Biometrics 66, 11–19
March 2010

DOI: 10.1111/j.1541-0420.2009.01234.x

Score Test for Conditional Independence Between Longitudinal Outcome and Time to Event Given the Classes in the Joint Latent Class Model

Hélène Jacqmin-Gadda,[1,2,*] Cécile Proust-Lima,[1,2] Jeremy M.G. Taylor,[3] and Daniel Commenges[1,2]

[1]INSERM, U897, 33076 Bordeaux, France
[2]Université Victor Segalen, 33076 Bordeaux, France
[3]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
[*]*email:* helene.jacqmin-gadda@bordeaux.inserm.fr

Summary. Latent class models have been recently developed for the joint analysis of a longitudinal quantitative outcome and a time to event. These models assume that the population is divided in G latent classes characterized by different risk functions for the event, and different profiles of evolution for the markers that are described by a mixed model for each class. However, the key assumption of conditional independence between the marker and the event given the latent classes is difficult to evaluate because the latent classes are not observed. Using a joint model with latent classes and shared random effects, we propose a score test for the null hypothesis of independence between the marker and the outcome given the latent classes versus the alternative hypothesis that the risk of event depends on one or several random effects from the mixed model in addition to the latent classes. A simulation study was performed to compare the behavior of the score test to other previously proposed tests, including situations where the alternative hypothesis or the baseline risk function are misspecified. In all the investigated situations, the score test was the most powerful. The methodology was applied to develop a prognostic model for recurrence of prostate cancer given the evolution of prostate-specific antigen in a cohort of patients treated by radiation therapy.

Key words: Joint model; Latent class model; Mixture model; Model diagnosis.

1. Introduction

Latent class models (LCMs) have been recently developed for the joint analysis of a longitudinal quantitative marker and a time to event (Lin et al., 2002, Lin, McCulloch, and Rosenheck, 2004; Proust-Lima, Joly, and Jacqmin-Gadda, 2009). Similar to random effect joint models (Henderson, Diggle, and Dobson, 2000), they have been used to develop prognostic tools for the event based on repeated measures of the marker (Lin et al., 2002; Proust-Lima and Taylor, 2009), to study natural history of chronic diseases (Proust-Lima et al., 2009), or to take informative dropout into account in longitudinal studies (Lin et al., 2004). LCMs assume that the population is divided in G latent classes characterized by different evolution profiles for the marker and different risk functions for the event. Typically, they combine a mixture model for longitudinal data (Verbeke and Lesaffre, 1996) and a proportional hazard model depending on the latent class. On the other hand, shared random effect models assume that the risk of event depends on the random effects from the mixed model describing the repeated measures of the marker. The likelihood of LCM is a sum over the classes and thus is easier to compute than the likelihood of a shared random effect model that includes integrals without closed form.

Joint LCMs are a good choice when the change over time of the marker is very heterogeneous and that this heterogeneity is associated with different level of risk for the event. This is the case in the study of prostate cancer recurrences after treatment by radiation therapy. Indeed, the posttreatment prostate-specific antigen (PSA) change over time was found to be very different between cured patients and patients in whom prostate cancer subsequently recurred. Typically, PSA level falls and remains at low level for cured patients whereas rising PSA levels are observed before recurrence (Proust-Lima et al., 2008). In such a case, the usual assumption in shared random effects model that the random slopes has a Gaussian distribution with common mean and variance in the overall population is not tenable. Thus, a joint LCM was proposed to describe the subpopulations of patients characterized by different risks of recurrence associated with different PSA profiles of change (Proust-Lima and Taylor, 2009). This work is motivated by the development of a prognostic tool for prostate cancer recurrence using a cohort of patients treated by radiation therapy for prostate cancer at the University of Michigan (the UM data set) (Taylor, Yu, and Sandler, 2005).

Selection of the number of classes and model checking are difficult issues for LCM. The key assumption of conditional independence between the marker and the event given the latent classes is difficult to evaluate because the latent classes are not observed. When parameters from LCMs are estimated, posterior class membership probabilities may be estimated using Bayes' theorem, allowing posterior classification of the subjects. It has been shown that the distribution of the outcomes conditional on the posterior classes tended toward their distribution conditional on the latent classes (Bandeen-Roche et al., 1997; Lin et al., 2002). Based on this result, the conditional independence between the marker and the time to event may be checked by conditioning on the posterior classification. The methods proposed in the literature differ in three ways. Firstly, some authors (Roy, 2003; Guo, Wall, and Amemyia, 2006) randomly allocated the N subjects to the G classes using the estimated posterior class-membership probabilities and then tested the dependence within each class. This leads to tests with decreasing power when the number of classes increases. Lin et al. (2002, 2004) performed a global test of dependence adjusting for the posterior classes that avoids the loss of power, and to account for the uncertainty of the classification, they carried out a weighted analysis on $N \times G$ pseudosubjects using the posterior probability as weight. Secondly, some authors (Lin et al., 2002, 2004) estimated a survival model to test if the risk of event depended on a function of the marker evolution when adjusted for the posterior class, whereas others (Roy, 2003; Lin et al., 2004) estimated a mixed model to test if the repeated measures of the marker depended on a function of the time to event after adjustment for the posterior class. Thirdly, in the literature, various functions of the marker or of the time to event are used as explanatory variables in the regression model. When the posterior analysis is carried out using a mixed model, the failure indicator or the ratio of the number of events observed over the follow-up time appear as the most obvious functions of the time to event to include as explanatory variables (Roy, 2003; Lin et al., 2004), but the censoring of time-to-event blurs the results and decreases the power of the test. In a survival analysis, Lin et al. (2004) tested the association between the risk of event and the last observed value of the marker. However, this is problematic when the marker measures are very sparse or irregular. Lin et al. (2002) introduced the marginal expectation of the marker at the current time as the explanatory variable, but this choice is also questionable as the marginal expectation is a weighted mean of the covariates, whereas the interest is mainly in the residual dependence between the event and the repeated measures adjusting for covariates and classes. It appears more sensible to introduce as explanatory variable the estimated subject-specific expectation or the empirical Bayes' estimates of the random effects.

