Investigating the Reliability and Validity of the Waterlow Risk Assessment Scale: A Literature Review

Breda Walsh¹ and Laura Dempsey²

Abstract
The aim of this review was to examine health literature on the reliability and validity of the Waterlow pressure sore assessment scale. A systematic review of published studies relating to the topic was conducted and literature was examined for its relevancy to the topic under investigation. Findings suggest that despite the availability of over 40 assessment tools, the Waterlow assessment scale is the most frequently used by health care staff. Research suggests that the Waterlow Scale is an unreliable method of assessing individuals at risk of pressure sore development with all studies indicating a poor interrater reliability status. Its validity has also been criticized because of its high-sensitivity but low-specificity levels.

Keywords
pressure ulcer, Waterlow Score, reliability, validity

Introduction and Background
A pressure ulcer also known as a pressure sore, bedsore, or decubitus ulcer in simple terms is caused “when an area of tissue is denied a flow of blood/
nutrients and a means of removing waste products, usually where the thickness of tissue between the surface of the skin and underlying bone is thin. This usually occurs when tissue is compressed for a length of time, when lying or sitting” (Waterlow, 2005a, p. S26). Pressure ulcers are costly both in terms of human suffering and resources. Gethin, Jordan-O’Brien, and Moore (2005) estimate that it costs a quarter of a million euro’s per annum to manage pressure ulcers in hospital and community settings across Ireland. However, the costs of litigation and quality of life of the individual, in terms of pain, depression, and social isolation are not included in these estimates (Moore, Jordan-O’Brien, & Barrett, 2006). Therefore, prevention by means of a risk-assessment scale is crucial. There are several benefits in using a validated risk assessment scoring system according to Waterlow (2005a). A scale’s ease-of-use helps to ensure that all patients are assessed in determining their “risk status” in developing ulcers. Through establishing a patients “risk status,” a scale ensures appropriate nursing care and cost-effective measures are taken in alleviating the risk. Furthermore, a scale records a patients care requirements and monitors their progress through regular reassessments.

Despite the availability of over 40 assessment scales, the Waterlow Scale is the most widely used (Thompson, 2005). It was developed in 1984/1985 for use in medical and surgical settings and displays predisposing factors which include the following: build/weight for height, continence, skin type, mobility, sex, age, and appetite. It also has a special risk category, divided into tissue malnutrition, neurological deficit, major surgery/trauma, and medication. Each of these categories is allocated a score, which when added together determine the patients “risk” status. A score of more than 10 is classified as “at risk,” between 15 to 20 as “high risk” and more than 20 as “very high risk.” Finally, the scale offers guidance on general preventative and treatment guidelines.

Patients experience pain, discomfort, suffering, and altered body image because of pressure ulcers and are time consuming and expensive in treatment regimes (Flanagan & Fletcher, 2003). The Waterlow Scale has been selected in determining if it is the most effective and reliable risk-assessment tool in identifying patients at risk, so that preventative measures may be introduced early and scarce resources directed appropriately. This topic is relevant to nursing as the Waterlow Scale is widely used by nurses in providing patient care and is the only familiar scale implemented within the clinical field, however it must be used accurately. This literature review aims to critically analyze the Waterlow Scale as a risk-assessment tool in identifying patients at risk of developing pressure ulcers.
Method

Reviewing literature provides a complete picture, which often remains partially hidden when a single piece of research is viewed in isolation (Aveyard, 2007). It also reviews and synthesizes evidence-based information to gain knowledge and improve nursing practice (Polit, Beck, & Hungler, 2001). A computerized literature search was conducted utilizing databases such as Applied Social Sciences Index and Abstracts, CINAHL, EMBASE.com, EBSCO, MEDLINE, ScienceDirect, SwetsWise Fulltext Journals, and Wiley Interscience Journals. Keywords used included the following: pressure ulcer, pressure sore, bedsore, decubitus ulcer, Waterlow, risk assessment, reliability, validity, and nurses. The Boolean operator term AND was used in broadening the search. The search was refined further by selecting only “full text” publication which reduced the volume of articles found in some databases. Many publications overlapped in databases whereas some databases provided solely summary articles. Publications which focused on pressure ulcer risk-assessment scales especially with reference to Waterlow and which were peer reviewed were utilized.

