A Bayesian Analysis of the Age of Occurrence of a Change in Prevalence of Activity Limitation Among Adults for the NHIS

Balgobin Nandram and **Yan Bai**

Department of Mathematical Sciences, Worcester Polytechnic Institute
100 Institute road, Worcester MA 01609
(balnan@wpi.edu, baiy@wpi.edu)

Jai Won Choi

Office of research and Methodology, National Center for Health Statistics, CDC
3311 Toledo Road, Hyattsville, MD 20782
(jwc7@cdc.gov)

June 30, 2005

Abstract

Data from the 1995 National Health Interview Survey (NHIS) indicate that for adults with chronic conditions, a significant change in prevalence of activity limitation typically occurs between the ages of 40 years and 70 years (i.e., the proportion of young adults with activity limitation is small and roughly constant with age and then it starts to change, roughly constant again). We use a hierarchical Bayesian model to detect the change of prevalence of a positive activity limitation status (ALS) for twelve domains formed by crossing education, sex and race. We obtain weighted binomial counts using a regression analysis on the sample weights which incorporate important survey features. Given the proportion of individuals in the population with positive ALS, we assume that the number of individuals with positive ALS at each age has a binomial probability mass function. The proportions across ages are different, and have the same beta distribution up to the change point (unknown), and the proportions after the change point have a different beta distribution. We consider two different analyses. The first considers a model for each domain individually and the second considers the twelve domains simultaneously in a single model to "borrow strength" as in small area estimation. In the first analysis, we use the Gibbs sampler to fit the model, and a computation of the marginal likelihoods, using an output analysis from the Gibbs sampler, provides the posterior distribution of the change point. In the second analysis, we use the Gibbs sampler to fit the joint posterior distribution of the twelve change points. This is a difficult problem because the joint posterior density requires numerical computation of a triple integral at each iteration. The other parameters of the process are obtained using data augmentation by a Metropolis sampler and a Rao-Blackwellization. We found that overall the age at which a significant change occurs is about 56 years, with some variation variation with education, sex and race.

Key Words: Change point, Gibbs sampler, Hierarchical Bayesian model, Marginal likelihood, Small area estimation.

1. Introduction

Data from the 1995 National Health Interview Survey indicate that for adults with chronic conditions, a significant change of activity limitation typically occurs between the ages of 40 years and 70 years. This information is important to medical practitioners for planning purposes. Important survey features are contained in the survey weights. The data we analyze consist of weighted binomial counts, obtained by using a logistic regression to incorporate the survey weights. We use a Bayesian methodology to detect the change point of a positive activity limitation status across twelve domains, formed by crossing education, sex and race. Each domain has its own change point, and there are two types of analysis. Although the data are very sparse in each domain, our first analysis is on each individual domain. The second analysis is on all domains simultaneously, permitting us to pool the data across the domains.

The National Health Interview Survey (NHIS) is an important source of information on the health of the U.S. population. One of the variables in NHIS is activity limitation status (ALS). Activity limitation among adults due to chronic conditions is a major health problem in United States. It may influence the quality of an individual's life, and it can cause socio-economic problems. Box plots of the proportions of adults with ALS by age show that activity limitation status changes some where between ages 40 and 70. The main issue we want to address in this study is to find the age that a significant change in ALS occurs.

Zimmer and House (2003) and Burchardt (2003) indicate that characteristics such as race, gender, and socioeconomic status (e.g., education and income) influence the probability of having functional disorders. Education is determined early in life and influence psychosocial mechanisms throughout life. It is highly correlated with income. However, education is more strongly predictive of onset of functional health problem such as activity limitation, while income is more predictive of course or progression. Activity limitation in our study is a

chronic health problem of adulthood, and is the outcome of a long process of development as a function of exposure to a wide range of social, psychological, and biomedical risk factors. Education indexes both the socio-economic position of individuals early in adulthood and a stock of human capital available to them from that time on. All these influence long-term patterns of exposure to and experience of major psychosocial and biomedical risk factors that cause activity limitation. Income, on the other hand, is usually measured in years, and thus reflects socioeconomic position and resources closer to the time of assessment of activity limitation. It is more strongly related to the resources available for the treatment or management of the health problem. Therefore, income more relates to the severity of the problem while education more strongly relates to the onset or existence of the health problem. So in our study, we also include variables like race, gender, and education.

Loss of activities of daily living will lead to disability. With activity limitation, an individual's life may be influenced both physically and psychologically (see Burchardt 2003). Because of the disability, they may encounter changes in paid and unpaid work, and therefore, have lower income and less benefit. On the other hand, they will need more help in their daily living and the cost of living will increase. These will result in higher risk of poverty. While they are having all these economical difficulties, they may also have social problems. Disability may prevent people from enjoying normal social life and make them socially excluded. The main issue in this study is activity limitation, which is different from disability. However, knowing what impact of disability would make the onset of activity limitation meaningful. Doctors can give their patients an indication about when they are near the onset of activity limitation, therefore patients can prevent activity limitation or be prepared for the consequence of it.

We have envisioned fitting two horizontal lines to the proportions of adults with ALS at each age by domain. The first line starts at 30 years and goes to a age k, unknown, and the other line some what higher and it goes from k to 80 years. We want to find the best pair of horizontal lines over all possible values of $k = 40, \ldots, 70$. That is, the selected k is the

value that best explains the data. Thus, our objective is to age k at which there is a change in prevalence in ALS. This problem falls in the area of change point models.

A sequence of random variables, $\underset{\sim}{x} = (x_1, \ldots, x_n)$ is said to have a single change-point at k if $x_1, \ldots, x_k \mid k, \theta_1 \overset{iid}{\sim} p_1(x \mid \theta_1)$ and $x_{k+1}, \ldots, x_n \mid k, \theta_2 \overset{iid}{\sim} p_2(x \mid \theta_2)$, where $p_1(x \mid \theta_1)$ and $p_2(x \mid \theta_2)$ are unknown probability 'density' functions in the same parametric family. The problem of estimating the location of the change-point k has been extensively studied for the past several decades. Smith (1975) proposed a Bayesian approach to make inference about the change-point based on the posterior probabilities of the possible change-points. In this model the joint distribution of $x_1, \ldots, x_n$ conditional on θ_1, θ_2 and the change-point k $(1 \leq k \leq n)$ is given by $p(x_1, \ldots, x_n \mid k, \theta_1, \theta_2) = \prod_{i=1}^{k} p_1(x_i \mid \theta_1) \prod_{i=k+1}^{n} p_2(x_i \mid \theta_2)$. With θ_1 and θ_2 unknown, assuming a prior distribution of k over the set of possible change points $p_0(k)$, $\sum_{k=1}^{n} p_0(k) = 1$, and a prior density of θ_1, θ_2, $p_0(\theta_1, \theta_2)$, Smith (1975) obtained

$$\begin{aligned} p(k \mid \underset{\sim}{x}) &\propto p(x_1, \ldots, x_n \mid k) p_0(k), \; k = 1, \ldots, n, \\ p(\theta_1, \theta_2 \mid k, \underset{\sim}{x}) &\propto p(x_1, \ldots, x_n \mid k, \theta_1, \theta_2) p_0(\theta_1, \theta_2), \end{aligned}$$

where $p(x_1, \ldots, x_n \mid k) = \int_{\theta_1} \int_{\theta_2} p(x_1, \ldots, x_n \mid k, \theta_1, \theta_2) p_0(\theta_1, \theta_2) d\theta_1 d\theta_2$. Smith (1975) made inference about k by computing the posterior modes of k, θ_1 and θ_2. Carlin, Gelfand and Smith (1992) explore the change-point problem using hierarchical Bayesian models further via the Gibbs sampler. Girón, Ginebra and Riba (2005) has an application of this approach, where they use a Bayesian multinomial change-point analysis to determine the authorship of a book.

Given the proportion of individuals in the population with positive ALS, we assume that the number of individuals with positive ALS at each age group has a binomial probability mass function. The proportions across age are different, and have the same beta distribution up to the change point (unknown), and the proportions after the change point have a different beta distribution. This model applies to each domain separately. Because the data are very sparse, the posterior probability of a change is roughly uniform over a large range of age,

making the results unreliable. In the second analysis, we use the Gibbs sampler to fit the joint posterior distribution of the twelve change points. This is a difficult problem because the joint density requires the numerical computation of a triple integral at each iteration.