To avoid posterior classification and choice of a particular function of the marker or of the time to event as covariate, Proust-Lima et al. (2009) have previously suggested testing conditional independence by comparing the mean of the conditional residual of the marker given the time to event between censored and uncensored subjects. A simulation study showed the validity and usefulness of this test but it is expected to have poor power under some form of conditional dependence, for instance, when the risk of event depends on the marker evolution rather than the mean level.

Another strategy is to compare the LCM with a more flexible model where residual dependence is possible between the event and the marker. A joint LCM with shared random effects (Beunckens et al., 2008) is a good alternative as it allows dependence through both the latent classes and the random effects. This alternative model remains convenient for heterogeneous population and the correlation structure is more flexible as the risk of event depends also on the quantitative random effects. However, estimation of such models is very difficult as it involves potential numerical problems from both models: numerical integration for the shared random-effect model and possible local maxima for the LCM. Thus, in this article, we develop a score test for the null hypothesis of independence between the marker and the event given the latent classes versus the alternative hypothesis that the risk of event depends on one or several random effects from the mixed model. As a score test, it only requires estimation of the null model (the LCM).

The next section describes the joint LCM whereas Section 3 presents the score test for conditional independence. A simulation study is performed in Section 4 to compare different versions of the score test with the previously proposed tests when the alternative is correctly specified and when it is misspecified. In Section 5, the methodology is applied to a data set to develop a prognostic model for the recurrence of prostate cancer given PSA measures.

2. The Joint LCM

2.1 *Class Membership Probability*

The LCM assumes that the population of N subjects can be divided into G unobserved subpopulations represented by latent classes. For each subject i, $i = 1, \ldots, N$, the latent class membership is denoted by a latent variable c_i, which equals g if subject i belongs to latent class g. The individual probability of belonging to class g is explained by covariates X_{i1} through a multinomial logistic regression:

$$\pi_{ig} = P(c_i = g) = \frac{e^{\xi_{0g} + X_{1i}^T \xi_{1g}}}{\sum_{l=1}^{G} e^{\xi_{0l} + X_{1i}^T \xi_{1l}}}, \tag{1}$$

where ξ_{0g} is the intercept for class g and ξ_{1g} is the q_1-vector of class-specific parameters associated with the q_1-vector of time-independent covariates X_{1i}. For identifiability, $\xi_{01} = 0$ and $\xi_{11} = 0$. Each latent class g is then characterized by a specific marker evolution and a specific risk of event described below.

2.2 *Marker Evolution*

The evolution of the longitudinal outcome is assumed to follow a linear mixed model (Laird and Ware, 1982) specific to each class g. Denoting Y_i the vector of marker measurements for subject i, the model may be written:

$$Y_i \,|_{c_i = g, b_{ig}} = Z_i(\mu_g + b_{ig}) + X_{2i}\beta_g + \epsilon_i, \tag{2}$$

with $\epsilon_i \sim \mathcal{N}(0, \Sigma_i)$ and $b_{ig} \sim \mathcal{N}(0, \omega_g^2 B)$, where $\omega_1 = 1$ and B is an unstructured $q \times q$-matrix; Z_i is the $n_i \times q$ matrix of covariates associated with the q-vector of random effects b_{ig}; Z_i can typically include 1 for a random intercept and possibly a polynomial function of time. We assume no overlap between variables included in Z_i and X_{2i}, so that no constraint is required on μ_g. Covariates in X_{2i} may be time dependent or not. Thus the conditional distribution of Y_i given $c_i = g$ is Gaussian with mean $E_{ig} = Z_i\mu_g + X_{2i}\beta_g$ and variance $V_{ig} = \omega_g^2 Z_i B Z_i^T + \Sigma_i$.

2.3 *Survival Model*

Let us define $(T_i,\, \delta_i)$, where T_i is the minimum between T_i^* the time to event and C_i the time of censoring. The indicator of event δ_i equals 1 if $T_i^* \leqslant C_i$ and 0 if $C_i < T_i^*$. We assume that a parametric proportional hazard model describes the risk of event in latent class *g*:

$$\lambda(t \mid c_i = g) = \lambda_{0g}(t; \zeta_g) e^{X_{3i}^T \gamma_g}, \tag{3}$$

where X_{3i} is a q_3-vector of covariates associated with the vector of parameters γ_g that may be specific to the latent classes, and $\lambda_{0g}(t;\, \zeta_g)$ is a parametric baseline risk function in latent class *g*.

