Inter-rater reliability of pressure ulcer staging: ordinal probit Bayesian hierarchical model that allows for uncertain rater response

Byron J Gajewski, University of Kansas Medical Center
Sara Hart, University of Kansas Medical Center
Sandra Bergquist-Beringer, University of Kansas Medical Center
Nancy Dunton, University of Kansas Medical Center

Available at: https://works.bepress.com/byron_gajewski/6/
Inter-rater reliability of pressure ulcer staging: Ordinal probit Bayesian hierarchical model that allows for uncertain rater response

Byron J. Gajewski¹,*,†, Sara Hart², Sandra Bergquist-Beringer² and Nancy Dunton²

¹Schools of Allied Health and Nursing, Center for Biostatistics and Advanced Informatics, The University of Kansas Medical Center, Kansas City, KS 66160, U.S.A.
²University of Kansas School of Nursing, Kansas City, KS 66160, U.S.A.

SUMMARY

This article describes a method for estimating the inter-rater reliability of pressure ulcer (PU) staging (stages I–IV) from raters in National Database of Nursing Quality Indicators (NDNQI) participating hospitals. The method models ordinal spanning data utilizing an ordinal probit Bayesian hierarchical model (BHM) across several hospitals in which raters monitor patient’s PUs. An ulcer that cannot be accurately assessed because the base of the wound cannot be seen is defined as unstageable. Our novel approach allows for an unstageable PU rating to be included in the analysis. We compare the ordinal probit BHM to an approximate random-effects (standard approach in the literature) model that assumes that the raw ordinal data are continuous. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: rater reliability; shrinkage; intraclass correlation coefficient; random-effects model; NDNQI

1. INTRODUCTION

Pressure ulcers (PUs) are a significant problem for hospitalized patients and those receiving nursing home and home health care. A PU is a wound of the skin and is usually caused by unrelieved pressure [1]. PUs most commonly form over bony prominences, when tissue becomes compressed between bone and tissues. Small blood vessels supplying nutrients and oxygen to the skin are shut off, tissue dies, and a PU forms [2]. For a bed-bound person, these areas include the sacrum, buttocks, trochanter, and heels. Other areas where PUs are frequently seen include the knees, ankles, shoulder blades, back of the head, and spine [2]. Persons with compromised skin can also
develop PUs from activities such as sliding down in a bed and chair, or medical devices placed against the skin. PUs are sometimes called decubitus ulcers or bedsores.

Staging guidelines were established by the National Pressure Ulcer Advisory Panel and the Agency for Healthcare Research and Quality [3, 4] according to the extent of tissue loss. A stage I PU is defined as an area of unbroken skin that is persistently red in lightly pigmented skin. In darkly pigmented skin, its color may differ from the surrounding area. Changes in skin temperature, tissue consistency and/or sensation when compared to surrounding skin may be noted. Stage II PUs involve partial thickness skin loss and look like a superficial abrasion or blister. A stage III PU involves loss of all skin tissues. Stage IV PUs also involve full thickness skin loss with destruction and damage to underlying muscle and bone. An ulcer that cannot be accurately assessed because the base of the wound cannot be seen is defined as an unstageable [3, 4]. Examples of unstageable PUs include those whose wound base is covered by dead tissue or a deep tissue injury with unbroken skin.

Persons who develop PUs experience unrelieved pain and social isolation, and may suffer complications such as cellulitis, bacteremia, or osteomyelitis requiring prolonged hospitalization [5–8]. Accurate identification and classification of PUs are important to determining appropriate treatment and evaluating quality improvement measures to prevent their occurrence.

The National Database for Nursing Quality Indicators (NDNQI) was established by the American Nurses Association (ANA) in 1998 to monitor outcome indicators in the acute care setting and to evaluate the impact of nurse staffing on hospital care. Data are collected currently on multiple nursing indicators including PUs. Participating hospitals transmit data to NDNQI quarterly. Data are then summarized and published in a quarterly report that allows participating facilities to compare their results with the results from previous quarters and with other hospitals across the nation having similar characteristics. As of August 2006, 1003 acute care hospitals from all 50 United States and the District of Columbia were members of NDNQI.

Inter-rater reliability, defined as the agreement between two or more raters observing the same thing, is the method used for determining the level of agreement between raters staging PUs. For the purposes of PU staging, reliability is most often evaluated by having multiple raters observe the same PU, either sequentially or simultaneously. When staging PUs for prevalence or incidence studies, reliability is crucial to documenting the accuracy and comparability of the data. The raters’ expertise, interpretation of clinical tools/scales, and clinical judgment all have the potential to influence the reliability of PU data.

To understand this further, we initiated a large study to investigate the reliability of the PU rating among NDNQI hospitals. For the overall study, the specific aims were to: (1) determine the reliability of PU staging from bedside evaluation of PUs; (2) determine the reliability of PU staging from web-based photographs; and (3) examine the association between bedside and web-based reliability assessments. The data reported in this paper address specific aim 1.

The key scientific question of interest in this paper centres on estimating the reliability of PUs from raters in hospitals belonging to the NDNQI database. Specifically, what is the reliability of the rating process in terms of classifying a PU into stages? One of the biggest challenges in addressing this question is dealing with a rating of ‘unstageable.’ In the past, researchers have dealt with this by categorizing the responses and assessing agreement with a kappa- or weighted-kappa statistic (e.g. [9, 10]). We take a different approach for assessing the reliability of PU ratings by modelling the ordinal data via a probit model [11]. We assess the reliability on the probit scale, on which the variables are assumed to be normally distributed. This probit approach allows for (1) the staging to be continuous rather than categorical and (2) an unstageable PU to be staged.
on the ordinal scale. We use the probit bounds to create a Bayesian hierarchical model (BHM) to estimate reliability within and across hospitals.