In Section 2 we describe the data on ALS in the NHIS. In Section 3 we show how to use a Bayesian model to estimate the change point for each domain separately. For many of the domains, the change points are very different. Thus, in Section 4 we attempt to "borrow strength" across the domains, using a procedure similar to the one in Section 3. However, this procedure is slightly more complex, because we need to integrate out all parameters to form a Gibbs sampler of the change points. Section 5 has a discussion.

2. Main Features of the Data

The National Health Interview Survey (NHIS) has been conducted every year since 1957 by the National Center for Health Statistics (NCHS) to measure an aspect of health status of the U.S. population (Adams and Marano 1995). Through this sample survey, NCHS conducts surveys on chronic and acute conditions, doctor visits, hospital episodes, disability, household and personal information, and other special aspects of health of the U.S. population.

The questionnaire is divided up into two major sections, core and supplemental. The core section includes items on household and personal information, basic health questions on conditions, doctor visit, hospital discharge and other supplemental information. The supplemental section includes questions about selected interests from the general public, encompassing a wide range of topics such as prescription medicine, hypertension, diabetes, high blood pressure, and HIV. The core section is administered annually and the supplemental section is administered as its need arises.

The data we used for this study comes from the 1995 National Health Interview Survey. The interviewed sample was has 41,824 households containing 102,467 persons. The range of age is from 0 to 99. Since we are only interested in the age at which a significant change in activity limitation occurs among adults with chronic conditions, we only use the data from

those whose age is from 30 to 80. Respondents in NHIS were asked to provide their activity limitation status during the interview; there are nonrespondents, but these are taken care of using survey weights.

The degree of activity limitation is divided in four categories: 1. Unable to perform major activity. 2. Limited in kind/amount of major activity. 3. Limited in other activities. 4. Not limited (includes unknowns). Here, major activity refers to activities like going to work, going to school, keeping house, etc. We have recoded the four levels of activity limitation status (ALS) into two levels (i.e., binary data) in two scenarios. In the first scenario, we let the binary variable be 1 if an adult is in the first category, and 0 otherwise; we will call this variable CAT1. In the second scenario, we let the binary variable be 1 if an adult is in either the first or the second category, and 0 otherwise; we will call this variable CAT2. We can also create a third binary variable which is 1 if an adult is in either the first, second or third category, and 0 otherwise; but this is similar to these two scenarios.

There are two levels of race (white and nonwhite). Gender also have two levels (male and female). In the original data, there are seven levels of education, and we have recoded it into three levels (pre-college, college and post-college). Each combination of race, gender, and education is considered as a domain, therefore, the dataset is divided into twelve domains and one can expect different change point for each domain. All data combined is considered as the 13^{th} domain. We will call these domains $D1, D2, \ldots, D13$ (e.g., $D1$: white males with pre-college education, $D2$: white females with pre-college education etc.)

We have looked at the boxplots of the proportions of positive ALS for the 12 domains by age for both scenarios, and we notice that there is a sudden drop of proportions of ALS around age 70. We believe that this is because many people with positive ALS are of poor health, and they passed away around age 70, therefore, the proportion of positive ALS drops. We will correct this effect using the sample weights.

Since the NHIS uses a multistage sample designed to sample the population of the United States, it is necessary to utilize a individual's basic weight for accurate analysis of the data.

The weight for each sampled individual is the product of four components.

(a) **Probability of selection.** The basic weight for each person is obtained by multiplying the reciprocals of the probabilities of selection at each step in the design: PSU, segment, and household.

(b) **Household nonresponse adjustment within segment.** In the NHIS, interviews are completed in about 94 percent of all eligible households. Because of household nonresponse, a weighting adjustment is required. The nonresponse adjustment weight is a ratio of the within-segment weighted number of sample households divided by the within-segment weighted number of actually interviewed households, both numbers exclusive of households with unknown black/Hispanic status. For segments with non-responding households of unknown black/Hispanic status, the previously mentioned factor was multiplied by the ratio of the number of segment households divided by the number of known status households. This adjustment reduces bias in an estimate to the extent that persons in the noninterviewed households have the same characteristics as the persons in the interviewed households in the same segment.

(c) **First-stage ratio adjustment.** The weight for persons in the nonself-representing PSU's is ratio adjusted to the 1990 population within four race-residence classes of the nonself-representing strata within each geographic region.

(d) **Poststratification by age-sex-race-ethnicity.** Within each of 88 age-sex-race-ethnicity cells, a weight is constructed each quarter to ratio adjust the first-stage population estimate based on the NHIS to an independent estimate of the population of each cell. These independent estimates are prepared by the U.S. Bureau of the Census and are updated quarterly.

In addition to the design and ratio adjustments included in the person basic weight, the person weight is further modified depending on the variable selected, the length of the recall

period, and the period of time for which the estimate is to be made. For a review of weighting methods, see Kalton and Florence-Cervantes (2003).

Let $\tilde{y}_{ij\ell}$ denote a binary variable for the ℓ^{th} adult, j years old, in the i^{th} domain; $\tilde{y}_{ij\ell} = 1$ if the individual has ALS, and $y_{ij\ell} = 0$ if the individual does not. Also, let $w_{ij\ell}$ denote the corresponding sample weights. Then, we construct a weighted proportion of individuals with ALS, $\hat{p}_{ij} = \sum_{\ell \in D_{ij}} w_{ij\ell}\tilde{y}_{ij\ell} / \sum_{\ell \in D_{ij}} w_{ij\ell}, \quad i = 1, \ldots, 12, \quad j = 30, \ldots, 80$, where D_{ij} are domains formed by race, sex, and education. We assume that the only variable discriminating among the individuals within a domain is age. Letting $\underset{\sim}{X}_{ij}$ denote a vector that contains an intercept and the value of sex, race, and education for each domain and each age, Next, we fit a logistic regression to the $\hat{p}_{ij}$, $\log[\hat{p}_{ij}/\{1-\hat{p}_{ij}\}] = \underset{\sim}{X}'_{ij}\underset{\sim}{\beta} + \varepsilon_{ij}$, where approximately the ε_{ij} are independent with $E(\varepsilon_{ij}) = 0$ and $Var(\varepsilon_{ij}) = \sum_{l \in D_{ij}} w_{ijl}^2 / \hat{p}_{ij}(1-\hat{p}_{ij}) \left(\sum_{l \in D_{ij}} w_{ijl}\right)^2 = 1/W_{ij}$. Thus, the least square estimator of $\underset{\sim}{\beta}$ is

$$\hat{\underset{\sim}{\beta}} = \left\{ \sum_i \sum_j W_{ij} \underset{\sim}{X}_{ij} \underset{\sim}{X}'_{ij} \right\}^{-1} \sum_i \sum_j W_{ij} \log\left[\frac{\hat{p}_{ij}}{1-\hat{p}_{ij}}\right] \underset{\sim}{X}_{ij}.$$

In the case in which $\hat{p}_{ij} = 0$ or 1, we add $1/2n_{ij}$ to both $\hat{p}_{ij}$ and $1-\hat{p}_{ij}$; see Cox (1970) for a similar adjustment for the empirical logistic transform. Finally, the main variable we analyze is the weighted binomial counts for each domain by age,

$$y_{ij} = [n_{ij} e^{\underset{\sim}{X}'_{ij}\hat{\underset{\sim}{\beta}}} / \{1 + e^{\underset{\sim}{X}'_{ij}\hat{\underset{\sim}{\beta}}}\}], \tag{1}$$

where [ ] is the nearest integer, and $y_{ij} = 0, 1, \ldots, n_{ij}$.

In Table 1 we present in the first column the domains sex (male, female), race (white, non-white), and education (low, medium, high) (e.g., domain 1 is white male with low education). The second column is the total number of respondents. Thus, in the second and fifth columns we have the number individuals with ALS among the respondents, and in the third and sixth columns we have the weighted binomial counts. We have drawn boxplots for the proportions of adults with positive ALS using the binomial weighted data, and we notice that the plots are smoother than those using the unweighted data and the sudden

drop around age 70 is eliminated; see Figure 1. Henceforth, we study the weighted binomial counts (1) for the two scenarios, CAT1 and CAT2, and the twelve domains of education, sex and race.

