Published in final edited form as:

J Stat Theory Pract. 2012 December ; 6(4): 725–744. doi:10.1080/15598608.2012.719805.

Analysis of Partially Incomplete Tables of Breast Cancer Characteristics with an Ordinal Variable

B. Nebiyou Bekele,

Gilead Sciences, Inc., USA

Luis E. Nieto-Barajas, and

Department of Statistics, ITAM, Mexico

Mark F. Munsell

The University of Texas M. D. Anderson Cancer Center, USA

B. Nebiyou Bekele: neby.bekele@gilead.com; Luis E. Nieto-Barajas: lnieto@itam.mx; Mark F. Munsell: mfmunsell@mdanderson.org

Abstract

Our goal is to model the joint distribution of a series of $4 \times 2 \times 2 \times 2$ contingency tables for which some of the data are partially collapsed (i.e., aggregated in as few as two dimensions). More specifically, the joint distribution of 4 clinical characteristics in breast cancer patients is estimated. These characteristics include estrogen receptor status (positive/negative), nodal involvement (positive/negative), HER2-neu expression (positive/negative), and stage of disease (I, II, III, IV). The joint distribution of the first three characteristics is estimated conditional on stage of disease and we propose a dynamic model for the conditional probabilities that let them evolve as the stage of disease progresses. The dynamic model is based on a series of Dirichlet distributions whose parameters are related by a Markov prior structure (called dynamic Dirichlet prior). This model makes use of information across disease stage (known as “borrowing strength”) and provides a way of estimating the distribution of patients with particular tumor characteristics. In addition, since some of the data sources are aggregated, a data augmentation technique is proposed to carry out a meta-analysis of the different datasets.

Keywords

Bayesian methods; breast cancer; CISNET; data augmentation; dynamic model; Markov Dirichlet sequence; meta-analysis

1 Introduction

In this article, our goal is to model the joint distribution of a series of 4 dimensional ($4 \times 2 \times 2 \times 2$) contingency tables in which some of the data are partially collapsed (i.e., aggregated in as few as 2 dimensions). More specifically, we estimate the joint distribution of 4 clinical characteristics known to be important in determining breast cancer outcome in cancer patients. These characteristics include estrogen receptor (ER) status (positive/negative), nodal involvement (positive/negative) and human epidermal growth factor receptor (HER2-neu) expression (positive/negative) as a function of stage of disease (I, II, III, IV). In particular, parameters that characterize patient characteristics conditional on the h th stage

are assumed to be dependent on model parameters conditional on the $(h - 1)$ th stage by means of a Markov structure generalizing Nieto-Barajas and Walker (2002) to the multivariate setting. By doing this we propose an approach capable of modeling trends, using information from the conditional probabilities across the different ordered stages.

Since these clinical characteristics are associated with breast cancer survival, National Cancer Institute policy makers (for example, Feuer, et al., 2006) are interested in characterizing the distribution of these disease markers as an aid in modeling future trends in breast cancer mortality. We should emphasize that in this research our goal is to help policy makers characterize the joint distribution of these four characteristics and not to predict survival from these characteristics.

An interesting aspect of this current research is that we will be making inferences based on four datasets with different levels of disaggregation. That is, three of the four datasets have been collapsed in one or two of the characteristics, and only one dataset will provide full information on the four characteristics. For example one dataset includes information only on disease stage and ER status, but not on the other two characteristics. To address this complication, we use a data augmentation technique similar to that presented by Tanner (1991), together with a fixed effect meta-analysis to combine the information from the different sources. The datasets represent a snap-shot of patient characteristics at diagnosis, therefore longitudinal data characterizing the natural history of the disease at the level of the individual patients are not available.

Probability models for contingency tables goes back to the log-linear models with Poisson likelihoods (e.g. Agresti, 2007) where the logarithm of the probability/intensity of each cell is modeled via row, column and cell random effects producing what is called the saturated model. This model makes use of lots of parameters. On the other hand, to propose a more parsimonious model, we concentrate on the vectors of conditional probabilities and impose a dependence structure to share information across rows (stages). Other approaches to impose dependence in a series of conditional probabilities are also available. For example, Veeramachaneni et al. (2005) proposed a hierarchical model that induces exchangeability in the conditional probability vectors. This approach, however, induces a symmetric dependence in any pair of vectors, ignoring the ordinal nature of the conditioning variable stage. An alternative approach, that does consider the ordering in the conditional probability vectors, is the multinomial logistic regression model (Agresti, 2007) with an appropriate parameterization and using stage as explanatory variable.

The outline of this paper is as follows. In Section 2 we describe the motivating example which gives rise to our data analysis problem. In Section 3 we describe the probability model for contingency tables and define the Markov Dirichlet sequence (a dynamic model based on a series of Dirichlet models whose parameters are related by a Markov prior structure), which is used as the prior distribution. Posterior characterization of the process is also derived. We address the computational aspects related to the implementation of our model in Section 4, and include a comparative analysis with alternative models using the M. D. Anderson dataset. In Section 5 we discuss our meta-analysis approach arising from having different sources of information at different levels of disaggregation. We carry out

the analysis of the four available datasets and assess the performance of the fixed effects meta-analysis via a simulation study. Finally Section 6 concludes with a discussion.

2 Motivating example

The National Cancer Institute (NCI) is interested in assessing the impact of cancer-control interventions on the incidence and risk of death from cancer in the general population. With this end in mind, the NCI has sponsored a group of investigators who use statistical modeling and simulation to increase understanding of the influence of treatment and preventative interventions (e.g., preventive administration of tamoxifen, mammography screening, chemotherapy treatment) on future population trends in the incidence and mortality of breast cancer. This group, known as the Breast Cancer Intervention and Surveillance Modeling Network (CISNET), has developed models to forecast trends and to potentially identify effective breast cancer control strategies (Berry et al., 2005; Feuer et al., 2006).

Part of the process of developing these models includes gathering and summarizing model inputs from multiple sources. One of the most important inputs used to forecast future breast cancer outcomes is the distribution of disease characteristics, including ER status, nodal involvement, HER2-neu status, and disease stage. Since these factors have been associated with survival, estimating the distribution of these characteristics among patients is an important first step in forecasting breast cancer mortality. Breast cancers that are ER positive tend to have better outcomes and are sensitive to drugs that specifically target estrogen receptors. For example, two drug classes that function in this manner are selective estrogen receptor modulators (e.g. tamoxifen) and aromatase inhibitors. Nodal involvement is a measure of disease severity and determines how threatening the cancer might be. HER2-neu expression is a measure of the amount of HER2 protein present on the surface of cancer cells. A high HER2-neu expression is indicative of an aggressive form of the disease. Cancers that express HER2-neu are categorized as HER2 positive.

Breast cancer specialists use the breast cancer stage to describe the extent of disease observed in a patient. The stage at diagnosis has been shown to be associated with disease recurrence and overall survival. The stage usually takes on integer values between 1 and 4, with 1 signifying the least invasive disease (disease confined locally to the breast) and 4 signifying the most invasive disease (metastatic disease). Stage is determined by assessing the size of a tumor, how deep it has penetrated, whether it has invaded adjacent organs, and whether it has spread to distant organs.

Obtaining relevant data that categorize patients on the basis of the above clinico-pathologic characteristics can be a challenge. The main U.S. repository for this type of data is the Surveillance, Epidemiology, and End Results (SEER) database maintained by the NCI. This is a publicly available database (<http://seer.cancer.gov>) setup primarily to track the incidence of cancer and survival after cancer diagnosis. To this end, SEER data are collected from cancer registries that cover over 20% of the U.S. population. While these data provide information on stage, ER status, and nodal involvement for most breast cancer patients in the registry, HER2-neu status has only recently been identified as associated with long term

breast cancer survival and, thus, has not been collected by SEER. Therefore, the data from SEER are collapsed across the HER2-neu positive/negative categories. While all patients have stage information available, an additional complicating factor is that a significant proportion of the data in the SEER registries provides information on either ER status or nodal status, but not both. Because of this, we supplemented the SEER data with a dataset of patients from The University of Texas M. D. Anderson Cancer Center (M. D. Anderson) that summarizes patients based on all of the desired variables: stage, ER status, nodal involvement, and HER2-neu status. The M. D. Anderson data was collected between 1997 and 2005. While the SEER datasets include data collected between 1973-2003, for the purposes of this analysis we only included data from patients diagnosed between 1990-2003 to ensure similar disease detection patterns. In summary, we will be using data from four sources: SEER data with information on stage, ER status, and nodal involvement; SEER data with information on stage and ER status; SEER data with information on stage and nodal involvement; and lastly M. D. Anderson data with information for all 4 characteristics.

