. (1977) Incidence of pressure sores in the Greater Glasgow Health Board. *Lancet* **2**, 548–550.

Bethell E. (2003) Controversies in classifying and assessing grade 1 pressure ulcers. *Journal of Wound Care* **12**, 33–6.

Bours J.J.W., Halfens R.J., Lubbers M. & Haalboom J.R.E. (1999) The development of a national registration form to measure the prevalence of pressure ulcers in the Netherlands. *Ostomy and Wound Management* **45**(11), 28–40.

Bridel J., Banks S. & Mitton C. (1996) The admission prevalence and hospital-acquired incidence of pressure sores within a large teaching hospital during April 1994 to March 1995. In: *Proceedings of the 5th European Conference on Advances in Wound Management* (Cherry G.W. *et al.*, eds), Macmillan Magazines, London, pp. 83–85.

Buntinx F., Beckers H., Briers A., De Keyser G., Flour M., Habex M., Nissen G., Raskin T. & De Vet H. (1995) Inter-observer variation in the assessment of skin ulcers. In: *Proceedings of the 5th European Conference on Advances in Wound Management* (Cherry G.W., Gottrup F., Lawrence J.C., Moffatt C.J. & Turner T.D., eds) Macmillan Magazines, London, pp. 44–46.

Byrt T., Bishop J. & Carlin J.B. (1993) Bias, prevalence and kappa. *Journal of Clinical Epidemiology* **46**(5), 423–429.

Clark M. & Cullum N. (1992) Matching patient need for pressure sore prevention with the supply of pressure redistributing mattresses. *Journal of Advanced Nursing* **17**(3), 310–316.

Clark M. & Watts S. (1994) The incidence of pressure sores within a National Health Service Trust hospital during 1991. *Journal of Advanced Nursing* **20**(1), 33–36.

Cullum N., Deeks J., Fletcher A., Long A., Mouneimne H., Sheldon T. & Song F. (1995) The prevention and treatment of pressure sores: how effective are pressure-relieving interventions and risk assessment for the prevention and treatment of pressure sores? *Effective Health Care Bulletin* **2**(1), 1–16.

Cullum N., Deeks J., Sheldon T.A., Song F. & Fletcher J. (2000) Beds mattresses and cushions for pressure sore prevention and treatment. *The Cochrane Library* [Issue 2, Update Software].

David J.A., Chapman R.G., Chapman E.J. & Lockett B. (1983) *An Investigation of the Current Methods used in Nursing for the*

Care of Patients with Established Pressure sores Northwick Park. Nursing Research Unit, UK.

Dealey C. (1991) The size of the pressure sore problem in a teaching hospital. *Journal of Advanced Nursing* **16**, 663–670.

EPUAP (1999) *Pressure Ulcer Treatment Guidelines*. European Pressure Ulcer Advisory Panel, Oxford.

Guggenmoos-Holzmann I. (1993) How reliable are chance-corrected measures of agreement? *Statistics in Medicine* **12**, 2191–2205.

Gunningberg L., Lindholm C., Carlsson M. & Sjoden P.O. (2000) The development of pressure ulcers in patients with hip fractures: inadequate nursing documentation is still a problem. *Journal of Advanced Nursing* **31**(*5*), 1155–1164.

Healey F. (1995) The reliability and utility of pressure sore grading scales. *Journal of Tissue Viability* **5**(4), 111–114.

International Council of Nurses (1997) *ICN on Regulation: Towards 21st Century Models*. ICN, Geneva.

International Council of Nurses (2003) *An Implementation Model for the ICN Framework of Competencies for the Generalist Nurse*. ICN, Geneva.

Lantz C.A. & Nebenzahl E. (1996) Behavior and interpretation of the statistic: Resolution of the two paradoxes. *Journal of Clinical Epidemiology* **49**(4), 431–434.

McGough A. (1998) Epidemiological issues in monitoring pressure damage. *Journal of Wound Care* **7**(4), 214–215.

Nelson J.C. & Pepe M.S. (2000) Statistical description of interrater variability in ordinal ratings. *Statistical Methods in Medical Research* **9**, 475–496.

Nixon J.E. (2001) *Predicting and Preventing Pressure Sores in Surgical Patients*, Unpublished PhD Thesis. University of Newcastle-upon-Tyne, Newcastle-upon-Tyne.

Nixon J., McElvenny D., Mason S., Brown J. & Bond S. (1998) A sequential randomised controlled trial comparing a dry visco-elastic polymer pad and standard operating table mattress in the prevention of post-operative pressure sores. *International Journal of Nursing Studies* **35**(4), 193–203.

Nixon J., Smye S., Scott J. & Bond S. (1999) The diagnosis of early pressure sores: report of the pilot study. *Journal of Tissue Viability* **9**(2), 62–66.

Nursing and Midwifery Council (2002a) *Code of Professional Conduct*. NMC, London.

Nursing and Midwifery Council (2002b) *Guidelines for Records and Record Keeping*. NMC, London.

Russell L. & Reynolds T.M. (2001) How accurate are pressure ulcer grades? An image-based survey of nurse performance. *Journal of Tissue Viability* **11**(1), 67–75.

Witkowski J.A. & Parish L.C. (1981) Histopathology of the decubitus ulcer. *Journal of the American Academy of Dermatology* **6**, 1014–1021.