Bayesian Analysis (2009) 4, Number 3, pp. 523–538

# Prediction of Pregnancy: A Joint Model for Longitudinal and Binary Data

Julie Horrocks* and Marianne J. van Den Heuvel†

**Abstract.** We consider the problem of predicting the achievement of successful pregnancy, in a population of women undergoing treatment for infertility, based on longitudinal measurements of adhesiveness of certain blood lymphocytes. A goal of the analysis is to provide, for each woman, an estimated probability of becoming pregnant. We discuss various existing approaches, including multiple t-tests, mixed models, discriminant analysis and two-stage models. We use a joint model developed by Wang et al. (2000), consisting of a linear mixed effects model for the longitudinal data and a generalized linear model (glm) for the primary endpoint, (here a binary indicator of successful pregnancy). The joint longitudinal/glm model is analogous to the popular joint models for longitudinal and survival data. We estimate the parameters using Bayesian methodology.

**Keywords:** joint model, mixed linear model, generalized linear model, longitudinal data, binary data

## 1 Introduction

The medical community defines infertility as the inability to establish a viable pregnancy over a twelve month period. About 8.5% of Canadian couples fall into this category and most of these seek medical assistance (Royal Commission on New Reproductive Technologies Final Report, 1994). It should be recognized that new reproductive technologies and assisted conception treatments do not cure infertility, but rather circumvent it. Twenty percent of infertility cases can be accounted for by sexually transmitted diseases, while smoking, delayed childbearing, harmful environmental agents, alcohol abuse, stress and eating disorders have also been shown to have roles in development of infertility. A significant number of cases of infertility cannot be attributed to any specific cause and these couples often access every available technology in an effort to achieve a family. The incidence and costs associated with the treatment of infertility have become issues of public interest over the past twenty years, particularly as the success rate hovers around 30%. Thus it is of both public and private interest for physicians and patients to have some indication of the chance of success in attempting a plan of treatment.

Successful pregnancy is dependent upon the maternal immune system recognizing and tolerating the growth and development of a fetus, an entity which is foreign to maternal cells. We focus here on $CD56^{bright}$ cells, a type of lymphocyte which is rare

*Department of Mathematics and Statistics, University of Guelph, Guelph, Ontario, Canada, mailto:jhorrock@uoguelph.ca

†Department of Biomedical Sciences, University of Guelph, Guelph, Ontario, Canada, mailto:mvandenh@uoguelph.ca

 DOI:10.1214/09-BA419

in blood and the non-pregnant uterus (representing less than 1% of circulating lymphocytes) but abundant in the uterus at the time of implantation and even more prevalent through the first three months of pregnancy. $CD56^{bright}$ cells are thought to have significant roles in promoting the development of a thick and well-nourished uterine lining by producing factors which encourage the growth of new maternal blood vessels. A reduction in the numbers of these cells is postulated to impair blood vessel development, resulting in implantation failure or inadequate development of the placenta.

In this article, we consider the problem of predicting the achievement of successful pregnancy, in a population of women undergoing treatment for infertility, based on longitudinal measurements of adhesiveness of $CD56^{bright}$ cells. Thus, for each woman, we have a set of continuous longitudinal measurements, and a primary endpoint (a single binary variable indicating whether the woman subsequently became pregnant). A goal of the analysis is to provide, for each woman, an estimated probability of becoming pregnant and an estimate of the precision of this estimate (preferably an interval estimate). This probability and associated interval could be used for counseling infertile couples about whether they could benefit from costly treatment for infertility or for providing couples in treatment with empirical data about their chances of success if they continue treatment.

The longitudinal measurements in this application are subject to measurement error, as well as other sources of variability. For instance, adhesion is measured by dropping dye-labeled lymphocytes from individual patients onto frozen tissue sections of murine uterine endometrium mounted on microscope slides, allowing these to interact for 30 minutes, washing away non-adherent cells and counting those lymphocytes that remain. Rather than count all the cells on a slide, a sample of high-power magnification microscope fields is taken and the number of cells in each field is counted and then averaged across all sampled fields on the slide. This process is subject to substantial error, arising both from the sampling and human error due to fatigue, inattention, etc. In addition, for each woman, there is day-to-day variability in the adhesive properties of her blood lymphocytes. We would like to make inference about the effect of the "true" adhesion measurements on the probability of pregnancy, after removing these sources of error. When the explanatory variable is measured with error, it is well known that regression estimates from conventional methods (e.g. linear regression, logistic regression) are biased.

To address these challenges, we use a parametric joint model, consisting of a linear mixed effects submodel for the longitudinal adhesion data and a generalized linear submodel for the primary endpoint. The two submodels are linked by the random coefficients which appear in both submodels. This model was developed by Wang et al. (2000). While these authors used frequentist methods to estimate the parameters, here we estimate the parameters using a Bayesian approach. The methodology is similar to Bayesian joint models for longitudinal and survival data espoused by Guo and Carlin (2004).