To get a preliminary understanding of the change point for each domain, we have used the expectation maximization (EM) algorithm; see Dempster, Laird and Rubin (1977). Letting $\hat{p}_j = y_{ij}/n_{ij}$ for domain i (note we drop the subscript i), we consider the simple model for each domain

$$\begin{array}{ll} \hat{p}_j \overset{iid}{\sim} \text{Normal}(\mu_1, \frac{\hat{p}_j(1-\hat{p}_j)}{n_j}) & j = 30, \ldots, k \\ \hat{p}_j \overset{iid}{\sim} \text{Normal}(\mu_2, \frac{\hat{p}_j(1-\hat{p}_j)}{n_j}) & j = k+1, \ldots, 80, \end{array} \tag{2}$$

where $\hat{p}_j$ is the proportion of adults with positive ALS who are j years old. Before the change point, the $\hat{p}_j$ are independently and identically normally distributed with mean μ_1 and variance $\hat{p}_j(1-\hat{p}_j)/n_j$. After the change point, the $\hat{p}_j$ are independently and identically normally distributed with mean μ_2 and variance $\hat{p}_j(1-\hat{p}_j)/n_j$.

$\hat{p}_j$'s follow the same distributions as before, except that the mean is μ_2. For the maximization step,

$$\hat{\mu}_1 = \sum_{j=30}^{k} n_j \hat{p}_j / \sum_{j=30}^{k} n_j, \quad \hat{\mu}_2 = \sum_{j=k+1}^{80} n_j \hat{p}_j / \sum_{j=k+1}^{80} n_j \tag{3}$$

and the expectation step

$$E(k \mid \hat{\mu}_1, \hat{\mu}_2, \hat{p}) = \sum_{k=40}^{70} k P(k \mid \hat{\mu}_1, \hat{\mu}_2, \hat{p}) \tag{4}$$

where

$$P(k \mid \hat{\mu}_1, \hat{\mu}_2, \hat{p}) \propto \exp\left[-\frac{1}{2} \left\{ \sum_{j=30}^{k} \frac{n_j(\hat{p}_j - \hat{\mu}_1)^2}{\hat{p}_j(1-\hat{p}_j)} + \sum_{j=k+1}^{80} \frac{n_j(\hat{p}_j - \hat{\mu}_2)^2}{\hat{p}_j(1-\hat{p}_j)} \right\} \right],$$

$k = 40, \ldots, 70$. The EM algorithm proceeds by iterating between (3) and (4). Note that we have the distribution of k as well.

In Table 1 we present the most plausible age where a significant change of prevalence occurs, and its associated probability by domain in the fifth and eight columns. Note that the probabilities are almost uniform over the 12 domains, with those corresponding to categories 1, 2 larger. Thus, these estimates for the age of change of prevalence are unreliable. Note

also that when all the domains are put together (Domain 13), the change point is much more certain at age 56 or 57.

3. Data Analysis for Each Domain Separately

In this section, we consider each domain separately. We describe a hierarchical Bayesian model, and we discuss inference for the change points.

3.1 Model and Posterior Density

Let y_j denote the number of adults with positive ALS who are j years old, and let n_j denote the total number of adults that are j years old in each domain. We assume that

$$y_j \mid \theta_j \overset{ind}{\sim} \text{Binomial}(n_j, \theta_j), \quad j = 30, \ldots, 80, \tag{5}$$

We assume that the prior distributions for all θ_j are conjugate $Beta$ distributions, but those before the onset and those after the onset have different parameters. Thus,

$$\begin{array}{ll} \theta_j \mid \mu_1, \tau \overset{iid}{\sim} \text{Beta}(\mu_1\tau, (1-\mu_1)\tau) & j = 30, \ldots, k \\ \theta_j \mid \mu_2, \tau \overset{iid}{\sim} \text{Beta}(\mu_2\tau, (1-\mu_2)\tau) & j = k+1, \ldots, 80. \end{array} \tag{6}$$

Note that for a parsimonious representation we have taken the same τ before and after the change point.

We believe that the change point occurs between the ages 40 and 70, so the range for k is from 40 to 70. A uniform prior distribution is assumed on k, such that

$$Pr(k = a_r) = 1/31, \quad r = 40, \ldots, 70, \quad a_r = r. \tag{7}$$

Finally, non-informative priors for the hyperparameters μ_1, μ_2 and τ,

$$\mu_1, \mu_2 \overset{iid}{\sim} \text{Uniform}(0,1), \quad p(\tau) = \frac{1}{(1+\tau)^2}, \qquad \tau \geq 0. \tag{8}$$

Note that μ_1, μ_2 and τ are independent.

Then, using Bayes' theorem, the joint posterior density of θ, μ, τ, and k is

$$P(\underset{\sim}{\theta}, \underset{\sim}{\mu}, \tau, k = a_r | \underset{\sim}{y}) \propto \frac{1}{(1+\tau)^2} \prod_{j=30}^{80} \theta_j^{y_j}(1-\theta_j)^{n_j - y_j} \times \prod_{j=30}^{a_r} \frac{\theta_j^{\mu_1\tau-1}(1-\theta_j)^{(1-\mu_1)\tau-1}}{B(\mu_1\tau,(1-\mu_1)\tau)} \prod_{j=a_r+1}^{80} \frac{\theta_j^{\mu_2\tau-1}(1-\theta_j)^{(1-\mu_2)\tau-1}}{B(\mu_2\tau,(1-\mu_2)\tau)}, \quad (9)$$

$r = 40, \ldots, 70$. It is convenient to collapse over θ and to condition on k. Thus, we can obtain the joint posterior of $\underset{\sim}{\mu}$ and τ given k as

$$P(\mu_1, \mu_2, \tau \mid \underset{\sim}{y}, k = a_r) \propto \frac{1}{(1+\tau)^2} \prod_{j=30}^{a_r} \left[\frac{B\left(y_j + \mu_1\tau, (n_j - y_j) + (1-\mu_1)\tau\right)}{B\left(\mu_1\tau, (1-\mu_1)\tau\right)} \right] \times \prod_{j=a_r+1}^{80} \left[\frac{B\left(y_j + \mu_2\tau, (n_j - y_j) + (1-\mu_2)\tau\right)}{B\left(\mu_2\tau, (1-\mu_2)\tau\right)} \right]. \quad (10)$$

We use the griddy Gibbs sampler to draw samples from (10).

The posterior distribution of k is

$$P(k = a_r \mid \underset{\sim}{y}) = \frac{P(k = a_r)P(\underset{\sim}{y} \mid k = a_r)}{\sum_{s=40}^{70} P(k = a_s)P(\underset{\sim}{y} \mid k = a_s)} = \frac{P(\underset{\sim}{y} \mid k = a_r)}{\sum_{s=40}^{70} P(\underset{\sim}{y} \mid k = a_s)}, \quad (11)$$

$r = 40, \ldots, 70$, where

$$P(\underset{\sim}{y} \mid k = a_r) = \prod_{j=30}^{80} \binom{n_j}{y_j} \int_{\underset{\sim}{\mu}} \int_{\tau} \frac{1}{(1+\tau)^2} \times \prod_{j=30}^{a_r} \left[\frac{B\left(y_j + \mu_1\tau, (n_j - y_j) + (1-\mu_1)\tau\right)}{B\left(\mu_1\tau, (1-\mu_1)\tau\right)} \right] \times \prod_{j=a_r+1}^{80} \left[\frac{B\left(y_j + \mu_2\tau, (n_j - y_j) + (1-\mu_2)\tau\right)}{B\left(\mu_2\tau, (1-\mu_2)\tau\right)} \right] d\underset{\sim}{\mu} d\tau. \quad (12)$$