In this work, we focus on the case where T_i^* is the time to a clinical event. Marker measurements after the event are excluded from the data set because the objective is to describe the link between the risk of event and the marker change over time preceding the event. In Section 6, we discuss the case where T_i^* is a time to dropout.

2.4 *Likelihood*

We denote θ the vector of parameters of the joint LCM defined by equations (1), (2), and (3) that includes ξ_{0g}, ξ_{1g}, μ_g, β_g, ω_g, ζ_g, γ_g for $g = 1, \ldots, G$, and the variance parameters from the matrices Σ_i and B. Parameter estimation is achieved by a maximum likelihood method for a given number of latent classes *G*. Using the conditional independence assumption between the marker and the time to event given the latent classes, the individual contribution of subject i to the likelihood is:

$$\begin{aligned} L_i(\theta) = \sum_{g=1}^{G} \pi_{ig} f(y_i \mid c_i = g; \theta) \lambda(T_i \mid c_i = g; \theta)^{\delta_i} \\ \times S(T_i \mid c_i = g; \theta), \end{aligned} \tag{4}$$

where $S(T_i \mid c_i = g;\, \theta)$ is the survival function, and the global log likelihood is $l(\theta) = \sum_{i=1}^{N} \log L_i(\theta)$.

3. Score Test for Conditional Independence

3.1 *Model under the Alternative Hypothesis*

We propose a score test for the null hypothesis of conditional independence between the time to event and the marker distribution given the classes versus the alternative of conditional independence given the classes and the random effects b_{ig}. Thus the joint LCM described in Section 2 is the null model, whereas the model under the alternative is a joint LCM with shared random effects defined by equations (1), (2), and the following survival submodel:

$$\lambda_a(t \mid c_i = g, b_{ig}; \theta, \eta) = \lambda_{0g}(t; \zeta_g) e^{X_{3i}\gamma_g + b_{ig}^T \eta}. \tag{5}$$

The vector of parameters η links the risk of event with the random effects from the heterogeneous mixed model. Under the null hypothesis, $H_0 : \eta = 0$, the risk of event depends on the marker evolution only through the latent classes.

The associated survival function under the alternative hypothesis H_a is:

$$\begin{aligned} S_a(t \mid c_i = g, b_{ig}; \theta, \eta) = \exp\Bigl\{ -\Lambda_0(t \mid c_i = g; \zeta_g) \\ \times \exp\Bigl(X_{3i}\gamma_g + b_{ig}^T \eta \Bigr) \Bigr\}. \end{aligned}$$

Under the assumption that $Y_i \perp T_i^* \mid c_i = g,\, b_{ig}$ and that missing measures of the marker before the event are missing at random, the individual contribution to the likelihood for this alternative model is given by:

$$\begin{aligned} L_{ai}(\theta, \eta) = \sum_{g=1}^{G} \pi_{ig} \int f(y_i \mid c_i = g, b_{ig}; \theta) \lambda_a(T_i \mid c_i = g, b_{ig}; \theta, \eta)^{\delta_i} \\ \times S_a(T_i \mid c_i = g, b_{ig}; \theta, \eta) f(b_{ig})\, db_{ig}, \end{aligned} \tag{6}$$

and the global log likelihood may be written:

$$l_a(\theta, \eta) = \sum_{i=1}^{N} \log L_{ai}(\theta, \eta).$$

The alternative model may be easily extended by replacing b_{ig}^T in equation (5) by any vector-valued function of the random effects or by allowing the parameter η to be dependent on the classes (η_g). In the latter case, H_0 would be $\eta_g = 0\ \forall g$. One could also be interested in dependence on the current individual deviation from the mean, which is a time-dependent variable (typically $b_{0ig} + b_{1ig}t$ in a model with random intercept and slope). In this case, the integral in the survival function would not always have an analytical solution and we could not obtain a general formula for the score test whatever the parametric choice for $\lambda_{0g}(t;\, \zeta_g)$. However, in the simulation study, we evaluate the power of the proposed score test against this misspecified alternative.

3.2 *Derivation of the Score Statistic*

In the following section, for brevity we delete explicit dependence on fixed parameters and we denote $f_g(y_i \mid b_{ig}) = f(y_i \mid c_i = g,\, b_{ig};\, \theta)$, and $f_g(y_i) = f(y_i \mid c_i = g;\, \theta)$. For the time-to-event submodel, we use $\lambda_{ag}(T_i \mid b_{ig})$, $\Lambda_{ag}(T_i \mid b_{ig})$ and $S_{ag}(T_i \mid b_{ig})$, respectively, for the risk function, the cumulative risk function, and the survival function given the random effects and the classes under H_a. When $\eta = 0$, dependence on b_{ig} and subscript *a* disappear leading to $\lambda_g(T_i)$, $\Lambda_g(T_i)$, and $\lambda_g(T_i)$.

