Clinician inter-rater reliability using a medical wound imaging system

Flowers C, Newall N, Kapp S, Lewin G, Gliddon T, Carville K, Martinelli D & Santamaria N

Abstract

The ability to determine the effectiveness of alternative treatments in the management of wounds is central to wound research and ultimately to the development of sound evidence on which to base clinical decision making. Measurement precision is therefore a critical factor in research which aims to determine the effects of treatment on healing rates, especially when the research is being conducted across sites and measurements are being made by different individuals. Additionally, as with any digital planimetric system, the quality of the measurement is dependent on the skill of the operator, therefore it is considered essential to measure the inter-rater reliability of the measuring system used.

This paper describes how the inter-rater reliability of the Alfred/Medseed Wound Imaging System, now known as the Advanced Medical Wound Imaging System (AMWIS) was established. Four clinicians involved in the trial were asked to independently trace (measure) the same 20 wound images using AMWIS. The data collected for each image included the total area and the various tissue characteristics for the wound. The results showed that there was high inter-rater reliability for the measures of total area, granulation and slough. Additional feedback suggested that better definitions and guidelines to assist in the differentiation and recording of tissue characteristics would be of value to the users.

Introduction

In 2006, the Royal District Nursing Service (RDNS) VIC and Silver Chain Nursing Association WA (Silver Chain) commenced a randomised controlled trial (RCT) to compare the clinical outcomes for clients living with a lower leg ulcer randomised to one of two antimicrobial dressings, cadexomer-iodine and a silver nanocrystalline impregnated dressing. The trial was funded by the Angior Family Foundation. The primary outcome measure was healing rate (expressed as a percentage change in the total surface area of the wound) and it was measured using the Advanced Medical Wound Imaging System (AMWIS). As healing rate was a critical outcome measure for the RCT, an inter-rater reliability assessment was planned to determine the comparability of different assessors' ratings.

The AMWIS software, developed by Santamaria and Clayton in 2000, was designed as a clinical and research system to enable clinicians to more accurately measure and document a wound using digital images and to assist in accurately determining the effectiveness of treatment [1]. The clinical applicability and planimetric reliability of the AMWIS software and its use in chronic wounds were subsequently described [2, 3]. More recently, AMWIS was demonstrated to enhance clinical outcomes for patients in remote locations by enabling remote expert wound consultation via its telemedicine functions [4]. The system has also been used as a valuable component of an integrated pressure ulcer prediction, prevention and management system [5].

The primary objective of the study was to determine the inter-rater reliability associated with the use of the AMWIS wound management software to analyse digital images of wounds. Inter-rater reliability is "the consistency of observations between two or more observers; often expressed as a percentage of agreement between raters or observers or a coefficient of agreement that takes into account the element of chance" [6]. The secondary project objective was to examine this reliability statistic to identify a level of confidence associated with the use of the AMWIS software as a measure in the RCT.

Methods

RCT imaging process

For clients participating in the RCT, a digital image was taken of the wound at baseline and every 2 weeks for as long as the wound was unhealed up to a maximum of 12 weeks, providing a possible seven images per trial participant. The images were analysed using the AMWIS program, whereby a digital wound image was uploaded to the program on the user's (usually a clinician) computer. The image was then

Charne Flowers

BA (Hons), GDipStat
Researcher, Helen Macpherson Smith Institute [1]

Nelly Newall *

RN, Clinical Research Coordinator
Silver Chain Nursing Association [2]

Suzanne Kapp

RN, MNSci
Clinical Nurse Consultant Wound Management
Helen Macpherson Smith Institute [1]

Gill Lewin

PhD MPH
Professor, Centre for Research on Ageing at Curtin
University of Technology and Research, Perth WA
Director, Silver Chain, Western Australia
Adjunct Senior Lecturer, Edith Cowan University,
Perth WA

Terry Gliddon

RN, MAppSci
Manager, Research & Development Department
Helen Macpherson Smith Institute [1]

Associate Professor Keryln Carville

RN PhD
Associate Professor Domiciliary Nursing
Silver Chain Nursing Association [2]
Curtin University, Perth WA

Daniel Martinelli

Research Assistant
Helen Macpherson Smith Institute [1]