We use Monte Carlo integration to perform the integration in (12). As an importance function, we take μ_1, μ_2 and τ to be independent with Beta distributions for μ_1 and μ_2 and a Gamma distribution for τ as

$$\mu_i \overset{iid}{\sim} \text{Beta}(\nu_i\phi_i, (1-\nu_i)\phi_i) \quad i = 1, 2, \quad \tau \sim \text{Gamma}(\alpha, \beta).$$

We obtain the parameters $(\nu_i,\ \phi_i)$ and $(\alpha,\ \beta)$ using the method of moments on the output of the Gibbs sampler $\underset{\sim}{\mu}^{(h)},\ \tau^{(h)},\ h = 1, \ldots, M \approx 1000$. Thus, we take

$$\nu_i = \frac{1}{M}\sum_{h=1}^{M} \mu_i^{(h)}, \qquad \phi_i = \nu_i(1-\nu_i)/\{\frac{1}{M-1}\sum_{h=1}^{M}(\mu_i^{(h)} - \frac{1}{M}\sum_{h=1}^{M}\mu_i^{(h)})^2\},$$

$$\alpha = \beta \frac{1}{M} \sum_{h=1}^{M} \tau^{(h)}, \quad \beta = \frac{1}{M} \sum_{h=1}^{M} \tau^{(h)} / \{ \frac{1}{M-1} \sum_{h=1}^{M} (\tau^{(h)} - \frac{1}{M} \sum_{h=1}^{M} \tau^{(h)})^2 \}.$$

3.2 Data Analysis

In Tables 2 and 3 we present the distributions of k for the weighted binomial counts by domain. For both scenarios it is difficult to tell where the change points are for the domains because the probabilities are small, and they spread out almost uniformly. With a few exceptions, the estimates are similar to those obtained for the EM algorithm on the approximate model; see Table 1. These exceptions correspond to domains D8, D10, D12, which correspond respectively to non-white females with college education, and females with post-college education; these are most sparse domains.

When only one category is used to define positive ALS, there are differences in change points for domains D8, D10, D12. For D8 the change point is at 61 years, probability .055, compared with 53 years, probability .062, for the EM algorithm; for D10 the change point is at 70 years, probability .058, compared with 52 years, probability .071, for the EM algorithm; and for D12 the change point is at 69 years, probability .054, compared with 57 years, probability .054, for the EM algorithm.

When two categories are used to define positive ALS, there are differences in change points for domains D10, D12. For D10 the change point is at 70 years, probability .048, compared with 41 years, probability .045, for the EM algorithm; and for D12 the change point is at 70 years, probability .054, compared with 54 years, probability .051, for the EM algorithm.

However, when the data are collapsed over education, sex and race (domain D13), one can have some confident in the change point. When only one category is used to define positive ALS, the change point lies in the interval (53, 61) with probability .951. When two categories are used to define positive ALS, the change point lies in the interval (53, 60) with probability .972. These results are similar to those obtained for the EM algorithm on the

approximate model; see Table 1. Thus, in both scenarios, there is reasonable confidence that a change in prevalence of ALS occurs in the interval (55, 60).

Finally, we note that the computation is time consuming. One way to reduce the time considerably is to use the reversible jump algorithm. When there are many competing models with different parameter spaces (e.g., dimensions) and there is uncertainty about the model itself creating a parameter which indexes the model, Green (1995) proposes a reversible Markov chain sampler that jumps between parameter subspaces of different dimensionality. We have experienced some difficulties using this algorithm. The estimates of the change points are mostly larger than those in Tables 2, 3 with a few of domains similar. Also for domain D13, when one category is used to define positive ALS, the change point lies in (52, 60) with probability .959; and when two categories are used to define positive ALS, the change point lies in (52, 60) with probability .957. These are similar to the results without using the reversible jump sampler.

However, there are some domains where the reversible jump algorithm gets stuck. When only one category is used to define positive ALS, the change points are at age 70 years with probability 1 for domains D10 and D12. When two categories are used to define positive ALS, the change points are at age 70 years for domain D10 and D12 with probabilities 1 and .695 respectively. However, when one category is used to define positive ALS, for domain D13 the change point lies in (52, 60) with probability .959; and when two categories are used to define positive ALS, for domain D13 the change point lies in (52, 60) with probability .957. These are similar to the results without using the reversible jump algorithm.

Can these conclusions change for the individual domains when the data are pooled? Here, we mean that each domain is treated in its own right, but there is a common stochastic process across these domains. We note that there is a problem in collapsing over education, sex and race as in domain D13; this ignores the heterogeneity in the data. Thus, while the answers for D13 seem reasonable, this heterogeneity in the data has not been taken into account. Next, we want to take the heterogeneity into account; this analysis falls in the area

of small domain estimation.

4. Adaptive Pooling of the Domains: Small Domain Analyses

In this section we model heterogeneity among the twelve domains. We assume that the parameters of the Binomial distribution follow a common stochastic process. This allows us to pool the domains adaptively (i.e., according to the sample size). This comes naturally under small area estimation. There are two parts in this section. In the first part, we assume that all domains have the same change point. Although this is an unrealistic assumption, it is used to facilitate the second part which eliminates this assumption to have different change points for the domains, a more realistic approach.

4.1 A Single Change Point for all Domains

As in the model in Section 2, the y_{ij} follow a Binomial distribution with parameters n_{ij} and θ_{ij}, and the θ_{ij} have Beta distributions, but the parameters are different for those before the change point and those after the change point. That is, for $i = 1, \ldots, 12$,

$$y_{ij}|\theta_{ij} \overset{ind}{\sim} \text{Binomial}(n_{ij}, \theta_{ij}) \qquad j = 30, \ldots, 80 \tag{13}$$

$$\theta_{ij} \mid \mu_1, \tau \overset{iid}{\sim} \text{Beta}(\mu_1\tau, (1-\mu_1)\tau) \qquad j = 30, \ldots, k$$

$$\theta_{ij} \mid \mu_2, \tau \overset{iid}{\sim} \text{Beta}(\mu_2\tau, (1-\mu_2)\tau) \qquad j = k+1, \ldots, 80. \tag{14}$$

A priori we take μ_1, μ_2 and τ to be independent with

$$\mu_1, \mu_2 \overset{iid}{\sim} \text{Uniform}(0,1), \quad p(\tau) = \frac{1}{(1+\tau)^2}. \tag{15}$$

Also, the prior distribution for k is uniform on $(40, 70)$,

$$Pr(k) = 1/31, \qquad k = 40, \ldots, 70. \tag{16}$$

Then the joint posterior distribution for all the parameters is

$$p(\underset{\sim}{\theta}, \mu_1, \mu_2, \tau, k|\underset{\sim}{y}) \propto \frac{1}{(1+\tau)^2} \prod_{i=1}^{12} \prod_{j=30}^{80} \left\{ \theta_{ij}^{y_{ij}} (1-\theta_{ij})^{n_{ij}-y_{ij}} \right\}$$

$$\times \prod_{i=1}^{12} \left\{ \prod_{j=30}^{k} \frac{\theta_{ij}^{\mu_1\tau-1}(1-\theta_{ij})^{(1-\mu_1)\tau-1}}{B(\mu_1\tau,(1-\mu_1)\tau)} \prod_{j=k+1}^{80} \frac{\theta_{ij}^{\mu_2\tau-1}(1-\theta_{ij})^{(1-\mu_2)\tau-1}}{B(\mu_2\tau,(1-\mu_2)\tau)} \right\}. \tag{17}$$

As in Section 2, we have collapsed over the θ_{ij} to fit this posterior density.