Many behavioural studies measure outcomes using ordinal data. Commonly, researchers analyse these data treating the ordinal responses as though they were continuous and normally distributed. To investigate this common assumption, we estimate the intra-class correlation coefficient (ICC) with a model using the raw ordinal data. The approximate model treats the ordinal data as normally distributed. If the normal approximation works well, then this simple model may be an adequate practical working model.

However, we will show that for our data, a more complicated ordinal BHM is more appropriate because it explicitly accounts for the discrete ordinal nature of the data. An additional advantage of our method is that it yields random-effect estimates for the individual hospitals that can be used to rank hospitals and identify where interventions are needed. These random effects are estimates of the ICC (reliability).

We present a Bayesian method for assessing the reliability. We discuss other statistical methods from the literature, the design, and data in Section 2. Subsequently, we describe the BHMs in Section 3 and provide their prior distributions. In Section 4, we discuss the computational issues. The results are in Section 5. We discuss the conclusions in Section 6.

2. GENERAL MODELLING AND DATA

2.1. Approaches

Our overall goal is to construct a model to estimate the reliability of rating PUs from patients in NDNQI hospitals. The aim of this section is to discuss a very basic model choice (random-effects model) before introducing the experiment and data. The basic method is well established and is a good statistical practice for analysing single reliability studies [12, 13].

The simplest rater reliability experiment generally involves \( p \) raters each rating a set of \( n \) subjects. In our case, the raters are nurses that are each rating the same set of PUs. Arguing that ordinal data are approximately normally distributed, one can attempt to analyse the data utilizing a random-effects one-way analysis of variance (ANOVA) model. Other names for the model include ‘variance components’ or ‘linear mixed’ models [14–18]. The key feature of such a model is that it provides between-subject and within-subject variance parameters. The ratio of the quantities ‘between-subject variance parameter’ and the ‘total variance parameter’ is the ICC. A perfect agreement among raters would lead to an ICC = 1.0 and really poor agreement would lead to an ICC = 0.0.

2.2. Simple model

A common model for reproducibility studies is the one-way random-effects model:

\[
y_{jk} = \mu + d_j + e_{jk}, \quad j = 1, 2, 3, \ldots, n; \quad k = 1, 2, 3, \ldots, p
\]

where \( y_{jk} \) refers to the \( k \)th rating performed on the \( j \)th subject (PUs), \( \mu \) is the overall mean, the subject effects \( d_j \) are \( \text{N}(0, \sigma_d^2) \), and the random errors \( e_{jk} \) are \( \text{N}(0, \sigma_e^2) \).

The ICC, \( \rho \), is defined as \( \rho = \sigma_d^2 / (\sigma_d^2 + \sigma_e^2) \) and is a measure of the reliability of the raters rating the same subject. A point estimate of \( \rho \) is \( \hat{\rho} = (F - 1) / [F + (p - 1)] \) where \( F \) is the \( F \)-statistic.
from the one-way ANOVA table for testing for significant between-subject variance. A \((1 - \alpha)\)100 per cent confidence interval is

\[
\left(\frac{F}{F_L} - 1\right) / \left(\frac{F}{F_L} + (p - 1)\right), \left(\frac{F}{F_U} - 1\right) / \left(\frac{F}{F_U} + (p - 1)\right)
\]

where \(F_L\) and \(F_U\) are, respectively, the \((\alpha/2)\)100 and \((1 - \alpha/2)\)100 per cent points in the cumulative \(F\)-distribution with \(n - 1\) and \(n(p - 1)\) degrees of freedom. A useful advancement from Giraudieu and Mary [19] is a normal approximation of the confidence interval for the ICC using the standard error s.e. \(\hat{\rho} = \sqrt{2[1 + (p - 1)\rho]}(1 - \rho) / \sqrt{np(p - 1)}\). The approximation is useful for designing future studies.

It is noted that an alternative to the above specification would be to fit a mixed model using restricted maximum likelihood estimates (REML) since using an ANOVA table essentially uses method of moment (MOM) estimates which could lead to negative \(\hat{\rho}\) (e.g. [20]). However, the classic simple random-effects model is useful for helping us summarize the data that we introduce next.

2.3. Data

For our PU study, it was decided for design purposes that we would focus on inference of rater reliability of an individual hospital. The research team decided that each hospital should have a goal for a minimum of \(n = 15\) PUs and \(p = 6\) raters. Some hospitals had more or less PUs and raters available. Using Giraudieu and Mary [19] formula, and assuming a correlation of \(\rho = 0.4\), the confidence interval of the ICC, for an individual hospital, of our proposed design has an expected half-width of 0.24.

After communicating design protocol to hospital participants, we directed the implementation of the reliability study. We currently have data from 20 hospitals. Hospitals deviated from our specification of \(n = 15\) PUs and \(p = 6\) raters. The summary statistics for the data are given in Table I. Among all of the PUs staged in this study, most tended to be stage II PUs (39 per cent). Ten per cent of the PUs were staged as unstageable. Hospital #2 raters classified 56 per cent of their PUs as unstageable and Hospital #20 raters classified 1 percentage of their PUs as unstageable. The high number of unstageable PUs can be problematic for data analysis purposes because of the wide range of uncertainty in the staging. We remedy the unstageable problem by dropping the unstageable PUs and creating estimates based on the one-way ANOVA model using the staged PUs only. For the moment, we treat the staged PUs as continuous.

To obtain point estimates and 95 per cent confidence intervals for the ICC in each hospital, we counted the number of PUs staged within each hospital \((n)\) and the average number of raters within each hospital \((p)\). The reason for this was to introduce the data and the basic reliability approach. The number of PUs rated within each hospital varied from 6 to 108 and the number of raters within each hospital varied from 3 to 9. The \(F\)-statistic from the one-way ANOVA was used to estimate the ICC. The estimated ICC varied from 0.06 to 1.00. The estimated standard error ranged from 0.01 to 0.16.

