Published in final edited form as:

Stat Med. 2008 August 30; 27(19): 3776–3788. doi:10.1002/sim.3279.

An Unified Modeling Framework for Metabonomic Profile Development and Covariate Selection for Acute Trauma Subjects

S. Ghosh^{1,*} and D. K. Dey²

¹Department of Mathematical Sciences, Indiana University Purdue University, Indianapolis

²Department of Statistics, University of Connecticut, Storrs

SUMMARY

Acute trauma is often associated with progressive deterioration of multiple organ systems in humans and is the leading cause of death in trauma care units. Identification of specific organ failure in non-invasive fashion and the contribution of different demographic factors on the casual progression of acute trauma is of supreme interest for successful diagnosis, prognosis or monitoring of trauma status. Recently electrospray ionization (ESI) and matrix assisted laser desorption (MALDI) time-of-flight (TOF) mass spectrometry have been used to identify biomarkers both in Proteomics and Metabonomics studies. Data sets generated from mass spectrometers in such studies are generally very large in size and thus require the use of sophisticated statistical techniques to glean useful information. In a recent development Ghosh et al. [1] suggested a unified semi-parametric approach to distinguish urinary metabolic profiles in a group of traumatic subjects from those of a control group consisting of normal individuals. In this study we have extended their approach by combining available covariate information in the development of metabonomic profile of acute trauma. We have shown that age is a statistically significant covariate across trauma and control group, thus pointing out the fact that prognosis of trauma may be acutely linked with subjects age.

1. Introduction

Acute trauma with associated hemorrhage, shock and sepsis, often produces a generalized inflammatory response resulting in organ dysfunction and multiple organ failure. Multiple organ failure is associated with high mortality rates and is among the leading causes of death among Americans between the ages of 1 and 44 [2]. It is also the leading cause of death in trauma intensive care units [3]. This is despite tremendous technological advances in the use of mechanical ventilation, invasive hemodynamic monitoring, vasoactive therapy, dialysis, and parenteral nutrition. The mechanisms underlying the progression from acute trauma to multiple organ failure remain poorly understood. An emerging consensus is that organ dysfunction caused by hemorrhage, ischemia or blunt tissue injury can lead to a clinical syndrome that, although variable in expression, has a common pathophysiologic basis [3]. Current methods of assessing trauma patient status and predicting outcome are based on a variety of anatomical and/or physiological scoring models. While being useful, these are labor intensive and do not allow for real-time analysis of patient status. Also the progression of trauma is highly interlinked with not only injury itself but also with various other cofactors associated with specific subject. In this study, we have developed a metabonomic

based approach using mass spectrometry to elucidate more about the covariate effect on traumatic disorder.

Metabolites are end products and by-products resulting from the various biosynthetic and catabolic pathways that exist in all living organisms. Historically, the qualitative and quantitative measurement of metabolites has been routinely used for a variety of diagnostic clinical procedures. Examples of these include serum and urine creatinine and urea concentrations, and serum glucose, lactate, and bilirubin concentrations. These metabolites are routinely quantified by automated instruments and their concentrations are correlated with one or more physiological parameters. With recent advances in analytical detection technologies and computer software, it is now possible to identify and quantify complex time-related metabolite profiles representing thousands or even tens-of-thousands of metabolites in one sample. The term metabonomics has been coined over the last 10 years for these types of metabolite profiling [4]. Most metabonomic analysis have been performed using high-resolution nuclear magnetic resonance spectroscopy. More recently mass spectrometry has been used to perform this analysis because of its ability to identify and quantify a greater number of metabolites at much lower concentrations [5,6,7,8].

With advent of Mass Spectrometry based technology a new direction in bioinformatics has emerged for the identification of biomarker patterns that can be used for diagnosis, prognosis or monitoring of disease status. The underlying hypothesis is that the mass spectral profile of patients will differ significantly than that of healthy individuals (controls). As might be expected, the methods used for assessing “*significant difference*” vary widely. Starting from different clustering algorithms [9], wavelet based feature extraction [10,11,12,13] and a plethora of other methods have been suggested in the recent past to understand this “*significant difference*”. However apart from usual bioinformatics pattern discovery none of them mention how to associate covariate information (if any) in the MS framework. Also the reason for the “*significant difference*” could come from multitude of sources for example the nature of injury that causes trauma itself and the other somatic/demographic variables. Hence along with the MS profile the covariate information may pertain more light on any biological discovery and if available they should not be ignored. Apart from the above several problems [14] regarding instrument sensitivity and reproducibility of the data set underlying these approaches remain also unresolved. A major concern expressed in the literature [15] is the calibration of the mass spectrometer. The data generated by mass spectrometry of urine, blood or serum is composed of the ions of compounds or fragments of compounds identified by their mass-to-charge ratio (m/z) and the abundance of each ion as depicted in Figure 1. A typical mass spectrum contains thousands of points (m/z values) while subject numbers are quite low. Due to the inherent analytical variability the same ions in replicates of a sample will have slightly different measured m/z values. Different binning/alignment [16,15] algorithms and clustering techniques [9] have been suggested to properly correlate ions between samples, prior to advance analysis. Recently Ghosh et al. [1] suggested a novel approach where they have tackled instrument variability as an integral part of model building process. They have also pointed out in the presence of accurate alignment algorithm, how their result could be used more efficiently, thus reducing another level of variation. However their main interest was developing metabonomic profile based on m/z values and corresponding relative intensity/abundance. For that reason associated covariate information was some what ignored. Since this is essentially modeling time to event phenomenon, a natural extension of time to event data modeling is to use Cox proportional hazard model and thus combine available covariate information in a semiparametric modeling framework. In the current paper we have considered various demographic as well as somatic variable/scale as possible covariates and used in conjunction with the realtime abundance data generated by mass spectrometer. We found some of the covariates forming natural cluster/group which one expects to see