## 2 Controlled Ovarian Hyperstimulation (COH) Study

Women of reproductive age who had been previously diagnosed as infertile were recruited to participate in this study by physicians in the Reproductive Endocrinology and Infertility Program at London Health Sciences Center, London, Ontario, Canada. The study was approved by the health sciences research ethics board at University of Western Ontario, and all recruited patients were fully informed of the potential risks and signed consent forms before participation. The subset of women discussed in this article was enrolled in a controlled ovarian hyperstimulation (COH) protocol for in vitro fertilization. These women underwent daily injections of follicle-stimulating hormones to increase the number of ova (egg)-containing follicles developing in their ovaries. The number and size of developing ova and the amount of hormones in their blood were monitored approximately every other day. When a sufficient number of ova (at least one) were of mature size, an injection of human choriogonadotropin was administered to effect their final maturation. The following day, the mature oocytes were aspirated from the ovaries using ultrasound guided needles, mixed with her partners sperm in a petri dish and allowed to incubate for 3 days. They were then assessed for fertilization and quality as evidenced by cell division and intact cell membranes. Up to three high quality embryos were then transferred back to the patient's uterus. Women subsequently underwent serum pregnancy tests at 18 days after embryo transfer. If the test was positive, they were examined by ultrasound at 40 days post-transfer to confirm the presence of a viable gestational sac and fetal heart.

Exclusion criteria included insufficient response to hormone therapy (such that no follicles were found to develop) or failure of the sperm to fertilize the eggs (so that no embryos were available for transfer). For statistical analysis, the day of embryo transfer was designated time zero, while the day of oocyte pickup and human choriogonadotropin treatment were designated analysis times -1 and -2, respectively. Blood samples taken before these days were assigned analysis times of -3, -4, -5, and -6. Because the menstrual cycle lengths of individual women differed, not all women had measurements at all time points. The blood samples collected during the monitoring stage were further analyzed (within 3 hours of collection) as follows. Lymphocytes were separated and $CD56^{bright}$ cells were dye-labeled using standard methods as previously described by van Den Heuvel et al. (2005b). $CD56^{bright}$ cells were then layered onto cryosections of mouse uterine tissues. After 30 min of rotation at 112 rpm in a cold chamber, non-adherent cells were rinsed off, and the tissue was fixed. Two independent researchers then counted the number of adherent $CD56^{bright}$ cells in 25 high power fields (x400). The average number of adherent cells, measured up to 6 times for each woman, became the longitudinal data in our analysis. Data were log transformed, which stabilized the variance and led to a more normal distribution.

## 3 Related Approaches

In this section, we survey existing methods of analysis for continuous longitudinal data collected on two groups of individuals. Assuming that all individuals are measured at

the same points in time, a simple approach would be to perform separate t-tests comparing the two groups (e.g. pregnant/not pregnant) at each time point, using Bonferroni or some other procedure to adjust for multiple comparisons. A more sophisticated and efficient approach (that does not require common measurement times across all individuals) is a linear mixed effects model, with the continuous measurement as the response variable, and with time and group membership and their interaction as explanatory variables. Correlation between continuous measurements over time from the same individual is accounted for by allowing the within-person correlation matrix to have non-zero off-diagonal elements. This model, introduced by Laird and Ware (1982), allows differing random slopes and intercepts for each person. This was the approach taken in van Den Heuvel et al. (2005b). Mixed-model F-tests indicated a significant difference between the two groups of women with respect to their longitudinal adhesion measurements.

For our application, while both the multiple t-tests and mixed model approaches allow us to say whether or not adhesion was different in the two groups, they do not allow us to predict whether a woman will become pregnant or to estimate the probability of pregnancy. In both approaches, the longitudinal data are treated as the response, and the binary variable (group membership) is treated as the explanatory variable. This is unpalatable from a causal point of view, because the binary variable (pregnant or not) is the object of inference and was measured later in time than the longitudinal adhesion data. A better approach would be to treat the binary variable as the response, and the longitudinal data as explanatory information.

Logistic regression is commonly used to predict a binary outcome from continuous and/or discrete explanatory variables measured at a single point in time. However, when the explanatory information is longitudinal, more complicated strategies are needed. For instance, a summary measure that collapses the longitudinal information over time (e.g. overall average, change score, slope, intercept, maximum, minimum, achievement of a threshold, measurement on a particular day, etc.) could be used as a covariate in a logistic regression. However, it may be difficult to decide which summary measure to use, and as there are many candidate measures, there is perhaps a problem of multiple comparisons. Furthermore, if the longitudinal measurements are non-linear in time, it may be difficult to come up with a single summary measure. Generalized Estimating Equations (GEE) can be used to estimate the effect of a time-varying covariate on a time-varying multi-dimensional response; in our application however, the response has a single dimension, so GEE is not appropriate.

If the goal is to merely classify individuals into groups, a classical approach is discriminant analysis, where each continuous longitudinal measurement is treated as a separate explanatory variable. This requires that the measurement times be the same for every person and furthermore does not exploit the correlation between multiple measurements made on the same individual over time.

Another possible approach is a two-stage model: first a linear mixed effects model is fit to the longitudinal data and predicted values of the random slopes and intercepts are computed; then a logistic regression model is fit to the binary outcome, using the

random slopes and intercepts as covariates. Other covariates can be included in both models. Although the two-stage model is easily fit with existing software, it ignores the error inherent in the predicted random effects which are used as covariates in the logistic regression model, and thus is not recommended. We will return to this model below.

While joint models for longitudinal and time-to-event data are ubiquitous and have spawned a vast literature, joint models for longitudinal and binary outcomes are less common. This is perhaps surprising, given the pervasiveness of binary outcomes, especially in medicine and epidemiology, where they are used to indicate the presence of a disease. However, many binary outcomes are formed by downgrading the information contained in time-to-event outcomes. For instance, a binary variable indicating whether a patient survives five years after the start of treatment is a common outcome in medicine, but the exact survival time (possibly censored) is more informative. In this situation, a joint model for longitudinal and time-to-event data would be more informative than a joint model for longitudinal and binary data.

