

**Cross-calibration of stroke disability measures:  
Bayesian analysis of longitudinal ordinal categorical data  
using negative dependence**

Giovanni Parmigiani, Heidi W. Ashih, Gregory P. Samsa,  
Pamela W. Duncan, Sue Min Lai, David B. Matchar

Running head: Stroke disability measures.

Giovanni Parmigiani is Associate Professor, Departments of Oncology, Biostatistics, and Pathology, Johns Hopkins University, Baltimore, MD. Heidi Ashih is Visiting Assistant Professor, Institute of Statistics and Decision Sciences, Duke University, Durham, NC. Gregory P. Samsa is Associate Professor, Center for Clinical Health Policy Research, Duke University Medical Center, Durham, NC, and Departments of Medicine, and Community and Family Medicine, Duke University Medical Center, Durham, NC. Pamela W. Duncan is Professor, Center on Aging, University of Kansas Medical Center, Kansas City, MO, and Department of Veteran Affairs Medical Center, Kansas City. Sue Min Lai is Associate Professor, Center on Aging, University of Kansas Medical Center, Kansas City, MO, and Department of Veteran Affairs Medical Center, Kansas City. David Matchar is Professor and Director, Center for Clinical Health Policy Research, Duke University Medical Center, Durham, NC, and Professor Department of Medicine, Duke University Medical Center, Durham, NC, and Center for Health Services Research in Primary Care, VA Medical Center, Durham, NC. This study was funded by the Center for Medical Effectiveness Research, Agency for Health Care Policy Research, Contract No. 282-91-0028. Work at the Department of Veteran Affairs Rehabilitation Research and Development and the University of Kansas Claude D Pepper Center was funded by the National Institute of Medicine (P60 AG 14635-02).

## Abstract

It is common to assess disability of stroke patients using standardized scales, such as the Rankin Stroke Outcome Scale (RS) and the Barthel Index (BI). The Rankin Scale, which was designed for applications to stroke, is based on assessing directly the global conditions of a patient. The Barthel Index, which was designed for more general applications, is based on a series of questions about the patient's ability to carry out 10 basic activities of daily living. As both scales are commonly used, but few studies use both, translating between scales is important in gaining an overall understanding of the efficacy of alternative treatments, and in developing prognostic models that combine several data sets.

The objective of our analysis is to provide a tool for translating between BI and RS. Specifically, we estimate the conditional probability distributions of each given the other. Subjects consisted of 459 individuals who sustained a stroke and who were recruited for the Kansas City Stroke Study from 1995 to 1998. Patients were assessed with BI and RS measures 1, 3 and 6 months after stroke. In addition, we included data from the Framingham study, in the form of a table cross-classifying patients by RS and coarsely aggregated BI.

Our statistical estimation approach is motivated by several goals: (a) overcoming the difficulty presented by the fact that our two sources report data at different resolutions; (b) smoothing the empirical counts to provide estimates of probabilities in regions of the table that are sparsely populated; (c) avoiding estimates that would conflict with medical knowledge about the relationship between the two measures and (d) estimating the relationship between RS and BI at three months after the stroke, while borrowing strength from measurements made at one and six months. We address these issues via a Bayesian analysis combining data augmentation and constrained semiparametric inference.

Our results provide the basis for (a) comparing and integrating the results of clinical trials using different disability measures, and (b) integrating clinical trials results into comprehensive decision model for the assessment of long term implications and cost-effectiveness of stroke prevention and acute treatment interventions. In addition, our results indicate that the degree of agreement between the two measures is less strong than commonly reported, and emphasize the importance of trial designs that include multiple assessments of outcome.

*Key words* Stroke, cerebrovascular disease, local odds ratio, totally positive dependence, longitudinal contingency tables, outcome measures, constrained MCMC.

# 1 Introduction

## 1.1 Background

Stroke is a common event that frequently disables its victims. Two of the most common measures of disability used in stroke studies are the Rankin Stroke Outcome Scale (RS) and the Barthel Activity of Daily Living (ADL) Index (BI). The Rankin Scale (Rankin 1957), which was designed for applications to stroke, uses a global approach to assessment in which disability categories are ordered from 0 (no symptoms) to 5 (severe disability). Currently, it is common to use the Modified Rankin Stroke Outcome Scale (van Swieten, Koudstaal, Visser, Schouten, and van Gigin 1988) described in Table 1. The Barthel ADL Index (Mahoney and Barthel 1965) was designed for measuring general disability in a broad spectrum of applications. Patients are asked about their ability to carry out 10 basic activities of daily living, listed in Table 2. For each item, patients receive a score of 0 if entirely unable to carry out the activity, a maximum score if able to perform the activity independently, and a partial score if able to perform the activity with assistance. Possible values are assigned in increments of 5. The totals can range from 0 to 100. The BI is often treated as comparable to the RS, and various authors have suggested benchmarks for the overall score. For example, one such benchmark is 0-55 (dependent); 60-85 (moderately disabled); 90-95 (slightly disabled); 100 (independent) (Duncan, Goldstein, Divine, Feussner, and Matchar 1992).

Although ad hoc translations between RS and BI are helpful for qualitative purposes, an explicit quantitative translation would permit comparisons among clinical trials using either disability measure, and would contribute to generating prognostic estimates of stroke outcome such as those generated by the Stroke Policy Model (Matchar, Samsa, Matthews, Ancukiewicz, Parmigiani, Hasselblad, Wolf, D'Agostino, and Lipscomb 1997). The objective of our analysis is to provide a

method for translating between the RS and BI. Specifically, we provide estimates of the conditional probability distributions of RS given BI and BI given RS. By weighting results according to these conditional probabilities one can electively translate research results based on BI and RS, while accounting for the additional uncertainty introduced by the conversion. A secondary goal of our analysis is to provide an overall measure of the degree of association between the measures.

## 1.2 Statistical Approach

Our data sources included 2 studies. The Kansas City Stroke Study provided 3 cross-tabulations of BI versus RS from data taken at 1, 3, and 6 months post-stroke. The Framingham Study provided one, coarsely grouped, cross-tabulation of BI versus RS at 3 months post-stroke. In our analysis we estimate the conditional probability distributions of each of the scores given the other, at its natural level of resolution. Our estimation approach is motivated by the following goals: (a) overcoming the difficulty presented by the different resolutions of our two data sources; (b) interpolating the empirical counts to provide estimates of probabilities in regions of the tables that are sparsely populated; (c) avoiding estimates that would conflict with medical knowledge about the relationship between the two scales and (d) estimating the relationship between RS and BI at three months after the stroke, while borrowing strength from measurements made at one and six months.