Nick Santamaria

RN, PhD
Professor of Acute & Ambulatory Care
School of Nursing & Midwifery,
Curtin University, Perth WA

Institution(s) where the research was undertaken:

1. Helen Macpherson Smith Institute of Community
   Health, Royal District Nursing Service
   31 Alma Road, St Kilda VIC 3182

2. Silver Chain Nursing Association
   6 Sundercombe Street, Osborne Park WA 6017

*Corresponding author

calibrated and subsequently measured by outlining the wound surface area and the tissue characteristics present within or surrounding the wound with a mouse or stylus. AMWIS provides a measurement in millimetres squared (mm^2) for each type of wound tissue outlined, i.e. epithelialisation, granulation, slough, necrosis, hypergranulation, infected area, undermining and surrounding tissue, plus a measure for total wound area. The measures for each tissue type are therefore dependent on the ability of the rater to outline accurately, and to identify and label the tissue characteristics in the image. The two RCT sites utilised different versions of the AMWIS software; RDNS used AMWIS V1.0 and Silver Chain used AMWIS V2.0. Though there are minor differences between the two AMWIS versions, both undertake surface area measurement in an identical manner.

Inter-rater testing

To ensure the ratings resulting from the interpretation of the digital images were consistent, an inter-rater reliability score was determined. That is, the degree of agreement between raters was assessed. This investigation was not intended to establish how accurately the true size of the wound was measured by raters, rather, how consistent the measurements were between users. Nor was the purpose of this investigation to comment on the use of AMWIS as a clinical tool. Rather, the purpose was to provide evidence as to the reliability of the measures utilised in the antimicrobial RCT.

The sample

A stratified random sample of 20 wound images was chosen from the images collected during the trial, where the strata represented each of the trial sites and 10 images were chosen from each site. Image selection was further stratified to include images from baseline (four images), 6 week (three images) and 12 week (three images) periods per site. Images from three different time points were included because baseline images were expected to represent the wound at its most severe when differences between tissue characteristics could be expected to be most evident and their outlines most easily distinguishable. Conversely, it was thought likely that in later photos, if the wound was healing, the task of distinguishing between tissue characteristics would be more difficult.

Care was taken not to include images of poor quality to avoid confounding the results by having an uncontrolled quality variable. Where a randomly selected client image was found to be of questionable quality, the next wound image was included in the sample. Image selection was conducted independently from the subsequent process of image ratings.

Data collection

Four raters independently used AMWIS to trace each of the images. Raters of varying nurse grade and experience at both sites were selected from individuals using AMWIS during the trial. All four raters were provided with the same 20 images. For reasons of client confidentiality, client identifying information was electronically hidden in all images. In doing so, the calibration arrow for two images was partially obscured. As such, raters were ultimately able to assess only 18 of the 20 images that were circulated.

A data collection form was provided to each rater to record the measurements (mm^2 and percentage) associated with the total area and the various tissue characteristics for the wound. Data were recorded against each of the numbered images. Additionally, free text space was provided for the rater to comment on each image and any difficulties they experienced in the measurement process. The rater was also asked to indicate if they had seen the wound first hand as it was thought that this could provide a degree of clinical validation that was not available to the other raters.

Statistical analysis

It was decided that because all the measurements were in mm^2 and hence the variables were continuous, Pearson's *r* could be calculated as a measure of inter-rater reliability amongst raters [7]. An alternative statistic for determining reliability is the intra-class correlation coefficient (ICC) [8, 9]. This is defined as "the proportion of variance of an observation due to between-subject variability in the true scores" [10]. The 'subject' in this instance represents the multiple raters of the wound images. The range of the ICC is, as with other correlation coefficients, between 1.0 and -1.0. The ICC will be high when there is little variation between the scores given to each item by the raters, e.g. if all raters give the same, or similar scores to each of the items [10]. As the product moment correlation uses standardised data and can therefore be insensitive to scale, the ICC test offers the advantage of taking account of the variance between raters. An alpha, two-way mixed effects model with absolute agreement was determined appropriate for this analysis [11]. Both the Pearson's *r* and ICC tests were performed for the total area and tissue characteristics and are presented in the results section.