However, in other applications the exact timing of the event is not of interest. For instance, for the application considered in this paper, the time of pregnancy within a single menstrual cycle is clinically irrelevant, as the time of conception is biologically determined to within a few days. Thus, a joint longitudinal/binary model is appropriate.

Joint models for longitudinal and time-to-event data are well-developed, and have many similarities to the joint longitudinal/generalized linear models (glm) considered here. Henderson et al. (2000) propose a class of joint models, based on a latent bivariate Gaussian process, with a semi-parametric submodel for an event time. Guo and Carlin (2004) illustrate the fitting of these models with the readily available software WinBUGS, and compare them to separate modeling of the longitudinal and time-to-event data.

Joint models for continuous longitudinal data and binary data have been used to model data subject to selection or informative dropout in both econometrics (dating back at least to Heckman (1979)) and biostatistics (for instance Wu and Bailey (1989); Wu and Carroll (1988)). In these papers, the binary variable is an indicator for possibly informative dropout and the main focus of inference is the parameters associated with the continuous longitudinal variable. Gueorguieva and Agresti (2001), building on the work of Catalano and Ryan (1992) develop a model for clustered binary and continuous outcomes. Here the effect of covariates on both the binary and continuous outcomes is the object of inference; the effect of the longitudinal data on the binary outcomes is not discussed. Longitudinal discriminant analysis has recently been discussed by several authors, including Tomasko et al. (1999), Marshall and Baron (2000), and Wernecke et al. (2004). Here the goal is to merely classify individuals into groups, and probabilities of group membership are not given. Roy and Khattree (2005) also discuss longitudinal discriminant analysis and give expressions for parameter estimates, including probabilities of group membership. However, expressions for standard errors are not given. Their method however requires that the measurement times be the same across individuals. The joint longitudinal/generalized linear model used in this paper was introduced by

Wang et al. (2000) and further developed by Huang and Wang (2001), Wang and Huang (2001); Li et al. (2004); Li et al. (2007b); and Li et al. (2007a).

With the exception of Guo and Carlin (2004), none of these authors used Bayesian methodology to estimate the model parameters. A recent paper that does use Bayesian methods is that by de la Cruz-Mesia and Quintana (2007). These authors classify observations into two groups based on a nonlinear random effects model and augment the model with predicted probabilities.

## 4 Joint Model

The joint model consists of two submodels: a mixed effects model for the continuous longitudinal data, and a generalized linear model for the primary outcome (which in our application is a binary indicator of pregnancy). Let $\boldsymbol{Y_i} = (Y_{i1}, Y_{i2}, \ldots, Y_{im_i})$ represent the continuous longitudinal measurements for individual $i$, $i = 1, 2, \ldots, n$, at measurement times $t_{i1}, \ldots, t_{im_i}$. The submodel for the longitudinal data can be written as

$$\boldsymbol{Y_i} = \boldsymbol{X_i}\boldsymbol{\alpha} + \boldsymbol{Z_i}\boldsymbol{U_i} + \boldsymbol{\epsilon_i}$$

where $\boldsymbol{\alpha}$ is a $p$ x 1 vector of unknown fixed-effects parameters, $\boldsymbol{X_i}$ and $\boldsymbol{Z_i}$ are known design matrices of dimension $m_i$ x $p$ and $m_i$ x $q$ respectively, $\boldsymbol{U_i}$ is a $q$ x 1 vector of unknown random-effects and $\boldsymbol{\epsilon_i} = (\epsilon_{i1}, \ldots, \epsilon_{im_i})$ is a vector of measurement errors for individual $i$. We assume that

$$\boldsymbol{U_i} \sim N(\mathbf{0}\ , \boldsymbol{\Sigma}) \quad \text{and} \quad \boldsymbol{\epsilon_i} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I_{m_i}})$$

and that $\boldsymbol{U_1}, \ldots, \boldsymbol{U_n}, \boldsymbol{\epsilon_1}, \ldots, \boldsymbol{\epsilon_n}$ are independent, where $\boldsymbol{\Sigma}$ is a $q$ x $q$ covariance matrix for the random effects and $\boldsymbol{I_k}$ indicates the $k$ x $k$ identity matrix. For example, letting the $j^{th}$ row of $\boldsymbol{Z_i}$ equal $(1, t_{ij})$ for $j = 1, \ldots, m_i$, so that $\boldsymbol{U_i} = (U_{i1}, U_{i2})^T$ corresponding to the random intercept and slope for subject $i$, and with

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

we obtain a random slopes and intercepts model (see Laird and Ware 1982).

The submodel for the primary endpoint, $R_i$, is assumed to be a generalized linear model, with $\mu_i = E(R_i)$, $g(\mu_i) = \boldsymbol{V_i}^T\boldsymbol{\beta} + \boldsymbol{U_i}^T\boldsymbol{\gamma}$ and $\text{Var}(R_i) = \Phi\nu(\mu_i)$, where g(.) is a known function, often referred to as a link function, $\boldsymbol{U_i}$ is the $q$ x 1 vector of random effects from the previous submodel, $\boldsymbol{V_i}$ is a vector of observed covariates, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of unknown parameters, $\nu(.)$ is a known variance function, and $\Phi$ is a scale parameter. The distribution of $R_i$ is assumed to belong to the exponential family of distributions. For our application, $R_i$ is a binary indicator of successful pregnancy and follows a Bernoulli($p_i$) distribution where $\mu_i = p_i$ is the probability that individual $i$ becomes pregnant and $\nu(p_i) = p_i(1 - p_i)$ with $\Phi = 1$. We use a logit link function, although probit and complementary log-log are common choices.