We handled (a) via data augmentation (Tanner and Wong 1987; Tanner 1991), now common both in missing data problems and multiresolution analyses. The main thrust of data augmentation in this context is to develop a statistical analysis based on data at full resolution from one study, and on the unobserved full resolution data that would have been obtained in the other study if the data had not been aggregated. The entries in the latter table are not known with certainty and are treated as unknown nuisance parameters.

Regarding (b) and (c), we addressed both via a constrained nonparametric approach that recognizes the natural negative dependence in the data, and captures it without making any restrictive parametric assumptions about the relationship among the cell probabilities. Specifically, we assume that any two-by-two sub-table will show association (in this case negative) among the scores. This condition also imposes smoothness of the cell counts.

Finally, we addressed (d) by postulating an autoregressive stochastic process for the cell probabilities. The probability table at each time is modeled as a constrained Dirichlet distribution whose mean is given by the probability table at the previous time point. The Dirichlet distribution also includes a dispersion parameter which, in our context, controls the average amount of variation of the table entries from one time interval to the next.

Computationally, the data augmentation, the constrained analysis, and the autoregressive matrix process all fit naturally in the framework of Markov Chain Monte Carlo (Gilks, Richardson, and Spiegelhalter 1996). Our implementation includes an innovative technical device to handle the constrained analysis of the contingency table.

## 2 Data Sources

### 2.1 Purpose

Recovery from stroke is typically rapid during the first thirty days post-stroke, than slows and reaches a plateau within three to six months (Duncan, Goldstein, Horner, Landsman, Samsa, and Matchar 1994). Our goal was not to describe this pattern of recovery *per se* but rather to develop a translation between two measures of functional status. Accordingly, we sought population-based studies of patients with stroke, reflecting heterogeneity in both BI and RS, in which BI and IS were

cross-classified at one or more time points. The Kansas City Stroke Study and the Framingham Study met these criteria, and are described in brief in the remainder of this Section.

## **2.2 Kansas City Stroke Study**

The participants in this study were 459 individuals who sustained a stroke and who were recruited for the Kansas City Stroke Study, or KCSS (Duncan, Lai, and Keighley 2000; Duncan, Jorgensen, and Wade 2000). Case ascertainment for the KCSS occurred from October of 1995 through March 1998. Participants were recruited from any of 12 hospitals in the greater Kansas City area. Eligible stroke patients were identified by 1) a review of daily admission records, 2) referrals from physicians, clinical nurse specialists and therapists on medical neurology, and rehabilitation units, and 3) review of discharge codes. To be accepted into this study, a subject had to have a stroke confirmed by clinical assessment and/or by a CT/MRI scan. A stroke was defined according to the World Health Organization Criteria as "rapid onset and of vascular origin reflecting a focal disturbance of cerebral function excluding isolated impairments of higher function and persisting longer than 24 hours." (World Health Organization 1993) Patients included adults with strokes of all severity. They were evaluated at enrollment and again at 1, 3, and 6 months post-stroke. All testing was performed by a study nurse or physical therapist.

Table 3 summarizes the basic demographic characteristics of the 459 patients. Table 4 displays the cross classification of patient by RS and BI at one, three and six months in the KCSS. In all cases both measurements were taken by the same rater, although more than one rater was used. Although recovery of function can continue for a year or more post-stroke, for most patients the majority of gains occur during the first 30 days, with the elements of functional status measured by the RS and BI having more or less stabilized at 3-6 months after the initial event.

Indeed, the follow-up period for many randomized trials of acute stroke treatments has been 3 months, arguing that this approximates functional recovery in the long-term. Although the marginal distribution of BI and RS may change over time, for example reflecting functional recovery over time, it is reasonable to assume that the conditional distribution of RS given BI (BI given RS) will be relatively stable, regardless of time since stroke. Also, as expected, little change is observed in Table 4 between 3 and 6 months.

### **2.3 The Framingham study**

In addition, to increase the sample size of measurements at 3 months, and to improve the generalizability of results, we incorporated a published 4x4 table cross-classifying patients by aggregate RS and BI. This is from a commonly cited study by Wolfe, Taub, Woodrow, and Burney 1991, investigating the relationship between the two scales in 50 patients diagnosed clinically as having suffered a stroke (as defined by WHO) at least 3 months previously. Three nurses independently rated each patient's RS and BI. The authors used a generalized linear model approach to account for rater variability, and estimated the joint probability distribution of RS and BI, accounting for rater effects. Their published estimates are reproduced in Table 5. Wolfe et al. aggregated the two scores into categories that would provide greater correspondence between the two measures. After aggregation, the association of the two measurements is high.

We applied the same aggregation to the 3-months data from the KCSS and compared it to Table 5. The marginal distributions differ substantially, with the KCSS including a lower percentage of highly disabled individuals. However, the conditional distributions are similar, suggesting that the two studies can be combined in developing conversion tables.

### 3 Methods

We use the notations  $\pi^{(t)}$  and  $\mathbf{x}^{(t)}$  to denote the matrices of cell probabilities and cell counts at time  $t$ . In the KCSS,  $t = 1, 2, 3$  correspond to 1, 3 and 6 months after stroke. We use subscripts  $r = 1, \dots, R$  and  $b = 1, \dots, B$  to remind of a combination of RS and BI, although the discussion of this section applies to any longitudinal sequence of rectangular ordinal categorical tables. We begin by describing our methodology for a single time point, omitting the  $(t)$  superscript, and then move to the longitudinal case. The main goal of our analysis is to obtain conditional distributions of RS given BI and BI given RS. Both of these distributions are functions of the cell probabilities  $\pi$  —the joint distribution of RS and BI. We will develop our modeling and inference approaches in terms of this joint distribution, and derive inferences on the conditionals from the joint. Because inference is performed using simulation, it is straightforward to carry out the necessary transformations.

#### 3.1 Negative Dependence via Local Odds Ratios

In our sampling scheme, only the table total is fixed. A simple and popular approach to making inferences on the joint probability distribution  $\pi$  is to assume a multinomial model for the cell counts. This would lead to maximum likelihood estimates equal to the empirical frequencies, and to Bayesian posterior inferences closely approximating empirical frequencies, at least under independent uniform priors on the cells. This approach is reasonable but not fully adequate to the situation at hand, where, many cells are empty or have small counts, and where some smoothness, recognizing the ordered structure of the categories and the likely negative dependence of the variables, is clinically plausible.