The score statistic for $H_0 : \eta = 0$ versus $H_a : \eta \neq 0$ is $U(\eta, \theta) = \frac{\partial l_a}{\partial \eta} = \sum_{i=1}^{N} \frac{1}{L_{ai}} \frac{\partial L_{ai}}{\partial \eta}$ computed for $\eta = 0$. Simple calculations show that:

$$\begin{aligned} &\frac{\partial \lambda_{ag}(T_i \mid b_{ig})}{\partial \eta} = \lambda_{ag}(T_i \mid b_{ig}) b_{ig} \quad \text{and} \\ &\frac{\partial S_{ag}(T_i \mid b_{ig})}{\partial \eta} = -S_{ag}(T_i \mid b_{ig}) \Lambda_{ag}(T_i \mid b_{ig}) b_{ig} \\ &\frac{\partial}{\partial \eta} \{ \lambda_{ag}(T_i \mid b_{ig})^{\delta_i} S_{ag}(T_i \mid b_{ig}) \} \\ &\qquad = \lambda_{ag}(T_i \mid b_{ig})^{\delta_i} S_{ag}(T_i \mid b_{ig}) b_{ig} \left[\delta_i - \Lambda_{ag}(T_i \mid b_{ig}) \right]. \end{aligned}$$

Thus,

$$\begin{aligned} \frac{\partial L_{ai}}{\partial \eta} = \sum_{g=1}^{G} \pi_{ig} \int f_g(y_i \mid b_{ig}) \lambda_{ag}(T_i \mid b_{ig})^{\delta_i} S_{ag}(T_i \mid b_{ig}) \\ \times [\delta_i - \Lambda_{ag}(T_i \mid b_{ig})] b_{ig} f(b_{ig})\, db_{ig} \end{aligned}$$

and

$$\begin{aligned}\left.\frac{\partial L_{ai}}{\partial \eta}\right|_{\eta=0} &= \sum_{g=1}^{G} \pi_{ig}\lambda_g(T_i)^{\delta_i} S_g(T_i)[\delta_i - \Lambda_g(T_i)] \\ &\quad \times \int f_g(y_i \mid b_{ig}) b_{ig} f(b_{ig}) db_{ig} \\ &= \sum_{g=1}^{G} \pi_{ig} f_g(y_i)\lambda_g(T_i)^{\delta_i} S_g(T_i)[\delta_i - \Lambda_g(T_i)] \\ &\quad \times \int b_{ig} f(b_{ig} \mid y_i) db_{ig}.\end{aligned}$$

Finally,

$$\begin{aligned}U(0,\theta) &= \sum_{i=1}^{N}\sum_{g=1}^{G} P(c_i = g \mid y_i, T_i, \delta_i)\big[\delta_i - \Lambda_g(T_i)\big] \\ &\quad \times E(b_{ig} \mid c_i = g, y_i). \qquad (7)\end{aligned}$$

The posterior expectation $E(b_{ig} \mid c_i = g,\, y_i)$ is computed using properties of the multivariate Gaussian distribution by $E(b_{ig} \mid c_i = g,\, y_i) = \omega_g^2 B Z_i^T V_{ig}^{-1}\, (y_i - E_{ig})$.

Formula (7) shows that the score statistic is an estimate of the covariance between the residuals from the survival model and the class-specific empirical Bayes' estimates of the random effects weighted by the posterior class-membership probability. One can carry out a univariate test for the dependence on a specific random effect (random intercept for instance) or a multivariate test evaluating the dependence on the vector of random effects. Thus, the vector b_{ig} in the last term in formula (7) includes only the random effects that are assumed to be associated with the risk of event under the alternative hypothesis.

3.3 *Asymptotic Variance of the Score Statistic*

When the parameters θ are known, the statistic $U(0,\, \theta)$ is asymptotically normally distributed with mean 0 and variance $I_{\eta\eta} = E\{-\frac{\partial^2 l_a}{\partial\eta\partial\eta^T} \mid_{\eta=0}\}$, which is approximated by the observed information matrix:

$$\mathrm{Var}_{as}(U) = -\left.\frac{\partial^2 l_a}{\partial\eta\partial\eta^T}\right|_{\eta=0}.$$

With simple calculations detailed in Web Appendix A, we obtain:

$$\begin{aligned}\mathrm{Var}_{as}(U) &= -\sum_{i=1}^{N}\sum_{g=1}^{G} P(c_i = g \mid y_i, T_i, \delta_i) \\ &\quad \times \{[\delta_i - \Lambda_g(T_i)]^2 - \Lambda_g(T_i)\} E\big(b_{ig} b_{ig}^T \,\big|\, c_{ig} = 1, y_i\big) \\ &\quad + \Bigg\{\sum_{g=1}^{G} P(c_i = g \mid y_i, T_i, \delta_i) \\ &\qquad \times [\delta_i - \Lambda_g(T_i)] E(b_{ig} \mid c_i = g, y_i)\Bigg\} \\ &\quad \times \Bigg\{\sum_{g=1}^{G} P(c_i = g \mid y_i, T_i, \delta_i) \\ &\qquad \times [\delta_i - \Lambda_g(T_i)] E(b_{ig} \mid c_i = g, y_i)\Bigg\}^T\end{aligned}$$

where $E(b_{ig} b_{ig}^T \mid c_i = g,\, y_i)$ is computed using properties of the multivariate Gaussian distribution by $E(b_{ig} b_{ig}^T \mid c_i = g,\, y_i) = \mathrm{Var}\ (b_{ig} \mid c_i = g,\, y_i) + E(b_{ig} \mid c_i = g,\, y_i)E(b_{ig} \mid c_i = g,\, y_i)^T$, where $\mathrm{Var}\ (b_{ig} \mid c_i = g,\, y_i) = \omega_g^2 B - \omega_g^2 B Z_i^T V_{ig}^{-1} Z_i B$.