The two submodels are linked by their common random effects $\boldsymbol{U_i}$. Consequently, the vector parameter $\boldsymbol{\gamma}$ measures the strength of the association between the two models. If $\boldsymbol{\gamma} = \mathbf{0}$, then the longitudinal data has no predictive ability for the primary endpoint.

### 4.1 Estimation

We use Bayesian methods to estimate the parameters of the joint longitudinal/glm model, as software is readily available. Specifically we used the software package WinBUGS, which estimates the posterior distribution of the parameters using Markov Chain Monte Carlo (MCMC) methods. Of course, prior distributions must be specified for the parameters of the model. We use proper but vague prior distributions, since prior knowledge is limited and elicitation is difficult for this application. These prior distributions have large variances, and so have little effect on the final model, which is largely determined by the data. For the regression parameters ($\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$), we use multivariate normal priors with mean $\mathbf{0}$ and a diagonal covariance matrix with large diagonal elements. In other words, we specified independent normal priors with large variances for the regression parameters. An inverse gamma prior was used for the measurement error variance $\sigma^2$. A multivariate Wishart distribution can be used for the covariance matrix $\boldsymbol{\Sigma}$ of the random effects, or, for models with a single random effect, an inverse gamma can be used for the variance of this random effect.

### 4.2 Two-Stage Model

A two-stage version of the model can be fit with readily available software for classical statistical methods, as follows. First, a linear mixed model is fit to the longitudinal data $\boldsymbol{Y_i}$, and predicted values $\widehat{\boldsymbol{U}}_{\boldsymbol{i}}$ for the random effects are computed. Then, the predicted values of the random effects are used as covariates in a generalized linear model for the primary endpoint. This two-stage method is essentially the regression calibration approach studied by Wang et al. (2000, section 4).

The longitudinal model can be fit, for instance, using SAS Proc Mixed or the Splus function nlme. Then the estimated best linear unbiased predictors (BLUPs) for the random effects can be included as covariates in a logistic regression, which can be fit using SAS Proc Logistic or Proc Genmod or using the Splus function glm. However this two-stage model is not recommended as error in the estimation of the random coefficients is not accounted for, and it can result in biased inference (see Wang et al. 2000). We have included the two-stage model here for comparison.

## 5 Analysis of COH data

In the COH protocol, adhesion of blood lymphocytes was measured up to 6 times, on 18 women undergoing in-vitro fertilization. Figure 1 shows log adhesion for the 7 women who became pregnant (top panel) and the 11 women who did not become pregnant (bottom panel). We present here a simplified model with only random intercepts and time as

the single covariate. A full analysis would likely include other covariates. While a random slopes and intercepts model would usually be expected to give better predictions, for our data such a model resulted in numerous trap errors in WinBugs. Investigations of the longitudinal submodel indicated that the variance of the random slopes is very close to zero, so that a random slopes and intercepts model is not supported by these data.

Let $Y_{ij}$ represent the log adhesion measurement on the $i^{th}$ individual ($i = 1, 2, \ldots, n$) at time $t_{ij}$, where $t_{ij}$ is the $j^{th}$ day before embryo transfer for the $i^{th}$ individual, and let $p_i$ represent the probability that the $i^{th}$ individual becomes pregnant.

A joint model for these data consists of a random-intercept submodel for the longitudinal adhesion measurements and a logistic model for the response:

$$Y_{ij} = \alpha_0 + \alpha_1 t_j + U_i + \epsilon_{ij}$$

$$\text{logit}(p_i) = \beta_0 + \gamma_1 U_i$$

where

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

and

$$U_i \sim N(0, \sigma_1^2).$$

In this study individuals were measured at approximately the same times in their menstrual cycle, so that $t_{ij} = t_j$ for $i = 1, 2, \ldots, n$; however some times were missing for some individuals. The $U_i$ are the random intercepts and the parameters to be estimated are $\alpha_0, \alpha_1, \sigma^2, \beta_0, \gamma_1$ and $\sigma_1^2$.

In WinBUGS, the normal distribution is parameterized with mean and precision parameters, where the precision is the inverse of the variance. For the parameters $\alpha_0, \alpha_1, \beta_0$ and $\gamma_1$, we used normal priors with mean 0 and precision 0.01 (ie variance 1/0.01=100). For variance parameters, denoted generically as $\sigma^2$, a common practice in WinBUGS is to define $\tau = 1/\sigma^2$ and specify that this has a gamma prior distribution, so that the prior distribution for $\sigma^2$ is inverse gamma. Both $1/\sigma^2$ and $1/\sigma_1^2$ were given gamma(0.1,0.1) prior distributions. We used the Brooks-Gelman-Rubin (BGR) diagnostic tool in WinBUGS to check that the model had converged after 20000 iterations.