For example, consider cell BI=90 and RS=1 at month 3, in Table 4. Estimates of cell probabilities close to the empirical frequencies in this and the neighboring

cells would lead to a clinically implausible overall result. To illustrate, focus on the two-by-two table formed by the patients with RS of 1 and 2 and BI of 85 and 90. Within this table, the odds of having a BI of 90 vs 85 are even in the RS=1 column and are 9 to 4 in the RS=2 column. So the odds in favor of a higher BI increase with RS, which is contrary to the fact that both scales measure disability, and do so inversely to one another.

Here we are interested in using this simple condition on the odds to produce inferences that a) avoid results that would be contrary to clinical knowledge and b) sharpen our estimate of cell probabilities in regions of the table where the counts are small. We are going to assume that any two-by-two table like the one defined above will show association (in this case negative) among the scales. More formally, we will consider all subtables of the form:

$$\begin{array}{cc} \pi_{r-1,b-1} & \pi_{r-1,b} \\ \pi_{r,b-1} & \pi_{r,b} \end{array}$$

for  $r = 2, \dots, R$  and  $b = 2, \dots, B$ . The local odds ratio for this table is defined as

$$\frac{\pi_{r-1,b-1} \pi_{r,b}}{\pi_{r-1,b} \pi_{r,b-1}}$$

We can impose a negative association between the two scales by constraining the prior to the set of  $\pi$ 's such that all the local odds ratios are smaller than one, or, equivalently, that

$$\pi_{r-1,b-1} \pi_{r,b} < \pi_{r-1,b} \pi_{r,b-1}. \tag{1}$$

If the association were positive, that is if the inequality were reversed, the condition would be the so-called Total Positivity of order 2 (Karlin 1969). Following Douglas, Fienberg, Lee, Sampson, and Whitaker 1991 we will refer to condition (1) as Total Negativity of order 2, or TN<sub>2</sub>.

For matrices of cell probabilities that satisfy the  $TN_2$  constraint, the likelihood function for a single table has the same form as the one based on multinomial sampling. Therefore, assuming a uniform prior in the region satisfying the constraint,

$$p(\pi|\mathbf{x}) \propto \begin{cases} \prod_{r,b} \pi_{rb}^{x_{rb}} & \text{if } \pi \text{ is } TN_2 \\ 0 & \text{if } \pi \text{ is not } TN_2 \end{cases}$$

where  $\mathbf{x}$  and  $\pi$  are the matrices of counts and cell probabilities, respectively.

There is an extensive literature on modeling dependence in  $I \times J$  contingency tables, reviewed by Etzioni, Fienberg, Gilula, and Haberman 1994. Among the conditions that have been proposed,  $TN_2$  is among the most directly interpretable and computationally tractable, as we will see later in the section. Compared to other notions of negative dependence,  $TN_2$  is strong, as it implies other well known notions of dependence such as stochastic dominance of the conditional distributions, for both sets of conditionals, and monotonicity of both the conditional expectation functions. See Douglas, Fienberg, Lee, Sampson, and Whitaker 1991 for a detailed discussion of dependence models and their relations to log-linear and generalized linear models, and Johnson and Albert 1999 for a general discussion of modeling relationships between ordinal measurements.

### 3.2 Sampling from the Posterior Distribution of $\pi$ Under the $TN_2$ constraint

Sampling from the posterior distribution of  $\pi$  for an individual table under the  $TN_2$  constraint requires a tailored approach because of the numerous constraints and the relatively small portion of the hypercube supporting  $\pi$  that satisfies those constraints. The following implementation is simple and proved reliable in our application. It is based on augmenting the table with one additional auxiliary cell that we term the “daemon cell”, and sampling each cell probability in turn, conditional on all other cells probabilities in the table, except the daemon’s. Computationally,

advantages of using the daemon cell over choosing an existing cell in the table include that: a) none of the  $TN_2$  constraints apply to the daemon cell probability, so that one only has to deal with the constraints that apply to cell probability currently sampled; and b) when the chain is away from convergence, the daemon cell probability acts as a conduit for redistributing mass from cells that have too much to cells that have too little. The daemon approach gives us the opportunity to tune the capacity of the conduit, which would be fixed, and likely to be small, if we chose a cell in the table.

Let  $x^*$  and  $\pi^*$  be the count and cell probability for the demon cell. Here  $\pi^*$  is an unknown parameter, and  $x^*$  can be set arbitrarily. The chosen value of  $x^*$  has some effect on the speed of convergence of the chain, and small values should be avoided. In particular, if  $n^*$  is small, the conduit provided by the daemon cell could be too narrow. Values around  $n/3$  to  $n/2$  have worked well in our application. Performance of the chain shows little sensitivity in that range.

Our first step is to reparameterize the cell probabilities so that they sum to one in the table including the daemon cell or augmented table. Formally, we define

$$q_{rb} = \pi_{rb} * (1 - \pi^*).$$

We then consider the one-dimensional full conditional distribution of  $q_{rb}$  given all other cell probabilities except  $\pi^*$ , under the constraint

$$\pi^* + \sum_{r,b} q_{rb} = 1.$$

This distribution is a truncated Beta and can be easily simulated. After the simulation, the portion of the augmented table corresponding to the original  $\pi$  can be renormalized to sum to one, and the daemon cell can be ignored.

The upper limit  $q_u$  and lower limit  $q_l$  of the full conditional distribution of  $\pi_{rb}$  are derived from the local dominance condition, by considering all the constraints

that apply to cell  $r, b$ . There are four such constraints for cells inside the boundary of the table, two for cells on the boundary, but away from the corners, and one for cells on the corners.

Formally, for cells inside the boundary, that is for  $1 < r < R$ , and  $1 < b < B$ , we have:

$$\begin{aligned} q_l &= \max \left\{ \frac{q_{r-1,b} q_{r,b+1}}{q_{r-1,b+1}}, \frac{q_{r+1,b} q_{r,b-1}}{q_{r+1,b-1}} \right\} \\ q_u &= \min \left\{ \frac{q_{r-1,b} q_{r,b-1}}{q_{r-1,b-1}}, \frac{q_{r+1,b} q_{r,b+1}}{q_{r+1,b+1}} \right\}, \end{aligned}$$

while for cells on the boundary:

$$\begin{aligned} \text{if } r = 1, b = 1 & \quad q_l = 0 & \quad q_u = \frac{q_{1,2} q_{2,1}}{q_{2,2}} \\ \text{if } r = 1, b = B & \quad q_l = \frac{q_{1,B-1} q_{2,B}}{q_{2,B-1}} & \quad q_u = 1 \\ \text{if } r = R, b = 1 & \quad q_l = \frac{q_{R,2} q_{R-1,1}}{q_{R-1,2}} & \quad q_u = 1 \\ \text{if } r = R, b = B & \quad q_l = 0 & \quad q_u = \frac{q_{R,B-1} q_{R-1,B}}{q_{R-1,B-1}} \\ \text{if } r = 1, b > 1, b < B & \quad q_l = \frac{q_{1,b-1} q_{2,b}}{q_{2,b-1}} & \quad q_u = \frac{q_{1,b+1} q_{2,b}}{q_{2,b+1}} \\ \text{if } r = R, b > 1, b < B & \quad q_l = \frac{q_{R,b+1} q_{R-1,b}}{q_{R-1,b+1}} & \quad q_u = \frac{q_{R,b-1} q_{R-1,b}}{q_{R-1,b-1}} \\ \text{if } r > 1, r < R, b = 1 & \quad q_l = \frac{q_{r-1,1} q_{r,2}}{q_{r-1,2}} & \quad q_u = \frac{q_{r+1,1} q_{r,2}}{q_{r+1,2}} \\ \text{if } r > 1, r < R, b = B & \quad q_l = \frac{q_{r+1,B} q_{r,B-1}}{q_{r+1,B-1}} & \quad q_u = \frac{q_{r-1,B} q_{r,B-1}}{q_{r-1,B-1}}. \end{aligned}$$

These constraints are much smaller in number than the constraints implied by stochastic dominance, adding to the computational convenience of the  $TP_2$  approach.

If a set of  $q$ 's satisfies the  $TP_2$  constraints above, that is if  $q_l < q_{rb} < q_u$ , than the subsequent  $q$ 's drawn in the Markov chain will also satisfy the constraints. Therefore, if the Markov chain is started at a value satisfying the constraint, all subsequent values will also. The closer the empirical counts are to satisfying the  $TN_2$  constraints, the faster the convergence of this Markov chain implementation will be. Autocorrelation of the chain will increase with the mass that the untruncated

Beta functions assign to the regions outside the truncation points. R software for fitting this model to an arbitrary rectangular table with either a  $TN_2$  or a  $TP_2$  constraint, is available from the author’s website at url

<http://www.jhsph.edu/Departments/Biostats/research/parmigiani.html>.

### 3.3 A Markov Model for Contingency Tables

Data is available at successive time points after a stroke, and individual’s disabilities are likely to be correlated over time, so we considered a first–order Markov process for the contingency tables. We specified a prior distribution for the cell probabilities at time 1,  $p(\pi^{(1)})$ , and a conditional prior distribution for the cell probabilities at time  $t$  given the previous times,  $p(\pi^{(t)}|\pi^{(t-1)}, \tau)$ ,  $t > 1$ . The parameter  $\tau$  controls the dependence in the conditional distribution, as described later.

A practical approach is to model cell probabilities as a Dirichlet distribution and assume that after time 1, the mean is given by the cell probability at the previous time point. The Dirichlet distribution also includes a dispersion parameter which, in our context, controls the average amount of variation that occurs in the table from one time interval to the next. Indicate by  $\mathcal{D}(\mathbf{a})$  a Dirichlet distribution with parameter vector  $\mathbf{a}$ . Then:

$$p(\pi^{(t)}|\pi^{(t-1)}) = \mathcal{D}(\pi^{(t-1)}\tau + 1) \tag{2}$$

Because the mean of  $\pi^{(t)}$  a posteriori is a weighed average of  $(\pi^{(t-1)} + 1)/\tau$  and the observed relative frequencies, with weight  $\tau/(\tau + n^{(t)})$  on the former,  $\tau$  is interpretable as a prior weight, or “prior sample size”. In our application,  $\tau$  reflects the degree of change in the joint distribution from one period to the next. This change is mainly attributable to improvements in the ability to carry out activities of daily living that are sufficient to increase BI, but not to decrease RS.

We assume that  $\tau$  is the same in both the 1–3 month interval and the 3–6 month interval. As the earlier months are likely to be associated with more rapid change in the distribution, the overall amount of change in the two intervals is likely to be similar. In a different application, a simple adjustment for the unequal spacings could be carried out by multiplying  $\tau$  by the length of the interval.

In our application,  $\pi^{(1)}$  is a priori uniform within the  $TN_2$  subspace. Analyses based on informative Dirichlet priors can proceed in the same manner. The prior on the autoregression parameter  $\tau$  is a normal with mean 50 and standard deviation 10, truncated to the range (10,200) and discretized in 5–unit intervals. In short longitudinal studies, inference will be sensitive to the prior distribution on  $\tau$ .

### 3.4 Sampling from the Overall Posterior Distributions

Parameter inference in our model uses a Markov Chain Monte Carlo (Gilks, Richardson, and Spiegelhalter 1996) method that alternates between 1) drawing a sample of the unobserved cross classification for the Framingham Study given the observed Table 5 and the cell probabilities, and 2) drawing the cell probabilities given the complete tables, the neighboring probabilities and the constraints, and 3) drawing the autoregression parameter  $\tau$ .

Our overall approach to sampling from the posterior distribution of the parameters  $\pi^{(t)}$  and  $\tau$  is as follows. We sample each table given the others. Table  $t$  depends on both table  $t - 1$  (via the prior on  $\pi^{(t)}$ ), and table  $t + 1$  (via the prior on  $\pi^{(t+1)}$ ). The term coming from the prior on  $\pi^{(t)}$  combines with the likelihood for  $\pi^{(t)}$  in a conjugate fashion. If one ignores the contribution of the prior on  $\pi^{(t+1)}$ , the method of Section 3.2 can be used to draw samples from  $\pi^{(t)}$ , after a minor adaptation to update the posterior hyperparameters. We use such samples to draw proposed values in a Metropolis update. Because the contribution of the prior on  $\pi^{(t+1)}$  is typically not overwhelming, this approach is efficient.

The full conditional distribution of the  $q_{rb}$  associated with  $\pi^{(t)}$  in the autoregressive case is:

$$q_{rb}^{n_{rb} + \tau \pi_{rb}^{(t-1)}} (1 - q_{rb})^{n^*(t) + \tau \pi^*(t-1)} [\pi_{rb}^{(t+1)}]^{q_{rb} \tau} [\pi^*(t+1)]^{(A - q_{rb}) \tau} \Gamma(q_{rb} \tau + 1) \Gamma((A - q_{rb}) \tau + 1),$$

for  $q_l < q_{rb} < q_u$  and where

$$A = 1 - \sum_{(r', b') \neq (r, b)} q_{r', b'}.$$

## 4 Results

Table 6 presents the posterior means of the cell probabilities. Comparing the estimated means to the empirical counts of Table 4, it is evident that the TN<sub>2</sub> condition contributed to smoothing the cell probabilities and to providing information, via a clinically compelling structural condition, about the sparsely populated cells in the table. IN this application, an additional factor contributing to smoothing of the cell probabilities is the coarseness of the categories of the cross-classification in the Framingham study.

