Inter-rater Reliability of Nursing Home Surveys: A Bayesian Latent Class Approach

Byron J Gajewski, University of Kansas Medical Center
Sarah Thompson
Nancy Dunton
Annette Becker
Marcia Wrona
Inter-rater reliability of nursing home surveys: a Bayesian latent class approach

Byron J. Gajewski¹,*,†, Sarah Thompson², Nancy Dunton², Annette Becker² and Marcia Wrona³

¹Schools of Allied Health and Nursing, University of Kansas, Kansas City, KS 66160, U.S.A.  
²School of Nursing, University of Kansas, Kansas City, KS 66160, U.S.A.  
³Department of Biometry, School of Allied Health, University of Kansas, Kansas City, KS 66160, U.S.A.

SUMMARY

In the U.S., federal and state governments perform routine inspections of nursing homes. Results of the inspections allow government to generate fines for findings of non-compliance as well as allow consumers to rank facilities. The purpose of this study is to investigate the inter-rater reliability of the nursing home survey process. In general, the survey data involves 191 binary deficiency variables interpreted as ‘deficient’ or ‘non-deficient’. To reduce the dimensionality of the problem, our proposed method involves two steps. First, we reduce the deficiency categories to sub-categories using previous nursing home studies. Second, looking at the State of Kansas specifically, we take the deficiency data from 1 year, and use Bayesian latent class analysis (LCA) to collapse the sub-categories to a binary variable. We evaluate inter-rater agreement using deficiency data from two separate survey teams on one facility, a matched-pair design. We evaluate the agreement of the two raters on binary data using the weights from the LCA. This allows a two-by-two contingency analysis using a Bayesian beta-binomial model. We elicit informative prior distributions from the nursing home providers. Together, with the experimental data, this provides a posterior distribution of the kappa agreement of the raters for nursing home deficiency citation data. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: beta-binomial; MCMC; kappa; Quality of Care; informative prior

1. INTRODUCTION

In response to public concerns about resident safety and quality of care, the nursing home industry has faced increasing regulatory mandates and monitoring. To participate in U.S. government medical programs (Medicare and Medicaid), nursing facilities must meet U.S.
federal regulations or conditions of participation as established by the Centers for Medicare and Medicaid Services (CMS). Although the regulation and survey process is federally mandated, the individual state agencies are responsible for administering the survey process.

Federal regulations require each nursing facility to undergo a survey once every 9–15 months to assess for compliance. The survey process is designed to regulate care in nursing homes, and facilities must be in compliance with 191 federal regulations in order to operate. The 191 regulations fall into several categories, specifically: Resident Rights; Quality of Life; Quality of Care; Resident Assessment; Services: Dietary, Pharmacy, Rehabilitation, Dental and Physician; Physical Environment; and Administration [1]. State or federal surveyors cite deficiencies when a facility fails to be in substantial compliance with each of the regulations. The definition of substantial compliance is vague and the rule is entirely up to the whims of the raters.

Each survey follows a standardized procedural format that allows some flexibility once a surveyor is inside a nursing facility. The standardized format consists of offsite preparation, a review of facility data, entrance conference, initial tour, resident sample selection, information gathering, information analysis, and exit conference. The information gathered refines the focus of the survey from one moment to the next. This flexibility contributes to a potential weakness in the consistency of the survey process, i.e. there are differences from one facility survey to another in the depth of information obtained and in which areas are examined.

The purpose of this study was to estimate the reliability of the survey process with matched-pair experimental design. Specifically, this involves sending two survey teams to assess the same facility. The first survey team, called the regular survey team (RST), evaluates the nursing facility on all 191 regulations and they evaluate the nursing home using their usual survey protocol. Mathematically this results in a vector of 0’s and 1’s, where a 0 represents a facility non-deficient in a certain category and 1 represents deficient category. The second survey team, called the simultaneous survey team (SST), also evaluates the same nursing facility on all 191 regulations using the same process as the RST. This is repeated for a total of \( n = 12 \) nursing facilities.

The key scientific question of interest centers on estimating the reliability of the survey process. Specifically, what is the reliability of the survey process in terms of classifying a nursing facility into deficient or non-deficient categories? We assume, from previous research [1], six of the 191 items correspond to Quality of Care sub-scale. However, there is still a dimension problem because there are \( 2^6 = 64 \) possible response patterns and only 12 nursing homes in our reliability data set. To remedy this issue, we assume a latent variable called ‘Quality of Care’ produces the outcome of these regulations. To perform the classification analytically, we utilize Bayesian latent class analysis (LCA). We may study other traits besides Quality of Care using the other 185 variables.

We perform the LCA on the deficiency data from all the nursing facilities in the State of Kansas (our design state). From this analysis, we obtain item categories and class proportions. These parameters allow us to estimate the class categories for each nursing facility in the experiment from the assessments by both the regular survey team and the simultaneous survey team, essentially reducing the problem to a two-by-two contingency table, with columns corresponding to the classification from the regular survey team and rows corresponding to classification from the simultaneous survey team. Using a beta-binomial model, we estimate the distribution of the probability cells from the contingency table. These probability cells provide us with a measure of agreement, kappa.
We present a Bayesian method for analysing the matched-pair experiment. We discuss the design, data, and other statistical methods from the literature in Section 2. Subsequently, we describe the main approach, consisting of a latent class analysis and beta-binomial model, in Section 3. In Section 4, we provide a discussion of the construction of the prior distributions utilized in this study. The prior is a key feature of our analysis because we utilize enthusiastic information from the nursing home providers. We demonstrate the resulting method on the category, ‘Quality of Care’, in Section 5, and carry it forward to several other categories. We discuss the conclusions in Section 6.

2. GENERAL MODELLING AND DATA

2.1. Design

The purpose of the simultaneous surveys was to evaluate the consistency of the survey process between two survey teams, all from the State of Kansas. Kansas consists of six survey regions. Each region has at least two trained survey teams, a quality improvement coordinator, and a regional manager.

The protocol for simultaneous surveys consists of having one in-region (RST) and one out-of-region team (SST) survey a facility together. The RST receives its name because it is already in the regular schedule to review the facility. The SST follows the RST in the same facility. The SST was selected randomly from another region and the RST directed the survey. The RST and SST members were not allowed to discuss their observations or conclusions with each other, and all SST and RST team analysis meetings were separate. The goal was to assess the agreement between the two teams exposed to the same environment at the same point in time. A research staff member observed each simultaneous survey to ensure that protocol was followed.