At each iteration of the Markov chain, WinBUGS calculates estimates of all the parameters and any missing values, as well as predicted random intercepts $\widehat{U}_i$, $i = 1, 2, \ldots, n$. The predicted probability of pregnancy for person $i$, $\widehat{p}_i$, can be calculated as

$$\widehat{p}_i = \exp(\widehat{\eta}_i)/(1 + \exp(\widehat{\eta}_i))$$

where

$$\widehat{\eta}_i = \widehat{\beta}_0 + \widehat{\gamma}_1 \widehat{U}_i.$$

The results from the model are shown in Table 1. There is evidence that the parameter $\gamma_1$ is different from 0, since the 95% credible interval does not include 0. Thus there is evidence that a linkage exists between the two models and that the longitudinal

adhesion data can be used to predict pregnancy. The fact that the sign of $\gamma_1$ is positive indicates that individuals with higher random intercepts for the adhesion measurements tend to have a higher probability of becoming pregnant. However the credible interval is very wide (1.2, 19.8) and this parameter is less well-estimated than the other parameters of the model. This could be due to complete or quasi-separation induced by the estimated $U$ parameters in the logistic regression model. Briefly, (quasi-) separation occurs when group membership can be (nearly) perfectly predicted by the model.

To check the fit of the model, we plotted residuals from the longitudinal submodel against time. For mixed models, numerous types of residuals have been defined. The marginal residuals are defined as

$$r_{ij}^m = y_{ij} - \boldsymbol{x_i^T}\widehat{\boldsymbol{\alpha}}$$

where $\boldsymbol{x_i^T} = (x_{ij1}, \ldots, x_{ijp})$ and conditional residuals (conditional on random effects) as

$$r_{ij}^c = y_{ij} - \boldsymbol{x_i^T}\widehat{\boldsymbol{\alpha}} - \boldsymbol{z_i^T}\widehat{\boldsymbol{U_i}}$$

where $\boldsymbol{z_i^T} = (z_{ij1}, \ldots, z_{ijq})$. Figure 2 shows plots of studentized residuals, which are the above residuals, divided by their respective estimated standard errors. These plots did not indicate any systematic lack of fit. As another approach, we tested the linear submodel for the longitudinal data against more complicated models. Models with higher-order polynomials (up to the fourth degree) for the mean response did not improve the fit over the linear model for the mean, as measured by AIC and BIC. A model that allowed different means at each time point also did not fit better than the simple linear model.

The longitudinal submodel assumes that individual-level random effects, $U_i$, have a normal distribution. It is well known (see for instance Verbeke and Lesaffre (1997); Heagerty and Kurland (2001)) that if this assumption is not satisfied, inferences can be compromised, and this problem also occurs with joint models (Li et al. 2004). We examined histograms and qqplots of the random effects and conducted a Shapiro-Wilks test which did not indicate signifcant deviation from normality (p=0.6725). For a joint model which does not assume any distribution for the random effects, see Li et al. (2004).

Furthermore, the model used here assumes that the measurement errors, $\epsilon_{ij}$, are independent and identically distributed (i.i.d.) from a normal distribution. While the assumption of a normal distribution is wide-spread in the literature and often suitable (perhaps after a transformation), the assumption of i.i.d. errors should be carefully checked. To test the i.i.d measurement error assumption, we fit a longitudinal submodel with AR(1) covariance structure. This did not fit the data as well as the random intercept model, as measured both by AIC and BIC. We also fit a model where the variance was allowed to be different in the two groups. Again this model did not fit better than the simple random intercepts model.

The joint model also produces predicted probabilities of becoming pregnant (see Table 1). Here the posterior mean is used as a point estimate of the predicted probability

and 95% credible intervals (CI) are also shown. The intervals are wide, but would presumably be narrower with a larger data set. If we take a predicted probability of 50% or higher as predicting pregnancy, then the predictive accuracy of the model was 83%, i.e. the pregnancy outcomes for 15 out of the 18 women were correctly predicted by the model. Specifically, 5 out of 7 women were correctly predicted to become pregnant, and 10 out of 11 were correctly predicted to remain not pregnant. These calculations were made by using the same data to fit and test the model and so are overly optimistic. As an alternative, cross-validation was performed by leaving one response out, re-fitting the model, and predicting the response for the omitted observation. With cross-validation, the predictive accuracy was estimated as 77.8%.

The model predictions can be used to counsel women who participated in the study, but did not achieve pregnancy during the study. For instance, the joint model predicts that the probability of pregnancy for woman 1 is 0.657 (95% CI=(0.143, 0.980)). On the basis of these results, woman 1 may decide to continue with fertility treatments since her point estimate is greater than 0.5 and the upper limit of her 95% CI is close to 1. On the other hand, woman 2 may decide to forgo future treatments as her point estimate is only 0.053 with 95% CI (0.000,0.313). However, the decision to continue with further treatments is very emotional and these model-based probabilities will be only a small factor in the decision-making.

The women in this study were already undergoing COH treatment for infertility. However the adhesion of $CD56^{bright}$ cells can also be studied in women who are not currently undergoing treatment for infertility (see van Den Heuvel et al. 2005a) but who may be deciding whether or not to attempt IVF treatment for infertility.

Table 1 also shows estimates and 95% confidence intervals from a two-stage model, fit using SAS Proc Mixed and Proc Logistic. The parameter estimates are similar for both the joint and two-stage approaches and the interval estimates for the mixed submodel are similar in width. However for most parameters of the logistic model, the two-stage approach gives narrower interval estimates. This is presumably because the two-stage model does not account for the error in the estimates of the random effects. As previously stated, the estimates from the two-stage model are known to be biased.