Posterior means are a useful summary of each cell probability's marginal distribution. However, the overall table has limitations as a synthetic representation of the joint distribution. While this is always the case when dealing with correlated parameters, in this case, one additional caveat applies. Even though each sampled set of cell probabilities satisfies the TN<sub>2</sub> constraints, the table made of the cell-specific averages does not necessarily satisfy them. An alternative summarization strategy that does not suffer from this limitation is the mode of the joint distribution. While the mode would be our recommended summarization strategy in smaller dimensional tables, in this application its evaluation is unfortunately not practical.

Tables 7 and 8 summarize the conditional distributions of RS given BI, and that of BI given RS, respectively. These are the distribution that are directly relevant in

translating results of clinical studies from one scale to the other.

To illustrate how the translation may be carried out, consider the results of a trial comparing a treatment to a control group, with samples of 100 patients in each group. Hypothetical distributions of RS in the two groups are shown in Table 9. Using Table 8 one can convert the two distributions in Table 9 in terms of BI. For example the 10 patients in the treatment group receiving a RS of 3 would be reallocated to the BI categories proportionally to the probabilities in the column labeled 3 in Table 8, and so forth. More generally, defining by  $\pi_{b|r}$  the conditional probabilities of BI given RS, any marginal distribution  $\pi_r$  of RS can be converted into BI by the law of total probability:

$$\pi_b = \sum_{r=0}^5 \pi_{b|r} \pi_r.$$

Figure 1 portrays the uncertainty in the cell estimates, by presenting samples of cumulative distribution functions (CDF) of representative conditional distributions of each of the two kinds. The samples also highlight the flexibility of the nonparametric approach adopted here in capturing the shape of the conditional distributions, and also the effects of the constraints, seen especially in the lower range of the distribution on the left. Using samples of curves such as those of Figure 1, is straightforward to develop probability intervals and standard deviations on the results of conversions such as those illustrated in the previous hypothetical example.

When investigating the relationship between two clinical classifications of the same patients using different scales, it is common to examine overall measures of association. Goodman and Kruskal developed the index  $\lambda$ , expressly for situations in which the goal is to predict the category for the row (column) variable from the category of the column (row) variable. The index  $\lambda$  is based on the proportional reduction in error (PRE) and measures the relative improvement in predicting one scale when the other is known, as is germane in this application. See Goodman

and Kruskal 1954 and Bishop, Fienberg, and Holland 1975 for discussion. Figure 2 summarizes the posterior distributions of the measures of association  $\lambda_{B|R}$  and  $\lambda_{R|B}$  between the two scales. Values of  $\lambda_{B|R}$  are significantly smaller than those of  $\lambda_{R|B}$ . This is partly to be expected, in view of the much larger number of BI categories. In this case additional problems arise. For example the conditional distribution of BI for RS category 4 is extremely diffuse.

## 5 Discussion

Measures of outcomes after stroke will typically be correlated but not perfectly so. The correlation results from the measures' common emphasis on functional status, particularly those components of functional status involving physical function. The divergence results from the focus on somewhat different elements of functional status, different resolution of the scales (e.g. the RS has fewer categories than the BI), different background of the respondents (e.g. the RS assessment is typically performed by a clinician, while the BI accommodates reports by the patient, clinician or proxy), different question formats, and so forth. Recognizing these divergences, some trials of acute stroke treatment have measured stroke outcomes using multiple scales, and assessed the efficacy of interventions using a global outcome measure, which simultaneously includes multiple dimensions (Tilley, Marler, Geller, Lu, Legler, Brott, Lyden, and Grotta 1996).

However, the vast majority of the literature reports either the RS or the BI but not both, and thus a methodology for translating the results of one scale into the other would be of practical value. In this article we developed an approach and estimated tables for comparing and integrating the results of clinical trials of stroke treatments using the RS and the BI, and for integrating clinical trials results into comprehensive decision model for the assessment of long term implications and

cost–effectiveness of stroke prevention and acute treatment interventions. While we combined two high–quality studies, we could

Our results indicate that substitution of one scale for the other at full resolution involves substantial uncertainty. In particular, prediction of BI values based on RS is not generally reliable because of the larger number of BI categories. The effect of this uncertainty on the practical ability to translate results from one scale to the other in both clinical trial and cost–effectiveness modeling applications will vary with the specific application, depending on the strength of the relationship between disability category and primary outcome of interest. For example, survival may depend mostly on aspects of the two scale that correlate well with one another (Wolfe, Taub, Woodrow, and Burney 1991) and may be translated reliably, while quality adjusted survival is more likely to be sensitive to specific aspects of the two scale that are difficult to translate.

We developed a tailored statistical approach to the analysis, that takes into account the negative dependence between the RS and the BI scale, the time variation of the cross-classification, and the different resolutions of our two data sources. We adopted a Bayesian implementation, because it provided a practical solution to many of the challenges posed by the application: combination of information from different sources, multiresolution analysis, estimation in presence of complex constraints, lack of reliable asymptotic approximations to the variance of the estimated parameters, and need to make inferences on complex functions of the parameters, such as conditional distributions and the  $\lambda$ 's.

In the joint distribution reported in Section 4, the variability in one scale given the other is within rater, and reflects the different degree to which patients can be classified into categories according to the two scales. The result of the conversion is interpretable as an estimate of the distribution that would have been obtained if

the same set of raters had measured BI instead of RS. In practice, one may be also be interested in modeling the variability across raters, arising from different raters applying the categorization differently. This can be critical when the goal of the conversion is combining evidence across trials, each of which uses a different set of raters. Depending on the study design, multi-level models could be used for this purpose (Johnson and Albert 1999).

A joint analysis of two related indices can be also useful in the framework of a single study in which both indices are collected, by providing additional power to detect intervention effects. In the two-arms case the null hypothesis could be that the joint distribution is the same across groups. An image display of, say, the cell-specific log odds ratios across groups, would provide a map of the intervention effects.