Finally, the posterior density of k is

$$Pr(k \mid y) = \frac{Pr(k=a_r)P(y \mid k=a_r)}{\sum_{s=40}^{70} P(k=a_s)Pr(y \mid k=a_s)} = \frac{Pr(y \mid k=a_r)}{\sum_{a_s=40}^{70} Pr(y \mid k=a_s)}, \tag{18}$$

$a_r = 40, \ldots, 70$, where

$$p(y \mid k) = \frac{1}{31} \prod_{i=1}^{12} \prod_{j=30}^{80} \binom{n_{ij}}{y_{ij}} \int_0^\infty \int_0^1 \int_0^1 \frac{1}{(1+\tau)^2} \times \prod_{i=1}^{12} \left\{ \prod_{j=30}^{k} \frac{B(y_{ij}+\mu_1\tau, n_{ij}-y_{ij}+(1-\mu_1)\tau)}{B(\mu_1\tau,(1-\mu_1)\tau)} \right.$$

$$\left. \times \prod_{j=k+1}^{80} \frac{B(y_{ij}+\mu_2\tau, n_{ij}-y_{ij}+(1-\mu_2)\tau)}{B(\mu_2\tau,(1-\mu_2)\tau)} \right\} d\mu_1 d\mu_2 d\tau. \tag{19}$$

We present the distributions of k in Table 4. Age 57 is the same chage point for both scenarios. When one category is used, the probability that change point occurs at 57 years is .237, but when both categories are used the probability is much larger, .547; otherwise the two distributions of k are very similar. For both scenarios most of the probability lies in the interval (53, 59). It gives some confidence that these results are also very similar to those obtained for domain D13 in Section 3. We extend this model to make inference about the individual domains in the next section.

4.2 Different Change Points for Each Domain

Now, we assume that each domain has a different change point. Let $\underset{\sim}{k} = (k_1, k_2, \ldots, k_{12})$ denote the vector of change points, where $k_i, i = 1, \ldots, 12$ denote the change point for the i^{th} domain and $\underset{\sim}{k}_{(i)}$ denotes the vector of change points for all the domains except the i^{th} domain. We assume that the k_i are identical and independently distributed.

Given $\underset{\sim}{k}$, the joint posterior distribution for $\underset{\sim}{\theta}$, μ_1, μ_2, and τ is

$$p(\underset{\sim}{\theta}, \mu_1, \mu_2, \tau | y, \underset{\sim}{k}) \propto \frac{1}{(1+\tau)^2} \prod_{i=1}^{12} \left\{ \prod_{j=30}^{80} \theta_{ij}^{y_{ij}} (1-\theta_{ij})^{n_{ij}-y_{ij}} \right.$$

$$\times \prod_{j=30}^{k_i} \left[\frac{\theta_{ij}^{\mu_1\tau-1}(1-\theta_{ij})^{(1-\mu_1)\tau-1}}{B(\mu_1\tau,(1-\mu_1)\tau)} \right] \prod_{j=k_i+1}^{80} \left[\frac{\theta_{ij}^{\mu_2\tau-1}(1-\theta_{ij})^{(1-\mu_2)\tau-1}}{B(\mu_2\tau,(1-\mu_2)\tau)} \right] \Bigg\}, \tag{20}$$

and collapsing over $\underset{\sim}{\theta}$, we get

$$\begin{aligned} p(\mu_1,\mu_2,\tau|\underset{\sim}{y},\underset{\sim}{k}) \quad \propto \quad & \frac{1}{(1+\tau)^2} \prod_{i=1}^{12} \{ \prod_{j=30}^{k_i} \left[\frac{B(y_{ij}+\mu_1\tau, n_{ij}-y_{ij}+(1-\mu_1)\tau)}{B(\mu_1\tau,(1-\mu_1)\tau)} \right] \\ & \times \prod_{j=k_i+1}^{80} \left[\frac{B(y_{ij}+\mu_2\tau, n_{ij}-y_{ij}+(1-\mu_2)\tau)}{B(\mu_2\tau,(1-\mu_2)\tau)} \right] \Bigg\}. \end{aligned} \tag{21}$$

Thus, again we use the griddy Gibbs sampler to fit this posterior density for each value of $\underset{\sim}{k}$.

Now we need to compute the posterior distribution for $\underset{\sim}{k}$. Unfortunately, this is a difficult problem, because our previous procedure is impractical as the computation is prohibitively expensive. This is true because we have to, a) run 31^{12} Gibbs samplers, b) compute 31^{12} marginal likelihoods, and c) as we have seen in Section 3 the reversible jump algorithm is unreliable for very small domains. To solve this problem, we can draw $p(\underset{\sim}{k}|\underset{\sim}{y})$ using a Gibbs sampler by drawing from $p(k_i|\underset{\sim}{k}_{(i)},\underset{\sim}{y})$. Here,

$$\begin{aligned} p(k_i|\underset{\sim}{y},\underset{\sim}{k}_{(i)}) \quad \propto \quad & \int_0^\infty \int_0^1 \int_0^1 \frac{1}{(1+\tau)^2} \prod_{i=1}^{12} \Bigg\{ \prod_{j=30}^{k_i} \frac{B(y_{ij}+\mu_1\tau, n_{ij}-y_{ij}+(1-\mu_1)\tau)}{B(\mu_1\tau,(1-\mu_1)\tau)} \\ & \times \prod_{j=k_i+1}^{80} \frac{B(y_{ij}+\mu_2\tau, n_{ij}-y_{ij}+(1-\mu_2)\tau)}{B(\mu_2\tau,(1-\mu_2)\tau)} \Bigg\} d\mu_1 d\mu_2 d\tau. \end{aligned} \tag{22}$$

A more convenient form for computation of the posterior density in (22) is

$$\begin{aligned} p(k_i|\underset{\sim}{y},\underset{\sim}{k}_{(i)}) \propto & \int_0^\infty \int_0^1 \int_0^1 \frac{1}{(1+\tau)^2} \prod_{j=30}^{k_i} \frac{B(y_{ij}+\mu_1\tau, n_{ij}-y_{ij}+(1-\mu_1)\tau)}{B(\mu_1\tau,(1-\mu_1)\tau)} \\ & \times \prod_{j=k_\ell+1}^{80} \frac{B(y_{ij}+\mu_2\tau, n_{ij}-y_{ij}+(1-\mu_2)\tau)}{B(\mu_2\tau,(1-\mu_2)\tau)} \\ & \times \prod_{\ell\neq i} \Bigg\{ \prod_{j=30}^{k_\ell} \frac{B(y_{\ell j}+\mu_1\tau, n_{\ell j}-y_{\ell j}+(1-\mu_1)\tau)}{B(\mu_1\tau,(1-\mu_1)\tau)} \\ & \prod_{j=k_\ell+1}^{80} \frac{B(y_{\ell j}+\mu_2\tau, n_{\ell j}-y_{\ell j}+(1-\mu_2)\tau)}{B(\mu_2\tau,(1-\mu_2)\tau)} \Bigg\} d\mu_1 d\mu_2 d\tau. \end{aligned} \tag{23}$$

Note then that in (23) one needs to compute a three-dimensional integral.

Frist, set $\underset{\sim}{k}_{(1)}$ to the starting value which we obtained from the model in Section 4.1. Using the importance function, obtained in Section 4.1, we can compute $p(k_1|\underset{\sim}{k}_{(1)}, y)$ by Monte Carlo integration and Gibbs sampler for each $k_1 = 40, \ldots, 70$. Once we get this conditional posterior 'density' of $k_1|\underset{\sim}{k}_{(1)}, y$, we can draw a random value from it and fix the k_1 in $\underset{\sim}{k}_{(2)}$ at this value. So in $\underset{\sim}{k}_{(2)}$, k_1 will be the sample drawn from $p(k_1|\underset{\sim}{k}_{(1)}, y)$ and $k_3, \ldots, k_{12}$ will still be the starting value we obtained from part 1. Now for the fixed k_1 and $\underset{\sim}{k}_{(1)}$, perform the Gibbs sample on $p(\underset{\sim}{\theta}, \mu_1, \mu_2, \tau|k_1, \underset{\sim}{k}_{(1)})$ at each value of k_2. At this point, k_2 is the starting value as we obtained in part 1. Obtain $p(k_2|\underset{\sim}{k}_{(2)}, y)$ as in the previous step, and draw a value for $k_2|\underset{\sim}{k}_{(2)}, y$. Now we have updated k_1, k_2. We continue the process to update $k_3, k_4, \ldots, k_{12}$ in the same manner. Repeat the entire process until we get a large sample $\underset{\sim}{k}^{(1)}, \ldots, \underset{\sim}{k}^{(M)}$. This is a sample from the posterior density of $p(\underset{\sim}{k} \mid y)$. Now we can construct 95% credible interval for each component $k_1, \ldots, k_{12}$.