There are strengths and weaknesses to the design of the simultaneous survey process. The primary strength is that surveyors are exposed to exactly the same conditions. One weakness is the SST did not have the ability to re-direct their exploration of potential problem areas, which would have the effect of making the two team assessments more similar and therefore increase the estimated reliability. They had to follow the RST. A second weakness is the familiarity of the RST with the facility is an issue since the SST comes from another region. A third weakness is that two teams in a facility at one time may have altered team behaviour. The RST may have felt scrutinized and identified more deficiencies than they might have otherwise. To help account for this, RST reports had to contain the usual amount of documentation and were subject to routine review by the regional manager. The SST may have felt competitive and identified more deficiencies than they would have otherwise.

2.2. Data

A key aspect of any solution to our defined problem allows summarization of the multivariate data into smaller, more manageable pieces. This is of particular importance given the number of U.S. federal deficiencies (F-tags), in this case a total of 191. State and federal agencies that study nursing facility deficiency data choose various approaches for data summary. A popular method is to simply add the number of deficiencies cited in the nursing facility. The range of deficiencies is between 0 and 191. This method is intuitively attractive because it is
so easy to interpret. The disadvantage of this approach is that nursing facility A may have ten deficiencies in ‘un-important’ F-tags, whereas nursing facility B may have ten deficiencies in ‘very important’ F-tags. The importance of particular F-tags may depend on the point of view of the provider or the consumer.

Our approach to the data summarization and modelling is different in two ways. First, we begin our enumeration by dichotomizing the F-tags. The dichotomized approach is consistent with other researchers [1]. Either the nursing facility is deficient in the particular F-tag or it is non-deficient. Second, we preserve the individuality of the F-tag category by modelling it as an individual binary data value. This approach still provides dimensionality issues because the 191 F-tags need reduction because we have too many variables for the given sample size. Initially, similar to other researchers, we reduce 191 binary values down to 35, from the work by Mullan and Harrington [1]. Using a weighted confirmatory factor analysis on the tetrachoric correlation matrix, the authors categorize the F-tags into seven major deficiency categories: Quality of Care (six F-tags), Abuse (four F-tags), Assessment (five F-tags), Rights (five F-tags), Environment (five F-tags), Nutrition (five F-tags) and Pharmacy (five F-tags). They assume, like us, that each F-tag is a binary value. They obtain factor loadings for each of the factors except they modify the inference for these loadings with a weighted likelihood. While factor analysis on the tetrachorics may not be ideal [2] it does seem to perform well in this case and allows for consistency with the work of Mullan and Harrington. Mullan and Harrington’s results make sense in the nursing home application and they have a large sample size of nursing facilities (N = 7557).

Their results show 191 F-tags can be reduced to 35 because they met the face validity and had a simple structure and each F-tag loaded similarly to F-tags within sub-scales [1]. Based on this analysis, states may increase reliability in the survey process by excluding some of these F-tags.

In the 2001 data there are 346 nursing facilities in the State of Kansas. To understand the prevalence of the 35 deficiencies during the most recent available database (2001), we report the percentage of facilities deficient according to F-tag in Table I. The prevalence varies from about 2 per cent deficient in an F-tag up to a prevalence of around one-third of the nursing facilities (F-tag 316 titled ‘treatment of incontinent bladder’ under the category of Quality of Care). Additionally, Quality of Care appears to be the most prevalent category in terms of average deficiencies by F-tag (17.8 per cent), followed by Assessment (11 per cent), Environment (10.2 per cent), Abuse (8.0 per cent), Nutrition (7.8 per cent), Pharmacy (6.2 per cent) and Rights (5.4 per cent) are the least prevalent.

The summary statistics for the experiment are presented for all 12 matched-pairs in Table I. For each F-tag item, the percentage of nursing homes cited between the RST and the SST appear to be fairly similar to one another. The per cent cited for most F-tags appears higher for the RST and the SST relative to the entire state. This slight increase may be because the survey teams were not blind to the experimental process thus they may raise their sensitivity to assigning a deficiency to a nursing facility. As we mention in the introduction, we wish to assess survey reliability in terms of each of the major categories. Therefore, we first consider inter-rater reliability for the category Quality of Care.

Table II presents the Quality of Care results as binary values with ‘0’ reflecting a non-deficient F-tag as assessed by the survey team and a ‘1’ as deficient. A naïve overall assessment of the reliability indicates poor results. Of the 12 facilities, only facility 2 and 7 match up precisely. We do not agree with this approach. This naïve overall assessment
Table I. Per cent of nursing facilities deficient according to F-tags (U.S. federal deficiencies) categorized into one of five major categories.

<table>
<thead>
<tr>
<th>Category</th>
<th>F-tag</th>
<th>Description</th>
<th>State $N = 346$</th>
<th>RST $n = 12$</th>
<th>SST $n = 12$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality of Care</td>
<td>F312</td>
<td>Addressing activities of daily living</td>
<td>22</td>
<td>67</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>F314</td>
<td>Pressure sores</td>
<td>24</td>
<td>58</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>F316</td>
<td>Treatment incontinent bladder</td>
<td>34</td>
<td>33</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>F325</td>
<td>Parameters of nutritional status</td>
<td>11</td>
<td>17</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>F327</td>
<td>Sufficient fluids for hydration</td>
<td>11</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>F353</td>
<td>Sufficient nursing staff</td>
<td>5</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td>Abuse</td>
<td>F221</td>
<td>No unnecessary physical restraints</td>
<td>17</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>F223</td>
<td>Free from abuse</td>
<td>1</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>F224</td>
<td>Treatment of resident property</td>
<td>2</td>
<td>0</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>F225</td>
<td>Alleged abuse or property violations</td>
<td>12</td>
<td>17</td>
<td>25</td>
</tr>
<tr>
<td>Assess.</td>
<td>F272</td>
<td>Comprehensive assessment RAI</td>
<td>11</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>F274</td>
<td>Assessment 14 days after change</td>
<td>7</td>
<td>0</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>F276</td>
<td>Quarterly review</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>F278</td>
<td>Accuracy of assessment</td>
<td>8</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>F279</td>
<td>Comprehensive care plan</td>
<td>26</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>Rights</td>
<td>F164</td>
<td>Privacy confidentiality</td>
<td>9</td>
<td>50</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>F166</td>
<td>Facility resolves grievances</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>F241</td>
<td>Enhance dignity</td>
<td>11</td>
<td>25</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>F242</td>
<td>Self-determination and participation</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>F246</td>
<td>Accommodation of needs</td>
<td>4</td>
<td>50</td>
<td>33</td>
</tr>
<tr>
<td>Environ.</td>
<td>F252</td>
<td>Safe clean comfortable home-like</td>
<td>6</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>F253</td>
<td>Housekeeping and maintenance</td>
<td>20</td>
<td>58</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>F254</td>
<td>Clean bed bath linens</td>
<td>1</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>F323</td>
<td>Resident free of accident hazards</td>
<td>22</td>
<td>58</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>F469</td>
<td>Effective pest control program</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Nutrition</td>
<td>F363</td>
<td>Menus meet RDA requirements</td>
<td>6</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>F364</td>
<td>Food prep methods nutritional</td>
<td>3</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>F366</td>
<td>Similar foods substitutes offered</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>F368</td>
<td>Frequency of meals</td>
<td>3</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>F371</td>
<td>Sanitary condition of foods</td>
<td>26</td>
<td>58</td>
<td>58</td>
</tr>
<tr>
<td>Pharm.</td>
<td>F329</td>
<td>Free of unnecessary drugs</td>
<td>16</td>
<td>50</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>F330</td>
<td>No antipsychotic drugs w/out diagnos.</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>F332</td>
<td>Facility med error rate less than 5 per cent</td>
<td>10</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>F333</td>
<td>No sig. med error rate</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>F426</td>
<td>Provide pharmacy services</td>
<td>2</td>
<td>8</td>
<td>0</td>
</tr>
</tbody>
</table>