3. BAYESIAN HIERARCHICAL MODEL AND PRIOR DISTRIBUTIONS

The one-way ANOVA random-effects model has several shortcomings. First, the treatment of the staged PUs as continuous is incorrect. Each of the staged outcomes is ordinal. A common approach
Table I. Summary statistics for the raw data among the 20 hospitals in this study.

<table>
<thead>
<tr>
<th>Hospital</th>
<th>0 (%)</th>
<th>I (%)</th>
<th>II (%)</th>
<th>III (%)</th>
<th>IV (%)</th>
<th>Un (%)</th>
<th>n</th>
<th>F</th>
<th>( \hat{\rho} )</th>
<th>( \rho_L )</th>
<th>( \rho_U )</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>0</td>
<td>60</td>
<td>23</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>1.31</td>
<td>0.06</td>
<td>0.00</td>
<td>0.50</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>11</td>
<td>14</td>
<td>2</td>
<td>17</td>
<td>35</td>
<td>10</td>
<td>5</td>
<td>150.26</td>
<td>0.97</td>
<td>0.93</td>
<td>0.99</td>
</tr>
<tr>
<td>3</td>
<td>51</td>
<td>33</td>
<td>12</td>
<td>0</td>
<td>5</td>
<td>11</td>
<td>5</td>
<td>24.21</td>
<td>0.82</td>
<td>0.68</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>44</td>
<td>9</td>
<td>18</td>
<td>32</td>
<td>26</td>
<td>5</td>
<td>11.82</td>
<td>0.73</td>
<td>0.61</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>14</td>
<td>22</td>
<td>43</td>
<td>5</td>
<td>2</td>
<td>14</td>
<td>12</td>
<td>4</td>
<td>3.82</td>
<td>0.41</td>
<td>0.17</td>
<td>0.69</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>27</td>
<td>57</td>
<td>10</td>
<td>2</td>
<td>27</td>
<td>6</td>
<td>8.93</td>
<td>0.57</td>
<td>0.44</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>22</td>
<td>24</td>
<td>4</td>
<td>32</td>
<td>17</td>
<td>15</td>
<td>8</td>
<td>128.81</td>
<td>0.94</td>
<td>0.90</td>
<td>0.97</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>78</td>
<td>18</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>16</td>
<td>6</td>
<td>11.21</td>
<td>0.63</td>
<td>0.48</td>
<td>0.78</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>11</td>
<td>78</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>8</td>
<td>6</td>
<td>Infinite</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>13</td>
<td>62</td>
<td>13</td>
<td>9</td>
<td>3</td>
<td>8</td>
<td>23.04</td>
<td>0.73</td>
<td>0.54</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>24</td>
<td>36</td>
<td>32</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>108</td>
<td>6</td>
<td>14.38</td>
<td>0.69</td>
<td>0.63</td>
<td>0.75</td>
</tr>
<tr>
<td>12</td>
<td>21</td>
<td>15</td>
<td>41</td>
<td>15</td>
<td>6</td>
<td>2</td>
<td>21</td>
<td>8</td>
<td>11.22</td>
<td>0.56</td>
<td>0.42</td>
<td>0.72</td>
</tr>
<tr>
<td>13</td>
<td>16</td>
<td>42</td>
<td>21</td>
<td>4</td>
<td>11</td>
<td>7</td>
<td>20</td>
<td>3</td>
<td>1.85</td>
<td>0.22</td>
<td>0.00</td>
<td>0.48</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>12</td>
<td>27</td>
<td>30</td>
<td>20</td>
<td>8</td>
<td>14</td>
<td>6</td>
<td>15.05</td>
<td>0.70</td>
<td>0.54</td>
<td>0.85</td>
</tr>
<tr>
<td>15</td>
<td>10</td>
<td>18</td>
<td>31</td>
<td>13</td>
<td>13</td>
<td>15</td>
<td>14</td>
<td>7</td>
<td>24.08</td>
<td>0.77</td>
<td>0.63</td>
<td>0.88</td>
</tr>
<tr>
<td>16</td>
<td>3</td>
<td>32</td>
<td>37</td>
<td>16</td>
<td>0</td>
<td>12</td>
<td>32</td>
<td>6</td>
<td>110.37</td>
<td>0.95</td>
<td>0.92</td>
<td>0.97</td>
</tr>
<tr>
<td>17</td>
<td>0</td>
<td>48</td>
<td>35</td>
<td>2</td>
<td>2</td>
<td>15</td>
<td>18</td>
<td>4</td>
<td>8.98</td>
<td>0.67</td>
<td>0.50</td>
<td>0.81</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>10</td>
<td>64</td>
<td>12</td>
<td>2</td>
<td>11</td>
<td>27</td>
<td>5</td>
<td>21.25</td>
<td>0.80</td>
<td>0.71</td>
<td>0.88</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
<td>12</td>
<td>51</td>
<td>14</td>
<td>6</td>
<td>14</td>
<td>31</td>
<td>6</td>
<td>8.88</td>
<td>0.57</td>
<td>0.44</td>
<td>0.70</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>21</td>
<td>36</td>
<td>28</td>
<td>15</td>
<td>1</td>
<td>15</td>
<td>9</td>
<td>40.33</td>
<td>0.81</td>
<td>0.71</td>
<td>0.91</td>
</tr>
<tr>
<td>All</td>
<td>9</td>
<td>26</td>
<td>39</td>
<td>10</td>
<td>6</td>
<td>10</td>
<td>347</td>
<td>7</td>
<td>10.44</td>
<td>0.57</td>
<td>0.54</td>
<td>0.61</td>
</tr>
</tbody>
</table>