## 6 Conclusions

In this article, we have used a joint longitudinal/glm model developed by Wang et al. (2000), to predict pregnancy in a group of women undergoing treatment for infertility, based on longitudinal adhesion measurements. This procedure has the potential to target therapy to couples who would have a high likelihood of success in treatment, while concurrently saving couples unlikely to benefit from the emotional and financial costs involved. While the results presented here are promising, further research is needed to study the sensitivity and specificity of the adhesion model on a broader cohort of women.

It is easy to envision similar applications. For instance, the incidence of type II

diabetes is increasing in western societies. It is largely preventable by early medical intervention which encourages healthy weight and activity levels. The joint model used here may be useful in monitoring waistlines and/or body mass index of overweight people and used to predict which individuals are at high risk of developing diabetes in the next year. Alternatively, the model could use serial blood pressure measurements to predict an individual's risk of stroke or heart disease.

In this application, the measurement times were approximately the same for all individuals, although some missing values occurred. In many applications, measurement times would vary across individuals but this would present no problem to either the joint or two-stage model, as the mixed linear submodel can handle unequal measurement times. Smoothing splines or functional data analysis could also be used for the longitudinal submodel. Registration of the longitudinal profiles is also of interest. It is possible that some women have shorter menstrual cycles than others, and so perhaps a time scale other than day of menstrual cycle would be more appropriate. Again functional data analysis may have something to offer in this respect.

We have estimated the parameters using a Bayesian approach. The advantage of a Bayesian analysis is that it provides exact inferences, versus the asymptotic approximations on which maximum likelihood estimation is based. Also it can easily be fit using available software (WinBUGS). Furthermore, specifying a prior distribution for the parameters gives the researcher the opportunity to incorporate any existing information into the model. If little or nothing is known about a parameter, a non-informative prior can be used.

This model has many similarities to joint models for longitudinal and time-to-event data. A binary variable can always be constructed by downgrading a time-to-event variable; for instance a binary variable indicating 5-year-survival can be constructed from a variable recording time from treatment until death. However the time of death is generally more informative and where possible this should be used. Of course, there are some applications, such as the one considered here, where a binary response is required.

| | Joint Model | | Two-Stage Model | |
|---|---|---|---|---|
| Parameter | Posterior Mean | 95% CI | Estimate | 95% CI |
| Linear Mixed Effects Submodel | | | | |
| $\alpha_0$ | 1.389 | (1.142, 1.650) | 1.395 | ( 1.145, 1.645) |
| $\alpha_1$ | 0.045 | (0.000, 0.105) | 0.047 | (-0.012, 0.107) |
| $\sigma_1^2$ | 0.137 | (0.054, 0.297) | 0.109 | ( 0.014, 0.204) |
| $\sigma^2$ | 0.136 | (0.092, 0.198) | 0.124 | ( 0.077, 0.170) |
| Logistic Submodel | | | | |
| $\beta_0$ | -0.810 | (-3.157, 1.149) | -0.564 | (-1.672, 0.545) |
| $\gamma_1$ | 7.599 | (1.134, 19.760) | 4.462 | (-0.228, 9.151) |
| $p_1$ | 0.657 | (0.143, 0.980) | 0.806 | ( 0.314, 0.974) |
| $p_2$ | 0.053 | (0.000, 0.313) | 0.049 | ( 0.003, 0.487) |

Table 1: Joint Model and Two-Stage Model for the Adhesion Data

## References

Catalano, P. J. and Ryan, L. M. (1992). "Bivariate Latent Variable Models for Clustered Discrete and Continuous Outcomes." *JASA*, 87: 651–658. 527

de la Cruz-Mesia, R. and Quintana, F. A. (2007). "A model-based approach to Bayesian classification with applications to predicting pregnancy outcomes from longitudinal beta-hCG profiles." *Biostatistics*, 8: 228–238. 528

Gueorguieva, R. V. and Agresti, A. (2001). "A Correlated Probit Model for Joint Modeling of Clustered Binary and Continuous Responses." *JASA*, 96: 1102–1112. 527

Guo, X. and Carlin, B. P. (2004). "Separate and Joint Modeling of Longitudinal and Event Time Data Using Standard Computer Packages." *The American Statistician*, 58: 1–9. 524, 528

Heagerty, P. J. and Kurland, B. F. (2001). "Misspecified maximum likelihood estimates and generalized linear mixed models." *Biometrika*, 88: 973–985. 531

Heckman, J. J. (1979). "Sample selection bias as a specification error." *Econometrica*, 47: 153–161. 527

Henderson, R., Diggle, P. J., and Dobson, A. (2000). "Joint Modelling of Longitudinal Measurements and Event Time Data." *Biostatistics*, 1: 465–480. 527

Huang, Y. and Wang, C. Y. (2001). "Consistent functional methods for logistic regression with errors in covariates." *JASA*, 96: 1469–1482. 528

Laird, N. M. and Ware, J. H. (1982). "Random-Effects Models for Longitudinal Data." *Biometrics*, 38: 963–974. 526, 528

Li, E., Wang, N., and Wang, N. (2007a). "Joint Models for a Primary Endpoint and Multiple Longitudinal Covariate Processes." *Biometrics*, 63: 1068–1078. 528

Li, E., Zhang, D., and Davidian, M. (2004). "Conditional estimation for generalized linear models when covariates are subject-specific parameters in a mixed model for longitudinal measurements." *Biometrics*, 60: 1–7. 528, 531

— (2007b). "Likelihood and pseudo-likelihood methods for semiparametric joint models for a primary endpoint and longitudinal data." *Computational Statistics and Data Analysis*, 51: 5776–5790. 528