between traumatic subjects and control subjects and thus pertaining useful information for efficient future identification and treatment of trauma subjects.

The rest of the paper is organized as follows. We have developed semiparametric framework for covariate inclusion in section 2. Choice of prior and development of posterior sampling schemes have been discussed in section 3. Section 4 explains the data set under consideration. We have presented data analysis result in section 5. This section also explains various convergence issue and other MCMC output related information. We have concluded the article with a brief discussion in section 6.

2. Semiparametric Framework

We begin with the raw mass spectrometer data set. Some of the very low intensities are removed to eliminate possible noise. Details of these are instrument specific and hence discussion on these are postponed for the time being and will be revisited in section 4. Suppose our data came from N subjects, where each subject could be either traumatic or control. Data for a single subject will have $r(\geq 1)$ replicates. Due to the inherent variability of the mass-spectrometer, intensities may get registered at slightly different positions even within the replicates of a single subject. In this paper we have used alignment algorithm developed by Kazmi et al. [17] to align the replicates first. It has been reported [17] that this algorithm does a near perfect job in aligning m/z values for such data set. Notably perfect alignment if possible improves signal to noise ratio, which in turn enhances confidence in estimation itself [1]. The detailed issues related to do this will be discussed in detail in section 4. We consider the average intensity produced by the alignment algorithm over all replicates for each registered m/z value. Let D^i denote a $n_i \times 2$ dimensional average

spectrum for the i -th subject, where $i = 1, 2, \dots, N$. For each spectrum $D^i(t) = \left[d_t, \{t\}_{t_{1_i}}^{t_{n_i}} \right]$ has two components. First component d_t indicates observed relative abundance/intensity at an associated position t , for $t = t_{1_i}, t_{2_i}, \dots, t_{n_i}$. For the computational flexibility we have log-transformed all the m/z values (t 's here). In the present context T denotes the time-of-flight for the particles detected in a mass spectrometer. Since a monotone function converts time-of-flight T to an m/z value, they have a one to one correspondence and we would not differentiate between them in our analysis. We can further assume that $f(t)$ denotes the probability density function of the stochastic process underlying T . The $h_0(t)$ is termed as intensity function, which denotes the instantaneous rate of ions hitting the detector at time t . This in turn denotes instantaneous rate of change in the density function $f(t)$, i.e.,

$h_0(t) = \lim_{\Delta t \rightarrow 0} \frac{F(t+\Delta t) - F(t)}{\Delta t(1 - F(t))}$ which is identical to the baseline hazard function we observe in survival analysis literature ($F(t)$ is the usual c.d.f.). If $h_0(t)$ is high, it indicates hitting rate or detection rate is also high. In particular $h_0(t)\Delta t$ is the approximate probability of hitting the detector in $(t, t + \Delta t]$, given that the ion has not arrived at the detector (hence not detected) up to time t . We will not differentiate between the event of arrival of the ions at the detector of the mass spectrometer and the actual detection (current flow). In reality there is a very tiny time lag between these two events. The function $h_0(t)$ has two desirable properties

$$h_0(t) \geq 0 \quad \text{and} \quad \int_0^\infty h_0(t) dt = \infty.$$

Our goal is to model $h_0(t)$ using mixture of betas. Note that at any instance t , the area under the curve $h_0(t)$ is given by $H_0(t) = \int_0^t h_0(u) du$. It is possible to calculate the approximate value of $H_0(t)$ from any given spectrum. For modeling $H_0(t)$, we will use the representation

theorem of Diaconis and Ylvisaker [18] for monotone functions on $[0, 1]$. First we consider the transformation

$$H^*(t) = \frac{H_0(t)}{H_0(t) + 1}.$$

Note that $H^* : \mathbb{R}^+ \rightarrow [0, 1]$ while $H_0 : \mathbb{R}^+ \rightarrow \mathbb{R}^+$. Clearly H^* is the cumulative distribution function. Having transformed the mean intensity into the interval $[0, 1]$ we approximate