3 Estimating proportions

3.1 Basic modeling

Consider the following three categorical variables with two categories each: ER status (0-negative, 1-positive), nodal status (0-negative, 1-positive), and HER2-neu status (0-negative, 1-positive). Let stage be a fourth categorical variable with 4 levels, say 1 to 4. Let $Y_{h,i,j,k}$ be the random variable that denotes the number of cancer patients in stage h , ER status i , nodal status j , and HER2-neu status k , with $h \in \{1, 2, 3, 4\}$, $i \in \{0, 1\}$, $j \in \{0, 1\}$, and $k \in \{0, 1\}$, respectively.

To reduce the notational burden associated with keeping up with four indices, we re-group the 3 dichotomous variables ER status, nodal status, and HER2-neu status into a single categorical variable with $8 = 2^3$ categories. We define $r = 1 + (4 \times i) + (2 \times j) + (1 \times k)$ for $i, j, k \in \{0, 1\}$, and will use $Y_{h,r}$ to denote $Y_{h,i,j,k}$ whenever possible. The index r (which we call the profile index) denotes a specific combination of characteristics with $r \in \{1, 2, \dots, 8\}$. Similarly, let $\theta_{h,r}$ be the cell probability associated with stage h and profile r .

It is natural to assume that the vector of outcomes in all category combinations, $\mathbf{Y} = \{Y_{h,r}\}$, follows a multinomial distribution with parameters N , the total number of patients, and cell probabilities $\boldsymbol{\theta} = \{\theta_{h,r}\}$. In notation, $\mathbf{Y} | \boldsymbol{\theta} \sim \text{Mult}(N, \boldsymbol{\theta})$ such that

$$\sum_{h,r} \theta_{h,r} = 1 \text{ and } \sum_{h,r} Y_{h,r} = N \text{ with probability one (w.p.1).} \quad (3.1)$$

In a Bayesian context the simplest setting is to assign a conjugate prior to $\boldsymbol{\theta}$, which in this case follows a Dirichlet distribution. If *a-priori* $\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{a})$, where $\boldsymbol{a} = \{a_{h,r}\}$ and $a_{h,r} > 0$, then it is well known (Bernardo and Smith, 2000) that *a-posteriori* $\boldsymbol{\theta} | \mathbf{y} \sim \text{Dir}(\boldsymbol{a} + \mathbf{y})$. Throughout this paper, both $\text{Mult}(\cdot, \cdot)$ and $\text{Dir}(\cdot)$ denote multivariate distributions in a saturated form according to conditions (3.1). Thus, posterior estimates for $\theta_{h,r}$ based on a quadratic loss function are

$$\hat{\theta}_{h,r} = \frac{\alpha_{h,r} + y_{h,r}}{a + N},$$

where $a = \sum_{h,r} \alpha_{h,r}$. If $\alpha_{h,r} \approx 0$, then $\hat{\theta}_{h,r}$ coincides with the maximum likelihood estimator.

In order to borrow strength across stage categories when estimating the parameter vector θ , we first note that the probability model for $\mathbf{Y}$ can be re-written in terms of conditional distributions. Assuming that $H = 4$ and $R = 8$, we let $\mathbf{Y}_* = (Y_{1*}, \dots, Y_{H*})$, where

$$Y_{h*} = \sum_{r=1}^R Y_{h,r} \text{ denotes the number of patients in stage } h, \text{ and we let } \theta_* = (\theta_{1*}, \dots, \theta_{H*}),$$

where $\theta_{h*} = \sum_{r=1}^R \theta_{h,r}$ is the marginal probability of being in stage h , $h = 1, \dots, H$.

Additionally, let $\mathbf{Y}_h = (Y_{h,1}, \dots, Y_{h,R})$ be the vector of different profiles at a given stage h . It is straightforward to show that the joint, distribution for the whole vector of profiles $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_H)$ can be expressed as

$$f(\mathbf{y}|\theta) = \text{Mult}_{H \times R}(\mathbf{y}|N, \theta), = \text{Mult}_H(\mathbf{y}_*|N, \theta_*) \prod_{h=1}^H \text{Mult}_R(\mathbf{y}_h|y_{h*}, \boldsymbol{\eta}_h), \quad (3.2)$$

where $\boldsymbol{\eta}_h = (\eta_{h,1}, \dots, \eta_{h,R})$, and

$$\eta_{h,r} = \frac{\theta_{h,r}}{\theta_{h*}}, \quad (3.3)$$

is the conditional probability of having profile r given that the patient is in stage h for $r = 1, \dots, R$ and $h = 1, \dots, H$.

The model originally parameterized by θ has been re-parametrized by $(\theta_*, \boldsymbol{\eta})$. This re-parameterization of the model allows us to make inference on θ_* and $\boldsymbol{\eta}$ separately given the form of the likelihood (3.2). The idea is to define a prior on the sequence of conditional probability vectors $(\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_H)$ that allows to share information across h . For this purpose, a natural approach is to propose an exchangeable prior as the one considered by Veeramachaneni et al. (2005). This model is specified hierarchically as $\boldsymbol{\eta}_h|\boldsymbol{\beta} \sim \text{Dir}(\boldsymbol{\beta})$ and $\boldsymbol{\beta} \sim f(\boldsymbol{\beta})$ where $f(\boldsymbol{\beta})$ is a hyper-prior distribution on $\boldsymbol{\beta} = (\beta_1, \dots, \beta_R)$, say $\beta_r \sim \text{Ga}(b_0, b_1)$ independently for $r = 1, \dots, R$. However, this hierarchical prior induces a symmetric dependence between any pair $(\boldsymbol{\eta}_h, \boldsymbol{\eta}_\ell)$. Since index h denotes the stage status, it might be desirable that the pair $(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ had a stronger dependence than say $(\boldsymbol{\eta}_1, \boldsymbol{\eta}_4)$.

To incorporate the ordinal nature of variable stage into the model, we could alternatively use a multinomial logistic regression model (Agresti, 2007) with a polynomial trend in terms of stage status h as explanatory variable. In particular, the quadratic case has the form

$$\log \left(\frac{\eta_{h,r}}{1 - \sum_{s=2}^8 \eta_{h,s}} \right) = \alpha_r + \beta_r h + \gamma_r h^2,$$

for $r = 2, \dots, 8$, with prior distributions for α_r and β_r . For instance, independent normals with mean μ_0 and precision τ_0 .

3.2 Dynamic modeling

Since the stage variable is in fact an ordinal variable, therefore we consider a natural evolution in the conditional probabilities of disease η_h across stages $h = 1, \dots, H$. This assumption is plausible since these conditional probabilities follow a smooth trend (see Figure 5.2). Taking advantage of this trend, we propose to borrow strength by considering a dynamic prior for the vectors of conditional probabilities $\eta_1, \dots, \eta_H$.

We extend the Markov beta sequence (i.e., a series of beta models whose parameters are linked via a Markov prior structure) proposed by Nieto-Barajas and Walker (2002), and develop a Markov: Dirichlet sequence for $\{\eta_h\}$ that is defined through a latent process of multivariate vectors $\{\mathbf{u}_h\}$ as follows: $\eta_1 \xrightarrow{\mathbf{u}_1} \eta_2 \xrightarrow{\mathbf{u}_2} \dots$ such that

$$\begin{aligned} \eta_1 &\sim \text{Dir}(\beta_1), \\ \mathbf{u}_h | \eta_h &\sim \text{Mult}(c_h, \eta_h), \\ \eta_{h+1} | \mathbf{u}_h &\sim \text{Dir}(\beta_{h+1} + \mathbf{u}_h), \end{aligned} \quad (3.4)$$

for $h = 1, \dots, H-1$, where the vectors of hyper-parameters $\beta_h = (\beta_{h,1}, \dots, \beta_{h,R})$ are such that $\beta_{h,r} > 0$ and c_h is a non-negative integer for $h = 1, \dots, H-1$. The multinomial conditional distribution of the, latent variables $\mathbf{u}_h$ is a convenient choice justified by conjugacy in a Bayesian updating, however other choices could also be used. The role of $\mathbf{u}_h$ is to induce dependence in the pair (η_h, η_{h+1}) .