The three percentages are from: 2001 state data, RST and SST. The F-tags categories are consistent with Mullan and Harrington. No state F-tag is above 40 per cent.

provides an inadequate difference or lack there of. For facility 1, both survey teams agree on the deficiency on two of the three most prevalent statewide F-tag items, those associated with activities of daily living (F312) and pressure sores (F314). The RST cites the facility in the F-tag associated with incontinent bladder (F316) and the SST does not. The SST however, cites the facility for a lack of sufficient nursing staff (F353). Are the results for the two teams different? A visual inspection of the team differences shows that the scoring on the same
Table II. Results for the RST and SST for F-tags (U.S. federal deficiencies) categorized under Quality of Care.

<table>
<thead>
<tr>
<th>Facility</th>
<th>F312</th>
<th>F314</th>
<th>F316</th>
<th>F325</th>
<th>F327</th>
<th>F353</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The columns represent F-tags F312, F314, F316, F325, F327, and F353, respectively, for the RST and SST. A value of ‘1’ indicates a facility is deficient for the corresponding F-tag according to the respective survey team. A value of ‘0’ indicates that the facility is non-deficient.

facilities are different. But we argue that both survey teams agree that this facility is deficient. This argument is formally analysed and discussed using statistical models.

2.3. Modelling approaches

The approach begins by collapsing the six F-tags data into two classes or one binary variable. The parameters from this model will provide data-driven weights. This allows the data to determine the weight of an individual F-tag that goes into the scoring, and classify each facility into deficient or non-deficient for the category Quality of Care.

The original goal of our research is to use a kappa agreement for inference. Traditionally the kappa is interpreted as a per cent agreement between two raters. Throughout the article we refer to it as the ‘kappa agreement’ and the term is exchangeable with the Greek letter \( \kappa \). For more on the kappa measure see References [3, 4]. In Reference [5], they advocate for the kappa agreement measure in medical studies.

However, when the data are clustered, adjustments for subject correlation are necessary [6–8]. In our case facilities are clustered. Thompson et al. [8] adjusts the data using a fully Bayesian approach.

In our approach, we adjust for the correlation within group by incorporating an LCA model [9, 10]. Our approach involves two steps. The first step is an LCA Bayesian approach which closely matches a routine in Reference [11]. The second step is a Bayesian beta-binomial model where we obtain a posterior distribution of the kappa agreement measure. An outline of the advantages of the LCA approach is in Reference [12]. By restating Larsen we communicate two reasons for utilizing LCA over other approaches: (i) none of the deficiency variables exactly measure a deficient nursing facility but rather each of the variables measures special aspects of a domain and (ii) individual deficiencies may misclassify themselves more often.
3. SPECIFIC MODEL AND PRIOR DISTRIBUTIONS

3.1. General model

Consider our model for the two data sets described previously. For initial illustration we present the model for the \( p = 6 \) Quality of Care F-tags.

3.1.1. The overall model. The model for the first data set, involving \( N = 346 \) nursing facilities from the entire State of Kansas in 2001, is defined by the F-tags described with \( x_{ik} \) consisting of 0’s and 1’s where \( x_{ik} = 0 \) says facility \( i \) is non-deficient in F-tag \( k \). The value \( x_{ik} = 1 \) says facility \( i \) is deficient for F-tag \( k \). We assume that these binary variables are modelled with the following Bernoulli distribution:

\[
x_{ik} | \delta_{T_i}, \delta_{T,k} \sim \text{Bern}(\delta_{T,k}), \quad i = 1, 2, \ldots, N, \quad k = 1, 2, \ldots, p
\]

where

1. \( T_i \) is the unknown two-class variable that is ‘1’ if the facility \( i \) is in the first latent class (non-deficient) and ‘2’ if it is in the second latent class (deficient). Its probability distribution is

\[
\Pr(T_i | \theta) = \begin{cases} 
1 - \theta, & T_i = 1 \\
\theta, & T_i = 2 
\end{cases}
\]

2. \( \theta \) is a scalar that represents the probability of a randomly selected facility classified into latent class ‘2’ (i.e. deficient).

3. \( \delta_{T,k} \) is the probability that a facility \( i \) classified into latent class \( T_i \) is deficient in the \( k \)th F-tag.