$H^*(t) = \sum_{l=1}^m \eta_l IB(\tilde{H}(t); r_l, s_l)$, where η_l are weights satisfying $\sum_{l=1}^m \eta_l = 1$, for all replicates and $IB(\cdot; r_l, s_l)$ denotes the incomplete beta function. Traditionally m is chosen as a small constant number [19]. Collecting the weights into $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_m)$ a Dirichlet prior can be assigned. The (r_l, s_l) are chosen to ensure that Beta c.d.f's have equally spaced means and are centered around $\tilde{H}(t)$. $\tilde{H}(t)$ is a suitable function which maps the time scale into $[0, 1]$. The related issue of choosing $\tilde{H}(t)$ needs to be resolved by selecting a plausible central function around which the H^* function is distributed. If we consider cumulative intensity function, parallel to the cumulative hazard function of the survival analysis literature [19]

then we can take $\tilde{H}(t) = \frac{\tilde{H}_0(t)}{1 + \tilde{H}_0(t)}$. Different centering distribution could be taken to model $\tilde{H}_0(t)$. However for the present data set it has been reported elsewhere [1] that Double exponential based model produces the best result. In the present context only Double Exponential based model has been considered for that reason. The Double Exponential distribution produces following functional forms,

$$f_0(t) = \frac{1}{2\delta} e^{-|\frac{t-\alpha}{\delta}|}, \tilde{H}_0(t) = \begin{cases} -\log \left[1 - \frac{1}{2} \exp \left\{ \frac{t-\alpha}{\delta} \right\} \right], & t \leq \alpha \\ \frac{t-\alpha}{\delta} + \log 2, & t \geq \alpha \end{cases}; \gamma = (\alpha, \delta) \quad \text{and} \quad \delta > 0.$$

The $\tilde{H}_0(t)$ is serving the role of a central function and a flexible choice of this will enable us to capture non-standard patterns in the profile in a microscopic sense. Note that

$$\begin{aligned} H^*(t) = \frac{H_0(t)}{H_0(t) + 1} &\Rightarrow H_0(t) = \frac{H^*(t)}{1 - H^*(t)} = \frac{\sum_{l=1}^m \eta_l IB(\tilde{H}(t); r_l, s_l)}{\left\{ 1 - \sum_{l=1}^m \eta_l IB(\tilde{H}(t); r_l, s_l) \right\}}, \\ h_0(t) = \frac{d}{dt} H(t) &= \frac{d}{dt} \left(\frac{H^*(t)}{1 - H^*(t)} \right) = \frac{\frac{d}{dt} \tilde{H}(t) \sum_{l=1}^m \eta_l IB(\tilde{H}(t); r_l, s_l)}{\left\{ 1 - \sum_{l=1}^m \eta_l IB(\tilde{H}(t); r_l, s_l) \right\}^2}, \end{aligned} \quad (1)$$

$$\text{where} \quad \frac{d}{dt} \tilde{H}(t) = \frac{\frac{d}{dt} \tilde{H}_0(t)}{\left(1 + \tilde{H}_0(t) \right)^2}.$$

2.1. Conditional Hazard Rate

In the present context intensity/hazard function depends upon both time and a set of covariates. The proportional hazard model introduced by Cox [20] for an individual with covariate x is a product of a baseline hazard ($h_0(t)$) and a non-negative function of the covariates $c(\mathbf{x}'\boldsymbol{\beta})$. The link $c(\cdot)$ is commonly log-linear link given by $c(\mathbf{x}'\boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta})$. The basic proportional hazard regression model is

$$h(t|x) = h_0(t) \exp(x' \beta).$$

Thus we can rewrite the likelihood as

$$\begin{aligned} f(t) &= h(t) \exp(-H(t)) \\ &= h_0(t) \exp(x' \beta) \exp\{-H_0(t) \exp(x' \beta)\}. \end{aligned} \quad (2)$$

2.2. Likelihood Construction

If we have n number of ions detected by the detector and t_j denotes the observed time of flight (TOF) for the j -th ion, then the likelihood is

$$L = \prod_{j=1}^n f(t_j) = \prod_{j=1}^n h(t_j) \exp(-H(t_j)).$$

In the present context observed quantities can be described like this; we observe d_t , the relative intensity or relative amount of current flow detected by the mass spectrometer at t -th m/z value. Note that the amount of current flow is proportional to the number of ions detected at any associated position and amount of charge on the ions. The observed m/z values (t) have time stamp at $t_{1i}, t_{2i}, \dots, t_{n_i}$, for the i -th subject. All the replicates of a single

observations are aligned first to produce $D^i(t) = [d_t, \{t\}_{t_{1i}}^{t_{n_i}}]$. The complete data likelihood for N subjects can be expressed as