Results

A total area score was provided by each rater for each wound image. The most common tissue characteristics selected by raters were slough and granulation. Inter-rater reliability results are therefore presented for the areas of these tissue characteristics as well as total area.

Two of the four raters chose to define wound areas in a select number of images using the epithelialisation tissue type. The other raters did not use this tissue category in any of their assessments. The variation between raters could reflect difficulty discerning this tissue type from a photograph, the individual's preference to define areas of epithelialisation, or the capacity and skill of the clinician to correctly identify and label areas of epithelialisation. Differences were also evident in relation to the identification of other tissue characteristics. Only one rater used the hypergranulation classification of wound tissue in a small number of images. Another rater identified surrounding tissue in some images. No other raters used the hypergranulation or surrounding tissue classification. It was not possible to calculate inter-rater reliability for tissue characteristics that had not been identified and measured by all the raters.

Two raters indicated they had seen one or more wounds first hand. As already discussed, their ability to trace the outline of the wound and different tissue characteristics within it, could have been influenced by their recall of the clinical presentation of the wound. One rater reported having seen one wound. Another rater recalled sighting the wounds depicted in four of the images. An indication that recall by the raters was open to some error was one rater's nomination of having sighted a wound which was actually from the alternate study site and, therefore, clearly had not been seen by that person.

Nonetheless, a separate analysis was conducted to compare the level of agreement achieved for those images where at least one clinician reported having seen the wound and those images where none of the clinicians reported having seen the wound. There was no statistical difference for the level of agreement between raters for total wound area, granulation or slough and whether the image was seen previously by a rater in the clinical setting. This indicates that clinical exposure to the wound had marginal influence on the measurement of wound characteristics in this study. Results are reported, therefore, for the entire sample of images.

Total area

High correlations were identified among individual raters for their assessment of the total area for the 18 images (Table 1). Given the fact that the estimation of total wound area depended on the clinician outlining the image which is then used for the software calculation, this finding suggests that clinicians in this study actually outlined the images in the same way. This is reflected in the ICC result which found a significant and high correlation between raters (ICC=0.958, p<0.001).

Granulation

Generally, there was a high level of correlation between raters in their assessment of granulation tissue for the 18 images (Table 2), though Rater 1 tended to present assessments of granulation tissue which were not significantly associated with the assessments of Raters 3 and 4. These differences were associated with Rater 1 opting to classify what other raters considered granulation tissue as epithelialisation tissue. Nevertheless, the ICC result still suggests that overall there was high agreement between raters when defining the granulation tissue areas of the 18 images (ICC=0.877, p<0.001).

Slough

Raters seemed to be able to discern areas of slough with high consistency. Significant and high correlations were identified between all raters for this wound area (Table 3). Again this result is supported by a high and significant ICC test result (ICC=.952, p<0.001) which confirms that there is a high inter-rater reliability when identifying the area of a wound which is slough.

Qualitative assessment of tracing

Raters' comments were considered for each image in combination with the subsequent measurements provided. Five key themes were identified which illuminated areas of uncertainty and difficulty experienced by raters. These themes related to the positioning of the calibration arrow, imaging technique and quality, distinguishing tissue characteristics, multiple wound areas, and calculation technique.

The positioning of the calibration arrow

The positioning of the calibration arrow was identified by raters as potentially causing inaccuracy when rating the image. Photo 5 (Figure 1) was identified as an example where the calibration arrow was not perpendicular to the angle that the photo was taken, thus potentially influencing the raters' calibration of the image. Arrow curvature is acknowledged by the authors as influencing how accurately the measurements reflect the true size of the wound – a particular concern when assessing wound size over time. In this investigation, unless raters attempted to compensate for curvature in their calibration of the arrow, any inaccuracy resulting from arrow curvature should not have affected inter-rater reliability.

Imaging technique and quality

Though images were screened for their quality before including them in the sample for assessment, raters still felt poor image quality interfered with their capacity to make an accurate judgement. Photo 6 (Figure 2) was identified as an example where the light/dark contrast, particularly at the top of the wound, made assessing the wound difficult.