The score statistic and the variance are computed by replacing θ by $\hat\theta$ obtained under H_0. Neglecting uncertainty due to the estimation of parameters θ, under the null hypothesis, the test statistic $U(0,\hat\theta)^T\,\mathrm{Var}(U)^{-1}U(0,\hat\theta)$ follows asymptotically a χ^2_m distribution, where m is the size of η.

3.4 *Robust Variance of the Score Statistic*

As the score statistic is a sum over the subjects $U(0,\hat\theta) = \sum_{i=1}^{N} U_i(0,\hat\theta)$, Freedman (2007) recommended to estimate $\mathrm{Var}(U(0,\hat\theta))$ using the empirical covariance matrix of the summands that may be more robust than the asymptotic estimate. More precisely:

$$\mathrm{Var}_{em}(U) = \sum_{i=1}^{N} U_i(0,\hat\theta)U_i(0,\hat\theta)^T - U(0,\hat\theta)U(0,\hat\theta)^T/N.$$

This variance estimate is easy to compute and accounts for the uncertainty due to parameter estimation, but it is an estimate of the variance of U at the true value η, whereas $\mathrm{Var}_{as}(U)$ is computed under H_0. If the variance of U increases under H_a, as for homogeneity score tests (Commenges et al., 1994), the test using $\mathrm{Var}_{em}(U)$ could be slightly less powerful. In the next section, we compare the test statistics using the two variance estimates.

4. Simulation Studies

4.1 *Objective and General Design*

The aim of the simulation study was to evaluate the type I error and the power of the score test when the alternative is correctly specified, that is when data are generated using a LCM with shared random effect, and when the alternative is misspecified. In both cases, we compared four versions of the score test with three tests previously proposed:

- The univariate score test for dependence on the random intercept using either the asymptotic variance $\mathrm{Var}_{as}(U)$ or the empirical variance $\mathrm{Var}_{em}(U)$;
- The multivariate score test for dependence on the vector of random effects using either the asymptotic variance or the empirical variance;
- The test proposed by Proust-Lima et al. (2009), which is a simple comparison of the means of the standardized conditional residuals of the marker given the event $(Y_i - E(Y_i \mid T_i,\, \delta_i))$ between censored and uncensored subjects; and
- Two Wald tests performed in a weighted linear mixed model adjusted for the posterior classes for testing dependence either on the failure indicator (denoted WMM1 for weighted mixed model) or on the failure indicator divided by the follow-up time (WMM2).

A few comparisons were also performed using random class allocation according to the posterior class-membership probabilities, but results were very close to those of the weighted analysis (results not shown). The sample size was either N=100 or N=500 and the marker was measured every year from time 0 to censoring or failure time. The censoring time was uniformly distributed between 3 and 12.

4.2 *Simulation Study 1: Shared Random Effect*

For the first set of simulations, data were generated according to the following joint LCM with shared random effects and two latent classes with probabilities $\pi_{i1} = \pi_{i2} = 0.5$ or $\pi_{i1} = 0.2$ and $\pi_{i2} = 0.8$:

$$Y_{ij}|_{c_i=g} = (\mu_{0g} + b_{0i}) + (\mu_{1g} + b_{1i}) \times t_{ij} + \epsilon_{ij}$$

and

$$\lambda(t)|_{c_i=g} = \zeta_{1g}\zeta_{2g}(\zeta_{1g}t)^{\zeta_{2g}-1}\exp(\eta_1 b_{0i} + \eta_2 b_{1i}).$$

Parameter values for the mixed model were those estimated for the intercept and the linear trend in a two-class model without covariate applied to the UM data. The class-specific parameters were $\mu_{01} = 2.8$, $\mu_{02} = 1.4$, $\mu_{11} = 1.3$, $\mu_{12} = 0.2$, $\sigma_\epsilon = 0.3$, and $b_i = \binom{b_{0i}}{b_{1i}} \sim \mathcal{N}(\binom{0}{0}, \left(\begin{smallmatrix} 0.5 & -0.03 \\ -0.03 & 0.12 \end{smallmatrix}\right))$. Parameters for the Weibull distribution were chosen to obtain about 70% and 30% of event in the two classes.

Under H_0, $\eta_1 = \eta_2 = 0$. When $\eta_1 > 0$ and $\eta_2 = 0$, the event distribution depends only on the random intercept, and thus on the mean level of the marker, whereas when $\eta_1 = 0$ and $\eta_2 > 0$, the event distribution depends only on the random slope. The former favors the three previously proposed tests and the univariate score test for dependence on the random intercept, which are all sensitive to the mean marker difference between censored subjects and cases, whereas the latter favors the bivariate score test for dependence on the two random effects. The estimated model was a joint LCM with two classes assuming conditional independence.

Results are presented in Table 1. For the four score tests, the type I error is close to the nominal value 5%. As expected, the power increases with the sample size, and when η_1 or η_2 increase. It is slightly lower when $\pi_{i1} = 0.2$ and $\pi_{i2} = 0.8$; this may be due to a lower rate of events in the whole sample (40% instead of 50%). The use of the empirical variance estimate instead of the asymptotic variance estimate has little impact on the results. When the risk of event depends only on the random intercept, the univariate score test for dependence on the random intercept is slightly more powerful than the bivariate score test. However, the power of the univariate score test is very bad compared to its bivariate counterpart when the risk of event depends on the random slope. Thus we recommend the use of the multivariate score test of dependence over the entire vector of random effects in any case.