An incremental search was also carried out, whereby the reference lists of many publications were traced in finding new relevant studies. Some nursing and medical internet sites were explored while interlibrary loan was also utilized in finding research papers which were difficult to locate. A manual search was undertaken in nursing and medical libraries for books and journals pertaining to the subject.

Findings

Reliability of the Waterlow Scale. According to Mac Donald (1995), because of current staffing levels and patient throughput, it is highly unlikely that a patient will be assessed by the same nurse each time their condition is evaluated. A risk-assessment scale must be reliable regardless of the assessor. Many studies have evaluated the interrater reliability, where two independent raters or observers assign the same ratings/values for an attribute being measured (Polit & Hungler, 1997) of the Waterlow Scale. This is crucial in determining the scales consistency and stability. The Waterlow Scale and the nurse using it influence a patient’s risk status of pressure ulcer development.

Watkinson (1996) examined the interrater reliability of the Douglas, Braden, and Waterlow Scales finding the reliability of all three scales to be poor. In assessing the Waterlow Scale, the assessors found it to be user friendly and unambiguous. However, this statement has been criticized by Pang and
Wong (1998) who found the scale time consuming. Watkinson (1996) believes that the variance in scores may have resulted because of different levels of knowledge by staff members. This study also found that the scale placed patients at a higher risk of pressure ulcer development than either the Braden or Douglas Scales. This outcome may be because of the fact that the Waterlow Scale includes age as a risk factor and this study was undertaken in an older patients’ department.

Similarly, Cook, Hale, and Watson (1999) assessed the interrater reliability of an adapted Waterlow Scale. The trial suggested that there were limitations to using this scale in the elderly care area, therefore, minor changes were made to the “continence,” “neurological deficit,” and “medication” subscales of the original scale. Interrater reliability was assessed using two statistical techniques (percentage of agreement and correlation), which demonstrated a weak-to-moderate degree of interrater reliability. Surprisingly, the results found that no patient received scores which placed them in a single-risk category. As evident in Watkinson (1996), Cook et al (1999) believed variance in the scores might have been because of differing levels of knowledge regarding the patients being assessed. However, all nursing staff shared the same patient information, but not all used this, which had implications for the risk category into where the patient was placed. This study also analyzed the subscales scores for each patient to ascertain if the variation in the nurse rater scores were caused by difficulties in a particular section of the scale. The categories of “sex,” “age,” and “minor surgery” showed a high degree of consensus in scores, but the raters varied most in the categories of “mobility” and “skin type.”

Certain aspects of the Waterlow Scale have been criticized of not having clear descriptions and the outcome of the assessment scale is dependent on the arbitrary judgment of the assessor (Pang & Wong, 1998). However this statement may be true of several assessment scales and cannot be generalized to the Waterlow Scale. Cook et al. (1999) recommend similar studies should be conducted on the Waterlow Scale itself, as the principles of assessment and scoring are the same for all tools. Kelly (2005) determined whether lack of reliability of the Waterlow Scale is due to different perceptions of the patient by the assessors, or to different interpretations on how to use the scale. Findings showed that the nurses tended to overrate (65%) rather than underrate (23%) the patient’s risk of developing a pressure ulcer. According to Kelly (2005), part of the problem may be due to the absence of the actual patient, nonetheless, the study provided the assessors with the exact same information. Furthermore, this study highlights that nurses are not using the scale for the purpose intended. Take for example, the “special-risk category” of the scale. The patient described
in Kelly’s (2005) case study had anemia as demonstrated by her prescription for iron tablets. Twenty-three of the nurses failed to acknowledge this and many failed in their understanding of terminal cachexia, despite using the scale daily. This study planned a follow-up of the nurses that participated after 6 months, in determining if interrater reliability had improved. An administration error highlights a limitation in this study, whereby the nurses contact details were lost and a follow-up posttest was unachievable. Kelly (2005) believes that this could form the focus for future study to see if active teaching on utilization of the scale improves interrater reliability.