$$\begin{aligned} L(\eta, \gamma) &= \prod_{i=1}^N \prod_{t=t_{1i}}^{t_{n_i}} (h(t) \exp(-H(t)))^{d_t} \\ &= \prod_{i=1}^N \prod_{t=t_{1i}}^{t_{n_i}} [h_0(t)]^{d_t} \prod_{i=1}^N \prod_{t=t_{1i}}^{t_{n_i}} \exp \left[d_t \left(x'_i \beta - H_0(t) e^{x'_i \beta} \right) \right] \\ &= \prod_{i=1}^N \prod_{t=t_{1i}}^{t_{n_i}} [h_0(t) \exp(-H_0(t))]^{d_t} \prod_{i=1}^N \prod_{t=t_{1i}}^{t_{n_i}} \exp \left[d_t \left(x'_i \beta + H_0(t) - H_0(t) e^{x'_i \beta} \right) \right]. \end{aligned} \quad (3)$$

Next we like to impose suitable priors on these parameters namely η , γ and β which completes Bayesian hierarchical setup. At this point we should mention that covariates play an important role not only in successful disease identification but also effect of the disease itself may be completely different from one set of covariate value to the other. This natural stratification of different biological cohort is very important and hence importance of covariates should not be undermined.

3. Prior Choices and Numerical Implementation

With the notation of section 2 the joint posterior distribution of the parameters of interest namely $\Theta = (\eta, \gamma, \beta)$ under suitable prior on η , γ and β is given by

$$\begin{aligned}
f(\eta, \gamma, \beta | \{t\}_{t=t_{1_i}}^{t_{n_i}}, h(t)) &\propto f(\eta | \psi_\eta) \\
&\times \prod_{i=1}^N \prod_{t=t_{1_i}}^{t_{n_i}} \exp \left[d_t \left(x'_{ij} \beta + H_0(t) - H_0(t) e^{x'_{ij} \beta} \right) \right] \\
&\times f(\gamma) \\
&\times f(\beta) \times \prod_{i=1}^N \prod_{t=t_{1_i}}^{t_{n_i}} \left[\frac{\frac{d}{dt} \tilde{H}(t) \sum_{l=1}^m \eta_l B e \left(\tilde{H}(t); r_l, s_l \right)}{\left\{ 1 - \sum_{l=1}^m \eta_l I B \left(\tilde{H}(t); r_l, s_l \right) \right\}^2} \exp \left(- \frac{\sum_{l=1}^m \eta_l I B \left(\tilde{H}(t); r_l, s_l \right)}{1 - \sum_{l=1}^m \eta_l I B \left(\tilde{H}(t); r_l, s_l \right)} \right) \right]^{d_t}.
\end{aligned}$$

Please note that here our main focus is on the covariates i.e., posterior distribution of β . Posterior distribution of η and γ can be expressed as

$$\pi(\nu | \beta, \text{Data}) \propto f(\nu) \times \prod_{i=1}^N \prod_{t=t_{1_i}}^{t_{n_i}} \left[\frac{\frac{d}{dt} \tilde{H}(t) \sum_{l=1}^m \eta_l B e \left(\tilde{H}(t); r_l, s_l \right)}{\left\{ 1 - \sum_{l=1}^m \eta_l I B \left(\tilde{H}(t); r_l, s_l \right) \right\}^2} \exp \left(- \frac{\sum_{l=1}^m \eta_l I B \left(\tilde{H}(t); r_l, s_l \right)}{1 - \sum_{l=1}^m \eta_l I B \left(\tilde{H}(t); r_l, s_l \right)} \right) \right]^{d_t}, \quad (4)$$

where ν could be either η or γ . We have taken similar prior choices and necessary transformation steps as discussed in Ghosh et al. [1] for these parameters. Notably it is also reported there that $m = 5$ as the optimal choice considering flexibility of the modeling as well ease of computation. For the present case we have nine covariates (originally 4 at different levels, as described in the next section). Posterior distribution of β can be expressed as

$$\pi(\beta | \text{Data}) \propto \prod_{i=1}^N \prod_{t=t_{1_i}}^{t_{n_i}} \exp \left[d_t \left(x'_{ij} \beta - H_0(t) e^{x'_{ij} \beta} \right) \right] \times f(\beta).$$

Since the posterior distribution is not available in closed form we have used Metropolis step within the Gibbs sampling, which requires acceptance of new proposal value β^* from proposal density with probability $\alpha = \min \{1, p\}$. The proposal distribution is normal