The three other tests have a lower power compared to the score test even when the risk of event depends only on the random intercept. When $\eta_1 > 0$, WMM1 is more powerful than WMM2 and the test comparing residual means. In the less favorable scenario of dependence on the random slope ($\eta_2 > 0$), these three tests have extremely low power, but the test of residual means is slightly better than the two tests based on posterior classification. The test of residual means is slightly conservative with $N = 100$ and $\pi_{i1} = \pi_{i2} = 0.5$ whereas WMM2 is conservative for $N = 500$ and $\pi_{i1} = \pi_{i2} = 0.5$. In any case, WMM2 is the least powerful.

The same set of simulations was conducted to evaluate the impact of misspecifying the baseline risk function. Data generation was unchanged but the estimated model assumed a stepwise constant baseline risk with three steps. This misspecification has little impact on the behavior (power and type I error) of the score test, which remains much more powerful than the other tests (see Web Table 1). An alternative set of simulations using a mixed model with a quadratic time trend and three random effects led to similar conclusions (Web Appendix B and Web Table 2).

Table 1

Estimated type I error and power over 500 simulations for several tests of conditional independence given the latent classes when data are generated by a two-LCM with shared random effects

Test	$N = 100$						$N = 500$					
η_1	0	0.3	0.5	1	0	0	0	0.3	0.5	1	0	0
η_2	0	0	0	0	1	2	0	0	0	0	1	2
					$\pi_1 = \pi_2 = 50\%$							
Univ. asy. score test	6.2	25.6	55.8	97.0	7.2	9.6	4.4	88.8	99.8	100	19.0	51.0
Univ. emp. score test	6.8	24.8	55.8	97.2	6.0	10.4	4.4	88.8	99.8	100	18.0	50.8
Biv. asy. score test	5.6	20.6	45.8	94.6	38.8	90.0	4.4	81.0	99.8	100	98.8	100
Biv. emp. score test	6.0	22.0	47.2	94.4	39.2	90.8	5.2	80.4	99.8	100	99.0	100
Residual test	2.8	10.0	25.8	64.6	3.4	4.8	5.6	55.2	89.6	100	4.4	35.6
WMM1	4.8	14.2	32.2	84.0	4.2	6.2	4.0	54.6	93.8	100	7.4	19.8
WMM2	3.6	6.8	14.6	42.8	3.6	3.2	2.6	13.8	35.6	72.2	3.4	5.8
					$\pi_1 = 20\%$ $\pi_2 = 80\%$							
Univ. asy. score test	5.2	22.6	47.8	95.0	7.6	9.2	3.8	73.4	99.2	100	12.2	35.6
Univ. emp. score test	5.4	22.2	49.6	94.6	8.6	8.8	4.2	73.0	99.2	100	11.8	36.0
Biv. asy. score test	4.0	16.2	37.8	90.2	32.6	78.9	3.6	64.8	98.6	100	95.2	100
Biv. emp. score test	4.0	16.8	40.0	90.6	34.0	80.7	3.6	65.4	98.6	100	94.8	100
Residual test	5.2	5.8	15.2	56.3	6.8	13.6	4.2	25.8	69.4	100	23.2	60.2
WMM1	5.0	13.6	37.8	86.0	3.8	6.1	3.8	51.6	95.8	100	4.6	11.2
WMM2	3.2	4.6	13.2	34.5	5.6	5.3	4.6	13.2	27.0	66.8	4.8	6.4

WMM1: Wald test of dependence on δ_i in a weighted mixed model.

WMM2: Wald test of dependence on δ_i/T_i in a weighted mixed model.

Table 2

Estimated power over 500 *simulations for several tests of conditional independence given the latent classes when a two-LCM was estimated on data generated by a three-LCM*

Class probabilities	$\pi_1 = 0.06$, $\pi_2 = 0.11$, $\pi_3 = 0.83$			$\pi_1 = \pi_2 = \pi_3$			$\pi_1 = 0.06$, $\pi_2 = 0.83$, $\pi_3 = 0.11$		
Scenarios	1	2	3	1	2	3	1	2	3
Rate of events (%)	18.5	18.5	19.5	61.9	61.9	43.5	75.0	75.1	35.7
Univ. asy. score test	61.0	34.8	48.6	100	65.8	75.4	99.6	52.6	44.2
Univ. emp. score test	59.8	36.4	41.2	100	67.6	75.6	99.6	53.2	47.4
Biv. asy. score test	94.2	78.0	70.6	100	100	88.0	99.6	80.2	63.2
Biv. emp. score test	96.2	84.8	56.8	100	100	88.6	99.6	80.4	68.6
Residual test	12.4	15.8	6.0	100	58.0	44.8	99.6	82.6	13.2
WMM1	60.4	36.0	31.6	100	65.2	53.6	99.4	46.6	22.4
WMM2	30.2	14.2	18.2	39.8	13.4	12.0	52.4	29.8	12.0

WMM1: Wald test of dependence on δ_i in a weighted mixed model.
WMM2: Wald test of dependence on δ_i / T_i in a weighted mixed model.