In particular, if the parameters β_h and c_h are common for all h in (3.4), that is $\beta_h = \beta$ and $c_h = c$ for $h = 1, \dots, H$, then the process $\{\eta_h\}$ becomes strictly stationary with marginals $\eta_h \sim \text{Dir}(\beta)$. The parameter c_h controls the degree of dependence between the vectors η_h and η_{h+1} . In fact, in the stationary case,

$$\text{Corr}(\eta_{h,r}, \eta_{h',r}) = \left(\frac{c}{b+c} \right)^{|h-h'|},$$

where $b = \sum_{r=1}^R \beta_r$. Thus inducing a stronger/weaker dependence for closer/faraway stages h and h' .

Moreover, construction (3.4) implies a linear form for the conditional mean of η_{h+1} given η_h . In other words, it is straightforward to show that

$$E(\boldsymbol{\eta}_{h+1}|\boldsymbol{\eta}_h) = \frac{\boldsymbol{\beta}_{h+1} + c_h \boldsymbol{\eta}_h}{b + c_h}.$$

Therefore our dynamic Dirichlet model defines, on average, a prior piecewise linear evolution among $\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_h$. This is certainly a nice feature of our model which equips it with greater flexibility for modeling both linear and non-linear trends.

When the number of profiles R is equal to 2, the Markov Dirichlet sequence reduces to the Markov beta sequence of Nieto-Barajas and Walker (2002).

Under model (3.2), the likelihood for $\boldsymbol{\eta}_h$ is proportional to $\prod_{r=1}^R \eta_{h,r}^{y_{h,r}}$. The posterior distribution can thus be characterized by the full conditional distributions $[\boldsymbol{\eta} | \mathbf{u}, \mathbf{y}]$ and $[\mathbf{u} | \boldsymbol{\eta}, \mathbf{y}]$ required for the implementation of a Gibbs sampler. The vectors $\boldsymbol{\eta}_h$ are conditionally independent given $(\mathbf{u}, \mathbf{y})$, and their distribution is given by

$$f(\boldsymbol{\eta}_h | \mathbf{u}_{h-1}, \mathbf{u}_h, \mathbf{y}) = \text{Dir}(\boldsymbol{\eta}_h | \boldsymbol{\beta}_h + \mathbf{u}_{h-1} + \mathbf{u}_h + \mathbf{y}_h), \quad (3.5)$$

for $h = 1, \dots, H$ with $\mathbf{u}_0 = \mathbf{0}$ w.p.1. The vectors of latent $\mathbf{u}_h$ variables are also conditionally independent given $\boldsymbol{\eta}$. Their distribution does not depend on the data $\mathbf{y}$ and is given by

$$f(\mathbf{u}_h | \boldsymbol{\eta}_h, \boldsymbol{\eta}_{h+1}) \propto \prod_{r=1}^R \frac{(\eta_{h,r} \eta_{h+1,r})^{u_{h,r}}}{\Gamma(\beta_{h+1,r} + u_{h,r}) \Gamma(u_{h,r} + 1)} I(\mathbf{u}_h' \mathbf{1} = c_h), \quad (3.6)$$

for $h = 1, \dots, H - 1$, where $\mathbf{1}$ denotes a vector of 1's.

Regardless of the prior we choose for $\boldsymbol{\eta}$ to estimate $\boldsymbol{\theta}$ we need to provide a prior for $\boldsymbol{\theta}_*$. We propose a conjugate analysis with prior distribution $\boldsymbol{\theta}_* \sim \text{Dir}(\boldsymbol{\delta})$. By looking at the likelihood from model (3.2), the posterior is given by $\boldsymbol{\theta}_* | \mathbf{y} \sim \text{Dir}(\boldsymbol{\delta} + \mathbf{y}_*)$. We can obtain estimates of $\boldsymbol{\theta}$ by considering its relationship with $(\boldsymbol{\theta}_*, \boldsymbol{\eta})$ given by (3.3).

4 Implementation

4.1 Computational aspects

Posterior inference for $\boldsymbol{\eta}$, when using the dynamic Dirichlet prior of Section 3.2 is achieved by implementing a Gibbs sampling scheme (Smith and Roberts, 1993) with full conditional distributions (3.5) and (3.6).

Sampling from (3.5) can be done directly using standard statistical packages. However, sampling from (3.6) requires a special treatment. This latter distribution is discrete, but the multivariate constraint $\mathbf{u}_h' \mathbf{1} = c_h$ makes the computations more difficult. We can simplify the computations by sampling from the univariate conditional distributions for each $u_{h,r}$, which are given by

$$f(u_{h,r} | \mathbf{u}_{-(h,r)}, \boldsymbol{\eta}) \propto \frac{\{\eta_{h,r} \eta_{h+1,r} / (\eta_{h,R} \eta_{h+1,R})\}^{u_{h,r}} I(u_{h,r} \leq c_h - \sum_{j \neq r}^{R-1} u_{h,j})}{\Gamma(\beta_{h+1,r} + u_{h,r}) \Gamma(u_{h,r} + 1) \Gamma(\beta_{h+1,r} + c_h - \sum_{r=1}^{R-1} u_{h,r}) \Gamma(c_h - \sum_{r=1}^{R-1} u_{h,r} + 1)}.$$

Since this is a discrete univariate distribution, we can sample from it directly evaluating the right hand side and normalizing.

The parameter c_h plays an important role in the estimation of $\boldsymbol{\eta}_h$'s (and thus in the estimation of $\boldsymbol{\theta}$) because it determines the degree of information that is shared between neighbors $\boldsymbol{\eta}_h$ and $\boldsymbol{\eta}_{h+1}$. In order to let the data help us determine the best value for c_h , we assign a Poisson hyper prior distribution, say $c_h \sim \text{Po}(\lambda)$, with a large value for λ if we want to be relatively non-informative. In this case the conditional posterior distribution for c_h has the form

$$f(c_h | \mathbf{u}_h, \boldsymbol{\theta}, \mathbf{y}) \propto \frac{\Gamma(\sum_{r=1}^R \beta_{h+1,r} + c_h) (\lambda \eta_{h,R} \eta_{h+1,R})^{c_h} I(c_h \geq \sum_{r=1}^{R-1} u_{h,r})}{\Gamma(\beta_{h+1,R} + c_h - \sum_{r=1}^{R-1} u_{h,r}) \Gamma(c_h - \sum_{r=1}^{R-1} u_{h,r} + 1)}.$$

To sample from this distribution we propose a Metropolis-Hastings step (Tierney, 1994)

with a proposal distribution $c_h^* - \sum_{r=1}^{p-1} u_{h,r} \sim \text{Po}(\kappa \lambda \eta_{h,R} \eta_{h+1,R})$, with $\kappa > 0$ as an overdispersion parameter.

4.2 Comparative analysis

To assess the advantages of our model with respect to other approaches, we consider the only source of information with full specifications of the 32 categories, which is the M. D. Anderson dataset presented in Table 4.1.

Prior specifications for the models are the following. For the dynamic Dirichlet prior we took $\boldsymbol{\beta}_h = (1, \dots, 1)$ and a hyper-prior distribution $c_h \sim \text{Po}(\lambda)$ with a range of values of λ for comparison. Note that $\lambda = 0$ implies $c_h = 0$ and $\mathbf{u}_h = \mathbf{0}$ w.p.1, meaning prior independence among vectors $\boldsymbol{\eta}_h$. The overdispersion parameter for implementing the Metropolis-Hastings step when sampling from c_h was $\kappa = 10$. For the exchangeable prior we took $\beta_r \sim \text{Ga}(1, 1)$ independently for $r = 1, \dots, 8$. For the multinomial logistic regression model we consider two options for the predictor, linear and quadratic as function of stage. The priors are $\alpha_r \sim \text{N}(0, 0.001)$, $\beta_r \sim \text{N}(0, 0.001)$ and $\gamma_r \sim \text{N}(0, 0.001)$ independently for $r = 2, \dots, 8$ with $\alpha_1 = \beta_1 = \gamma_1 = 0$ for identifiability of the parameters. In all cases we considered $\boldsymbol{\theta}_*$ coming from a uniform distribution a-priori, that is, $\boldsymbol{\theta}_* \sim \text{Dir}(1, \dots, 1)$.