Approaches to multiresolution analysis using multiple imputations or MCMC approaches to recreate the lowest resolution in all subsets of the problem have been previously explored. For example Heitjan and Rubin 1990 considered it in the context of age heaping. We are not aware of existing Bayesian approaches to inference and computing in contingency tables under the  $TN_2$  or similar constraints. However, alternative Bayesian approaches to modeling dependence in ordinal categorical data are numerous (Erkanli, Stangl, and Müller 1993; Cowles, Carlin, and Connett 1996; Johnson 1996; Johnson and Albert 1999). A common strategy is to postulate the existence of individual-level latent continuous variables, whose binning leads to the observed ordinal categories, and to express conditional models in terms of the latent variables. in this vein Cargnoni, Müller, and West 1997 utilize a latent dynamic linear model to model a sequence of one-dimensional multinomial tables over time. Advantages of the approach developed here over those based on latent variables include that a) there is no need to model/simulate latent variables at the

individual patient level; computing takes place in much lower dimensions, and it is expedited and stabilized; and 2) model parameters are more directly related to contingency tables are more natural to interpret. For example in our framework it is simpler to incorporate clinical requirements such as the  $TN_2$  constraint, and to assign meaningful vague prior distribution, such as our uniform within the  $TN_2$  constraint. A disadvantage in other applications is that the latent variables may themselves carry important scientific information.

Ordered categorical scales for measuring disability and severity of disease are common in all areas of medicine (Potparic and Gibson 1993), and arise from many clinical trials and epidemiological studies. Scales are constantly revised and improved. Also, general scales like the BI coexist with disease-specific ones. As a result, the issue of cross-calibration of different, monotonically related, scales is commonly encountered. More generally, inference on cell probabilities of contingency tables when there is knowledge about the monotonicity of the relationship between the two variables being classified is common. Our strategy is relatively simple computationally, compared to alternative full Bayesian analyses, it is straightforward to interpret and communicate to a non-statistical audience, and it is especially attractive in sparse tables because the “shingle” effect of the constraints, which induces a smooth fit without the need for parametric assumptions. Our approach to sampling from contingency tables under a constraint using a daemon cell is applicable irrespective of the type of constraint considered and could be useful outside the area of categorical data analysis. Our general strategy is scalable to contingency tables of moderately high dimension.

## References

- Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland (1975). *Discrete Multivariate Analyses: Theory and Practice*. MIT Press.
- Cargnoni, C., P. Müller, and M. West (1997). Bayesian forecasting of multinomial time series through conditionally Gaussian dynamic models. *Journal of the American Statistical Association* 92, 640–647.
- Cowles, M. K., B. P. Carlin, and J. E. Connett (1996). Bayesian Tobit modeling of longitudinal ordinal clinical trial compliance data with nonignorable missingness. *Journal of the American Statistical Association* 91, 86–98.
- Douglas, R., S. E. Fienberg, M.-L. T. Lee, A. R. Sampson, and L. R. Whitaker (1991). Positive dependence concepts for ordinal contingency tables. pp. 189–202. Institute of Mathematical Statistics.
- Duncan, P. W., L. B. Goldstein, G. W. Divine, J. R. Feussner, and D. B. Matchar (1992). Measurement of motor recovery following stroke: Outcome assessment and sample size requirements. *Stroke* 23, 1084–1089.
- Duncan, P. W., L. B. Goldstein, R. D. Horner, P. B. Landsman, G. P. Samsa, and D. B. Matchar (1994). Similar motor recovery of upper and lower extremities after stroke. *Stroke* 25, 1181–1188.
- Duncan, P. W., H. S. Jorgensen, and D. T. Wade (2000). Outcome measures in acute stroke trials: A systematic review and some recommendations to improve practice. *Stroke* 31(6), 1429–1438.
- Duncan, P. W., S. M. Lai, and J. Keighley (2000). Defining post-stroke recovery: implications for design and interpretation of drug trials. *Neuropharmacology* 39(5), 835–841.
- Erkanli, A., D. Stangl, and P. Müller (1993). A Bayesian analysis of ordinal data

- using mixtures. In *ASA Proceedings of the Section on Bayesian Statistical Science*, pp. 51–56.
- Etzioni, R. D., S. E. Fienberg, Z. Gilula, and S. J. Haberman (1994). Statistical models for the analysis of ordered categorical data in public health and medical research. *Statistical Methods in Medical Research* 3, 179–204.
- Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (Eds.) (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.
- Goodman, L. A. and W. H. Kruskal (1954). Measures of association for cross classifications. *Journal of the American Statistical Association* 49, 732–764.
- Heitjan, D. F. and D. B. Rubin (1990). Inference from coarse data via multiple imputation with application to age heaping. *Journal of the American Statistical Association* 85, 304–314.
- Johnson, V. E. (1996). On Bayesian analysis of multirater ordinal data: An application to automated essay grading. *Journal of the American Statistical Association* 91, 42–51.
- Johnson, V. E. and J. Albert (1999). *Ordinal Data Modeling*. Berlin: Springer.
- Karlin, S. (1969). *Total positivity*. Stanford, California: Stanford University Press.
- Mahoney, F. I. and D. W. Barthel (1965). Functional evaluation: the barthel index. *Maryland State Medical Journal* 14, 61–65.
- Matchar, D. B., G. P. Samsa, J. R. Matthews, M. Ancukiewicz, G. Parmigiani, V. Hasselblad, P. A. Wolf, R. B. D’Agostino, and J. Lipscomb (1997). The stroke prevention policy model (SPPM): Linking evidence and clinical decisions. *Annals of Internal Medicine* 127(8S), 704–711.
- Potparic, O. and J. Gibson (1993). *Dictionary of Clinical Tests: A Concise Guide to Tests, Scales and Scores in Medicine*. Boca Raton: CRC Press–Parthenon

Publishers.

- Rankin, J. (1957). Cerebral vascular accidents in patients over the age of 60: II. prognosis. *Scottish Medical Journal* 2, 200–215.
- Tanner, M. A. (1991). *Tools for Statistical Inference – Observed Data and Data Augmentation Methods*, Volume 67 of *Lecture Notes in Statistics*. New York: Springer-Verlag.
- Tanner, M. A. and W. H. Wong (1987, June). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* 82(398), 528–550.
- Tilley, B., J. Marler, N. Geller, M. Lu, J. Legler, T. Brott, P. Lyden, and J. Grotta (1996). Use of a global test for multiple outcomes in stroke trials with application to the national institute of neurological disorders and stroke t-pa stroke trial. *Stroke* 27(11), 2136–2142.
- van Swieten, J., P. F. Koudstaal, M. C. Visser, H. J. A. Schouten, and J. van Gigin (1988). Interobserver agreement for the assessment of handicap in stroke patients. *Stroke* 19, 604–607.
- Wolfe, C. D. A., M. S. C. Taub, E. J. Woodrow, and P. G. J. Burney (1991). Assessment of scales of disability and handicap for stroke patients. *Stroke* 22, 1242–1244.
- World Health Organization (1993). Proposal for multinational monitoring and determinants in cerebrovascular disease. Monica Project WHO/MNC82, World Health Organization.