4.3 *Simulation Study 2: Misspecified Alternatives*

The first aim of this simulation study was to compare the power of the tests to detect an inadequate number of classes in the estimated models. Data were generated using a three-class model without shared random effect ($\eta_1 = \eta_2 = 0$), whereas the estimated model had only two latent classes. Thus, when estimating the two-class model on the simulated data, the conditional independence assumption between the longitudinal marker and the time to event does not hold (because the risk of event is different in the three classes) but the correlation structure is different from the one assumed under the alternative hypothesis when developing the score test.

To be close to the application data, we considered only samples of size $N = 500$ and a scenario with very heterogeneous class-membership probabilities was simulated. In scenario 1, the parameters for the survival model, for the mixed model, and for the class membership probabilities were chosen to mimic the three-class model estimated on the UM data set in the application (using only the intercepts and linear trends in the mixed model). The percentages of events were 98%, 83%, and 4% in class 1, 2, and 3, respectively, and intercepts and slopes were $\mu_{01} = 3.0$ and $\mu_{11} = 2.9$ in class 1, $\mu_{02} = 2.7$ and $\mu_{12} = 1.03$ in class 2, and $\mu_{03} = 1.4$ and $\mu_{13} = 0.17$ in class 3. In this scenario, the three classes had very different risk functions and evolution profiles. Then, two other scenarios were simulated, reducing either the difference between evolution profiles for the marker (scenario 2) or the difference in the risks of event (scenario 3). In scenario 2, the mixed model parameters were $\mu_{01} = 3.0$ and $\mu_{11} = 2.0$ in class 1, $\mu_{02} = 2.7$ and $\mu_{12} = 1.03$ in class 2, and $\mu_{03} = 2$, and $\mu_{13} = 0.5$ in class 3; in scenario 3, the percentages of events were 83%, 33%, and 4% in class 1, 2, and 3, respectively. In the three scenarios, the variance parameters were identical to the first simulation study and we evaluated three combinations of class membership probabilities:

- $\pi_{i1} = 0.06$, $\pi_{i2} = 0.11$, $\pi_{i3} = 0.83$: heterogeneous probabilities corresponding to values estimated on the application data set with the three-class model;
- $\pi_{i1} = 0.33$, $\pi_{i2} = 0.33$, $\pi_{i3} = 0.33$: homogeneous probabilities; and
- $\pi_{i1} = 0.06$, $\pi_{i2} = 0.83$, $\pi_{i3} = 0.11$: heterogeneous probabilities where the class with median risk and median change over time has the largest probability.

Results are presented in Table 2. All the tests are more powerful for scenario 1, where the discrimination among the three classes is the largest both for the risk of event and the marker evolution. In all the situations except one, the bivariate score test is the most powerful with increasing power when the class sizes are homogeneous and thus the number of events is higher. The superiority of the bivariate score test is particularly clear in the situation mimicking the application data ($\pi_{i1} = 0.06$, $\pi_{i2} = 0.11$, $\pi_{i3} = 0.83$, and event rate about 19%) because the power of the other tests is greatly reduced when the rate of event is low. The behavior of the test comparing the residual means highly depends on the data structure: its power is dramatically low for the data sets that mimic the application ($\pi_{i1} = 0.06$, $\pi_{i2} = 0.11$, *and* $\pi_{i3} = 0.83$) and it is the highest when the median class is the largest ($\pi_{i1} = 0.06$, $\pi_{i2} = 0.83$, *and* $\pi_{i3} = 0.11$).

Another simulation study was performed for the misspecified alternative defined by a dependence of the risk of event on the current subject-specific deviation from the mean (data generated with $\lambda_g(t) = \lambda_{0g} \exp(\eta b_{0i} + b_{1i} \times t)$ with an exponential baseline risk). In this case also, the two score tests for the alternative $\lambda_g(t) = \lambda_{0g}(t) \exp(\eta b_{0i})$ or $\lambda_g(t) = \lambda_{0g}(t) \exp(\eta_1 b_{0i} + \eta_2 b_{1i})$ were more powerful than the other tests (Web Table 3).

5. Application

We considered data from a prospective cohort that included patients treated by external beam radiation therapy (EBRT) for localized prostate cancer at the UM between 1988 and 2004 (Taylor et al., 2005). After the end of EBRT, patients were followed up until clinical recurrence of prostate cancer or last contact with repeated measures of PSA, a biomarker of progression of prostate cancer. Clinical recurrence was defined as any of the following: distant metastases, nodal recurrence, or any palpable or biopsy-detected local recurrence 3 years or later after radiation; any local recurrence within 3 years of EBRT if the most previous PSA was >2 ng/ml; and death

Table 3

Comparison of five latent class joint models estimated on the UM data set with one to five latent classes

# classes	Log-likelihood	BIC	Score test[*]	WMM1[**]
1	−2877.2	5852.5	170.4 ($p < 0.001$)	7.9 ($p < 0.001$)
2	−2595.1	5331.2	30.7 ($p < 0.001$)	3.1 ($p < 0.01$)
3	−2529.6	5243.1	12.1 ($p < 0.01$)	1.34 ($p = 0.18$)
4	−2502.4	5231.5	5.6 ($p = 0.2$)	1.47 ($p = 0.14$)
5	−2490.9	5251.5	4.4 ($p = 0.3$)	1.92 ($p = 0.06$)

[*] Trivariate score test for conditional independence (with asymptotic variance).