The dynamic model was implemented in Fortran and the exchangeable and multinomial logistic in Win-BUGS (Spiegelhalter et al., 2003). In all cases the Gibbs samplers were run for 200,000 iterations with a burn-in of 20,000, keeping one observation of every 10^{th} after burn-in to reduce the autocorrelation in our post-burn-in retained samples. The final number of generated samples used to compute the estimates were 18,000. As illustration, Figure 4.1 shows convergence diagnostics for θ_{11} when using the dynamic Dirichlet prior with $\lambda = 10$. The mixing of the chain is adequate and the ergodic means stabilize for the number of

simulations considered. The autocorrelation function shows almost null dependence after the thinning (taking one realization of every 10^{th}), and the probability histogram reiterates the good mixing by showing no lack of samples in any region.

To compare among the different models we compute two goodness of fit statistics, named log-pseudo marginal likelihood (LPML) statistic (Sinha and Dey, 1997) and the L-measure (Ibrahim and Laud, 1994). The LPML is a model selection criterion based on the conditional predictive ordinates (CPO). More explicitly, for a sample $y_1, \dots, y_n$, if we denote by $D^{(-i)}$ the data with the i^{th} observation removed, then for each i , $CPO_i = f(y_i | D^{(-i)})$ and

$LPML = \sum_{i=1}^n \log(CPO_i)$. On the other hand, the L-measure is a goodness of fit statistic based on a function of the variance and bias of the predictive distribution of each y_i , that is $L(y) = \sum_{i=1}^n \text{Var}(y_i^F | y) + \nu \sum_{i=1}^n \{E(y_i^F | y) - y_i\}^2$ where y_i^F is the predicted value for y_i and $\nu \in (0, 1)$ is a weighting term which determines a trade-off between bias and variance. The best fitting model is that with the largest LPML statistic and the smallest L-measure. Table 4.2 summarizes the values of these two statistics for the different models.

From Table 4.2 we can see that the multinomial logistic model with linear predictor provides the worst fitting according to the LPML statistic. Its predictive variance is the smallest ($\nu = 0$), however its bias is huge compared to the other models, making its L-measure for $\nu = 0.5, 1$ to be the largest. The fitting improves when considering a quadratic predictor, increasing the value of the LPML statistic and reducing the variance plus bias ($\nu = 0.5$ and 1). The exchangeable prior improves on the LPML statistic with respect to the two multinomial logistic regression models and reduces the bias (L-measure for $\nu = 0.5$ and 1). On the other hand, the best fit is attained with the dynamic Dirichlet prior with $\lambda = 10$. It has the largest LPML statistic, its predictive variance is better than the exchangeable and multinomial quadratic logistic. Moreover, its predictive variance combined with the bias (L-measure for $\nu = 0.5$ and 1) is the smallest among all models. Therefore the dynamic Dirichlet prior with $\lambda = 10$ yields the best fit for this dataset.

To have a visual idea of the estimates obtained under the different models, Figure 4.2 contains the evolution along the 4 stages of posterior point estimates (posterior means) of the conditional probabilities $\eta_{h,r}$ for the 8 profiles. It is clear from the graph that two of the point estimates follow close paths. These correspond to the dynamic model (solid line) and the hierarchical model (dashed line). Although the point estimates look similar, the overall variance and bias is a lot smaller with the dynamic model. This can not be appreciated in the graph but in the L-measures reported in Table 4.2. The dotted-dashed line in Figure 4.2 corresponds to the point estimates obtained from the multinomial logistic model. These are clearly more distinct to those obtained with the other two competing models.

Another way of comparing the performance of the dynamic model is by examining the lengths of the 95% posterior credible intervals for $\theta_{h,r}$. Table 4.3 summarizes the ratios of the lengths obtained with the dynamic Dirichlet prior with $\lambda = 10$ and those obtained with an independent prior (dynamic prior with $\lambda = 0$). These ratios range from 0.69 to 1.02, where the greatest improvement is obtained for those cells where none or few data were observed (stage 1, profiles 3, 4 and 8, and stages 3 and 4). Therefore, the dynamic model produces

more precise estimates by allowing the conditional probabilities by stage to be positive dependent and thus allowing borrowing of strength in the estimates.

5 Meta-analysis

5.1 Fixed effects meta-analysis in collapsed tables

As mentioned in the motivating example of Section 2, data that categorized patients using all four characteristics of interest were not easy to find. The ideal scenario would be to have a dataset in which cancer patients have been classified in each of the 4 stages $\times$ 8 profiles (status combinations) = 32 categories. The only dataset that satisfied this requirement was provided by M. D. Anderson. This dataset consists of a total of 1,632 post-menopausal women and is presented in Table 4.1.

In addition to the M. D. Anderson dataset, data on post-menopausal women obtained from the SEER database are also available. While a valuable source of information, the SEER datasets do not provide information regarding HER2-neu status, and for some patients information on ER status or nodal status is missing. Thus, there are three types of datasets from SEER. The first only provides information on ER and nodal status and contains 58,085 records; the second only includes nodal status and contains 12,019 records; the third only accounts for ER status and contains 8,125 records. These three datasets are summarized in Table 5.1.

Because of the difficulty in obtaining complete data sets of the kind considered in this paper, a key assumption of our model is that conditional on a profile, the incidence of breast cancer is the same in both the SEER and M. D. Anderson datasets. Thus, our modeling shares aspects of fixed effects meta-analysis, which also assumes that underlying parameters across various datasets are homogeneous. This differs from the alternative random-effects meta-analysis (Tudur Smith and Williamson, 2007) which assumes that parameters across datasets differ but are linked via a hierarchical structure. To ensure the reasonableness of this assumption we compared the prevalence of HER2-neu from the published literature to the prevalence of this biomarker in the M.D. Anderson dataset. De Laurentiis et al. (2005) found in their meta-analysis on the interaction of HER2-neu expression and response to endocrine treatment that the overall rate of HER2-neu expression was between 19 – 35% for studies included in their analyses. Choi et al. (2003) studied the associations between HER2-neu expression and various biomarkers and found the prevalence of HER2-neu expression to be approximately 16% in Caucasian Americans. Earlier research has put the HER2-neu over-expression rate in the range of 20 – 30% of human breast carcinomas (Di Fiore et al, 1987; Slamon et al 1987; Hudziak et al, 1987; Slamon et al 1989). Since the M.D. Anderson HER2-neu over-expression rate is 20% we assert that this rate is within the expected range.

To carry out our fixed effect meta-analysis we recall that the setting in Section 3 assumes that the data have been recorded for all 4×8 categories, as is the case for the M. D. Anderson dataset. In order to incorporate the three SEER datasets we must first make some adjustments. Consider the SEER 1 dataset for which the HER2 categories are not available. With the notational setting that uses indices (i, j, k) , the observations consist of $Y_{h,0,0,*}$,

$Y_{h,0,1,*}, Y_{h,1,0,*}, Y_{h,1,1,*}$, where * means that category has collapsed. Assuming the data came from model, (3.2), the induced likelihood for η_h is

$$\text{lik}(\eta_h | \mathbf{y}_h) \propto \prod_{i,j} (\eta_{h,i,j,0} + \eta_{h,i,j,1})^{y_{h,i,j,*}}.$$

If we try to use this likelihood to update the dynamic Dirichlet prior described in Section 3.2, the posterior full conditional for η_h will not have a simple expression.