$p(\beta) = N_9 \left(\theta^{(t-1)}, \widehat{\sigma}^2 I \right)$, where initial estimates are obtained through Nelder-Mead downhill simplex algorithm. For this proposal it can be shown that

$$\log(p) = \sum_{i=1}^N \left[\left(x'_{ij} \beta^* - x'_{ij} \beta \right) \times \sum_{t=t_{1_i}}^{t_{n_i}} d_t \right] - \sum_{i=1}^N \left[\left(e^{x'_{ij} \beta^*} - e^{x'_{ij} \beta} \right) \times \sum_{t=t_{1_i}}^{t_{n_i}} d_t H_0(t) \right] + \log \left(\frac{f(\beta^*)}{f(\beta)} \right) + \log \left(\frac{p(\beta)}{p(\beta^*)} \right). \quad (5)$$

The sampling scheme implemented is a variation of the grouped Gibbs sampler [21,22], which requires drawing:

1. $\theta \sim p\left(\theta|\gamma, \beta, D = \bigcup_{i=1}^r D^i\right)$, with acceptance probability

$$\alpha = \min\left\{1, \frac{p(\theta|\gamma, D) N_4(\theta_0, \widehat{\Sigma})}{p(\theta_0|\gamma, D) N_4(\theta, \widehat{\Sigma})}\right\}$$
and then transform back $\theta \mapsto \eta$, where

$$\theta_l = \log\left(\frac{\eta_l}{1 - \eta_l}\right) \text{ for } l = 1, \dots, 4,$$
2. $\alpha \sim p(\alpha|\eta, \beta, \delta, D)$,
3. $\delta \sim p(\delta|\eta, \beta, \alpha, D)$,
4. $\beta \sim p(\beta|D)$, with acceptance probability as stated in the equation (5).

Under this setup, parameters are updated sequentially and monitored for convergence following post burn-in period. We mention at this point that log-transformation of the m/z value is done only for greater numerical stability of the MCMC algorithm.

4. Data Description and Data Preparation: Trauma Data Set

Mass spectrometry is used for identifying compounds based on the mass-to-charge (m/z) ratio of their pseudomolecular and, possibly fragmentation of ions generated in an ion source. The intensity of these ions can be used to quantify the amount of compounds present. The output from the mass spectrometer consists of a series of paired m/z and intensity values. In this paper our main goal is to explore the effect of different covariates in determining trauma status through development of metabonomic profile using mass spectrometer. We have retrospectively analyzed urine samples collected from 6 normal healthy individuals and 6 patients with different degree of acute trauma. Each urine sample was analyzed in triplicate ($r=3$) in Q-TOF2 mass spectrometer using lisinopril as an internal standard. The mass spectral data was first processed using the smoothing and centroid algorithms of MassLynx software (Micromass). The data used in the present study were generated by eliminating all ions that were less than or equal to three times the instrument noise after baseline subtraction and dividing the intensity of each ion by the concentration of creatinine in the respective urine sample. The normalization of the ion intensities to the creatinine concentration was necessary to compensate for urine dilution in different individuals. As described earlier (see section 1), due to the inherent analytical variability the same ions in replicates of a sample may have slightly different measured m/z values when analyzed using different run order. This introduces additional variability and thus need proper alignment of the Mass spectrometry output before any fruitful analysis can be done. We have used alignment algorithm presented in Kazmi et al [17] to align each subject using all of its replicates (namely three). It has been reported [17] that this algorithm has linear runtime complexity (i.e., much faster) and produces output which is perfectly aligned. The output of this alignment algorithm is a single representative aligned mass spectra for each

subject of the form $D^i(t) = \left[d_t, \{t\}_{t_{1i}}^{t_{n_i}} \right]$, where d_t indicates average (over the three replicates) relative abundance/intensity at an associated position t , for $t = t_{1i}, t_{2i}, \dots, t_{n_i}$ for the i -th individual. Our hypothesis is that if controls behave ideally then all of them should have same ions. However as we can perceive this may not be possible due to two reasons. First, inherent biological variation among controls and secondly unavoidable instrumental variability. We took a conservative approach by considering only those ions in 12 subjects, which are present among all control in the aligned data set. This results in great reduction of number of m/z values among all samples (approximately only 10% value remained). We have four covariates information available for this data set, namely age, sex, race and SIRS (Systemic Inflammatory Response Syndrome) score. All covariate information is presented

in table I. A group indicator variable (0 if traumatic and 1 otherwise) is also introduced. Except age, other four are categorical variables. We have four different category of race namely Caucasian, Black, Asian and Hispanic. Three indicator variables are created with Caucasian as baseline variable. SIRS score can vary between 0 - 4, 0 being ideal and then increasing order more towards trauma depending upon severity. We have observed three different values of SIRS score for the current data set and created two indicator variables taking SIRS score equals to 0 as baseline. So we have total 8 covariates, one for sex and age each, three for race, two for SIRS score and one for group level indicator. An interaction variable between group and age is also introduced to capture differential effect of age between trauma and control group. Notably considerable emphasis was put in the past [23,24] to find significance of age on trauma status and its progression. The final data set has nine covariates.