The distance the photo was taken from the wound was also highlighted as potentially influencing the assessment of the images. In the case of Photo 1 (Figure 3), the distance of the photo from the wound was thought to have influenced the image quality and the raters' capacity to accurately assess the wound dimensions and characteristics. These concerns were repeated in relation to Photo 15 (Figure 4), in which the quality of the image was reported to be reduced when raters zoomed in to conduct the wound tracing.

Distinguishing tissue characteristics

The difficulty some raters experienced in selecting between tissue characteristics such as granulation, epithelialisation, hypergranulation and slough, as evidenced by their measurements, was also reflected in their comments. Photo 1 (Figure 3) consistently caused the most difficulty for raters.

Table 1. Pearson's r correlations for total area.

	Rater 1	Rater 2	Rater 3	Rater 4
Rater 1	1	0.974*	0.997*	0.710*
Rater 2	0.974*	1	0.975*	0.840*
Rater 3	0.997*	0.975*	1	0.720*
Rater 4	0.710*	0.840*	0.720*	1

* Correlation is significant at the 0.01 level (2-tailed)

Table 2. Pearson's r correlations for granulation tissue.

	Rater 1	Rater 2	Rater 3	Rater 4
Rater 1	1	0.873*	0.497	0.353
Rater 2	0.873*	1	0.889*	0.976*
Rater 3	0.497	0.889*	1	0.987*
Rater 4	0.353	0.976*	0.987*	1

* Correlation is significant at the 0.01 level (2-tailed)

Table 3. Pearson's r correlations for slough.

	Rater 1	Rater 2	Rater 3	Rater 4
Rater 1	1	0.759*	0.858*	0.760*
Rater 2	0.759*	1	0.869*	0.911*
Rater 3	0.858*	0.869*	1	0.962*
Rater 4	0.760*	0.911*	0.962*	1

* Correlation is significant at the 0.01 level (2-tailed)

The total wound area ranged from 911mm^2 to 4792mm^2 and raters commented that they "found this very hard to differentiate slough/epithelium", and that it was "very hard to tell where the initial wound boundaries were". The image was reported as not being clear enough to enable tissue type recognition "as central slough also looks like pale base". While not all images were found to be problematic for all raters, the comments about Photo 1 in particular suggest that many of the difficulties experienced by raters related to distinguishing different tissues types within an image.

Multiple wound areas

As some wounds healed, they separated into multiple wound areas. This presented a challenge for the raters as they were uncertain which area(s) to trace and whether or not to combine measurements of different areas. Photo 7 (Figure 5) is an example. One rater traced only the largest single wound area, resulting in a total wound area of 305mm^2, compared to other tracers who included all wounds for a total wound area ranging from 543mm^2 to 623mm^2. One rater commented "Thought I should only 'AMWIS' the biggest wound but due to proximity [I] recorded all three". Thus a question remains as to the issue of determining how many wounds are present (i.e. one or more than one) as healing occurs. This may be more to do with clear definitional rules than a clinical judgement as to whether one or multiple wounds are present.

Figure 1. Photo 5.

Figure 2. Photo 6.
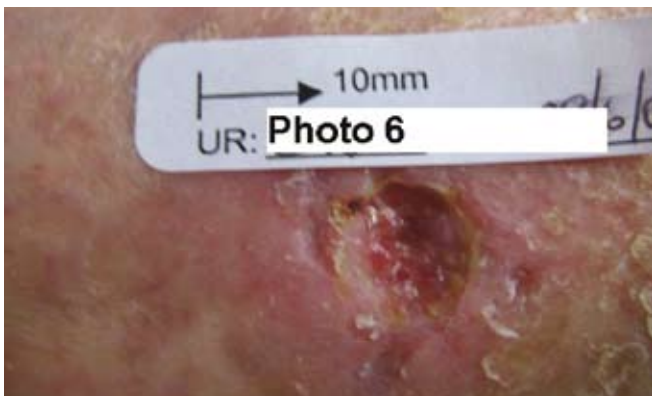
Calculation technique