Score	Interpretation
0	no symptoms
1	no significant disability (i.e., minor symptoms which do not interfere with the patient abilities to carry out their usual activities of daily living)
2	slight disability (i.e., unable to carry out some previous activities, but able to look after own affairs without assistance)
3	moderate disability (i.e., requiring some assistance with activities, but able to walk without assistance)
4	moderately severe disability (i.e., unable to walk without assistance, unable to attend to bodily needs without assistance)
5	severe disability (i.e., bedridden, incontinent, requiring constant nursing care and attention).

Table 1: Verbal description of the six categories defining the Modified Rankin Stroke Outcome Scale (RS).

Item	Maximum Score
feeding	10
transferring	15
grooming	5
toileting	10
bathing	5
walking	15
stairs	10
dressng	10
bowel continence	10
bladder continence	10

Table 2: The 10 activities of daily living considered by the Barthel ADL Index (BI).

GENDER	214 Male, 245 Female
MEAN AGE	70.5 (+/- 11.4)
RACE	366 White, 78 Black, 15 Other
TIME SINCE STROKE (MEAN)	8.5 Days (+/- 3.6)
STROKE SEVERITY (ORPINGTON)	51 Major, 227 Moderate, 179 Minor

Table 3: Demographic characteristics of the 459 patients enrolled in the Kansas City Stroke Study

	Month 1							Month 3							Month 6						
Rankin	0	1	2	3	4	5	Rankin	0	1	2	3	4	5	Rankin	0	1	2	3	4	5	
Barthel							Barthel							Barthel							
0	0	0	0	0	1	10	0	0	0	0	0	1	7	0	0	0	0	0	0	5	
5	0	0	0	0	0	9	5	0	0	0	0	0	2	5	0	0	0	0	1	2	
10	0	0	0	0	3	1	10	0	0	0	0	0	2	10	0	0	0	0	1	1	
15	0	0	0	0	2	1	15	0	0	0	0	5	0	15	0	0	0	0	3	0	
20	0	0	0	0	5	3	20	0	0	0	0	3	0	20	0	0	0	0	3	2	
25	0	0	0	0	7	1	25	0	0	0	0	7	0	25	0	0	0	0	4	1	
30	0	0	0	0	7	0	30	0	0	0	0	6	0	30	0	0	0	0	4	0	
35	0	0	0	0	8	0	35	0	0	0	0	7	0	35	0	0	0	1	3	0	
40	0	0	0	2	14	0	40	0	0	0	0	3	0	40	0	0	0	0	4	0	
45	0	0	0	0	4	0	45	0	0	0	0	3	0	45	0	0	0	0	2	0	
50	0	0	0	1	8	0	50	0	0	0	1	6	0	50	0	0	0	2	3	0	
55	0	0	0	1	9	0	55	0	0	0	0	5	0	55	0	0	0	3	5	0	
60	0	0	1	5	9	0	60	0	0	0	3	6	1	60	0	0	0	3	4	0	
65	0	0	1	5	3	0	65	0	0	0	4	3	0	65	0	0	1	0	5	0	
70	0	0	1	12	3	0	70	0	0	1	11	8	0	70	0	0	0	6	1	0	
75	0	0	1	19	6	0	75	0	0	0	5	0	0	75	0	0	0	12	2	0	
80	0	0	1	18	3	0	80	0	0	6	12	0	0	80	0	0	0	9	0	0	
85	0	0	4	26	0	0	85	0	2	4	21	0	0	85	0	0	3	18	0	0	
90	1	0	7	24	1	0	90	1	2	9	23	0	0	90	0	3	11	13	0	0	
95	1	4	31	13	0	0	95	1	4	24	20	0	0	95	2	6	35	16	0	0	
100	2	17	62	11	0	0	100	7	44	72	9	0	0	100	11	57	62	11	0	0	

Table 4: Cross classification of patients by BI (rows) and RS (columns) at one, three, and six months after stroke in the Kansas City Stroke Study

		Rankin			
		0,1,2	3	4	5
Barthel	<b>0-35</b>	0	0	6	25
	<b>40-65</b>	0	5	21	3
	<b>70-90</b>	5	16	2	0
	<b>95-100</b>	17	0	0	0

Table 5: Estimated joint probability distribution of patients by RS (rows) and BI (columns) at three months after stroke, as reported by Wolfe et al (1991). Counts are normalized to a total of 100 patients. The sample size in the study was 50.

		Rankin					
		0	1	2	3	4	5
	0	0	0	1	12	75	255
	5	0	0	1	11	55	118
	10	0	0	1	6	27	41
	15	0	1	6	32	110	75
	20	0	1	8	35	105	52
	25	0	2	14	57	149	58
	30	0	3	16	60	144	47
	35	0	4	16	58	125	35
	40	0	4	16	54	105	25
Barthel	45	1	4	15	47	87	18
	50	1	8	29	85	139	25
	55	1	8	25	72	102	16
	60	3	14	44	117	138	18
	65	4	19	54	136	112	13
	70	8	31	86	207	128	12
	75	7	27	69	155	49	4
	80	13	46	111	232	44	3
	85	23	75	170	328	41	2
	90	38	112	241	402	36	1
	95	79	213	422	395	24	1
	100	339	818	1390	235	8	0

Table 6: Estimated joint distribution of RS and BI at three months after stroke, for 10,000 hypothetical patients. Entries correspond to posterior means of MCMC samples, rounded to sum to 10,000.

		Rankin					
		0	1	2	3	4	5
Barthel	0	0	0	30	338	2179	7453
	5	1	0	71	597	2985	6346
	10	0	0	121	846	3617	5416
	15	1	23	259	1418	4932	3367
	20	2	50	382	1740	5221	2605
	25	6	81	490	2023	5326	2074
	30	9	111	579	2210	5347	1744
	35	15	147	688	2442	5255	1453
	40	23	188	794	2622	5146	1227
	45	32	228	880	2763	5045	1052
	50	45	276	996	2967	4845	871
	55	61	342	1132	3186	4558	721
	60	85	428	1317	3500	4119	551
	65	120	556	1605	4025	3316	378
	70	160	667	1829	4382	2701	261
	75	235	862	2226	4979	1572	126
	80	295	1020	2470	5177	974	64
	85	365	1164	2659	5133	646	33
90	460	1352	2899	4842	430	17	
95	700	1874	3723	3486	212	5	
100	1214	2931	4983	842	30	0	

Table 7: Estimated conditional distributions of RS given BI, for 10,000 hypothetical patients in each BI category.