Each of these regression coefficients (β 's) are estimated through MCMC. First 10,000 iterations are thrown out as burn-in period, then an additional 50,000 iterations are obtained out of which we have accepted only every 50-th iteration, creating 1000 samples from the posterior distribution. This was used to compute all necessary statistics namely posterior mean, 95% credible interval and other measures. Every MCMC chains are tested for convergence by Geweke [25] convergence diagnostic and autocorrelation plot through R software. Several combinations of hyper-parameter values were tried. We report the results for prior distribution of the location parameter as $N(2, 10)$, for scale parameter as $IG(2, 10)$, and for each β parameter as $N(0, 20)$ so that the priors do not drive the inference.

5. Data Analysis and Significance of the Covariates

Tables II shows the posterior median for different model parameters and their 95% credible interval along with other percentile points. It is evident from this table that only Age Group interaction is marginally significant, while all other covariates are non significant. However we need to look carefully at the posterior distribution of all these model parameters. We have drawn highest density regions for all of the nine covariates and it looks clear from the density plot that some of the posterior distributions are bimodal. We have recomputed to figure out highest posterior density (HPD) credible interval in table III along with posterior mode(s). Four of the covariates have bimodal posterior distribution as depicted in figure 2. This seems very intuitive considering the natural groupings that may occur depending upon the binary nature of these categorical variables. From the table III looking at 90% HPD credible interval, covariates such as Age, Group, RB, RA, RH, SF, SIRS1 and SIRS4 are not statistically significant. At 90% confidence level only significant covariate is Age and Group interaction, which captures differential effect of age between trauma and control group. Clinically it has been hypothesized [26, 27] that age is an important factor along with the severity of the wound in the outcome of trauma. We have also analyzed the effect of age separately in trauma and control group. Posterior distribution for two different group is plotted in figure 3. 90% HPD credible interval for age in the trauma group is $[-2.29, 0.29]$ and $[-0.8, 28.58]$ for the control group. It follows that though age is not significant when analyzed separately in trauma and control group, it is an important covariate when considered as differential effect between the two groups.

Notably three of the covariates (RB, SF and SIRS4) have clear bimodal posteriors and they are indeed binary in nature (see Figure 2). Concerning the bimodality of these posteriors a conditional approach can be taken in which we can divide the data set into two groups. To illustrate the procedure, suppose we take SIRS4 covariate which is defined as SIRS4=1 if SIRS score is 4 and 0 otherwise. Suppose we perform SIRS4 is fixed to 1 for one subgroup and 0 to other subgroup and we perform conditional analysis of other covariates within the subgroup. Though this will remove the bimodality of the variable SIRS4, will not

necessarily do so for other binary covariates. To solve this a tree based approach can be taken in which we can divide the whole data set into maximal possible subgroups depending upon three categorical covariates. This will essentially create $8(= 2^3)$ subgroups all together. In our data analysis this solves bimodality issues (though not presented in this paper). However with very few subjects (only 12 in the present context), allocation per group is rather marginal (2-3 subjects per group only) and thus MCMC convergence becomes very unstable within some of the subgroups. We postulate that this could be avoided in future studies if sufficient number of subject allocation per subgroup is possible. We would like to thank an anonymous reviewer for pointing this out.

6. Conclusions

In this paper we have demonstrated usefulness of the covariate information in the development of metabonomic profile of acute trauma subjects. The model developed earlier by Ghosh et al. [1] was enhanced to include and decide significance of covariates between trauma and control group. Since main focus of this work is on determining significance of covariates associated with trauma following metabonomic profile, we have not shown any work related to abnormal presence/absence of ion intensity in mass spectrometer. Prediction of trauma status depending upon the abundance of ion intensity was shown with great details in Ghosh et al. [1]. One question still remains is whether proportional model framework is most appropriate in incorporating covariate information. We would like to note that Cox proportional model simplifies likelihood as we can separate out covariate information and mixture of beta part easily. This offers great convenience in MCMC computation when compared to additive hazard rate models. We also emphasize and realize the fact that along with m/z profile developed earlier based on the mixture of beta distribution, if we could incorporate covariate information, then there is always a better chance of prognosis and diagnosis of the trauma.

Acknowledgments

Funding for this research was partially provided by the National Institute of Health (R01 ES011630), a UCONN Faculty Large Grant and the Office of Naval Research (N00014-00-1-0792, N00014-99-1-0905 and N00014-99-1-0606). We would like to thank Grant V. Bochicchio and R. Adams Cowley of University of Maryland for their help in data collection. Special thanks also go to our collaborators Dr. David Grant and Dr. Dennis Hill both from the University of Connecticut.