Note: For the PU stage, the data are summarized as 0 (no PU), I–IV staged PUs, and unstageable. For the purposes of F-statistics and ICCs, we took out the ‘unstageables.’ The number of PUs is n, the number of raters is p, F is the one-way ANOVA F-statistic, \( \hat{\rho} \) is the estimated ICC, \( \rho_L \) and \( \rho_U \) are the lower and upper 95 per cent intervals, and SE is the standard error.

in applied research is to argue that the central limit theorem allows normal approximations of the distribution of the point estimates to be normal. However, in our data, this is a difficult argument to defend because of small sample sizes. For example, in Hospital #1 there are only five raters and six rated PUs. A second shortcoming is illustrated in Hospital #9 for which the within-subject variance is estimated to be 0, the estimated ICC is equal to 1 and the corresponding 95 per cent interval is between 1 and 1. This is an obvious problem since no nursing researcher would believe that a hospital would have raters that perfectly agree on the rating of PUs—it is scientifically unreasonable. The point estimates, based only on the data from that single hospital, suffer from small sample sizes, and approaches borrowing strength across hospitals would likely perform much better. The benefit of borrowing strength is well documented in the literature (e.g. [21, 22]). A third shortcoming of the model is that the model does not include unstageable data. We would like to include the unstageable data in the analysis. It is believed, clinically, that unstageable PUs are actually between stages III and IV because total loss of skin is necessary before an obstruction can occur. Thus if three raters rate a PU as stage I and a fourth rates it as unstageable, one will falsely inflate the rater agreement if the unstageable data are not included in the analysis.

The literature suggests an alternative for addressing the first shortcoming by utilizing an ordinal regression model with either a probit or logistic link function (e.g. [23]). In this approach, one can argue that there is a latent continuous variable that is realized as ordinal outcomes. Thus allows researchers to correctly model the fact that we do not observe a continuous outcome. Using
maximum likelihood estimation and related frequentist-based theory, one can obtain point estimates and intervals for ICC.

There is much relevant work on probit models for rater agreement. There are two review papers [24, 25] that are helpful starts to the literature. In particular, there is literature coming from the tradition of latent structure analysis. Uebersax and Grove [26] model the distribution as a mixture of two normal distributions, which, among other things, is a way to model a single, skewed distribution. Uebersax [27] considered a semi-parametric or arbitrary latent trait distribution. These papers also use various constraints on rater thresholds to identify components of disagreement due to rater differences in bias and category definitions, and consider rater differences in measurement error.

A Bayesian alternative to the ordinal analysis, with good small sample properties, is an efficient simulation algorithm presented by Albert and Chib [28]. However, as in the one-way ANOVA model, hospitals with perfect agreement would be problematic as the maximum likelihood estimators would not exist and the posterior distribution would not be proper. The Bayesian approach can remedy this problem if the researchers have a proper prior distribution. But defining a proper prior distribution requires good prior information which may not be available.

Fortunately, the Bayesian paradigm has a good solution to this problem by making use of the information in the other hospitals via a hierarchical model. Essentially, imagine that all of the different ICCs from the hospitals are random draws from a distribution. Then, the posterior distribution of the individual hospital ICCs will be a combination of the information supplied from their own hospital and information from all of the other hospitals. This is an ideal way to remedy the shortcoming found in ‘perfect’ hospitals. The other hospitals are combined to give a realistic point and interval estimate of the ICC in the individual hospitals.

The hypothetical BHM does not directly remedy the issue of our desire to include unstaged PUs. However, the concept of data augmentation argued in Albert and Chib [28] can be extended in such a way that the unstageable PUs can be included in the analysis.

In the next section, we define an ordinal probit BHM that addresses all of the three shortcomings previously mentioned. By treating the latent values from the unstaged items as being left-censored, we can propose a solution for including these data previously dropped from the analysis.

3.1. Bayesian hierarchical model

Our ordinal probit BHM is fit to the data from 20 hospitals. The variable $y_{ijk}^*$ is a latent variable representing a continuous score for the $k$th rater from the $i$th hospital on the $j$th PU. The $y_{ijk}^*$ are unknown; however, given the observed ordinal data $y_{ijk}$, the distribution of $y_{ijk}^*$ follows a truncated normal distribution [28]. Therefore, we define a continuous, normal-based model on the latent variables for the $i$th hospital, $j$th PU, and the $k$th rater as

$$y_{ijk}^* = \mu + d_{ij} + e_{ijk}, \quad i = 1, 2, 3, \ldots, N; \quad j = 1, 2, 3, \ldots, n_i; \quad k = 1, 2, 3, \ldots, p_{ij}$$

where

- $\mu$ is the overall mean rating across all hospitals, PUs, and raters;
- $d_{ij} \sim N(0, \sigma_d^2)$ is the random effect of the $j$th PU within $i$th hospital;
- $e_{ijk} \sim N(0, \sigma_e^2)$ is the random effect of the $k$th rater within the $j$th PU within $i$th hospital; and
- $N$ is the number of hospitals in the study, $n_i$ is the number of PUs within hospital $i$, and $p_{ij}$ is the number of raters in hospital $i$ rating PU $j$.

This model allows the ICC to vary by hospital. Specifically, the ICC in the $i$th hospital is $ho_i = \frac{\sigma_d^2}{\sigma_d^2 + \sigma_e^2}$.