Edwards (1995) assessed the interrater reliability of the Waterlow Scale by undertaking a cross-sectional observation survey of 40 elderly patients with/without ulcers in a community setting. The findings failed to display high levels of reliability and in line with other studies overpredicted ulcer development. Categories where disagreement occurred the most were “skin type,” “build/weight for height,” and “mobility,” again indicating that the score is affected by a lack of adequate definitions in these categories. Furthermore, this study highlights that the elderly population proves difficult to obtain scores at the low end of the range, that is, a female over 80 years of age already has a score of 7 before any other risk factors are taken into account. This indicates a limitation of the scale that perhaps the risk thresholds are too high for these patients.

From examining the literature in relation to the reliability of the Waterlow Scale, especially in terms of its interrater reliability, the conclusion is that its reliability is poor. Studies suggest that variance of patient’s scores in determining their “at risk” status is because of a lack of clear descriptions within the categories of the scale, and to different levels of knowledge held by professionals. It is crucial that all nursing staff are fully educated on how the scale is intended to be utilized. Furthermore, the data set in these studies was small and recommends for further work to be undertaken with larger sample sizes.

**Validity of the Waterlow Scale.** The validity of the Waterlow Scale was examined by Pancorbo-Hidalgo, Garcia-Fernandez, Lopez-Medina, and Alvarez-Nieto (2006) who reviewed the effectiveness of risk-assessment scales used for pressure ulcer prevention. A total of 7 studies were analyzed on the Waterlow Scale’s validation and concluded that the scale has good pressure ulcer risk prediction capacity, offers a high-sensitivity (82.4%) but a low-specificity score (27.4%). Other studies that stress the scale’s high-sensitivity and low-specificity level include Pang and Wong (1998), Schoonhoven et al. (2002), and Balzer, Pohl, Dassen, and Halfens (2007).

The predictive validity of these tools has also been challenged by Franks et al. (2003) and Guy (2007) who demonstrated that they may overpredict
risk, incurring cost implications as expensive, preventative equipment is employed when it may not be necessary; or underpredict risk so that someone assessed as not being at risk develops a pressure ulcer. Gould, Goldstone, Kelly, and Gammon (2004) found the Waterlow Scale underpredicted risk in 36% of assessments, whereas, overprediction occurred in only 3.4% of cases.

Although the Waterlow Scale has been reported to be of use in a variety of clinical specialties, some suggest that the scale is not suited to all groups, for example, wheelchair users (Anthony, Barnes, & Unsworth, 1998). Furthermore, many studies suggest that some of its subscales are responsible for its poor predictive validity. The “mobility,” “continence,” “neurological deficit” and “medication” subscales have created problems, in assessing a patient due to lack of clarity (Cook et al., 1999; Lyne, Papanikolaou, & Lycett, 1999). Another criticism of Waterlow’s subscales surrounds the issue of gender. Sexes are weighted unequally, however the assumption that females are at a higher risk than males is not supported (Papanikolaou, Clark, & Lyne, 2002).

A prospective observational study of three intensive care units (ICU) in Sydney undertaken by Boyle and Green (2001) revealed that the Waterlow Scale has both a high false positive rate and a poor predictor of pressure ulcer occurrence. The study highlighted that many factors of the scale discriminate against ICU patients, as there is little variance in this population for the factors that address mobility, appetite, and urinary continence. Yet Weststrate, Hop, Aalbers, Vreeling, and Bruining (1998) disagree, claiming that by using this scale in ICU nursing staff are provided with reliable information at an early stage and are warned about a patient’s risk category in developing an ulcer, therefore introducing preventative measures early.