In the literature the model is known as latent class analysis (LCA) [9]. It is latent because of the several layers of hidden parameters: \( \theta, T_i \). It is a class model because there are two classes that correspond to a deficient or non-deficient facility under the category Quality of Care. Because the \( x_{ik} \)’s are the observed data, the likelihood is written as

\[
L(\theta, \delta, T | x_i) = \prod_{i=1}^{N} L_i
\]

where

\[
L_i = \begin{cases} 
(1 - \theta)\delta_{11} \cdots \delta_{1p} (1 - \delta_{11})^{1-x_{i1}} \cdots (1 - \delta_{1p})^{1-x_{ip}}, & T_i = 1 \\
\theta \delta_{21} \cdots \delta_{2p} (1 - \delta_{21})^{1-x_{i1}} \cdots (1 - \delta_{2p})^{1-x_{ip}}, & T_i = 2 
\end{cases}
\]

The top part of the likelihood corresponds to the first class (non-deficient) and the bottom part to the second class (deficient). Notice that when \( \theta \) and \( \delta \) are known we estimate the class participation of a new facility with deficiency vector \( x^* \) via the formula

\[
E(T^* | \theta, \delta, x^*) = \frac{\sum_{T^*=1}^{2} T^* L(\theta, \delta, T^* | x^*)}{\sum_{T^*=1}^{2} L(\theta, \delta, T^* | x^*)}
\]  \hspace{1cm} (1)

Copyright © 2005 John Wiley & Sons, Ltd.  \hspace{1cm} Statist. Med. 2006; 25:325–344
3.1.2. The study design model. The facilities in the experimental design are introduced in the second data set, involving \( n = 12 \) nursing facilities. Each nursing facility receives deficiency scores from the RST and the SST. The F-tag deficiency results from the RST for the \( j \)th nursing facility for the \( k \)th F-tag are reported in \( x_{jk} \). Similarly, the results for the SST are reported in \( y_{jk} \). We utilize \( j \) instead of \( i \) to distinguish the study design model from the overall model. Using the posterior distribution for \( \hat{\delta}_0 \) from the overall model, we take the observed deficiencies values from the RST and the SST and plug them into equation (1). This results in a posterior distribution of class membership from both survey teams for each of the 12 facilities. The expectation of these values are called \( \hat{T}_{1j} = E(E(T_{1j}|\theta, \hat{\delta}, x_j)) \) and \( \hat{T}_{2j} = E(E(T_{2j}|\theta, \hat{\delta}, y_j)) \). The next step is to take these estimates and use them to classify into deficient and non-deficient nursing facilities via the formula:

\[
t_{cj} = \begin{cases} 
0 & \text{if } \hat{T}_{cj} < 1.5 \\
1 & \text{if } \hat{T}_{cj} \geq 1.5
\end{cases}
\]

where \( c = 1 \) for the RST classification and \( c = 2 \) for the SST classification. We choose 1.5 as the threshold because it is halfway between the two classes 1 and 2. These variables (\( t_{1j} \) and \( t_{2j} \)) are then linked with the following beta-binomial model:

\[
t_{2j}|t_{1j} \sim \text{Bern}((1 - t_{1j})\beta + t_{1j}\alpha), \quad j = 1, 2, 3, \ldots, 12
\]

where

1. \( \alpha \) is the scalar associated with the probability that SST scoring deficient, \( t_{2j} = 1 \), conditional on the RST scoring deficient, \( t_{1j} = 1 \).
2. \( \beta \) is the scalar associated with the probability that SST scoring deficient, \( t_{2j} = 1 \), conditional on the RST scoring non-deficient, \( t_{1j} = 0 \).

Notice that the model describes the distribution of the SST conditional on the value score for the RST. Note that the prevalence parameter (\( \theta \)) is estimate from the overall model so it reflects the prevalence from the entire State of Kansas.

3.1.3. The overall model and the study design model together. The conceptualization for the two models together is summarized in Figure 1(a). The contingency table shows marginal and joint distributions for the deficiencies cited from the RST and the SST. From this table we can form a measure of agreement utilizing the kappa agreement measure:

\[
\kappa = \frac{(\Pi_0 - \Pi_e)}{(1 - \Pi_e)} \tag{2}
\]

where

\[
\Pi_0 = (1 - \beta)(1 - \theta) + \alpha \theta \quad \text{and} \quad \Pi_e = \{(1 - \beta)(1 - \theta) + (1 - \alpha)\theta\}(1 - \theta) + \{\beta(1 - \theta) + \alpha \theta\}\theta
\]
Figure 1. The parameter $\theta$ represents the prevalence rate of deficient facilities in the State of Kansas. The parameter $\alpha$ represents the probability that the SST categorizes the facility as deficient conditional on the RST classifying the same facility as deficient. The parameter $\beta$ represents the probability that the SST categorizes the facility as deficient conditional on the RST classifying the same facility as non-deficient. The $\kappa$ parameter is a function of prevalence rate and the conditional probabilities. $t_{1j}$ and $t_{2j}$ are class variables where ‘0’ is a non-deficient facility and ‘1’ is a deficient facility: (a) Contingency table with joint and marginal probabilities; (b) latent class analysis model for all nursing facility deficiency data in the State of Kansas. The parameter $\delta_{1j}$ represents the prevalence of F-tag $j$ in non-deficient nursing facilities. The parameter $\delta_{2j}$ represents the prevalence of F-tag $j$ in deficient nursing facilities; and (c) beta-binomial model for RST and SST.

Figure 1 also summarizes the model in graphical form. Figure 1(b) reflects the graphical model for the overall model or deficiency data from all 346 facilities. The direction of the model arrows is from class, to parameters, to observed data. This allows for the estimation of the $\delta$'s and $\theta$. In Figure 1(c) the graphical model summarizes the estimation process for the design model. The arrows in this case are reversed. The data and the parameters are assumed to be known, producing point estimates of the classification for each facility. The $\delta$'s and class probability $\theta$ have posterior distributions and the $x$'s and $y$'s are observed from the data set. The estimated class membership also produces distributions for the model parameters' $\alpha$ and $\beta$. Together, the results produce the posterior distribution of the kappa agreement measure.

There is an important subtle detail worth emphasizing. In the design we did not utilize the entire $N = 346$ state facilities. However, there is information about class membership probabilities in all state facilities for calculating the posterior distribution of the kappa agreement.
measure. To remedy this issue, we calculate the posterior distribution of class membership probability \( \theta \) using the entire state’s data. By utilizing the overall model’s posterior of \( \theta \), we increase the precision in the posterior distribution.

### 3.2. Priors

The priors for the overall model are all independent uniform distributions: \( \theta \sim \text{Beta}(1,1) \) and \( \delta_{kj} \sim \text{Beta}(1,1) \). The diffuse nature of these prior distributions is not too concerning considering the large number of facilities used to estimate posterior distributions. More specifically, the observed likelihood will drive the inference in the main model.

In contrast, more care is taken in defining the prior distributions for the parameters for the study design model. Using the notation presented in Figure 1, we see that \( \alpha = \Pr(t_{2i} = 1|t_{1i} = 1) \) and \( \beta = \Pr(t_{2i} = 1|t_{1i} = 0) \). This restricts us to eliciting information first about how accurate the SST is at categorizing a deficient facility given that the RST has designated the facility as deficient and second the prevalence of SST at categorizing as a deficient facility given that the RST has designated as non-deficient (\( \beta \)).