		Rankin					
		0	1	2	3	4	5
Barthel	0	0	0	4	42	414	3115
	5	0	0	5	40	307	1436
	10	0	0	3	23	150	494
	15	0	4	21	116	611	917
	20	1	7	28	128	581	638
	25	3	16	50	207	828	710
	30	5	22	57	218	801	575
	35	7	25	60	212	692	421
	40	9	28	59	196	583	306
	45	10	28	55	173	480	220
	50	25	57	105	312	772	306
	55	26	55	93	261	567	198
	60	55	103	161	428	765	225
	65	78	135	197	495	619	155
	70	146	227	316	757	708	150
	75	141	194	254	568	272	48
	80	254	329	405	848	242	35
85	450	537	622	1201	229	26	
90	736	809	880	1470	198	17	
95	1530	1531	1543	1446	134	7	
100	6524	5893	5082	859	47	1	

Table 8: Estimated conditional distributions of BI given RS, for 10,000 hypothetical patients in each RS category. The conditional distributions are dispersed, indicating that conversion may involve significant uncertainty.

	Rankin					
	0	1	2	3	4	5
Treatment	0	40	50	10	0	0
Control	0	0	0	10	40	50

Table 9: Observed distributions of RS in the hypothetical trial.

	Barthel																					
	0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100	
Treatment	0.1	0.1	0	0.2	0.3	0.5	0.6	0.6	0.6	0.6	1.1	0.9	1.6	2	3.2	2.6	4.2	6.5	9.1	15.3	49.8	
Control	17.3	8.4	3.1	7.1	5.6	7.1	6.3	5.1	4.1	3.2	4.9	3.5	4.6	3.7	4.3	1.9	2	2.2	2.3	2	1.1	

Table 10: Estimated distributions of BI in hypothetical trial. Each entry is computed from the marginal distribution  $\pi_r$  using the estimated conditional probabilities of BI given RS and the law of total probability.

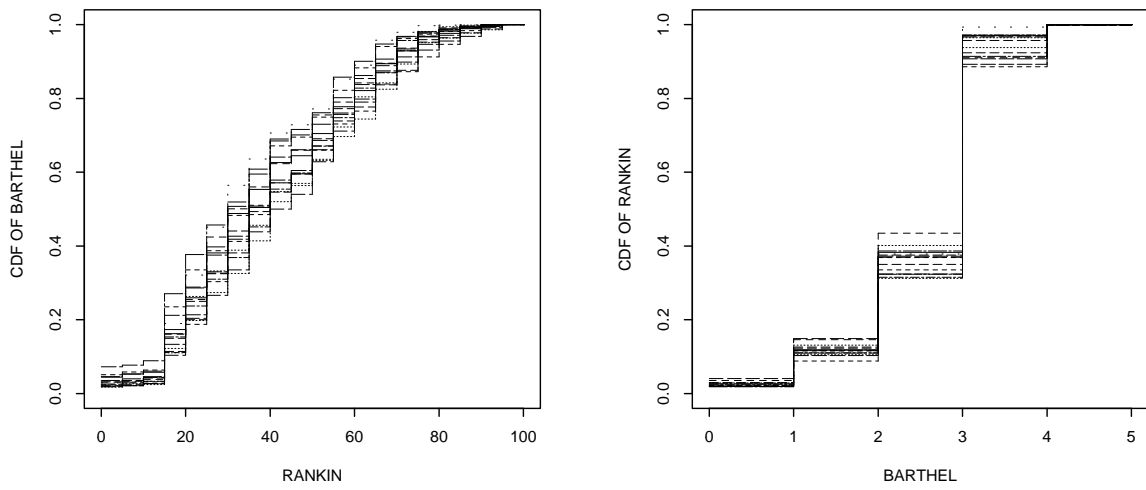


Figure 1: Posterior samples of cumulative distribution functions of RS given  $BI = 80$  (left) and  $BI$  given  $RS = 4$  (right). Aside from the  $TN_2$  condition, no constraints are posed on the jumps of these CDF's. Some of the lower probabilities show higher uncertainty than those closer to .5. Also, higher uncertainty is associated with jumps in less smooth parts of the CDF.

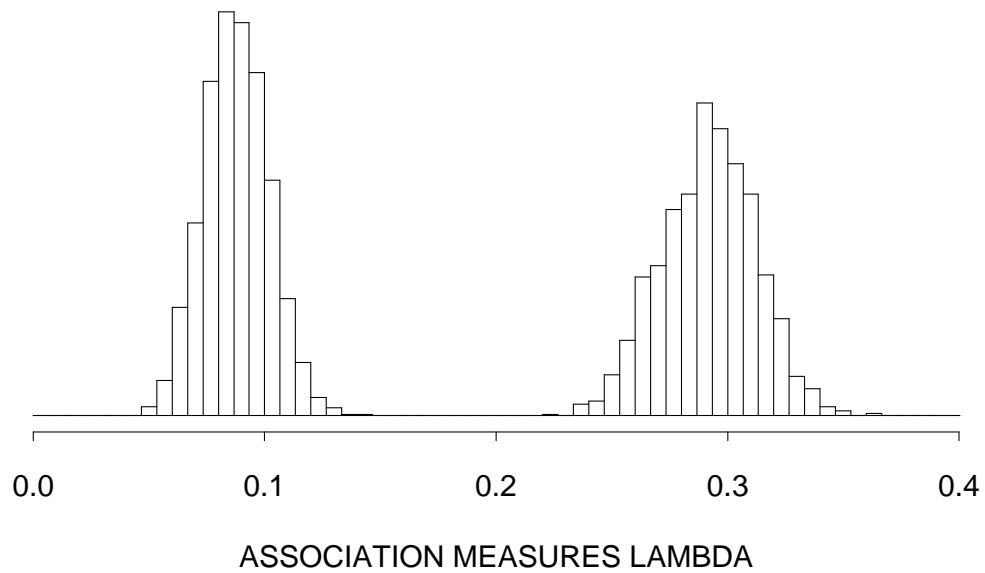


Figure 2: Samples from the posterior distributions of association measures  $\lambda$  for  $P(B|R)$  (left) and  $P(R|B)$  (right). Distributions concentrate on relatively low values. The difference in the two indices is attributable to the much larger number of BI categories as well as some diffuse conditional distribution, such as that of BI for RS category 4.