Some raters identified that they had manually summed the individual tissue characteristics within the wound to ensure that they added up to 100% of the total area. The high inter-rater reliability determined by this investigation, however, suggests that the influence of this on the results was minimal.
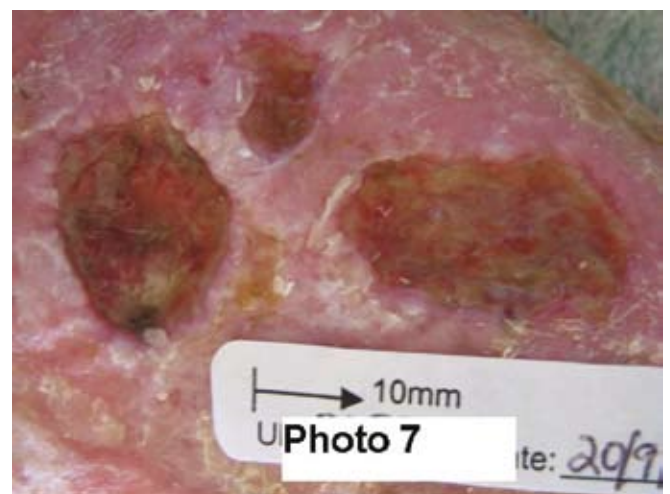
Figure 3. Photo 1.

Figure 4. Photo 15.

Figure 5. Photo 7.

The images with the greatest variation between the smallest total area rating and the largest total area rating were Photos 1, 6 & 7. Each of these images was identified as having a number of difficulties associated with their measurement, i.e. difficulty in determining wound edges, poor image quality, and the need to decide whether to trace multiple wounds.

Discussion

The three most commonly and consistently used dimensions of the AMWIS software were total area, granulation and slough. For these three areas, high inter-rater reliability was found, particularly for total area and slough. The identification of epithelialisation and hypergranulation tissues was infrequent and inconsistent, a problem which tended to impact on the measurement of granulation tissue. The practice of delineating surrounding skin tissue also varied between raters. However, as surrounding tissue is calculated in addition to the total area size and percentage, there was no impact on other tissue ratings or the total wound area.

Although the inter-rater reliability of AMWIS for the more commonly selected tissue characteristics has been demonstrated by this study, future use of digital imaging may need to be accompanied by more instruction and definitional guidelines in relation to epithelialisation and hypergranulation.

In this small study it could not be determined whether the difficulty in recognising and defining less common tissue characteristics was related to lack of clinical skill, poor image quality or inadequate guidelines in the use of AMWIS; indeed it may be a combination of all these things. The fact that the difficulties arose in less frequently occurring wound and healing responses and not just in 'inferior' images does suggest that all these areas will need to be addressed in future evaluation of the use of imaging for clinical and research purposes. However, the difficulties associated with identifying certain tissue characteristics may not be resolvable by the provision of specific instructions in programs such as AMWIS – clinical skill enhancement and the use of alternative strategies will also need to be tried and tested.

Two themes were evident in the qualitative feedback provided by raters about the difficulties they experienced using AMWIS in this study. The first was poor image quality which included the dark/light contrast in the image, the distance the photo was taken from the wound and the subsequent need to zoom in to conduct the measurement, which further reduced image quality. The second related to their uncertainty about how to measure multiple wound areas.

Strategies to address issues of image quality could include more instruction/training in the correct distance away from the wound to take the photo and in ensuring satisfactory lighting (an issue which can be especially problematic in the home environment). Guidance on tracing images with multiple wound areas and possibly clearer and more detailed instructions could minimise inconsistencies between raters and improve inter-rater reliability.

Additionally, the impact of clinical validation should not be underestimated. In practice, the wound measurement process may be undertaken *in situ* and in view of the wound. Whilst this may not necessarily improve the image quality upon which the tracings are conducted, the presence of the wound may influence the accuracy with which the clinician can delineate tissue characteristics.

Placement of the calibration arrow, though considered to have minimal impact on the reliability results in this investigation, is another area for improvement, particularly if differences in wound size across time are not to be falsely estimated or obscured by variations in this practice. Clear instructions, good education and diligent implementation of guidelines should result in improvement.