Both state surveyors and nursing home providers were skeptical of our design, believing it allowed the RST and SST to ‘artificially agree’. They asserted that the SST did not have the ability to re-direct their exploration of potential problem areas because they had to follow the RST. The surveyor’s believed agreement should be denoted between 70 and 80 per cent and the provider’s believed agreement should be denoted between 90 and 100 per cent. We call the agreement skeptical in the sense that the providers were afraid that the design would produce this artificial agreement. From a statistical point of view we will refer to the prior distribution from surveyors and providers as ‘enthusiastic’ information because the prior indicates strong agreement. The term ‘enthusiastic prior’ is from recent Bayesian literature [13]. Its definition labels the prior distribution as being close to a favourable outcome. In our case this means that the agreement between RST and SST is strong. The opposite is a pessimistic prior which would indicate a prior reflecting a lack of agreement between RST and SST.

This information is a basis for the construction of the prior distribution for \( \alpha \) and \( \beta \). We define the priors to have a beta distribution that does not dominate the posterior distribution. We do not want the information provided in the prior to exceed the information provided in the observed data. Therefore, we restrict the ‘prior sample size’ [14] to be less than six for both \( \alpha \) and \( \beta \). This will correspond to the prior information to be no greater than 12 (the size of the design). There are two types of priors we utilize in the analysis:

- \( \alpha \sim \text{Beta}(2.8, 1.2) \), an enthusiastic prior that is fairly diffuse having a mean 0.70 and \( \beta \sim \text{Beta}(1.2, 2.8) \), an enthusiastic prior that is fairly diffuse having a mean 0.30. Both distributions have prior sample size of 4.
- \( \alpha \sim \text{Beta}(5.5, 0.5) \), a very enthusiastic prior that is informative, having a prior mean 0.92 and \( \beta \sim \text{Beta}(0.5, 5.5) \), a very enthusiastic prior that is informative, having a prior mean 0.08. Both distributions have prior sample size of 6.

Both priors are fairly enthusiastic with the second being more informative.
4. COMPUTATION

The overall model is a two class LCA model. To check the adequacy of fit for two classes we initially fitted two and three class models and calculated the Bayesian information criteria (BIC). The definition for BIC we use is $-2 \times \text{loglikelihood} + p \ln(N)$ where $p$ is the number of variables and $N$ is the sample size. We evaluate the likelihood at the posterior median of model parameters, which is extremely close to the maximum likelihood estimates because of the diffuse priors and large sample sizes [15]. We chose the model with lowest BIC for final inference.

To obtain posterior distributions we performed Bayesian estimation of the models using Markov chain Monte Carlo (MCMC), implemented in WinBUGS [16]. We ran the LCA with data from all the states’ nursing facilities first (overall model). Then, utilizing the data from the RST and SST for the 12 nursing facilities and the posterior distributions from the overall model, we ran a second MCMC to obtain the distribution of the kappa agreement measure (study design model). The study design model is a straightforward application of MCMC since all complete conditional distributions are easy to generate. 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. We summarize the posterior distributions with means and standard deviations or 95 per cent credible intervals (CrI) with posterior medians. We present the WinBUGS code for the Quality of Care model in Appendix A.

5. RESULTS

5.1. The estimation of model parameters

5.1.1. The overall model. The BIC for all categories was smaller for a two class model than a three class model. Specifically, the BIC for two and three classes, respectively, for Quality of Care, Abuse, Assessment, Rights, Environment, Nutrition and Pharmacy is 1722.5 and 1748.9; 724.0 and 749.5; 1033.6 and 1048.0; 706.0 and 732.6; 970.4 and 1003.1; 803.0 and 829.9; and 743.3 and 774.9. Therefore a two class model is adequate for final inference.

In Table III we summarize the posterior distributions for the overall model parameters for the category Quality of Care. The higher $\delta$ values are associated with the deficient class. Therefore, the deficient facilities are in the second class because of the higher prevalence of deficiencies. The probability a randomly selected nursing facility is deficient is on average 0.24 $(\hat{\theta})$. The biggest difference in F-tag probabilities between classes occurs in F316 (incontinent bladder). However, when taking the ratio of mean $\delta$’s of deficient and non-deficient parameters, F353 holds the largest relative risk, 20. To obtain a feel for model results, we use Table III to approximate equation (1). We treat the posterior means of all parameters as Bayesian estimates and estimate equation (1) with

$$
\hat{E}(T^*|\theta, \hat{\delta}, \hat{x}^*) = \frac{\sum_{T^*=1}^2 T^* \ln(\hat{\delta}, \hat{\theta}, T^*|\hat{x}^*)}{\sum_{T^*=1}^2 \ln(\hat{\theta}, \hat{\delta}, T^*|\hat{x}^*)}
$$

(3)
Table III. Summary of posterior distributions for the overall model applied to Quality of Care.

<table>
<thead>
<tr>
<th>Label</th>
<th>Node</th>
<th>Mean</th>
<th>SD</th>
<th>2.5 percentile</th>
<th>Median</th>
<th>97.5 percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-deficient</td>
<td>$1 - \theta$</td>
<td>0.76</td>
<td>0.04</td>
<td>0.67</td>
<td>0.77</td>
<td>0.84</td>
</tr>
<tr>
<td>Deficient</td>
<td>$\theta$</td>
<td>0.24</td>
<td>0.04</td>
<td>0.16</td>
<td>0.23</td>
<td>0.33</td>
</tr>
</tbody>
</table>

Non-deficient

F312 | $\delta_{11}$ | 0.11 | 0.02 | 0.07           | 0.11   | 0.16            |
F314 | $\delta_{12}$ | 0.12 | 0.03 | 0.07           | 0.11   | 0.17            |
F316 | $\delta_{13}$ | 0.20 | 0.03 | 0.14           | 0.20   | 0.27            |
F325 | $\delta_{14}$ | 0.05 | 0.01 | 0.02           | 0.05   | 0.08            |
F327 | $\delta_{15}$ | 0.04 | 0.02 | 0.01           | 0.04   | 0.07            |
F353 | $\delta_{16}$ | 0.01 | 0.01 | 0.00           | 0.01   | 0.03            |