Contract/grant sponsor: Publishing Arts Research Council; contract/grant number: 98-1846389

REFERENCES

1. Ghosh S, Grant DF, Dey DK, Hill DW. A Semiparametric Modeling Approach for the Development of Metabonomic Profile and Bio-Marker Discovery. *BMC Bioinformatics*. 2008; 9:38. [PubMed: 18215299]
2. Anderson RN, Smith BL. Deaths: leading causes for 2001. *National Vital Statist. Rep.* 2003; 52(9): 1–85.
3. Kuhls DA, Malone DL, McCarter RJ, Napolitano LM. Predictors of mortality in adult trauma patients: the physiologic trauma score is equivalent to the Trauma and Injury Severity Score. *Journal of American Coll. Surg.* 2002; 194:695–704. 2002.
4. Nicholson JK, Lindon JC, Holmes E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*. 1999; 29:1181–1190. [PubMed: 10598751]
5. Jansen JJ, Huub CJH, Hans FMB, Greef JVD, Smilde AK. Analysis of longitudinal metabolomics data. *Bioinformatics*. 2004; 20:2438–2446. [PubMed: 15087313]
6. Plumb RS, Stumpf CL, Gorenstein MV, Castro-Perez JM, Dear GJ, Anthony M, Sweatman BC, Connor SC, Haselden JN. Metabonomics: the use of electrospray mass spectrometry coupled to

- reversed-phase liquid chromatography shows potential for the screening of rat urine in drug development. *Rapid Commun. Mass Spectrom.* 2002; 16:1991–1997. [PubMed: 12362392]
7. Plumb RS, Stumpf CL, Granger JH, Castro-Perez J, Haselden JN, Dear GJ. Use of liquid chromatography/time-of-flight mass spectrometry and multivariate statistical analysis shows promise for the detection of drug metabolites in biological fluids. *Rapid Commun. Mass Spectrom.* 2003; 17:2632–2640. [PubMed: 14648901]
 8. Plumb RS, Granger J, Stumpf C, Wilson ID, Evans JA, Lenz EM. Metabonomic analysis of mouse urine by liquid-chromatography-time of flight mass spectrometry (LC-TOFMS): detection of strain, diurnal and gender differences. *Analyst.* 2003; 128:819–842. [PubMed: 12894816]
 9. Tibshirani R, Trevor H, Balasubramanian N, Scott S, Gongyi S, Albert K, Quynh-Thu L. Sample classification from protein mass spectrometry, by “peak probability contrasts”. *Bioinformatics.* 2004; 20:3034–3044. [PubMed: 15226172]
 10. Randolph TW, Yasui Y. Multiscale Processing of Mass Spectrometry Data. *Biometrics.* 2006; 62:589–597. [PubMed: 16918924]
 11. Kwon DW, Tadesse MG, Sha N, Pfeiffer RM, Vannucci M. Identifying Biomarkers from Mass Spectrometry Data with Ordinal Outcome. *Cancer Informatics.* 2006; 3:19–28. [PubMed: 19455232]
 12. Randolph TW, Tasui Y. Multiscale processing of mass spectrometry data. *Biometrics.* 2006; 62:589–597. [PubMed: 16918924]
 13. Morris J, Combes RK, Baggerly K, Kobayasi R. Feature extraction and quantification for mass spectrometry data in biomedical application using the mean spectrum. *Bioinformatics.* 2005; 21:1764–1775. [PubMed: 15673564]
 14. Combes RK, Morris SM, Hu J, Edmonson RS, Baggerly AK. Serum proteomics profiling - a young technology begins to mature. *Nature Biotechnology.* 2005; 23:291–292.
 15. Yutaka Y, McLerran D, Bao-Ling A, Marcy W, Thornquist M, Ziding F. An Automated Peak Identification/Calibration Procedure for High-Dimensional Protein Measures From Mass Spectrometers. *Journal of Biomedicine and Biotechnology.* 2003; 4:242–248.
 16. Ghosh S, Dey D, Hill D, Grant FD. Statistical Approach to Metabonomic Analysis of Rat Urine Following Surgical Trauma. *Journal of Chemometrics.* 20:87–98.
 17. Kazmi AS, Ghosh S, Shin D-G, Hill DW, Grant FD. Alignment of high resolution mass spectra: Development of a heuristic approach for metabolomics. *METABOLOMICS.* 2006; 2:75–83.
 18. Diaconis P, Ylvisaker D. Quantifying prior opinions. *Proceedings of the Bayesian Statistics.* 1985; 2:133–156.
 19. Gelfand A, Mallick BK. Bayesian analysis of proportional hazards model built from monotone functions. *Biometrics.* 1995; 51:843–852. [PubMed: 7548703]
 20. Cox DR. Regression models and life tables. *Journal of the Royal Statistical Society.* 1972; 21:411–421.
 21. Geman S, Geman D. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Trans. on Pattern Anal. and Mach. Intel.* 1987; 6:721–741.
 22. Gelfand AE, Smith AFM. Sampling-based approaches to calculating marginal densities. *Journal of American Statistical Association.* 1990; 85:398–409.
 23. Watters JM, Redmond ML, Desai D, March RJ. Effects of age and body composition on the metabolic responses to elective colon resection. *Ann Surg.* 1990; 212:213–220. [PubMed: 2198002]
 24. Johnston CJ, Rubenfeld GD, Hudson LD. Effect of Age on the Development of ARDS in Trauma Patients. *Chest.* 2003; 124:653–659. [PubMed: 12907556]
 25. Geweke J, Bernardo JM, Berger JO, Dawid AP, Smith AFM. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. *Proceedings of the Bayesian Statistics.* 1992; 4
 26. McGraw BL, Cole RR. Pediatric maxillofacial trauma. Age-related variations in injury. *Arch. Otolaryngol. Head Neck Surg.* 1990; 116:41–46. [PubMed: 2294939]
 27. Munro PT, Smith RD, Parke TRJ. Effect of patients' age on management of acute intracranial haematoma: prospective national study. *BMJ.* 2002; 325:1001. [PubMed: 12411357]