In Tables 5,6 we see that the probabilities at the change points are much more elevated. Domains D1, D2, D3, D6 and D8 have change points near to 40 years; Domains D4, D5 and D7 have their change points somewhere in the middle of the age range; and Domains D9, D10, D11 and D12 have their change points near 70 years. This is true for both scenarios, CAT1 and CAT2. This suggests that the change points roughly increases with education levels (i.e., adults with higher level of education tend to have a change in prevalence of ALS much later than less educated adults).

Thus, we have collapsed over sex and race to produce one categorical variable education with three levels (pre-college, college, post-college). We present the results in Table 7. Now, for CAT1 (CAT2) we see that adults with pre-college education have change of prevalence around age 51 (55) years, adults with college education have change of prevalence around age 43 (49) years, and adults with post-college education have change of prevalence around age 60 (54) years. The results for the two scenarios are somewhat different. What intervals have a probability of .95 by education? First consider CAT1. For adults with pre-college education it is (43, 57); for adults with college education it is (40, 56); and for adults with

post-college education it is (54, 68). Second consider CAT2. For adults with pre-college education it is (47, 61); for adults with college education it is (42, 61); and for adults with post-college education it is (47, 63). Thus, when variation is taken into account, there appears to be very little difference among the three levels of education.

5. Discussion

Our main contribution is in a hierarchical Bayesian model which includes all domains with separate change points. The Gibbs sampler is used to fit the joint posterior distribution of the twelve change points. This is a difficult problem because the joint posterior density requires the numerical computation of a triple integral at each iteration. This model gives better estimates of the change points (more confident in the estimates) than a model that that fits each domain separately. This model that is fit separately to each domain does not take care of the similarity among the domains (too much heterogeneity).

The two scenarios (different definitions of positive ALS) do not make a large difference in terms of detection of the age at which there is a change in prevalence. There are differences among the domains formed by crossing education, sex and race, but it is difficult to trust these results. However. we have found that overall the age at which a significant change occurs is about 56 years. This is similar to the answers when all the domains are treated as a single one (i.e., education, race and sex are collapsed).

Although we did not present the results, the other parameters of the process are obtained using data augmentation by a Metropolis sampler and a Rao-Blackwellization. Another piece of work we did is that we also fit all our models with the unweighted binomial counts. While the results are different, they are not too different.

One caveat to our work is that the weighted binomial counts are sums of correlated binary variables. Thus, these counts do not really have binomial distributions. For simplicity, survey samplers incorporate sampling weights using weighted averages, as we did; this is defective. In future, we will look at this problem on correlated binomial counts.

References

Adams, P.F. and Marano, M.A. (1995), "Current Estimates from the National Health Interview Survey," *National Center for Health Statistics. Hyattsville, MD*, Vital Health Statistics, Series 10, 193.

Burchardt, T. (2003), "Being and becoming: Social exclusion and the onset of disability," *London, Case Report: Centre for Analysis of Social Exclusion, London School of Economics.*

Carlin, B.P. and Gelfand, A.E. and Smith, A.F,M. (1992), "Hierarchical Bayesian analysis of change point problems," Applied Statistics, 41, 389-405.

Cox, D.R. (1970), *Analysis of Binary Data*, Chapman and Hall.

Dempster, A. and Laird, N. and Rubin, D. (1977), "Maximum likelihood from incomplete data via EM algorithm," *Journal of the Royal Statistical Society*, Series B, 39, 1-38.

Girón, J. and Ginebra, J. and Riba, A. (2005), "Bayesian analysis of a multinomial sequence and homogeneity of literary," *The American Statistician*, 59, 19-30.

Green, P.J. (1995), "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," Biometrika, 82, 711-732.

Kalton, K. and Flores-Cervantes, I. (2003), "Weighting methods," *Journal of Official Statistics*, 19, 81-97.

Smith, A.F.M. (1975), "A Bayesian approach to inference about a change-point in a sequence of random variables," *Biometrika*, 63, 407-416.

Zimmer, Z. and House, J. S. (2003), "Education, income and functional limitation trasitions among American adults: Contrasting onset and progression," *International Journal of Epidemiology*, 32, 1089-1097.

Table 1: Number of respondents (res), number of individuals with ALS, weighted binomial counts (WBC), change point (CP) and its probability in parentheses by category (cat) and domain (D)

		CAT1			CAT2		
Domain	Res	ALS	WBC	CP	ALS	WBC	CP
D1	1361	242	257	51 (.104)	329	360	59 (.154)
D2	263	76	67	63 (.088)	96	82	58 (.093)
D3	1503	206	235	59 (.140)	366	398	54 (.135)
D4	322	67	72	56 (.084)	119	108	59 (.102)
D5	5814	494	559	53 (.156)	854	894	51 (.201)
D6	1129	162	167	51 (.069)	229	230	50 (.079)
D7	7023	450	577	54 (.135)	999	1117	54 (.190)
D8	1538	163	195	53 (.062)	304	316	52 (.085)
D9	6012	203	277	48 (.107)	463	522	51 (.162)
D10	934	48	63	52 (.071)	89	99	41 (.045)
D11	5969	236	211	59 (.072)	519	489	52 (.152)
D12	1236	57	61	57 (.054)	116	128	54 (.051)
D13	33104	2404	2741	56 (.446)	4483	4473	57 (.794)

NOTE: CAT1: "Limited in major activity"; CAT2 : "Limited in major activity" and "Limited in kind/amount of major activity" are added together. The domains are formed by crossing education (low,medium, high), sex (male, female) and race (white, non-white). The change points are obtained by an approximate model, and the EM algorithm is used to fit it. Domain D13 is formed by collapsing over education, sex and race.

Table 2: Distributions for the change point k for the 13 domains using the logistic weighted data, where only "Limited in major activity" is considered as positive ALS.

Age	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13
40	.008	.027	.001	.002	.001	.032	.000	.011	.015	.029	.014	.023	.000
41	.010	.021	.001	.002	.002	.030	.001	.013	.018	.027	.017	.023	.000
42	.011	.040	.002	.005	.003	.028	.001	.018	.021	.028	.019	.021	.000
43	.013	**.082**	.003	.011	.005	.032	.002	.022	.026	.030	.020	.020	.000
44	.018	.038	.004	.008	.008	.034	.005	.025	.031	.028	.023	.021	.000
45	.020	.060	.006	.011	.013	.029	.008	.026	.040	.025	.028	.021	.000
46	.025	.043	.008	.025	.018	.033	.012	.033	.048	.026	.029	.020	.000
47	.029	.027	.013	.038	.021	.034	.018	.037	.055	.024	.034	.020	.000
48	.029	.042	.019	.055	.031	.034	.023	.036	**.061**	.024	.038	.021	.000
49	.036	.034	.028	.110	.035	.036	.028	.036	.056	.023	.035	.021	.002
50	.051	.029	.027	.069	.051	.042	.033	.040	.052	.025	.033	.022	.005
51	**.073**	.036	.023	.046	.071	**.049**	.036	.042	.053	.027	.036	.022	.012
52	.065	.031	.032	.049	.076	.043	.045	.040	.049	.030	.035	.023	.024
53	.056	.026	.046	.077	.091	.043	.065	.043	.051	.028	.032	.024	.053
54	.053	.034	.058	.068	.081	.044	.079	.043	.043	.026	.036	.025	.080
55	.050	.047	.057	**.135**	**.094**	.038	.077	.043	.036	.026	.033	.026	.079
56	.064	.044	.062	.122	.076	.045	.084	.036	.038	.026	.040	.028	.125
57	.061	.051	.059	.057	.063	.041	.093	.036	.040	.027	.043	.031	.152
58	.061	.039	.066	.036	.050	.034	**.106**	.038	.041	.028	.050	.035	.135
59	.057	.037	.090	.029	.041	.037	.098	.047	.040	.028	**.062**	.042	**.234**
60	.054	.045	**.091**	.015	.047	.026	.079	.046	.028	.030	.051	.038	.067
61	.038	.031	.078	.009	.034	.028	.053	**.055**	.029	.032	.044	.047	.026
62	.028	.032	.054	.005	.027	.029	.026	.051	.028	.036	.032	.043	.005
63	.026	.033	.041	.003	.024	.022	.013	.041	.021	.041	.023	.040	.001
64	.025	.022	.036	.003	.015	.024	.007	.042	.018	.039	.027	.054	.000
65	.015	.012	.024	.002	.010	.024	.004	.031	.016	.039	.029	.050	.000
66	.008	.010	.023	.002	.006	.022	.003	.027	.012	.041	.028	.046	.000
67	.006	.008	.019	.002	.003	.023	.001	.017	.009	.048	.029	.043	.000
68	.005	.009	.014	.001	.001	.023	.001	.011	.009	.049	.026	.043	.000
69	.003	.006	.011	.001	.001	.021	.000	.007	.008	.053	.026	**.055**	.000
70	.002	.005	.005	.001	.000	.019	.000	.007	.006	**.058**	.028	.053	.000

NOTE: See note to table 1.