Deficient

F312 | $\delta_{21}$ | 0.56 | 0.07 | 0.43           | 0.56   | 0.71            |
F314 | $\delta_{22}$ | 0.67 | 0.08 | 0.51           | 0.67   | 0.83            |
F316 | $\delta_{23}$ | 0.80 | 0.06 | 0.67           | 0.80   | 0.91            |
F325 | $\delta_{24}$ | 0.32 | 0.07 | 0.21           | 0.32   | 0.46            |
F327 | $\delta_{25}$ | 0.35 | 0.06 | 0.23           | 0.34   | 0.48            |
F353 | $\delta_{26}$ | 0.20 | 0.05 | 0.11           | 0.20   | 0.31            |

The parameter $\theta$ represents the prevalence rate of deficient facilities in the State of Kansas. The parameter $\delta_{1k}$ represents the prevalence of F-tag $k$ in non-deficient nursing facilities. The parameter $\delta_{2k}$ represents the prevalence of F-tag $k$ in deficient nursing facilities. The best discrepancy between deficient and non-deficient nursing facilities appears to be F316, as there is a 0.60 difference in their mean prevalence.

A value of $\hat{E}(T^*|\theta, \delta, \chi^*) > 1.5$ corresponds to a deficient nursing facility and non-deficient otherwise. For a nursing facility with one deficient F-tag the $\hat{E}(T^*|\theta, \delta, \chi^*)$ is no larger than 1.12, demonstrating that all nursing facilities with one deficiency are classified as non-deficient. At the other extreme, when a nursing facility is deficient in five of the six Quality of Care F-tags, $\hat{E}(T^*|\theta, \delta, \chi^*)$ is no smaller than 1.99 demonstrating these nursing facilities will classify as deficient. It turns out that only some nursing facilities with a sum of the two deficient F-tags classify as deficient, depending on which deficiencies are cited. For example, if F312 and F326 are the only deficient F-tags in the nursing facility, then it is non-deficient. But if F327 and F353 are the only deficient F-tags the facility is deficient. All of the facilities with three deficient F-tags or more classify as deficient.

In Table IV we summarize the posterior distribution of equation (1) applied to the study design data. Most of the results indicate clear classification into deficient or non-deficient nursing facilities as their posterior mean is close to 1 or 2 with a small standard deviation. One exception is that the class variable for nursing facility 4 has a posterior mean of 1.47 (SD = 0.15) for the RST and a posterior mean of 1.57 (SD = 0.14) for the SST. We report this as a difference in classification between teams. The other exception is that nursing facility 9 has a posterior mean of 1.46 (SD = 0.17) for the SST, classified as non-deficient but close to a deficient classification. Model parsimony is gained by preserving the definition of two classes and not utilizing a continuum in the analysis.
Table IV. Results for the study design model for F-tags (U.S. federal deficiencies) categorized under Quality of Care.

<table>
<thead>
<tr>
<th>Facility</th>
<th>x₁</th>
<th>y₁</th>
<th>x₂</th>
<th>y₂</th>
<th>x₃</th>
<th>y₃</th>
<th>x₄</th>
<th>y₄</th>
<th>x₅</th>
<th>y₅</th>
<th>x₆</th>
<th>y₆</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>/</td>
<td>1</td>
<td>/</td>
<td>1</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>/</td>
<td>1</td>
<td>/</td>
<td>1</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>1</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>1</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>1</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>1</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>/</td>
</tr>
</tbody>
</table>

\(E(T^*|\theta, \delta, \gamma^*)\)

The \(x\)'s and the \(y\)'s represent the F-tags F312, F314, F316, F325, F327, and F353. A value of ‘1’ indicates a deficient facility for the corresponding F-tag according to the respective survey team. A value of ‘0’ indicates that the facility is non-deficient. The posterior distribution from the state’s model classifies each facility into a deficient or non-deficient nursing facility according to the class assignment variables T1 and T2.

5.1.2. The study design model. In Table V we summarize the posterior for the design model parameters for (A) enthusiastic non-informative prior and (B) the enthusiastic informative prior for several categories of F-tags. Focusing on Quality of Care with an enthusiastic non-informative prior, we see that we estimate \(\alpha\) as 0.88, reflecting moderate-to-fairly consistent for the probability of a deficiency by the SST conditional on an RST assigning a deficiency. We estimate \(1 - \beta\) as 0.68, reflecting less consistency for the probability of a non-deficiency by the SST conditional on an RST assigning a non-deficiency. The standard deviations are fairly high as the sample size of the experiment is 12. Their posterior distributions and the posterior distribution of \(\theta\) are plugged into the kappa measure showing an estimate for \(\kappa\) is 0.44 with close to symmetric CrI (0.12–0.78). Another way to look at this result is that despite providers’ ‘optimism’, there is only a 0.025 probability that the difference between
Table V. Posterior summaries for model parameters for all categories defined by Mullan and Harrington.

<table>
<thead>
<tr>
<th>Category</th>
<th>Mean (SD)</th>
<th>κ (per cent)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>θ</td>
<td>α</td>
</tr>
<tr>
<td>(A) $\alpha \sim \text{Beta}(2.8, 1.2)$ and $\beta \sim \text{Beta}(1.2, 2.8)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QOC</td>
<td>0.24 (0.04)</td>
<td>0.88 (0.10)</td>
</tr>
<tr>
<td>Abuse</td>
<td>0.03 (0.02)</td>
<td>0.70 (0.20)</td>
</tr>
<tr>
<td>Assessment</td>
<td>0.22 (0.04)</td>
<td>0.65 (0.16)</td>
</tr>
<tr>
<td>Rights</td>
<td>0.04 (0.02)</td>
<td>0.40 (0.17)</td>
</tr>
<tr>
<td>Environment</td>
<td>0.11 (0.06)</td>
<td>0.88 (0.10)</td>
</tr>
<tr>
<td>Nutrition</td>
<td>0.04 (0.02)</td>
<td>0.88 (0.10)</td>
</tr>
<tr>
<td>Pharmacy</td>
<td>0.09 (0.04)</td>
<td>0.56 (0.20)</td>
</tr>
<tr>
<td>(B) $\alpha \sim \text{Beta}(5.5, 0.5)$ and $\beta \sim \text{Beta}(0.5, 5.5)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QOC</td>
<td>0.96(0.06)</td>
<td>0.21(0.11)</td>
</tr>
<tr>
<td>Abuse</td>
<td>0.92(0.10)</td>
<td>0.08(0.06)</td>
</tr>
<tr>
<td>Assessment</td>
<td>0.83(0.12)</td>
<td>0.17(0.09)</td>
</tr>
<tr>
<td>Rights</td>
<td>0.61(0.15)</td>
<td>0.10(0.15)</td>
</tr>
<tr>
<td>Environment</td>
<td>0.96(0.05)</td>
<td>0.12(0.09)</td>
</tr>
<tr>
<td>Nutrition</td>
<td>0.96(0.06)</td>
<td>0.12(0.09)</td>
</tr>
<tr>
<td>Pharmacy</td>
<td>0.78(0.15)</td>
<td>0.09(0.06)</td>
</tr>
</tbody>
</table>