While improving the quality of the images is a feasible and achievable strategy to improve inter-rater reliability, achieving consistent ratings when tissue characteristics are dispersed or 'marbled' within the wound requires further consideration – see for example Photo 9 (Figure 6). Although the impact of marbling on the total wound area rating

Figure 6. Photo 9.

has been demonstrated to be minimal, it does influence how accurately the rater can identify the proportion of the wound represented by particular tissue characteristics. The importance of this is illustrated when one considers, for example, wound bed preparation and the need to have short-term wound management goals. Whilst the overall objective may be to heal the wound, the short-term goal may be to autolytically debride the wound. In this instance, there is a need to accurately account for a reduction in tissue characteristics such as slough and necrotic tissue.

This is both a user issue and a system issue. Users need to firstly identify and distinguish the tissue characteristics, and then, using the digital imaging software, separate and quantify the tissue characteristics. Comments such as: "[It was]… too time consuming to [separately measure] AMWIS tissue characteristics therefore I did the total surface area then 'guesstimated' the slough and granulation" indicate how clinicians adopt alternate strategies rather than attempting to trace the tissues types within a wound as intended by the program. This was especially a problem when confronted with an image of a wound with dispersed or 'marbled' appearance.

This finding presents a challenge to developers of wound imaging and measurement software as well as to service managers and researchers. Devising ways and means of ensuring that clinicians accurately photograph and measure wounds, particularly 'tricky' wounds, needs to be a priority if the utility of the software in both clinical and research applications is to be maximised.

Limitations

There are at least four ways in which the present study might be considered to be limited. Firstly, as the study looks only at inter-rater reliability, no comment can be made about the intra-rater reliability of AMWIS, for which the same raters would need to conduct repeat ratings of the same images. Nor has it considered reliability when reviewing a sequence of images of the same wound. Secondly, as raters completed their assessments remotely and in most instances had not seen the wound first hand, it is not possible to comment regarding the merit of AMWIS as a clinical tool. Thirdly, as the dimensions of the wound were not established separate to the AMWIS measures, the validity of the AMWIS wound size measures as compared to true wound size cannot be determined. Lastly, the small number of raters involved could be seen as a limitation of the study.

Conclusions

This study showed high inter-rater reliability for the measures of total area, granulation and slough when AMWIS was used to measure 18 randomly selected photographs of leg ulcers. The qualitative data collected indicated that the inter-rater reliability of AMWIS would be further increased by improving instruction to users in how to; use the equipment to ensure good quality images, place the calibration arrows correctly and measure when there are multiple wound sites.

The need for better definitions and guidelines to assist in the differentiation and recording of tissue characteristics has been suggested. Digital imaging provides a means for clinicians and researchers to obtain accurate measure of wound healing rates. However, the technology relies on and does not replace good clinical judgement.

References

1. Santamaria N & Clayton L. Cleaning up: the development of The Alfred/Medseed wound imaging system. Collegian 2000; **7(4)**:14-18.

2. Santamaria N, Austin D & Clayton L. Multi-site trial and evaluation of the Alfred/Medseed Wound Imaging System prototype. Prim Intent 2002; **10(3)**:119-124.

3. Austin D & Santamaria N. Digital imaging and the chronic wound: clinical application and patient perception. J Stomal Ther Aust 2002; **23(4)**:24-29.

4. Santamaria N, Carville K, Ellis I & Prentice J. The effectiveness of digital imaging and remote wound consultation on healing rates in chronic lower leg ulcers in the Kimberley region of Western Australia. Prim Intent 2004; **12(2)**:62-70.

5. Santamaria N, Carville K, Prentice J, Ellis I & Ellis T. Clinical IT in Aged Care Product Trial. Final Evaluation Report for the Pressure Ulcer Research, Intervention, Management and Evaluation (PRIME) System in Aged Care Trial. Commonwealth Department of Health and Ageing, 2006.

6. Beanland C, Schneider Z, LoBiondo-Wood G & Haber J. Nursing Research: Methods, Critical Appraisal and Utilisation. Australia: Harcourt Australia, 2000, p.580.

7. Bland JM & Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; **1**:307-310.

8. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971; **76(5)**:378-382.

9. Shrout P & Fleiss JL. Intraclass correlation: uses in assessing rater reliability. Psychol Bull 1979; **86**:420-428.

10. Everitt B. Making Sense of Statistics in Psychology. Oxford: Oxford University Press, 1996.

11. Nichols DP. Choosing an intraclass correlation coefficient. SPSS Keywords 1998; No. 67.