Figure 1.
A sample ES/TOF mass spectrometry spectrum.

Figure 2.

Posterior density plot for all nine covariates. HPD credible intervals are shown as colored (shaded) horizontal bar in each plot. Green bar (top most and most lightly shaded) for 50%, red bar (in the middle) for 90% and blue bar (lowest and most darkly shaded) indicating 95% HPD credible interval

Figure 3. Posterior density plot for Age effect in Control (left) and Trauma (right) group. HPD credible intervals are shown as colored (shaded) horizontal bar in each plot. Green bar (top most and most lightly shaded) for 50%, red bar (in the middle) for 90% and blue bar (lowest and most darkly shaded) indicating 95% HPD credible interval.

Table 1

Covariate Data

Patient ID	Age	Sex	Race	SIRS	Control ID	Age	Sex	Race	SIRS
DG1	19	F	Caucasian	4	DG62	48	M	Caucasian	0
DG3	61	F	Caucasian	4	DG63	36	M	Caucasian	0
DG12	28	M	Hispanic	4	DG64	41	M	Caucasian	0
DG31	88	M	Hispanic	4	DG66	27	M	Asian	0
DG41	20	M	Black	0	DG70	51	F	Caucasian	0
DG51	47	M	Black	1	DG72	49	F	Asian	0

Table II

Posterior Percentiles

Variable	2.5%	5%	25%	Median	75%	97.5%
Age*Group	-5.0311	-3.5568	-1.4339	-0.7642	-0.2799	0.0053
Group	-7.2176	-5.8586	-3.7103	-2.8155	-1.638	3.2012
Age	-4.6492	-3.2171	-1.3523	-0.7225	-0.3073	0.0252
RB(1 if black, 0 ow)	-22.6693	-18.5657	-4.1276	6.8520	19.0196	28.4941
RA(1 if asian, 0 ow)	-5.8132	-2.1543	8.9757	15.7111	23.1270	39.5602
RH(1 if hispanic, 0 ow)	-18.9399	-15.3852	-5.7278	1.5810	8.6897	23.6926
SF(1 if female, 0 ow)	-8.1517	-4.8753	10.6250	22.0421	34.3926	44.9174
SIRS1(1 if SIRS=1, 0 ow)	-18.5895	-16.0109	-6.3276	0.6933	8.3665	23.2918
SIRS4(1 if SIRS=4, 0 ow)	-8.8384	-5.1650	11.2368	21.7692	34.0053	43.5677

Table III

HPD Credible Intervals (95%, 90% and 50%) and Posterior Mode(s)

Variable	95%	90%	50%	Mode
Age*Group	[-4.13, -3.96]U[-3.24, 0.4]	[-2.45, -0.035]	[-0.82, -0.01]	-4.05, -0.245
Group	[-7.20, 3.44]	[-5.83, 0.11]U[0.41, 2.93]	[-3.97, -1.94]	-3.1, 1.34
Age	[-3.26, 0.41]	[-2.33, 0.29]	[-0.83, -0.02]	-0.319
RB(1 if black, 0 ow)	[-20.58, 30.36]	[-15.99, 28.76]	[-4.54, 4.83]U[14.56, 24.60]	-0.79, 19.68
RA(1 if asian, 0 ow)	[-6.08, 38.88]	[-2.22, 35.55]	[7.90, 21.91]	16.363
RH(1 if hispanic, 0 ow)	[-18.74, 24.22]	[-16.33, 19.76]	[-5.65, 8.78]	1.852
SF(1 if female, 0 ow)	[-13.05, 50.50]	[-7.55, 46.62]	[11.49, 21.25]U[28.79, 40.13]	34.844, 16.825
SIRS1(1 if SIRS=1, 0 ow)	[-25.87, 31.24]	[-18.78, 23.26]	[-7.62, 6.63]	-0.6205
SIRS4(1 if SIRS=4, 0 ow)	[-13.87, 48.82]	[-6.49, 45.95]	[12.05, 21.16]U[29.14, 39.53]	16.56, 35.076