To simplify the posterior analysis, we rely on a data augmentation technique similar to Tanner (1991). Let $(Z_{h,i,j,0}, Z_{h,i,j,1})$ be two latent variables such that $Z_{h,i,j,0} + Z_{h,i,j,1} = Y_{h,i,j,*}$. We can then augment the likelihood by assuming that $(Z_{h,i,j,0}, Z_{h,i,j,1})$ have been observed and obtain

$$\text{lik}(\eta_h | \mathbf{y}_h, \mathbf{z}_h) \propto \prod_{i,j} \eta_{h,i,j,0}^{z_{h,i,j,0}} \eta_{h,i,j,1}^{y_{h,i,j,*} - z_{h,i,j,0}},$$

which has the same form as the likelihood with full information, and therefore the η_h parameters can be straightforwardly updated. The latent variables $(Z_{h,i,j,0}, Z_{h,i,j,1})$ can then be simulated from their full conditional distribution,

$$(Z_{h,i,j,0}, Z_{h,i,j,1}) | Y_{h,i,j,*}, \eta_h \sim \text{Mult}_2 \left(y_{h,i,j,*}; \frac{\eta_{h,i,j,0}}{\eta_{h,i,j,0} + \eta_{h,i,j,1}}, \frac{\eta_{h,i,j,1}}{\eta_{h,i,j,0} + \eta_{h,i,j,1}} \right),$$

in accordance with our saturated notation, although $Z_{h,i,j,0} | Y_{h,i,j,*}, \eta_h$ has a binomial distribution and $Z_{h,i,j,1} = y_{h,i,j,*} - Z_{h,i,j,0}$ w.p.1. The role of the latent $Z_{h,i,j,k}$ variables is to simplify the posterior analysis by, imputing observations as part of the Markov chain Monte Carlo estimation process, and thus inform unobserved outcomes.

A data augmentation procedure analogous to the one outlined above is used to augment the likelihoods obtained from the SEER 2 and SEER 3 datasets. It is worth mentioning that the previous data augmentation technique is only possible if we have at least one dataset with information in all categories or some other source of prior information regarding the model parameters. Without a full dataset (like the M. D. Anderson dataset) or prior information it is impossible to estimate the unobserved categories since these parameters would otherwise be unidentified.

Let l for $l = 1, \dots, 4$ be the dataset indicator, where the first three correspond to the SEER 1 to 3 datasets and $l = 4$ corresponds to the M. D. Anderson dataset. Since we have assumed that, conditional on profile, the incidence of breast cancer is the same for the four datasets, then it follows that $\mathbf{Y}^{(l)}$, $l = 1, \dots, 4$ are conditionally independent, given θ , and come from the same probability model (3.2). Additionally, the information contained in the four

datasets is assumed to come from different patients. Thus the posterior analysis is carried out in the following way:

Define latent variables $\mathbf{Z}^{(l)}$ such that:

- For dataset 1,

$$z_{h,i,j,k}^{(1)} | \mathbf{y}^{(1)}, \boldsymbol{\eta} \sim \text{Mult}_2(y_{h,i,j,*}^{(1)}; \eta_{h,i,j,0}^{(1)}, \eta_{h,i,j,1}^{(1)}),$$

$$\text{where } \eta_{h,i,j,k}^{(1)} = \eta_{h,i,j,k} / \sum_k \eta_{h,i,j,k}.$$

- For dataset 2,

$$z_{h,i,j,k}^{(2)} | \mathbf{y}^{(2)}, \boldsymbol{\eta} \sim \text{Mult}_4(y_{h,*,j,*}^{(2)}; \eta_{h,0,j,0}^{(2)}, \eta_{h,0,j,1}^{(2)}, \eta_{h,1,j,0}^{(2)}, \eta_{h,1,j,1}^{(2)}),$$

$$\text{where } \eta_{h,i,j,k}^{(2)} = \eta_{h,i,j,k} / \sum_{i,k} \eta_{h,i,j,k}.$$

- For dataset 3,

$$z_{h,i,j,k}^{(3)} | \mathbf{y}^{(3)}, \boldsymbol{\eta} \sim \text{Mult}_4(y_{h,i,*,*}^{(3)}; \eta_{h,i,0,0}^{(3)}, \eta_{h,i,0,1}^{(3)}, \eta_{h,i,1,0}^{(3)}, \eta_{h,i,1,1}^{(3)}),$$

$$\text{where } \eta_{h,i,j,k}^{(3)} = \eta_{h,i,j,k} / \sum_{j,k} \eta_{h,i,j,k}.$$

- For dataset 4,

$$z_{h,i,j,k}^{(4)} = y_{h,i,j,k}^{(4)} (\text{w.p.1.})$$

Using the latent variables, the likelihood for $\boldsymbol{\eta}$ has the form

$$\text{lik}(\boldsymbol{\eta} | \mathbf{z}) = \prod_{h,i,j,k} (\eta_{h,i,j,k})^{\sum_{l=1}^4 z_{h,i,j,k}^{(l)}}.$$

Assuming the dynamic model prior for $\boldsymbol{\eta}$, the posterior full conditional distributions are:

$$f(\boldsymbol{\eta}_h | \mathbf{u}_{h-1}, \mathbf{u}_h, \mathbf{z}) = \text{Dir}_8 \left(\boldsymbol{\eta}_h | \boldsymbol{\beta}_h + \mathbf{u}_{h-1} + \mathbf{u}_h + \sum_{l=1}^4 \mathbf{z}_h^{(l)} \right),$$

where $\mathbf{z}_h^{(l)} = \{z_{h,i,j,k}^{(l)}\}$ is a vector of dimension 8 and $f(\mathbf{u}_h | \boldsymbol{\eta}_h, \boldsymbol{\eta}_{h+1})$ is given in (3.6). Finally, by assigning a conjugate prior, for $\boldsymbol{\theta}_* = (\theta_{1,*}, \dots, \theta_{4,*})$ as in Section 3, the posterior distribution using the four datasets becomes

$$\theta_* | \mathbf{z} \sim \text{Dir}_4 \left(\boldsymbol{\delta} + \sum_{l=1}^4 \mathbf{y}_*^{(l)} \right),$$

where $\mathbf{y}_*^{(l)} = (y_{1*}^{(l)}, \dots, y_{H*}^{(l)})$ and $y_{h*}^{(l)} = \sum_{i,j,k} y_{h,i,j,k}^{(l)}$ is the number of patients in stage h from study l .

5.2 Analysis of the four datasets

We now carry out a meta-analysis of the four datasets outlined in Section 2, that is, the M. D. Anderson dataset of Table 4.1 and the three SEER datasets of Table 5.1. Recall that the only source of information about HER2-neu status is the M. D. Anderson dataset, and the size of this dataset is small compared with that of the SEER datasets (1632 versus 58085, 12019 and 8125 respectively). Therefore, it might be appropriate to consider the prevalence of HER2-neu status, obtained from other studies and discussed at the beginning of Section 5, in the form of an informative prior.

Let $\theta_{**1} = \sum_{r \in \{2,4,6,8\}} \theta_{hr}$ denote the probability of HER2-neu positive. Considering the prevalence obtained by De Laurentiis et al. (2005), 19 – 35%, as a benchmark, we want that $E(\theta_{**1}) = 0.27$ (middle point) and $P(0.19 < \theta_{**1} < 0.35) = 0.95$. If we further distribute the remaining probability uniformly among the other categories, it is not difficult to show that a $\text{Dir}(\boldsymbol{\alpha})$ prior for $\boldsymbol{\theta}$ with $\alpha_{hr} = 89.2/16$ for $r = 1, 3, 5, 7$ and $\alpha_{hr} = 33/16$ for $r = 2, 4, 6, 8$ satisfies the required conditions.

We implemented our fixed effects approach using the dynamic model under two prior scenarios. A relative non-informative scenario with uniform prior specifications: $\boldsymbol{\delta} = (1, \dots, 1)$, $\boldsymbol{\beta}_h = (1, \dots, 1)$; and an informative scenario with prior specifications: $\boldsymbol{\delta} = (1, \dots, 1)$, $\boldsymbol{\beta}_{hr} = 89.2/16$ for $r = 1, 3, 5, 7$ and $\boldsymbol{\beta}_{hr} = 33/16$ for $r = 2, 4, 6, 8$. In both cases we took $c_h \sim \text{Po}(\lambda)$, $h = 1, 2, 3$, for a set of values $\lambda \in \{0, 10, 20, 50, 100\}$ to compare. Again, the overdispersion parameter κ was taken to be 10 for the Metropolis-Hastings proposal distribution. The Gibbs sampler was run for 200,000 iterations with a burn-in of 20,000 and keeping one realization of every 10^{th} after the initial burn-in period.