Many studies suggest that the Waterlow Scale needs changes. A study by Anthony, Clark, and Dallender (2000) attempted to optimize the scale’s ability in predicting individual’s vulnerability to developing pressure ulcers. The participating individual’s vulnerability to pressure ulcer development was assessed using the Waterlow Scale and skin inspection on admission. These reassessments of vulnerability were repeated 7 and 14 days post admission. The findings showed marked differences between the samples within the two settings. Patients in one clinical setting were older, had a higher rate of emergency admissions, showed higher systolic blood pressure, and a higher rate of fecal incontinence. Furthermore, 101 individuals from the first setting were reported to be bedfast in contrast to only eight in the second setting. Unsurprisingly, of the 69 individuals who developed ulcers, 60 of these occurred within the first setting. This data interestingly highlights that despite the subjects recruited in the first setting having a higher vulnerability to developing ulcers and having been given additional preventative care, they
developed more ulcers than those recruited in the second group. This ques-
tions the effectiveness of Waterlow as a prevention method, and highlights
that items on the scale may be less useful in prediction than others for a
particular patient group. It found that complex nonlinear solutions were less
successful in assessing risk than a simple linear model based on 5 of the
11 variables of the scale (appetite, continence, skin, surgery, and age), which
have a significant predictive ability. However, the study is not definitive and
requires further research with a larger sample. Scope for improvement exists
but can only be accomplished by the adoption of methods that are new in this
field although well tested elsewhere.

Papanikolaou et al. (2002) conducted a secondary analysis of Anthony,
Reynolds, and Russell (2000) study from which alternative models of pressure
ulcer risk assessments have been developed and evaluated. The study high-
lighted that by using the Brier Score, a simplified model of the Waterlow
Scale, was found to have a better predictive accuracy than the original scale.
Furthermore, the sensitivity and specificity level of the simplified scale reached
85%, suggesting that changing such factors as weight or scoring has the poten-
tial to improve its predictability. This belief that a scale based on a subset of
Waterlow data plus additional information performs better than a scale based
solely on all the Waterlow risk factors, is further emphasized by Papanikolaou,
Lyne, and Lycett (2003). However, both of these studies have some limitations.
The stochastic models presented are based on current theoretical and empirical
knowledge of pressure ulcer risk factors, and the results presented are not yet
generalizable and external validation methods need to be applied.

Some of Waterlow’s categories are responsible for its poor predictive
validity. Studies suggest that the scale is not suited to all patient groups and
yet it is a popular tool. A simplified version is required to improve its predic-
tive ability. Furthermore, the Waterlow Scale is highly overpredictive in
terms of its high-sensitivity and low-specificity level. However, according to
Flanagan (1995), determining true sensitivity and specificity would involve
withholding preventative interventions and allowing tissue breakdown in
vulnerable patients, which would be ethically unacceptable. Therefore, a risk-
assessment scale’s predictive value can never be calculated without exposing
an unfortunate group of patients to no intervention (Maylor, 2006).

Discussion

Research suggests that the Waterlow Scale is an unreliable method of
assessing individuals at risk of pressure sore development with all of the
studies examined indicating a poor interrater reliability status. Several authors
claimed that variance in scores among the raters was because of different levels of knowledge (Cook et al., 1999; Watkinson, 1996). Nonetheless, all raters were given the same information regarding the patient’s status. These studies suggest that Waterlow’s poor reliability lies in its subscales, especially within the “mobility,” “skin type,” and “build/weight for height” categories because of a lack of clear clarification. In 2005, Waterlow revised the scale, updated the guidelines, and clarified many aspects. Future examination will highlight the reliability and validity of the 2005 version. Interestingly, Kelly (2005) highlighted the nurse’s lack of understanding of the scale, despite its daily use, and their tendency to ignore (i.e., terminal cachexia) rather than ask for clarification.

In terms of the validity of the Waterlow Scale, all studies with the exception of one (Gould et al., 2004) found the scale to display high-sensitivity but low-specificity levels. This introduces Waterlow’s major criticism of overprediction, perhaps, caused by its “blanket coverage” (Wardman, 1991). Edwards (1996) believes the scale performs better as a diagnostic rather than a predictive tool. Although it contains a category for skin assessment, it will almost inevitably achieve high sensitivity. Another point highlighted by Edwards (1996) is its inappropriateness in dealing with the elderly. This is because of high weighting given to age within the score.