Table 3: Distributions for the change point k for the 13 domains using the logistic weighted data, where both "Limited in major activity" and "Limited in kind/amount of major activity" are considered as positive ALS.

Age	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13
40	.001	.007	.001	.009	.000	.018	.000	.015	.004	.037	.004	.034	.000
41	.002	.007	.001	.007	.000	.022	.000	.021	.006	.040	.006	.031	.000
42	.003	.015	.002	.007	.001	.028	.000	.024	.009	.036	.009	.032	.000
43	.005	.034	.003	.007	.002	.027	.000	.028	.013	.033	.011	.032	.000
44	.007	.027	.006	.010	.005	.025	.001	.037	.019	.030	.017	.034	.000
45	.010	.046	.009	.011	.010	.027	.002	.043	.024	.029	.021	.035	.000
46	.015	.043	.013	.012	.016	.031	.004	.052	.030	.031	.028	.033	.000
47	.016	.040	.020	.020	.023	.040	.008	.050	.040	.030	.036	.031	.000
48	.023	.068	.025	.025	.033	.044	.016	.049	.042	.029	.045	.027	.000
49	.025	.068	.032	.022	.043	.045	.029	.051	.041	.028	.047	.027	.000
50	.025	**.068**	.035	.025	.065	.054	.049	.048	.050	.030	.060	.027	.001
51	.034	.045	.056	.028	.096	.049	.066	.048	.062	.029	.065	.026	.003
52	.037	.047	.059	.023	**.098**	.049	.079	.055	.062	.032	**.073**	.028	.008
53	.053	.055	.065	.039	.096	.040	.104	**.057**	.062	.028	.056	.027	.023
54	.069	.041	**.086**	.058	.093	.044	.111	.045	.060	.025	.051	.029	.057
55	.091	.032	.076	.053	.081	.042	.102	.047	.057	.023	.055	.028	.060
56	.075	.036	.081	.083	.081	.051	**.131**	.044	.059	.025	.051	.028	.142
57	.087	.052	.066	**.093**	.071	**.065**	.096	.040	.067	.028	.054	.028	**.304**
58	.087	.055	.061	.061	.055	.050	.071	.051	**.069**	.026	.044	.027	.168
59	**.113**	.034	.060	.075	.044	.038	.052	.050	.059	.028	.043	.028	.187
60	.073	.024	.045	.052	.036	.032	.032	.036	.042	.030	.046	.029	.031
61	.055	.022	.045	.065	.020	.033	.021	.032	.036	.033	.035	.030	.013
62	.032	.015	.043	.050	.012	.031	.013	.021	.029	.036	.037	.031	.002
63	.021	.019	.035	.032	.009	.023	.006	.015	.020	.034	.029	.032	.000
64	.019	.018	.034	.035	.005	.024	.004	.013	.012	.034	.025	.034	.000
65	.011	.014	.018	.031	.002	.016	.002	.010	.008	.034	.016	.036	.000
66	.006	.014	.011	.022	.001	.012	.001	.008	.006	.034	.013	.037	.000
67	.003	.015	.006	.019	.000	.012	.000	.004	.005	.036	.008	.039	.000
68	.002	.020	.003	.014	.000	.011	.000	.003	.004	.040	.007	.040	.000
69	.001	.011	.002	.007	.000	.010	.000	.001	.002	.043	.005	.045	.000
70	.001	.010	.001	.007	.000	.008	.000	.001	.001	**.048**	.004	**.054**	.000

NOTE: See note to Table 1.

Table 4: Distributions for the change point k for the two scenarios, CAT1 and CAT2, using weighted binomial counts with the same change point for all domains

k	CAT1	CAT2
50	.001	.000
51	.004	.000
52	.012	.003
53	.049	.013
54	.086	.046
55	.095	.040
56	.217	.155
57	**.237**	**.547**
58	.098	.102
59	.181	.091
60	.017	.003
61	.003	.001

NOTE: CAT1: Only "Limited in major activity" is considered as positive ALS; CAT2: Both "Limited in major activity" and "Limited in kind/amount of major activity" are considered as positive ALS.

Table 5: Distributions for the change point k for the 12 domains using the revised Bayesian hierarchical model and weighted binomial counts for the first scenario, CAT1.

Age	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
40	**.579**	.170	**.252**	.067	0	**.644**	0	**.288**	0	0	0	0
41	.269	.131	.183	.064	0	.222	0	.174	0	0	0	0
42	.107	.178	.164	.070	.002	.082	0	.195	0	0	0	0
43	.026	**.223**	.112	.083	.005	.029	0	.147	0	.002	0	0
44	.008	.086	.084	.063	.007	.020	0	.079	0	0	0	0
45	.004	.084	.072	.066	.011	.001	0	.048	0	0	0	0
46	.002	.054	.032	.084	.032	0	0	.029	0	0	0	0
47	.003	.014	.037	**.091**	.050	0	0	.021	0	0	0	0
48	0	.025	.026	.079	.071	0	0	.014	0	0	0	0
49	.001	.015	.025	.086	.092	.002	0	.003	0	0	0	0
50	.001	.008	.007	.057	.110	0	0	.001	0	.004	0	0
51	0	.006	.002	.048	.160	0	0	0	0	.006	0	0
52	0	.004	.001	.031	0.12	0	.001	.001	0	.012	0	0
53	0	.001	.002	.026	**.125**	0	.003	0	0	.016	0	0
54	0	.001	0	.032	.077	0	.004	0	0	.012	0	0
55	0	0	.001	.028	.068	0	.010	0	0	.005	0	0
56	0	0	0	.018	.033	0	.025	0	0	.009	0	.001
57	0	0	0	.005	.020	0	.049	0	0	.011	0	0
58	0	0	0	.002	.005	0	.102	0	0	.011	0	.001
59	0	0	0	0	.009	0	.133	0	0	.010	0	.002
60	0	0	0	0	.001	0	.132	0	0	.019	0	.007
61	0	0	0	0	.002	0	**.147**	0	0	.017	0	.008
62	0	0	0	0	0	0	.132	0	0	.017	0	.010
63	0	0	0	0	0	0	.119	0	0	.018	0	.019
64	0	0	0	0	0	0	.057	0	0	.039	0	.012
65	0	0	0	0	0	0	.032	0	0	.057	0	.036
66	0	0	0	0	0	0	.034	0	.008	.102	.003	.059
67	0	0	0	0	0	0	.010	0	.020	.147	.015	.126
68	0	0	0	0	0	0	.006	0	.083	.141	.044	.215
69	0	0	0	0	0	0	.004	0	.223	**.179**	.225	.195
70	0	0	0	0	0	0	0	0	**.666**	.166	**.713**	**.309**

Table 6: Distributions for the change point k for the 12 domains using the revised Bayesian hierarchical model and weighted binomial counts for the second scenario, CAT2.