Table 5.2 presents the LPML and the L-measure statistics for the uniform and informative prior scenarios and for a range of values of the prior mean of c_h , λ . In general, considering the two goodness of fit statistics, we can appreciate a better fit for larger values of λ . This behavior is consistent in both uniform and informative scenarios. However, the informative scenario shows a slightly better fit in the LPML statistic and a remarkable decrement in the L-measure statistic in comparison with that of the uniform prior. The best fit is therefore obtained by the informative scenario with $\lambda = 100$.

A concern that could arise here is why the best fit is obtained with $\lambda = 100$ and not with $\lambda = 10$ that provided the best fit in the analysis of the M. D. Anderson dataset of the previous section. The answer is based on the sample size and the structure of the dynamic Dirichlet prior. Considering the (conditional) posterior distribution of η_h , given in (3.5), we can see that the latent vectors $\mathbf{u}_h$ play the role of auxiliary data and the posterior impact is relative to the sample value $\mathbf{y}_h$. Therefore, since this meta-analysis involves the information of 79,861 patients and the M. D. Anderson data set has only 1,632 patients, it is understandable that a larger value of λ , and thus c_h and $\mathbf{u}_h$, is required for a better fit.

In order to better understand the mechanism of how the dynamic Dirichlet prior borrows strength across the stages, Figure 5.1 includes box-plots of posterior draws from the parameters c_h , when $c_h \sim \text{Po}(100)$ *a-priori*, for $h = 1, 2, 3$, and the informative prior scenario. In Section 2.2 we saw that the larger the value of c_h , the stronger the dependence between η_h and η_{h+1} is, with independence when $c_h = 0$. As can be seen from Figure 5.1, posterior distributions of the c_h 's are located away from zero, implying that the data support the borrowing of strength across stages. On the other hand, it is worth noting that the data are relatively uninformative regarding c_1 , since the posterior is almost identical to the prior. However, for c_2 and c_3 there is some learning from the data, adjusting the distribution to larger values.

Figure 5.2 shows the posterior estimates of the conditional probabilities $\eta_{h,r}$ across the stages. Recall that these values represent conditional probability of having profile r given that the patient is in stage h . The big dots correspond to the point estimates (posterior means), the vertical lines denote the length of 95% credible intervals, and the dotted lines are included to visually assess the changes across the stages. Note that the same vertical scale was kept in all graphs for comparison. From here it can be seen that profiles are naturally clustered in two groups defined by the nodal involvement status. When there is no nodal involvement (profiles 1, 2, 5, and 6) there is a decreasing trend in the proportions as the stage increases. On the other hand, when the breast cancer patient has nodal involvement (profiles 3, 4, 7, and 8) patients are relatively more likely to be diagnosed with a more severe disease.

The estimated marginal distribution of the stages, θ_* , are given by the following 95% posterior credible intervals (in percentage):

$\theta_1^* \in (51.9, 52.6)$, $\theta_2^* \in (36.4, 37.0)$, $\theta_3^* \in (6.8, 7.1)$ and $\theta_4^* \in (3.9, 4.2)$. The first two stages of breast cancer are the most probable, which shows evidence of early diagnosis of the disease.

The stage/profile proportions $\theta_{h,r}$ were also estimated. These estimators are summarized in Table 5.3, obtained with the best fitting model of Table 5.2. It is clear that the category corresponding to stage 1 / profile 5 (positive ER status, negative nodal status, and negative HER2-neu status) has the largest proportion with a value of 0.3941. The minimum proportion corresponds to stage 1 / profile 8 (positive nodal involvement, positive ER status, and positive HER2-neu status) with a value of 0.00004. Having precise estimates for these proportions is important for planning policies and health strategies to help end the disease.

5.3 Simulation study

The meta-analysis just described is based on the assumption that the incidence probabilities for the stage/profile combinations are the same for all datasets (i.e., it is based on a fixed effects assumption). This fixed effects meta-analysis might not be appropriate when the incidences for different studies are not the same. In order to assess what the implications are when the fixed effects assumption is not satisfied we performed a simulation study.

Let us consider a data structure similar to the breast cancer patients studied in this article. More precisely, let us consider L studies such that for each study $l = 1, \dots, L$ we have $\mathbf{Y}_l | \boldsymbol{\theta}_l \sim \text{Mult}(N, \boldsymbol{\theta}_l)$ independently, with $\mathbf{Y}_l$ a matrix of dimension $H \times R$, N the number of multinomial trials or sample size and $\boldsymbol{\theta}_l$ the cell probabilities for the study. Let us further assume that $\boldsymbol{\theta}_l$ is taken from a population of probabilities, which in this case we suppose $\boldsymbol{\theta}_l \sim \text{Dir}(a \boldsymbol{\theta}_0)$ independently for $l = 1, \dots, L$, where $\boldsymbol{\theta}_0$ is the population mean vector of probabilities and a is a precision parameter. In fact, marginally $E(\theta_{l,h,r}) = \theta_{0,h,r}$ and $\text{Var}(\theta_{l,h,r}) = \theta_{0,h,r}(1 - \theta_{0,h,r})/(a + 1)$ for all l . The parameter a controls the variability of $\boldsymbol{\theta}_l$ around $\boldsymbol{\theta}_0$. The larger the value, of a , the more similar $\boldsymbol{\theta}_l$ of $\boldsymbol{\theta}_0$ becomes. In the limit case when $a \rightarrow \infty$ then each study l will have the same cell probabilities $\boldsymbol{\theta}_0$, that is, all studies will have fixed effects.

We simulated $L = 4$ datasets from the above construction with $H = 4$ and $R = 8$, and took several values of the precision parameter a , including 100, 1000, 10000, 100000 and ∞ , to produce datasets ranging from highly different cell probabilities to the same cell probabilities among studies. We considered the same sample size N for all studies $l = 1, \dots, 4$ and examined the effects of varying this number by setting $N = 5000, 10000, 20000$. Moreover, we also took two different values for the population mean of probabilities, an unbalanced and a balanced case. The unbalanced case is similar to the estimated breast cancer probabilities in Section 5.1, i.e.,

$$\boldsymbol{\theta}_0^{(1)} = \frac{1}{100} \begin{pmatrix} 6 & 4 & 1 & 1 & 20 & 5 & 1 & 1 \\ 3 & 2 & 5 & 4 & 7 & 2 & 10 & 3 \\ 1 & 1 & 2 & 3 & 2 & 1 & 4 & 2 \\ 1 & 1 & 1 & 1 & 1 & 1 & 2 & 1 \end{pmatrix}.$$

For the balanced case $\boldsymbol{\theta}_0^{(2)}$ we took a uniform distribution, i.e., each $\theta_{0,h,r}^{(2)} = 1/32$. A last consideration is that simulated datasets $l = 1, 2, 3$ were collapsed in the same fashion as the SEER 1–3 datasets.

With the previous scenarios we reproduce the main features of the real data at hand, but with the advantage that we know the true values for all parameters. For each data setting we implemented our fixed effect meta-analysis model with the prior specifications that gave the best fit in Section 5.1. That is, $\boldsymbol{\delta} = (1, \dots, 1)$, $\boldsymbol{\beta}_h = (1, \dots, 1)$ and $c_h \sim \text{Po}(100)$, for $h = 1, 2, 3$. For each simulation setting the experiment was replicated 100 times. For each experiment we computed two measures to assess the quality of the posterior estimates with respect to the true population mean of probabilities $\boldsymbol{\theta}_0$. The first measure is the overall mean coverage

of 95% posterior credible intervals of the 32 cell probabilities. The second measure is the mean square error (MSE) of the cell probability estimates. The average overall coverage measure for the 100 replications of the experiment is reported in Table 5.4. The MSE was also averaged for the 100 experiments, and to have a point of comparison, these values were also reported relative to the averaged MSE of the case when $a \rightarrow \infty$. The MSE and the relative mean square errors (RMSE) are also reported in Table 5.4.