The parameter $\theta$ represents the prevalence rate of deficient facilities in the State of Kansas. The parameter $\alpha$ represents the probability that the SST categorizes the facility as deficient conditional on the RST classifying the same facility as deficient. The parameter $\beta$ represents the probability that the SST categorizes the facility as deficient conditional on the RST classifying the same facility as non-deficient. The $\kappa$ parameter is a function of prevalence rate and the conditional probabilities. Part A displays the posterior results assuming an enthusiastic non-informative prior distribution, and Part B displays the results for the posterior for the enthusiastic informative prior distribution.

observed and chance agreement is above 78 per cent. Focusing on Quality of Care for the enthusiastic informative prior we estimate $\alpha$ as 0.96 and $1 - \beta$ as 0.79. Additionally, the standard deviations are lower than the non-informative case. The estimate for $\kappa$ is 0.63 with a less symmetric CrI (0.30–0.90), saying that there is a 0.025 chance the difference between observed and chance agreement is above 90 per cent. We also give the results for the six other F-tag categories in Table V. We point out that $\alpha$ and $\beta$ reflect better agreement under the non-informative prior for Environment and Nutrition than Quality of Care but their kappas are smaller. This occurs because the prevalence, $\theta$, is much smaller. It is very difficult to achieve kappa agreement for a rare event [5].

We elaborate on this point in the context of the experiment. In Figure 2, for the enthusiastic non-informative prior, we summarize the posterior distribution for the experiment for all categories, as a function of their respective estimated $\theta$. We overlap the prior distribution and ‘perfect agreement’. The ‘perfect agreement’ is the theoretical case where the teams match perfectly in their categorization of the nursing facility. There are two things to notice. First, the prior distribution is actually fairly diffuse, $\kappa$ can be as low as zero and as high as 0.80. When observing the data, the kappas are tighter, and all but Abuse and Environment have lower 97.5 percentile than their prior distributions. However, their distributions
Figure 2. Summary of distributions for kappa measures. The enthusiastic non-informative prior distribution assumes 70 per cent match for SST results when the RST classifies the nursing facility as deficient. The prior also assumes a 70 per cent match of the SST results when the RST classifies the nursing facility as non-deficient. The prior distribution is equivalent in information to eight facilities. We summarize all distributions with 2.5, 50 and 97.5 percentile. The posterior distribution updates the prior distribution based on the experimental results with the 12 facilities. The posterior 50 percentile for all categories is far below the prior 50 percentile. Therefore, despite providers' belief in near ‘perfect agreement’, the data moves towards poor agreement. For comparative purposes, we present the theoretical posterior summary for the case of perfect agreement.

are quite different from the ‘perfect agreement’ case. For example, in Quality of Care, the posterior distribution for the median is just under 0.40, but for ‘perfect agreement’ the median is 0.60.

There are more pronounced differences in the informative case that we report in Figure 3. The prior distributions are much narrower. Additionally, the posterior distribution is similar to the prior at the 2.5 percentile, but is drastically different for the median and 97.5 percentile. For example, the median posterior kappa for Quality of Care is around 0.5, whereas the prior median is around 0.8. This is a drop of 0.3 from prior to posterior information. The 97.5 percentile goes from 0.99 (prior) to 0.82 (posterior). When comparing the posterior to ‘perfect agreement’ distribution, the 97.5 percentile is 0.82 (posterior) which is below the median of ‘perfect agreement’.
Figure 3. Summary of distributions for kappa measures. The enthusiastic informative prior distribution assumes 92 per cent match for SST results when the RST classifies the nursing facility as deficient. The prior also assumes a 92 per cent match of the SST results when the RST classifies the nursing facility as non-deficient. The prior distribution is equivalent in information to 12 facilities. We summarize all distributions with 2.5, 50 and 97.5 percentile. The posterior distribution is equivalent in information to 12 facilities. We summarize all distributions with 2.5, 50 and 97.5 percentile. The posterior distribution updates the prior distribution based on the experimental results with the 12 facilities. The posterior 50 percentile for all categories is far below the prior 50 percentile. Therefore, despite providers' belief in near 'perfect agreement', the data moves towards poor agreement. For comparative purposes, we present the theoretical posterior summary for the case of perfect agreement.

Even with very strong prior information from providers, the kappa agreement is lower than ideally desired. This is true not only on the absolute scale, but when comparing to an experiment with 'perfect agreement' in the sense of the experimental results.

6. CONCLUSION

Most researchers utilize subsets of nursing home deficiency data in one of two ways. For example, suppose the researcher is interested in the items related to Quality of Care. One method, as mentioned earlier, would be to sum the number of deficiencies among the six variables. In this case researchers would measure inter-rater reliability using interclass
A BAYESIAN ANALYSIS OF INTER-RATER RELIABILITY

341

correlation coefficient (ICC) defined from a normal random effects model. Another method is to treat each of the six binary variables individually and measure their kappa agreement measure for all six. A better approach is something in between the sum value and the raw dichotomization. This approach is the LCA approach to modelling. The advantage of our LCA approach is that we dichotomize a nursing facility into deficient or non-deficient for Quality of Care as a whole. This is important because of the subjectivity involved in rating a nursing facility. LCA is a parsimonious approach to modelling and is realistic in the sense that researchers value the dichotomization of nursing facilities into deficient or non-deficient. Notice that we could have used a sum score (or equivalently a Rasch model) or a two parameter logistic model if different weightings were desired [17]. However, the LCA provides researchers with a big picture of agreement using the kappa agreement. And of course, researchers may construct subsets of the data to study other types of measures of the facility (Environment, Nutrition, etc.). Therefore, the researchers may utilize all 191 items as long as the prevalence is relatively high (most F-tags do not have prevalence above 1.5 per cent).