Marshall, G. and Baron, A. E. (2000). "Linear discriminant models for unbalanced longitudinal data." *Statistics In Medicine*, 19: 1969–1981. 527

Roy, A. and Khattree, R. (2005). "On discrimination and classification with multivariate repeated measures data." *Journal of Statistical Planning and Inference*, 134: 462–485. 527

Tomasko, L., Helms, R. W., and Snapinn, S. M. (1999). "A Discriminant Analysis Extension To Mixed Models." *Statistics In Medicine*, 18: 1249–1260. 527

van Den Heuvel, M., Horrocks, J., Bashar, S., Taylor, S., Burke, S., Hatta, K., Lewis, J., and Croy, B. (2005a). "Menstrual Cycle Hormones Induce Changes in Functional Interactions Between Lymphocytes and Endothelial Cells." *Journal of Clinical Endocrinology and Metabolism*, 90: 2835–2842. 532

van Den Heuvel, M., Horrocks, J., Hatta, K., Burke, S., Bashar, S., Junkins, M. E., Evans, S. S., Croy, B. A., and Tekpetey, F. R. (2005b). "Periovulatory Increases in Tissue Homing Potential of Circulating CD56$^{bright}$ Cells Are Associated With Fertile Menstrual Cycles." *Journal of Clinical Endocrinology and Metabolism*, 90: 3606–3613. 525, 526

Verbeke, G. and Lesaffre, E. (1997). "The effect of misspecifying the random effects distribution in linear mixed effects models for longitudinal data." *Computational Statistics and Data Analysis*, 23: 541–556. 531

Wang, C. Y. and Huang, Y. (2001). "Functional methods for logistic regression on random-effect-coefficients for longitudinal measurements." *Statistics and Probability Letters*, 53: 347–356. 528

Wang, C. Y., Wang, N., and Wang, S. (2000). "Regression Analysis When Covariates Are Regression Parameters of a Random Effects Model." *Biometrics*, 56: 487–495. 523, 524, 528, 529, 532

Wernecke, K. D., Kalb, G., Schink, T., and Wegner, B. (2004). "A Mixed Model Approach to Discriminant Analysis with Longitudinal Data." *Biometrical Journal*, 46: 246–254. 527

Wu, M. C. and Bailey, K. R. (1989). "Estimation and Comparison of changes in the presence of informed censoring: Conditional linear models." *Biometrics*, 45: 939–955. 527

Wu, M. C. and Carroll, R. J. (1988). "Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process." *Biometrics*, 44: 175–188. (Corr: V45 p1347; V47 p357). 527

**Acknowledgments**

The authors wish to thank Anne Croy for support and encouragement, and M.Sc. student Azim Bhamani who worked on earlier versions of the two-stage model. This work was funded by the National Sciences and Engineering Research Council of Canada (NSERC Discovery Grant 261497-03) and Canadian Institutes of Health Research through a Collaborative Health Research Project (CHRP Grant 299110-04: A Study on the Importance of Synchronized Changes in Natural Killer Cells and Endometrium in Promotion of Human Pregnancy).