One way to connect the ordinal $y_{ijk}$ to the latent $y_{ijk}^*$ is through the following function:

$$y_{ijk} = \begin{cases} 
0 & \text{if } y_{ijk}^* < \tau_1 \\
1 & \text{if } \tau_1 \leq y_{ijk}^* < \tau_2 \\
2 & \text{if } \tau_2 \leq y_{ijk}^* < \tau_3 \\
3 & \text{if } \tau_3 \leq y_{ijk}^* < \tau_4 \\
4 & \text{if } \tau_4 < y_{ijk}^* 
\end{cases}$$

for a series of cut-point parameters $\tau_1, \tau_2, \ldots, \tau_4$. The 0–4 correspond to ‘No PU,’ and Stages I–IV, respectively. Note that the cut points are subjected to an ordering constraint ($\tau_1 < \tau_2 < \tau_3 < \tau_4$).

The piece missing in this specification is the ‘unstageable’ ratings. If the rater rates the PU as ‘unstageable’ then we know that $\tau_3 \leq y_{ijk}^*$ and $y_{ijk}^*$ is either III or IV. This assumption is based on third author’s expert clinical opinion. More specifically, that total loss of skin (stage III–IV) is necessary before an obstruction can occur. Statistically this is a censoring issue and can be dealt with, computationally, on the latent variables, using data augmentation [29, 30].

However, there is a debate regarding this assumption in the nursing literature. We might not really know that all unstageable ratings are between III and IV. For example, it might not be reasonable to think that a hospital with over 50 per cent unstageables had that many staged III–IV. Maybe some nurses are indecisive and give this rating for many PUs. Therefore, as discussed later, we assess the sensitivity of this assumption.

As mentioned previously, the error of $y_{ijk}^*$ is normally distributed and this type of modelling is known in the literature as probit modelling.

The unknown portions of this model include the usual model parameters as well as $y_{ijk}^*$’s and $\tau_k$’s. The next step in the modelling process is to link the hospital-level parameters via a hierarchical distribution. There are at least two ways to model these parameters hierarchically. One natural way to model the parameters is via the variance components. By placing semi-conjugate priors on the variance components, we allow the hospitals to borrow strength from each other. However, as discussed in the introduction, we are primarily interested in inference on the ICCs. Therefore, we choose to link hospitals hierarchically by placing a distribution on $\rho_i$. But because there are two variance components defining $\rho_i$, we also place a hierarchical distribution on $\sigma_e^2_i$. Since the ICC is on the line from 0 to 1, one family of distributions to consider is the Beta family of distributions. A natural choice for the prior of $\sigma_e^2_i$ is the semi-conjugate distribution which is a family of inverse gamma distributions. More specifically,

- $\rho_i \sim \text{Beta}(\alpha, \beta)$;
- $\sigma_e^2_i \sim \text{IG}(\chi + 1, \eta)$.

The $\alpha$ parameter represents the prior numerator for the ICCs and $\alpha + \beta$ represents the prior denominator for the ICCs. The parameters $\chi + 1$ and $\eta$ represent the scale and shape parameters from the prior on $\sigma_e^2_i$. The individual hospital ICCs are $\rho_i = \frac{\sigma_d^2}{\sigma_d^2 + \sigma_e^2}$ and the expected value
of the Beta leads to an expected value of the population ICC of $\theta = \frac{\alpha}{\alpha + \beta}$. Note that in our implementation we do not actually fix the parameters $\alpha$, $\beta$, $\gamma$, and $\eta$, but specify hyperpriors for these parameters.

The above model assumes that the ICCs for each of the hospitals are random draws from the same ‘super population.’ This assumes that the hospitals are exchangeable. That is, there is little information about why hospitals would have different reliabilities (ICC). At this point, we are simply trying to understand what the state of reliability is among all of the hospitals in NDNQI. This is done utilizing the BHM. In future studies, we would like to understand why hospitals differ in their ICC by utilizing predictor variables such as hospital contextual variables (e.g. bed size) or rater information (e.g. percentage of raters that are wound care specialists) to help predict these differences and assess the assumption of exchangeable hospitals. Note that incorporating rater information would require tracking raters and define an extra random effect for the raters.

Johnson and Albert [11] summarize the current theory and practice of ordinal Bayesian modelling where they also discuss computational strategies. However, in our approach, we handle the probit model a little differently. The Johnson and Albert approach can be applied to our data since the unstaged PUs just span several ordinal values (censored). To extend their algorithm to left censoring, the latent values will be drawn as truncated normals where the truncated ranges for censored values would span more than two neighbouring cut points. We use WinBUGS as a computational tool for implementation. To program the approach in WinBUGS, we use a similar technique found in Mwalili et al. [31]. In our program, we account for the likelihood of unstageable observations. To clearly illustrate this, consider the likelihood $(L_{ijk})$ in the following:

$$L_{ijk} = \begin{cases} 
\Pr(y_{ijk}^* < \tau_1) & \text{if } y_{ijk} = 0 \\
\Pr(\tau_1 < y_{ijk}^* \leq \tau_2) & \text{if } y_{ijk} = 1 \\
\Pr(\tau_2 < y_{ijk}^* \leq \tau_3) & \text{if } y_{ijk} = 2 \\
\Pr(\tau_3 < y_{ijk}^* \leq \tau_4) & \text{if } y_{ijk} = 3 \\
\Pr(\tau_4 < y_{ijk}^*) & \text{if } y_{ijk} = 4 
\end{cases}$$

The piece missing in this likelihood specification is the ‘unstageable’ ratings. If the rater rates the PU as ‘unstageable’ then we know that the likelihood is $L_{ijk} = \Pr(\tau_3 < y_{ijk}^*)$. Essentially, the problem can be viewed as right censored for stage IV and unstageable PUs; left censored for no PU (0); and interval censored for the rest.

We next need to specify the prior distribution for the unknown parameters before we are able to calculate a posterior distribution.

### 3.2. Priors

In order to run the analysis, there are several parameters that need prior distributions. The parameters needing prior information includes a parameter $\mu$ that describes the overall mean of the latent variable; $\gamma$ and $\eta$ that are parameters for the variance of the error; $\alpha$ and $\beta$ are the parameters for the ICCs; and the cut-point $\tau_k$’s. Some of the parameters can be given vague priors and others require subjective input.