From Table 5.4 we see that the sample size N has an important impact on the performance of the posterior estimates. For $N = 5000$ the overall coverage is a lot lower than it should be (95%) for all values of a . We recall that the data comes from a multinomial with 32 cells, and thus with $N = 5000$ each cell has 156 counts on average, which can be considered small. If we add the additional collapsing of three of the four datasets then the posterior estimates behave badly. For larger values of N , say 10000 and 20000, the coverage of 95% posterior credible intervals is as desired for $a \rightarrow \infty$ and for $a = 100000$. For $a = 10000$ the coverage is not as desired but acceptable, but for smaller values of a , 1000 and 100, the posterior behavior in terms of the coverage is too low. In terms of the MSE and RMSE, the performance of the posterior estimates is acceptable for $a = 10000$. For $a = 100$ and $a = 1000$ the RMSE become too big. The same conclusions can be obtained from the unbalanced $\theta_0^{(1)}$ and balanced $\theta_0^{(2)}$ population means. Perhaps smaller MSE are obtained for the unbalanced case.

In summary, it can be said that the fixed effects meta-analysis proposed here provides acceptable results when small or no variability is present in the cell probabilities of the different studies. Otherwise, a random effects meta-analysis (Hunter and Schmidt, 2000) would be more appropriate. However, a random effects meta-analysis for data following a complex data structure like the one in this paper (including, collapsed categories, multinomial outcomes, etc.) has never been proposed. In addition, such a model may not be feasible due to a lack of identifiability induced by the collapsed nature of some of the categories.

6 Conclusions

In this paper we have introduced an appealing way of estimating the distribution of breast cancer patients' characteristics. By assuming a natural trend of disease characteristics across stages, we proposed a dynamic Dirichlet prior to model this evolution while borrowing strength across stages. In consequence we produce better (more precise) estimates and achieve a better fit to the data. The reduction in length for the posterior credible intervals ranges from 0% to 30% with respect to the independence model.

We believe that having better estimates for the distribution of breast cancer patients by profiles (status combination) will produce more reliable breast cancer outcome forecasts and contribute to a better understanding of the disease.

The problem of having different datasets with different levels of disaggregation was addressed appropriately by a fixed effects meta-analysis together with a data augmentation technique. This is a standard approach when dealing with missing data from various sources

(Tanner, 1991). Posterior inference was made possible via an easy-to-implement Gibbs sampler. A Metropolis-Hastings step was required only when the parameter c_h was assigned a hyper prior distribution.

Based on the simulation study the results of our fixed effects meta-analysis are valid under an assumption of weak or no heterogeneity in the profile probabilities. An alternative modeling strategy could be to consider a random effects meta-analysis of the four datasets. Taking care of the collapsed nature of some datasets and the inclusion of the dynamic Dirichlet prior would be challenging and potentially not possible due to a lack of parameter identifiability in the partially observed tables.

Alternative applications of the dynamic Dirichlet prior are possible. For instance, consider a competing risk survival problem with $P(J = j) = p_j$, the probability of dying from specific cause j , for $j = 1, \dots, m$. If there are three dose levels, say d_x , for $x \in \{1, 2, 3\}$, we can borrow strength across dose levels by considering a dynamic Dirichlet prior $\mathbf{p}_{d_1} \xrightarrow{\mathbf{u}_1} \mathbf{p}_{d_2} \xrightarrow{\mathbf{u}_2} \mathbf{p}_{d_3}$ with latent variables $\mathbf{u}_1$ and $\mathbf{u}_2$. On the other hand, the model can also be used for time series analysis of multivariate proportions that add up to one. These and other possible applications are worth pursuing.

Acknowledgments

Dr. Bekele's work was supported in part by NIH grant 3U01 CA088278. Dr. Nieto-Barajas's work was partially supported by a grant from *The Fulbright-García Robles Program*. Mr. Munsell's work was supported by NIH contract 1HHSN2612007005. We also thank Shana Palla and Graciela Nogueras-Gonzalez for assistance in data management of the SEER and M. D. Anderson datasets and Dr. Richard Theriault for giving us permission to use the M. D. Anderson data.

References

- Agresti, A. An introduction to categorical data analysis. Wiley; Hoboken: 2007.
- Bernardo, JM.; Smith, AFM. Bayesian Theory. Wiley; New York: 2000.
- Berry DA, Cronin KA, Plevritis SK, Fryback DG, Clarke L, Zelen M, Mandelblatt JS, Yakovlev AY, Habbema JD, Feuer EJ. Effect of screening and adjuvant therapy on mortality from breast cancer. The New England Journal of Medicine. 2005; 353:1784–1792. [PubMed: 16251534]
- Choi DH, Shin DB, Lee MH, Lee DW, Dhandapani D, Carter D, King BL, Haffty BG. A Comparison of Five Immunohistochemical Biomarkers and HER-2/neu Gene Amplification by Fluorescence In Situ Hybridization in White and Korean Patients with Early-Onset Breast Carcinoma. Cancer. 2003; 8:1587–1595. 9bf. [PubMed: 14534873]
- De Laurentiis M, Arpino G, Massarelli E, et al. A meta-analysis on the interaction between HER-2 expression and response to endocrine treatment in advanced breast cancer. Clinical Cancer Research. 2005; 11:4741–4748. [PubMed: 16000569]
- Di Fiore PP, Pierce JH, Fleming TP, et al. Overexpression of the human EGF receptor confers an EGF-dependent transformed phenotype to NIH3T3 cells. Cell. 1987; 51:1063–1070. [PubMed: 3500791]
- Feuer, et al. The Impact of Mammography and Adjuvant Therapy on U.S. Breast Cancer Mortality (1975–2000): Collective Results from the Cancer Intervention and Surveillance Modeling Network. JNCI Monographs. 2006; 36:1–126.
- Hudziak RM, Schlessinger J, Ullrich A. Increased expression of the putative growth factor receptor p185HER2 causes transformation and tumorigenesis of NIH3T3 cells. Proceedings of the National Academy of Science. 1987; 84:7159–7163.

- Hunter JE, Schmidt FL. Fixed Effects vs. Random Effects Meta-Analysis Models: Implications for Cumulative Research Knowledge. *International Journal of Selection and Assessment*. 2000; 8:275–292.
- Ibrahim JG, Laud PW. A predictive approach to the analysis of designed experiments. *Journal of the American Statistical Association*. 1994; 89:309–319.
- Nieto-Barajas LE, Walker SG. Markov beta and gamma processes for modelling hazard rates. *Scandinavian Journal of Statistics*. 2002; 29:413–424.
- Sinha D, Dey DK. Semiparametric Bayesian analysis of survival data. *Journal of the American Statistical Association*. 1997; 92:1195–1212.
- Slamon DJ, Clark GM, Wong SG, et al. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science*. 1987; 235:177–182. [PubMed: 3798106]
- Slamon DJ, Godolphin W, Jones LA, et al. Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. *Science*. 1989; 244:707–712. [PubMed: 2470152]
- Smith AFM, Roberts GO. Bayesian computations via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*. 1993; 55:3–23.
- Spiegelhalter, D.; Thomas, A.; Best, N.; Lunn, D. WinBugs User Manual: Version 1.4. 2003. <http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/manual14.pdf>
- Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov). SEER * STAT Database: Incidence - SEER 9 Regs Public-Use, Nov 2005 Sub (1973-2003), National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2006, based on the November 2005 submission.
- Tanner, MA. Tools for statistical inference: Observed data and data augmentation methods. Springer; New York: 1991.
- Tierney L. Markov chains for exploring posterior distributions. *Annals of Statistics*. 1994; 22:1701–1728.
- Tudur Smith C, Williamson PR. A comparison of methods for fixed effects meta-analysis of individual patient data with time to event outcomes. *Clinical trials*. 2007; 6:621–630. [PubMed: 18042571]
- Veeramachaneni, S.; Sona, D.; Avesani, P. Proceedings of the 22nd International Conference on Machine Learning. ACM Press; New York: 2005. Hierarchical Dirichlet model for document classification; p. 928-935.

Figure 4.1.

Convergence diagnostics of θ_{11} , using the dynamic Dirichlet prior with $\alpha = (1, 1, 1, 1)$ and $\beta = (1, \dots, 1)$. M.D. Anderson dataset only. Chain trace (top left), Ergodic means (top right), autocorrelation function (bottom left) and probability histogram (bottom right).

Figure 4.2.

M. D. Anderson dataset. Posterior point estimates of $\eta_{h,r}$ for three different models.

Dynamic Dirichlet prior (solid line), hierarchical prior (dashed line) and quadratic multinomial logistic model (dotted-dashed line). In all cases the lines are included to visually assess the changes across stages. Note that the scale is different in each panel to better appreciate the differences.

Figure 5.1.