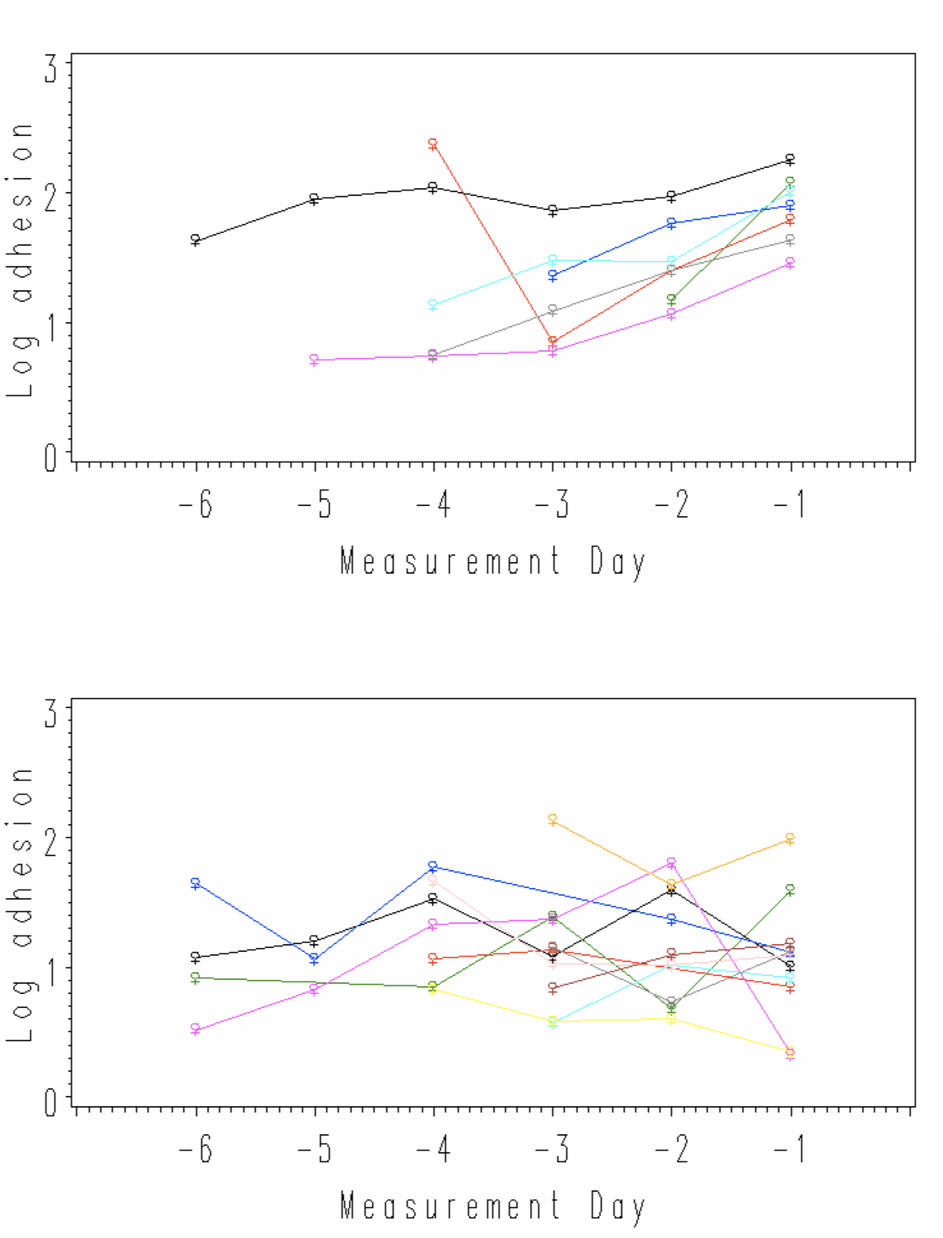


Figure 1: Log adhesion of CD56$^{bright}$ cells versus measurement day for women who became pregnant (top) and those who did not (bottom).

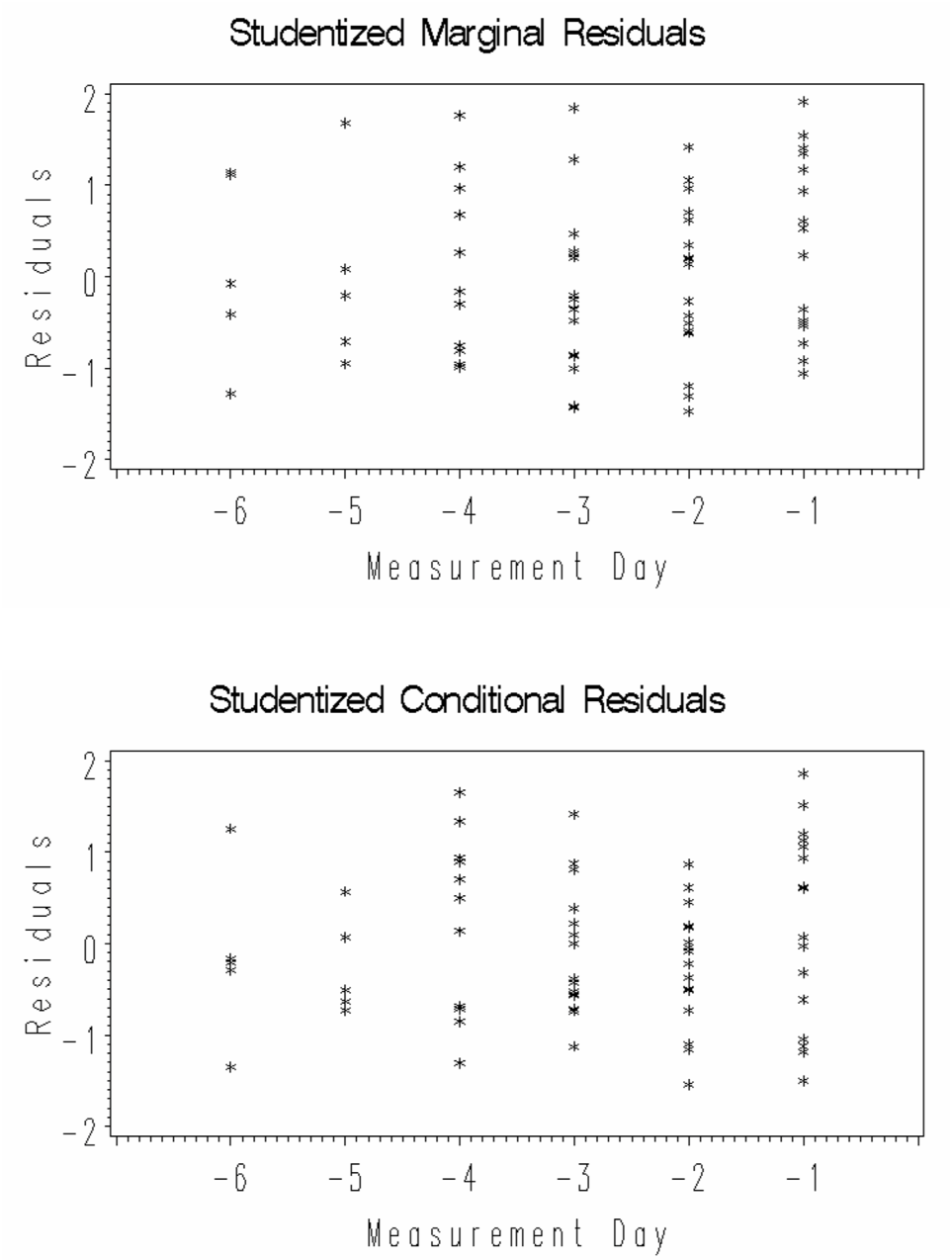


Figure 2: Studentized Marginal Residuals (top) and Studentized Conditional Residuals (bottom).