The parameters that can have vague priors start with the cut points are non-informative with $\tau_k \sim N(0, 100^2)$ when $k = 1, 2, 3, 4$. This allows the data to determine reasonable distances between
the cut points. We are assuming that the ICCs on the hospitals are exchangeable but in a vague kind of way a priori. For the hyper parameters on the hierarchical model, we set \( \alpha, \beta \sim \text{exp}(10) \), \( \chi \sim \text{exp}(2) \), and \( \eta \sim \text{exp}(1) \) corresponding to 5, 50, and 95 percentiles for \( \sigma_e \) of 0.13, 0.5, and 1.6, respectively, and for \( \rho_i \) of 0.01, 0.50, and 0.99. This leads to a vague prior for the ICCs, but an informative prior for the variance of the errors.

For identifiability purposes, we set \( \mu = 0 \). Together, this allows for our model to be identifiable. Typically, to accomplish this, one sets \( \sigma = 1 \) because the scale is established by both the cut-point parameters and \( \sigma \). One of the two must have a specification. However, we wish to allow the error to vary from hospital to hospital; therefore, we simply define a tight prior to set an approximate scale across all of the hospitals.

3.3. Comment on the Bayesian hierarchical model

To understand how the Bayesian hierarchical promotes the borrowing of strength, consider the complete conditional posterior distribution of \( \rho_i \) from hospital \( i \), conditional on all other parameters in the study. For simplicity, assume that hospital \( i \) has two raters. The log-posterior distribution for \( \rho_i \) is equal to

\[
\log \sum_j \{ \text{MVN}(\begin{bmatrix} z_{ij1}^* \\ z_{ij2}^* \end{bmatrix} | \mathbf{0}, \mathbf{R}) \} + (\alpha - 1) \log(\rho_i) + (\beta - 1) \log(1 - \rho_i) + \text{constant}
\]

where MVN is multivariate normal with standardized latent variables for raters 1 and 2 \( (z_{ijk}^* = (\gamma_{ijk} - \mu) / \sigma_e) \) with mean vector \( \mathbf{0} \) and correlation matrix

\[
\mathbf{R} = \begin{bmatrix} 1 & \rho_i \\ \rho_i & 1 \end{bmatrix}
\]

The first part of the log-posterior distribution is a function of the observed data from hospital \( i \). The information from the other hospitals is embedded in \( \alpha \) and \( \beta \) since their posterior distributions are proportional to the \( \rho_i \)'s from all of the hospitals. This illustrates how the posterior distribution for the ICC in hospital \( i \) is a weighted average of the information within hospital and the information across hospitals (borrows strength).

4. GENERAL COMPUTATION

4.1. Main model

A Bayesian estimation of the model parameters was performed using Markov chain Monte Carlo (MCMC), implemented in WinBUGS [21, 32–35]. Interested readers may contact the first author for the WinBUGS code used in this paper. Following a burn-in of 5000 iterations, the posterior distributions were monitored over a further 10 000 iterations of the MCMC. The length of burn-in and monitoring was sufficient to achieve convergence as assessed by trace and autocorrelation plots. With 10 000 iterations \( \rho_{13} \) (Hospital #13, full model) has the largest MCMC error (0.003). We are satisfied with this error rate. Increasing the iterations to 100 000, for example, would reduce the MCMC error to 0.001.
Several models were examined before choosing the final model. The relative fit of each of the models is assessed using Deviance Information Criterion (DIC [34]). The four variations of the final model include:

(M1) **Null model**: Set the variance of the random effect of the \( j \)th PU within \( i \)th hospital to be equal to 0 (\( \rho_i = 0 \)) and set the variance of the random effect of the \( k \)th rater within \( i \)th hospital to be equal across all hospitals (\( \sigma_e^2 = \sigma_r^2 \)). This is a model stating that there is an ICC = 0. This is a naive model but illustrates the relative merit of the more appropriate models.

(M2) **Equal variance components**: Set the variance of the random effect of the \( j \)th PU within \( i \)th hospital to be equal across all hospitals (\( \sigma_e^2 = \sigma_r^2 \)) and the variance of the random effect of the \( k \)th rater within \( i \)th hospital to be equal across all hospitals (\( \sigma_e^2 = \sigma_r^2 \)).

(M3) **Unequal variance components**: Full model.

We assess the adequacy of the fit of the final model, the model with the lowest DIC, using a Bayesian chi-squared test for goodness of fit [36]. The Bayesian chi-squared test for goodness of fit is conceptually like its classic parent except in this case the goodness of fit is calculated at each of the iterations of the MCMC. Therefore, each of the iterations produces a \( p \)-value conditional on the particular draw from the posterior distribution. Instead of calculating a single goodness-of-fit \( p \)-value, we obtain a posterior distribution of the \( p \)-value. We summarize the posterior distribution using usual summary statistics. Johnson’s notation can explain the computational details. First, at each iteration, the number of observations of \( y^*_{ijk} \) at each iteration of the MCMC that fall into the interval (\( \frac{\phi^{-1}(\mu_i) - d_{ij}}{\sigma_e} \) + \( \frac{\phi^{-1}(\mu_i) - d_{ij}}{\sigma_r} \) is the standard normal quantile function. For a sample size of 2753, 2753 4 = 24 equal probability bins for \( a_u \) were selected [36]. See Johnson [36] for further details of calculating the chi-squared test statistic. We define a good fit when the 5 percentile of the \( p \)-value of the chi-squared test statistic is above 0.01 and the 95 percentile is below 0.99. Models not fitting this criterion ‘must therefore be attributed to either dependence between sampled values of the chi-squared statistic, or lack of fit’ [36]. Therefore, chi-squared values in this range can lead to a properly specified conditional-independent model.