Box-plots of posterior draws for parameters c_h , $h = 1, 2, 3$ in the meta-analysis of the SEER 1 – 3 and M. D. Anderson datasets, when using the dynamic model prior with $\alpha = (1, 1, 1)$, $\beta_{hr} = 89.2/16$ for $r = 1, 3, 5, 7$, $\beta_{hr} = 33/16$ for $r = 2, 4, 6, 8$ and $c_h \sim \text{Po}(100)$.

Figure 5.2.

Meta-analysis for SEER 1 – 3 and M. D. Anderson datasets. Posterior estimates of $\eta_{h,r}$ using the dynamic model prior with $\alpha = (1, 1, 1, 1)$, $\beta_{hr} = 89.2/16$ for $r = 1, 3, 5, 7$, $\beta_{hr} = 33/16$ for $r = 2, 4, 6, 8$ and $c_h \sim \text{Po}(100)$. The big dots represent the mean and the vertical lines denote 95% credible intervals. The dotted lines are included to visually assess the changes across the stages and the vertical scale was kept the same in all the graphs for comparison.

M. D. Anderson dataset. Number of breast cancer patients per stage and profile. The profile indicators have been provided in the $\{1, \dots, 8\}$ scale and in the $\{0, 1\}^3$ scale.

Table 4.1

Stage/Profile	1 (0,0,0)	2 (0,0,1)	3 (0,1,0)	4 (0,1,1)	5 (1,0,0)	6 (1,0,1)	7 (1,1,0)	8 (1,1,1)
1	74	44	0	1	381	38	2	0
2	79	48	50	27	124	30	295	51
3	26	14	44	29	17	9	117	23
4	8	2	17	13	10	2	46	11

Source: Department of Breast Medical Oncology, M. D. Anderson Cancer Center, 1997-2005.

Table 4.2

Model comparison for M. D. Anderson dataset only. Log-pseudo marginal likelihood (LPML) and L-measure statistics.

Model	LPML	L-measure $\nu = (0, 0.5, 1)$
Dynamic ($\lambda = 0$)	-99.64	(1748, 1786, 1823)
Dynamic ($\lambda = 10$)	-97.76	(1757, 1771, 1786)
Dynamic ($\lambda = 20$)	-99.63	(1779, 1823, 1868)
Dynamic ($\lambda = 50$)	-104.17	(1774, 1898, 2021)
Dynamic ($\lambda = 100$)	-108.33	(1792, 2153, 2514)
Exchangeable	-100.36	(1796, 1809, 1823)
Linear M.Logistic	-201.02	(1667, 7443, 13219)
Quadratic M.Logistic	-114.38	(1774, 2518, 3263)

M. D. Anderson dataset. Ratio (dynamic Dirichlet prior with $\lambda = 10$ / independence prior) between the lengths of 95% credible intervals for $\theta_{h,r}$.

Table 4.3

Stage/Profile	1 (0,0,0)	2 (0,0,1)	3 (0,1,0)	4 (0,1,1)	5 (1,0,0)	6 (1,0,1)	7 (1,1,0)	8 (1,1,1)
1	1.00	1.01	0.69	0.85	1.01	1.00	1.00	0.73
2	1.00	1.01	1.00	0.99	1.02	1.01	1.00	1.00
3	0.98	0.96	0.97	0.97	1.01	0.96	0.99	0.99
4	0.94	0.88	0.96	0.95	0.92	0.87	0.95	0.94

Three SEER datasets. First: ER/Nodal status, second: nodal status only, third: ER status only. The profile indicators have been provided in the $\{1, \dots, 8\}$ scale and in the $\{0, 1\}^3$ scale.

Table 5.1

SEER 1			SEER 2			SEER 3		
Stage/Profile	(1,2)(0,0,*)	(3,4)(0,1,*)	(5,6)(1,0,*)	(7,8)(1,1,*)	(1,2,5,6)(*,0,*)	(3,4,7,8)(*,1,*)	(1,2,3,4)(0,*,*)	(5,6,7,8)(1,*,*)
1	5114	1	25062	3	7029	3	535	3461
2	2355	3043	6574	11339	1661	2413	243	976
3	77	966	180	2460	66	567	344	616
4	34	222	99	556	27	253	525	1425

Source: SEER Stat database: Incidence. November 2005 Submission (1973-2003).

Table 5.2

Model comparison for the 4 datasets. Log-pseudo marginal likelihood (LPML) and L-measure statistics for $\nu = 0, 0.5, 1$. Uniform and informative prior scenarios.

Dynamic Model	Uniform prior on η		Informative prior on η	
	LPML	L-measure $\nu = (0, 0.5, 1)$	LPML	L-measure $\nu = (0, 0.5, 1)$
$\lambda = 0$	-1986.25	(1629748, 5090744, 8551739)	-1982.32	(1491282, 4944137, 8396992)
$\lambda = 10$	-1984.73	(1607316, 5069904, 8532493)	-1980.51	(1520391, 4977969, 8435546)
$\lambda = 20$	-1983.91	(1547403, 5010297, 8473190)	-1979.19	(1505916, 4961273, 8416630)
$\lambda = 50$	-1981.10	(1583920, 5047389, 8510858)	-1978.81	(1464219, 4928101, 8391984)
$\lambda = 100$	-1980.78	(1581001, 5048321, 8515641)	-1977.23	(1345974, 4815165, 8284356)

Table 5.3

Meta-analysis results. Estimated percent of breast cancer patients in categories, using the dynamic model prior with $\alpha = (1, 1, 1, 1)$, $\beta_{hr} = 89.2/16$ for $r = 1, 3, 5, 7$, $\beta_{hr} = 33/16$ for $r = 2, 4, 6, 8$ and $c_h \sim \text{Po}(100)$ considering the information from SEER 1 – 3 and M. D. Anderson datasets. Posterior means in percentage.

Stage/Profile	1 (0,0,0)	2 (0,0,1)	3 (0,1,0)	4 (0,1,1)	5 (1,0,0)	6 (1,0,1)	7 (1,1,0)	8 (1,1,1)
1	5.61	3.07	0.01	0.005	39.62	3.93	0.02	0.004
2	2.44	1.37	3.09	1.61	8.79	1.65	15.15	2.59
3	0.15	0.08	1.19	0.69	0.28	0.11	3.79	0.67
4	0.13	0.04	0.6	0.37	0.33	0.09	2.11	0.42

Table 5.4

Simulation study results for different combinations of θ_0 , a and N . Reported are coverage, mean square error (MSE) and relative mean square error (RMSE).

a	N	$\theta_0^{(1)}$				$\theta_0^{(2)}$			
		Coverage	MSE	RMSE	Coverage	Coverage	MSE	RMSE	RMSE
∞	5000	53.31	6.90×10^{-6}	1	52.90		7.71×10^{-6}	1	
	10000	53.96	7.28×10^{-6}	1.05	52.37		8.01×10^{-6}	1.04	
	10000	50.15	8.81×10^{-6}	1.27	49.25		9.75×10^{-6}	1.26	
	1000	33.50	2.37×10^{-5}	3.44	32.68		2.59×10^{-5}	3.36	
	100	12.59	1.73×10^{-4}	25.07	14.12		1.79×10^{-4}	23.25	
∞	10000	97.00	3.66×10^{-6}	1	96.62		4.09×10^{-6}	1	
	10000	96.71	3.72×10^{-6}	1.02	96.18		4.35×10^{-6}	1.06	
	10000	90.96	5.31×10^{-6}	1.45	92.31		5.82×10^{-6}	1.42	
	1000	58.96	2.19×10^{-5}	5.99	59.12		2.41×10^{-5}	5.90	
	100	23.06	1.63×10^{-4}	44.75	24.09		1.85×10^{-4}	45.38	
∞	20000	94.62	1.83×10^{-6}	1	94.59		2.05×10^{-6}	1	
	20000	93.46	1.91×10^{-6}	1.05	93.21		2.23×10^{-6}	1.09	
	10000	80.75	3.68×10^{-6}	2.01	81.34		3.99×10^{-6}	1.95	
	1000	44.41	1.95×10^{-5}	10.65	43.22		2.28×10^{-5}	11.13	
	100	16.56	1.67×10^{-4}	91.52	15.53		1.92×10^{-4}	93.81	