We have developed a Bayesian method for estimating inter-rater reliability in nursing facility deficiency data. The Bayesian modelling strategy provides a posterior distribution of the kappa agreement measure. The posterior distribution is a function of the distribution of three parameters: $\theta$, $\alpha$ and $\beta$. The posterior distribution of the prevalence parameter, $\theta$, is provided by the deficiency data from the entire State of Kansas. Using an LCA Bayesian approach, the data also provides us the posterior distribution of $\delta_{Tk}$, the probability that facility $i$ is deficient for the $k$th F-tag, assuming they are classified into latent class $T_k$. Using this information, we obtain estimates of the latent classification of each facility as assessed by the regular survey team (RST) and the simultaneous survey team (SST).

We also note that there is a seemingly large impact of the prior distribution on the results. This is because of the relatively small sample size of $n = 12$ in the experimental design. Thus the prior distribution from the state surveyors and nursing home providers has a large impact on the analysis. Researchers skeptical about this opinion can incorporate their own prior distribution and see the impact it has on the analysis.

One can apply the method to other types of inter-rater reliability data when there are several binary outcomes per experimental unit. In the current model we assume there are two latent classes (deficient and non-deficient) and that we observe binary data. We may generalize this in at least two ways. First, we can extend the method to several latent classes, beyond deficient and non-deficient, resulting in a larger contingency table. Second, we may extend the raw data to have more than binary categories. For example, surveyors further categorize a deficient nursing facility by its severity.

Our example focuses on long-term care, specifically for the elderly. Providing quality care for an ageing society is a difficult business, and finding a good nursing facility can be a daunting task for a consumer. Based on the judgement of a survey team, the U.S. government provides information about the quality of facilities and posts them at the website www.medicare.gov [18]. Our study investigates the accuracy of these judgements. Overall, the results indicate low-to-moderate consistency if pass/fail decisions are made. Consequently, the consumer should exercise caution when interpreting nursing home deficiency data. Other types of investigation of the facility will aid the consumer in a better understanding of the facilities’ potential to care for an ageing loved one [19].

Since each facility is surveyed at least once a year, interesting future work is to study the stability of the latent classes longitudinally. This assesses the stability of the classification of
nursing home facilities. Other possible future work is to combine different domains (Quality of Care, Abuse, etc.) rather than analysing them separately.

Current and future analyses will help direct state agencies in understanding the temporal stability of nursing homes as well as help guide the nursing home survey process to become a more reliable inspection routine.

In closing, the federal government requires facilities to be in substantial requirement for each of the 191 individual deficiencies. Since we are measuring agreement in a category, we are proposing a new definition of compliance and a way of measuring agreement.

APPENDIX A: WinBUGS Code

A.1. The overall model (described in Section 3.1.1)

# The first two loops perform the LCA for facilities 'i' and F-Tags 'k'
#Input: x: n facilities by p F-tags
# C: number of latent classes
# xstar: RST F-tags
# ystar: SST F-tags
# prior[]: prior for theta
#Output: Posterior distributions for
#T[i]: Category for facility i
#post.prob[c,i]: posterior probability that facility i is in class c
#delta[c,k]: prevalence of F-tag k in latent class c
#EPstarx[j]: Equation 3 for RST
#EPstary[j]: Equation 3 for SST

for (i in 1:n)
{
T[i] ~ dcat(theta[1:C])
Tp[i] ~ dcat(theta[1:C])
for (c in 1:C)
{
    post.prob[c,i] <- equals(T[i],c)
}
for (k in 1:p)
{
    x[i,k] ~ dbern(delta[T[i],k])
}
for (c in 1:C)
{
    for (k in 1:p)
    {
        delta[c,k] ~ dbeta(1,1)
    }
}
theta[1:C] ~ ddirch(prior[])

#predictions
#regular survey team [x]: The 'xstar' is what is observed for the regular team and so we calculate 't1' here

for (j in 1:12)
{
    for (k in 1:p)
    {
        lPstar1x[j,k]<-log(delta[1,k])*xstar[j,k]+log(1-delta[1,k])*(1-xstar[j,k])
        lPstar2x[j,k]<-log(delta[2,k])*xstar[j,k]+log(1-delta[2,k])*(1-xstar[j,k])
    }
    Pstar1x[j]<-exp(sum(lPstar1x[j,]))*theta[1]
    Pstar2x[j]<-exp(sum(lPstar2x[j,]))*theta[2]
}

#simultaneous survey team [y]: The 'xstar' is what is observed for the regular team and so we calculate 't2' here

for (j in 1:12)
{
    for (k in 1:p)
    {
        lPstar1y[j,k]<-log(delta[1,k])*ystar[j,k]+log(1-delta[1,k])*(1-ystar[j,k])
        lPstar2y[j,k]<-log(delta[2,k])*ystar[j,k]+log(1-delta[2,k])*(1-ystar[j,k])
    }
    Pstar1y[j]<-exp(sum(lPstar1y[j,]))*theta[1]
    Pstar2y[j]<-exp(sum(lPstar2y[j,]))*theta[2]
}

A.2. The design model (design model described in Section 3.1.2)

#These loops calculate the Beta Binomial Model and the kappa agreement from their results

for (j in 1:12)
{
    T2[j]~dbern(P2[j])
}
\[ P2[j] \leftarrow \alpha \cdot T1[j] + \beta \cdot (1 - T1[j]) \]

\[ \theta \sim \text{dbeta}(a, b) \]

\[ P1 \leftarrow \theta \]

\[ \alpha \sim \text{dbeta}(c, d) \]

\[ \beta \sim \text{dbeta}(e, f) \]

\[ P11 \leftarrow (1 - \beta) \cdot (1 - P1) \]

\[ P12 \leftarrow \beta \cdot (1 - P1) \]

\[ P21 \leftarrow (1 - \alpha) \cdot P1 \]

\[ P22 \leftarrow \alpha \cdot P1 \]

\[ P0 \leftarrow P11 + P22 \]

\[ Pe \leftarrow (P11 + P12) \cdot (P11 + P21) + (P21 + P22) \cdot (P12 + P22) \]

\[ \text{Kappa} \leftarrow \frac{(P0 - Pe)}{(1 - Pe)} \]

ACKNOWLEDGEMENTS

We thank C. Shane Reese for his review of our paper and Mary Gajewski for editorial assistance. We also thank Linda Wendling for stimulated conversation regarding different deficiency weighting schemes. We thank referees and editor for their review and help in making this a better manuscript. Partial funding for all authors is from a grant by the Kansas Department on Aging (KDOA).

REFERENCES