Note that Johnson’s chi-squared test statistic is different from the one proposed by Gelman et al. [37]. They are different because Johnson’s approach requires categorizations and Gelman et al. do not. Additionally, Johnson has a distribution for the \( p \)-values whereas Gelman et al. have one \( p \)-value for the entire model.

4.2. **Impact of Bayesian hierarchical model and staging**

The results of our data modelling suggest that the Bayesian analysis is preferable to a simple non-Bayesian analysis. However, in a statistical methodology paper, it is helpful to disentangle these issues. One approach is to use only the staged ulcers to compare the models. This identifies any benefits of avoiding the normality assumption for ordinal data. Once that comparison is made, it is possible to compare different assumptions for the unstageable ulcers to see their effect on the inferences. Therefore, we will use several models for comparisons:

(M4) **Classic approach: staged data**: Utilizing the observed unstaged data, we model with the frequentist approach, which is non-hierarchical, for generating 90 per cent intervals for the ICC for each hospital (Table I). To
obtain an estimate for the population of hospitals, we treat each realization as coming from one super population. This is too liberal as the population 90 per cent interval will be much shorter than the interval supplied by any of the BHMs.

(M5) *Approximate BHM: staged data.* This model is a BHM on the same observed staged data. This allows us to see the impact that shrinkage has on the 90 per cent intervals.

(M6) *Ordinal BHM: staged data.* A fully ordinal Bayesian hierarchical model for the staged data allows us to understand the impact of that the ordinal data model has on analysis.

(M7) and (M8) *Ordinal probit hierarchical model: staged and unstaged data.* These models are the same as M6 except that we add the unstaged data and model the unstaged data in two different ways. Model M7 assumes that the unstaged values are from the latent variable $\tau_1 \leq y_{ijk}^*$, which assumes that the unstaged latent values are between stages I and IV. Model M8 (same as M3) assumes that the unstaged values are from the latent variable $\tau_3 \leq y_{ijk}^*$, which assumes that the unstaged latent values are between stages III and IV (this is considered the ‘final model’).

5. RESULTS

5.1. The estimation of model parameters

The distributions of the model parameters are estimated using a Gibbs sampler in WinBUGS. Inspection of the trace plots and autocorrelation plots for all model parameters indicated quick convergence to the stationary distribution.

When comparing the first three models, model M3 has the lowest DIC:

<table>
<thead>
<tr>
<th>Model</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>4041.4</td>
</tr>
<tr>
<td>M2</td>
<td>3684.1</td>
</tr>
<tr>
<td>M3</td>
<td>3638.0</td>
</tr>
</tbody>
</table>

A 90 per cent interval and median for the chi-squared test statistic in M3 is 18.7–47.8 and 31.2, respectively, corresponding to 90 per cent interval and median $p$-values of 0.23–0.99 and 0.85, respectively. This is a reasonable fit.

The summary statistics for the final model include 5, 50, and 95 percentiles for the smallest (ranked by posterior median) $\sigma_{d_0} = (0.68, 1.05, 1.69)$; the largest $\sigma_{d_3} = (3.45, 5.01, 7.82)$; the smallest $\sigma_{e_{16}} = (0.41, 0.53, 0.65)$; and the largest $\sigma_{e_{13}} = (2.42, 2.97, 3.71)$. The intervals for the $\tau_k$’s are $\tau_1 = (-4.91, -4.24, -3.15)$, $\tau_2 = (-1.46, -1.20, -0.91)$, $\tau_3 = (1.51, 2.34, 2.81)$, and $\tau_4 = (3.56, 5.08, 6.10)$. For the hyper parameters, the summary statistics are $\alpha = (3.49, 7.38, 14.67)$, $\beta = (1.03, 1.88, 3.34)$, $\chi = (2.00, 2.11, 2.45)$, and $\eta = (1.47, 2.04, 2.78)$.

The posterior summaries for the final BHM with non-informative priors for the ICCs are presented in Figure 1—rank ordered by the posterior median. We also overlay the point estimates.
of the ICC calculated with the raw data (non-Bayesian and pre-probit). The BHMs tend to estimate the $\rho_i$’s much higher than the ICC with the raw data (M4). The intervals for the BHMs do not overlap the ICC calculated from the raw data in Hospitals #1, #4, #6, #8, #9, #11, and #13. This discrepancy occurs for many reasons. First, the raw data ignore the fact that these values are not continuous. Second, the raw data model ignores the fact that some of the raters are rating a PU as unstageable. These data should not be eliminated in the analysis. Therefore, using the raw data produces an ICC that is a very rough estimate of the reliability. We will isolate the effect of these reasons after addressing the scientific question.

5.2. Addressing the scientific questions

Using the BHMs, we see a clear difference in the PU rating reliability among hospitals. Hospital #2 is almost perfect and Hospital #13 is moderate. However, in Hospital #13, there is a lot of variation in this estimate because only three raters rated 30 PUs. The 95 per cent credible interval (CrI) indicates that there is a reasonable chance that the ICC is as high as 0.67. None of the hospital ICCs is ‘significantly’ below 0.50. Hospitals #2, #7, and #16 are near perfect.

The posterior distribution of the expected population ICC ($\theta = \alpha / (\alpha + \beta)$) has a 90 per cent interval of 0.73–0.85 with a median of 0.80. We can see that it is most likely that the lower ICCs are unusual for the population of hospitals.