From an economic perspective, overprediction is a serious limitation of the scale because scarce and expensive resources are being given to patients who do not require them, which in turn is preventing preventative measures being given to those who are at risk. A European prevalence study conducted in 2002 revealed that no more than 10% of the patients at risk received adequate prevention care (Vanderwee et al., 2007a). Furthermore, many professional carers fail to reevaluate the risk status of patients. Therefore, many patients are nursed on pressure-relieving systems for longer than necessary, whereas others never benefit from such interventions because of a lack of resources (Flanagan, 1997). Clearly, it is evident that the Waterlow Scale is not suited to all groups (Anthony et al., 1998; Boyle & Green, 2001) and perhaps it is time to introduce a simplified model. Adjustments would promote effectiveness and improve preventative care (Anthony, Reynolds, & Russell, 2000; Papanikolaou et al., 2002; Papanikolaou et al., 2003), yet further research is merited.

Interestingly, an element that frequently appeared in various studies was the necessity to use a risk-assessment scale in conjunction with clinical judgment (Cook et al., 1999; Shukla et al., 2008; Thompson, 2005). Risk calculators were never designed to replace clinical judgment but to provide guidance in allocation of resources appropriately. Sharp et al. (2000) found that 79% of
clinical staff assessed did not use a pressure ulcer risk-assessment scale, but relied on a range of practice procedures and risk indicators in determining a patient’s risk potential of developing an ulcer, many of which were outdated practices. What this study failed to investigate was how many patients were/were not identified, as being at risk through the use of clinical judgment solely and whether these patients developed ulcers or not?

From examining the literature, it is evident that the Waterlow Scale has poor interrater reliability and validity. Like all risk-assessment scales, it has limitations and should be used with caution. According to Thompson (2005), there are over 40 different risk-assessment tools and clearly, Waterlow is not the best. Nonetheless, it does offer guidance especially to inexperienced staff, so that preventative measures may be introduced early and differentiates between “at risk,” “high risk,” and “very high risk” status (Shukla et al., 2008). From a nursing perspective, understanding the scale is crucial so that it correctly identifies those individuals at risk of pressure ulcers. Ideally, what is needed is the existence of a sole risk-assessment scale, suited to all patients regardless of their specialty. This in turn would ensure familiarity and reduce confusion among all raters. However, this may be an unrealistic goal because of the diversity of patients (Flanagan, 1997; MacDonald, 1995).

A risk-assessment tool is not self-explanatory and training is needed (Cook et al., 1999). Education is required in reducing pressure ulcers through study days, seminars, and ward meetings. Over time, this will increase Waterlow’s reliability (Waterlow, 1996). It is recommended that clinical nurse managers frequently assess their nursing staff’s ability in correctly using the scale and ensure that Waterlow’s (2005a) revised tool is used as the most up-to-date version. In addition, it is suggested by Webster, Gavin, Nicholas, Coleman, and Gardner (2010) as a result of the disparity of agreement as to whether the Waterlow assessment scale is indeed beneficial in clinical practice, a suitably powered, randomized controlled trial is urgently needed to provide definitive evidence about the usefulness of the Waterlow Scale compared with other screening tools and with clinical judgment.

**Conclusion**

One must not solely rely on the “pencil and paper” exercise in assessing an individual’s risk of pressure ulcer development (Gould et al., 2004), it must be used in conjunction with clinical expertise (Waterlow, 2005a). The Waterlow Scale has demonstrated poor interrater reliability, high-sensitivity, and low-specificity levels. Nonetheless, it remains a popular tool
used by nurses daily in patient care delivery, and will remain popular until a more effective, specific, and accurate mode of assessment is introduced.

**Declaration of Conflicting Interests**

The author(s) declared no potential conflicts of interest with respect to the authorship and/or publication of this article.

**Funding**

The author(s) received no financial support for the research and/or authorship of this article.

**References**


Bios

Breda Walsh, RGN, BNS, is a registered general nurse in theatre in the Mater Hospital Dublin, Ireland.

Laura Dempsey, RGN, BNS, MSc (Nursing), PG Dip CHSE, RNT, is a lecturer in the School of Nursing and Midwifery at the National University of Ireland, Galway, Ireland.