Age	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
40	**.432**	.076	**.566**	**.318**	0	**.357**	0	**.426**	0	0	0	0
41	.306	.051	.212	.182	0	.270	0	.262	0	0	0	0
42	.163	.110	.103	.119	0	.216	0	.152	0	0	0	0
43	.060	**.183**	.055	.076	0	.090	0	.084	0	0	0	0
44	.015	.100	.036	.083	0	.036	0	.038	0	0	0	0
45	.010	.129	.017	.044	.001	.011	0	.022	0	0	0	0
46	.007	.105	.006	.038	.004	.011	0	.013	0	0	0	0
47	.002	.061	.002	.048	.007	.006	.001	.002	0	0	0	0
48	.003	.068	.002	.034	.013	0	.005	.001	0	0	0	0
49	.001	.050	.001	.019	.031	.001	.013	0	0	0	0	0
50	.001	.028	0	.010	.065	0	.039	0	0	.002	0	0
51	0	.018	0	.008	.108	0	.089	0	0	0	0	0
52	0	.010	0	.003	.140	.002	.109	0	0	0	0	0
53	0	.007	0	.008	**.153**	0	.152	0	0	0	0	0
54	0	0.002	0	.006	.128	0	.137	0	0	0	0	0
55	0	.001	0	.001	.105	0	.140	0	0	0	0	.001
56	0	.001	0	.003	.108	0	**.161**	0	0	.003	0	.001
57	0	0	0	0	.068	0	.076	0	0	.010	0	.006
58	0	0	0	0	.037	0	.045	0	0	.007	0	.005
59	0	0	0	0	.016	0	.020	0	0	.015	0	.006
60	0	0	0	0	.013	0	.009	0	0	.022	0	.007
61	0	0	0	0	.002	0	.004	0	0	.029	0	.010
62	0	0	0	0	.001	0	0	0	0	.065	0	.022
63	0	0	0	0	0	0	0	0	0	.076	0	.032
64	0	0	0	0	0	0	0	0	0	.082	0	.046
65	0	0	0	0	0	0	0	0	0	.072	0	.128
66	0	0	0	0	0	0	0	0	.002	.070	.003	.132
67	0	0	0	0	0	0	0	0	.024	.066	.005	**.171**
68	0	0	0	0	0	0	0	0	.091	.104	.042	**.171**
69	0	0	0	0	0	0	0	0	.255	.171	.204	0.132
70	0	0	0	0	0	0	0	0	**.628**	**.206**	**.746**	0.130

Table 7: Distributions for the change point k for the three domains (levels of education) using the revised Bayesian hierarchical model and with weighted binomial counts.

Age	CAT1			CAT2		
	Low	Median	High	Low	Median	High
40	.010	.069	0	.001	0.180	0
41	.020	.061	0	0	.023	0
42	.020	.069	0	.002	.024	.003
43	.029	**.138**	0	.004	.034	0
44	.042	.084	0	.005	.045	.005
45	.055	.104	0	.013	.046	.010
46	.058	.072	0	.013	.056	.013
47	.069	.048	0	.026	.059	.030
48	.066	.061	0	.036	**.082**	.032
49	.066	.056	.002	.040	**.082**	.042
50	.090	.036	.002	.058	.065	.059
51	**.109**	.043	.002	.055	.059	.082
52	.088	.024	.012	.059	.051	.090
53	.064	.016	.013	.082	.065	.096
54	.048	.029	.029	.091	.047	**.119**
55	.050	.015	.027	**.094**	.040	.090
56	.036	.026	.038	.087	.036	.094
57	.029	.023	.060	.082	.045	.052
58	.024	.010	.073	.089	.036	.046
59	.009	.002	.106	.082	.023	.046
60	.009	.009	**.140**	.050	.021	.023
61	.005	.003	.128	.013	.017	.023
62	.003	.001	.087	.010	.009	.014
63	0	.001	.068	.006	.004	.016
64	0	0	.047	.001	.004	.010
65	0	0	0.049	0	.005	.003
66	0	0	.037	0	.003	0
67	0	0	.027	0	0	0
68	.001	0	.022	0	.001	.001
69	0	0	.023	.001	0	0
70	0	0	.008	0	0	0

NOTE: CAT1: Only "Limited in major activity" is considered as positive ALS. CAT2: Both "Limited in major activity" and "Limited in kind/amount of major activity" are considered as positive ALS.

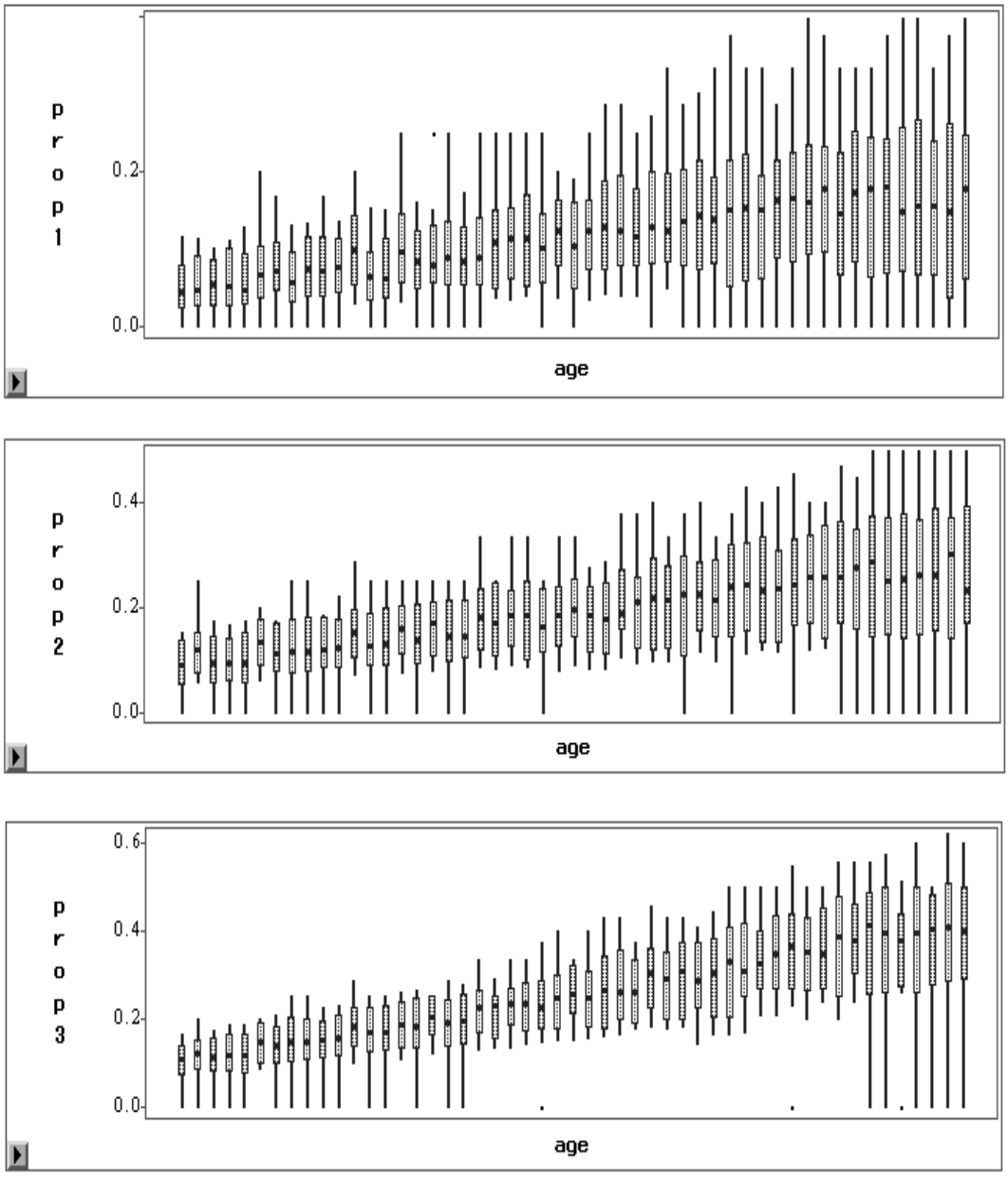

Figure 1: Boxplots of the proportions of positive ALS for the 12 domains by age using the weighted binomial counts; **Top panel**: Unable to perform major activity; **Middle panel**: Unable to perform major activity and Limited in major activity/amount of major activity; **Bottom panel**: Unable to perform major activity, Limited in major activity/amount of major activity and Limited in other activities