Figure 1. Summarization using 5, 50, 95 percentiles for the ICCs in hospitals from the fully probit ordinal Bayesian hierarchical model (●) and the point estimates utilizing the raw data assuming ordinal data are normally distributed and utilizing a one-way ANOVA at each individual hospital (○).
5.3. Results for impact of Bayesian hierarchical model and staging

In Figures 2–4, we compare the models that vary because of the data used and the assumptions. Some of the models were run on the raw data with only the staged ulcers (M4, M5), then assuming the data are non-normal (M6), and the other models were run using all of the data including unstaged PUs (M7–M8). Figure 2 best describes the impact of the BHM. The 90 per cent interval from M4 to M5 indicates that the shrinkage from the BHM is necessary. The classic model (M4) has too short an interval and underestimates the population ICC.

The population ICC estimated from M5 to M8 does not change as drastically. Therefore, to understand these model differences, consider the individual hospital ICC intervals in Figures 3 and 4. In Hospital #3 (Figure 3), the model that assumes PUs are staged III–IV has a lower-point estimate (wider interval) than their Bayesian competitors. This seems like an odd result until one inspects the raw data. Of the 11 PUs rated in Hospital #3, two have one rater rating as unstaged. In both PUs, the other raters rated no PU. Therefore, the drastic differences in models 7 and 8 are due to the fact that the assumption of unstageables being between III and IV is inconsistent with the other ratings. This is correctly reflected in model 8. In Hospital #9 (Figure 4), the point estimates are the same except that M5 is high and has too short of an interval. The intervals M6
Figure 3. The 5, 50, 95 percentiles for the ICCs in Hospital #3 using M4–M8: M4 = classic approach: staged data; M5 = approximate Bayesian hierarchical model: staged data; M6 = ordinal Bayesian hierarchical model: staged data; M7 = ordinal probit hierarchical model, staged and unstaged data (assumed I–IV); and M8 = ordinal probit hierarchical model, staged and unstaged data (assumed III–IV).

and M7 are longer than M8. In addition, M4 indicates that the 90 per cent interval is 1–1. The BHM uses shrinkage to correctly reflect that this is too optimistic as no scientist would believe a hospital has perfect reliability.

In summary, the effect of the modelling is drastic when one moves from classic inference to a BHM. The effect is less drastic when bringing in the ordinal modelling and the unstaged PUs. However, the effect is not negligible when looking at the results from the point of view of the individual hospitals.

5.4. Practical considerations

Some limitations of this study are worth mentioning. First, if the goal is to describe rater agreement, then the model might miss an important feature of the data as we ignore the disagreement among raters about whether or not a PU is unstageable. In our approach the only disagreement between a rating that a PU is stage III, say, and a rating that the PU is unstageable is the difference between a rating of a III and a rating of III or IV (using NPUAP and AHRQ assumptions). It does not account for the fact that one rater believed the information was sufficiently clear to allow for staging and one rater did not.

However, our goal is to determine the reliability of ratings data. Therefore, disagreement about whether or not a PU can be rated is not an important source of error. In this case, the goal is to understand the error in the ratings and how this can affect inferences hospitals are making based
on these ratings. Rater agreement is only a means to understanding the measurement error. In this case, the variability in the rating values is the only information that is necessary and the source of variability is not important.

We think the goal of reliability studies is to understand measurement error and support correct inferences from ratings data. Thus, we think this model and approach addresses the problem of interest.

The model in this paper could be used in interpreting ratings data for the NDNQI database. We can use the results from the BHM to estimate the measurement error in the staging of PU data. This measurement error can be estimated by calculating the posterior-predictive distribution of a rating from a randomly selected rater in hospital \( i \), conditional on observing a staged or unstaged PU from another randomly selected rater. This measurement error can be incorporated into the reports that NDNQI supplies participating hospitals. This can properly incorporate the uncertainty in the judgement across raters in the same hospital.

6. CONCLUSION

In this paper, we propose and implement an ordinal probit BHM for assessing the reliability of staging PUs. The approach assumes that there is a continuous latent variable \([38–40]\) for each
rating of a PU and is realized in one of the five stages: no PU, stages I–IV. A caveat in the staging is that some raters view the PU as ‘unstageable’ possibly due to obstruction of the view of the PU. The BHM accounts for this uncertain rating. We call this the exact Bayesian model.

If one considers our approach, at an individual hospital, after integrating out the rater effects, it leads to a special case of the multivariate probit model used by Chib and Greenberg [41]. The difference is that we assume that the correlation among raters is constant within hospital—the correlations are set to be equal. One could consider extending Chib and Greenberg’s model to be a hierarchical model using similar techniques as in this paper. However, by proper specification of the random effects our modelling has natural links to the linear mixed model literature, and Chib and Greenberg has more of a multivariate analysis flavour. Using data augmentation techniques, both techniques can simultaneously handle missing and censored data values.

A competing approach for the exact Bayesian approach was based on the raw PU staging. We use the summary statistics of the point and interval estimates from the one-way ANOVA’s $t$-statistic. However, this model assumes that the ordinal data are approximately continuous. Many researchers in the past have made this assumption by arguing that the central limit theorem provides sufficient statistics to summarize the distribution in each unit (in our case, a hospital). However, a combination of the ordinal data and the fact that we have an uncertain rating (unstaged PU) leads to the model providing inference that is too conservative. More specifically, the model based on raw data underestimates the ICC relative to the Bayesian model.

One could also consider, in the future, extending the model to include ‘rater’ as a random effect. In this case, the extra variance component could be modelled through ICCs.

The simplicity of the raw data model is attractive since it can allow for relatively easy design of future studies. Our analysis suggests it is preferable to use a fully BHM in our rater reliability study. Other researchers may want to consider this probit ordinal BHM in other medical studies such as tumour rating. Our paper offers guidance as to when the extra work of the exact Bayesian model is necessary and when the approximate model should be avoided.

REFERENCES

8. Szor J, Bourguignon D. Description of pressure ulcer pain at rest and at dressing change. Journal of Wound, Ostomy and Continence Nursing 1999; 26:115–120.