[**] Wald test for dependence on the failure indicator δ_i in a weighted linear mixed model adjusted for posterior classes.

from prostate cancer. The aim of this analysis was to distinguish profiles of evolution of posttreatment PSA associated with different risks of clinical recurrence after adjustment for pretreatment prognostic factors. Such analysis may help to define tools for early detection of patients at high risk of recurrence based on repeated measures of PSA (Proust-Lima and Taylor, 2009).

Patients were included in the analysis if (1) they had a prostate cancer with clinical stage T1–4 and neither positive nodes or metastases, (2) they had at least 1 year follow-up without clinical recurrence after end of EBRT, (3) they did not receive any salvage androgen deprivation therapy during the follow-up, and (4) they had at least two repeated measures of PSA before the end of follow-up.

Repeated measures of PSA were logarithmically transformed using $Y_{ij} = \ln\ (PSA_i(t_{ij}) + 0.1$ ng/ml) to satisfy the Gaussian assumption from the mixed model. The endpoint of interest was the first clinical recurrence, so that all the PSA measures collected after the end of treatment and before this point were included. Previous analyses of this data set (Proust-Lima et al., 2008) showed that posttreatment evolution of ln(PSA(t)+0.1) exhibited a decline in the first years after EBRT well fitted by the function of time $f_1(t) = ((1 + t)^{-1.5} - 1)$, and a subsequent stable or increasing (long-term) linear trend. Thus the mixed model for PSA evolution was defined by

$$Y_{ij}|_{c_i=g} = (\mu_{0g} + b_{0ig}) + (\mu_{1g} + b_{1ig})f_1(t) + (\mu_{2g} + b_{2ig})t + \epsilon_{ij}$$

with $\epsilon_i \sim \mathcal{N}(0, \sigma^2 I_{n_i})$ and $b_{ig} = (b_{0ig}, b_{1ig}, b_{2ig})^T \sim \mathcal{N}(0, \omega_g^2 B)$.

The time-to-event model was a proportional hazard model with class-specific Weibull baseline risk functions and three pretreatment prognostic factors as covariates: Gleason score (a scale that measures grades of prostate cancer) in three categories (7 and 8–10 versus 2–6), T-stage category (3–4 versus 1–2), and the pretreatment level of PSA transformed to ln(PSA + 0.1).

The sample included 459 subjects with a median of 8 (interquartile range [IQR] = [5–12]) repeated measures of PSA and a median follow-up of 5.16 (IQR = [2.68–7.69]) years. During the follow-up, 74 patients (16.1%) underwent a clinical recurrence with a median time to recurrence of 2.77 (IQR = [1.87–4.41]) years. The mean pretreatment PSA in the logarithm scale was 2.18 (SE = 0.90), 41 (8.9%) patients had a clinical T-stage of 3 or 4 and, respectively, 173 patients (37.7%) and 34 patients (7.4%) had a Gleason score of 7 and above 7.

Five joint LCMs from one to five classes were estimated, and for each of them, Table 3 displays the log likelihood, the Bayesian information criterion (BIC) (Schwartz, 1978), and two tests for conditional independence: the trivariate score test using the asymptotic variance, and the Wald test for dependence on the failure indicator in a weighted linear mixed model adjusted for posterior classes. According to the BIC, we retained the four-class model. The score test decreased continuously from one to five classes and was nonsignificant for the models with four and five classes. Thus the assumption of conditional independence was not rejected for the selected model with four classes. In contrast, the test statistic WMM1 decreased from one to three classes, became not significant for three classes but then slightly increased till five classes. However, the simulation study showed that this test was less powerful. We can also point out that the test comparing the means of the conditional residuals was never significant. This was consistent with the simulations because the class-specific initial levels were close in the four classes and the class membership probabilities were small for the three classes with high recurrence risk.

The four estimated class membership probabilities were $\pi_1 = 1.78\%$, $\pi_2 = 4.4\%$, $\pi_3 = 11.1\%$, and $\pi_4 = 82.6\%$. The estimated mean evolutions in the four classes are presented in Figure 1A and the associated survival functions in Figure 1B. Class 4 had a very low risk of recurrence and posttreatment PSA evolution was characterized by a low level at end of EBRT followed by a short-term decrease and long-term stability over 8 years. Classes 1, 2, and 3 corresponded to three groups of patients with increasing risk of recurrence associated with an increase of PSA level from 1 year of treatment.

Adjusted for the four latent classes, pretreatment PSA level was no longer associated with the risk of recurrence ($\beta = 0.047$, $SE = 0.13$) whereas T-stage above 2 ($RR = 2.0$, $p = 0.05$) and Gleason score of 7 ($RR = 2.9$, $p < 0.01$) or above 7 ($RR = 3.0$, $p = 0.03$) remained independent risk factors of recurrence.

6. Discussion

We have proposed a score test for the basic assumption of joint LCMs, that is, the conditional independence assumption between the time to event and the repeated measures of the marker given the latent classes. This procedure avoids posterior classification required by other methods. As a score test, it has theoretical validity and is simple to compute because computations are performed under the null hypothesis of conditional independence. In addition, the test statistic has a meaningful interpretation as the covariance between the residuals from the survival model and the class-specific empirical Bayes' estimates of the random-effects weighted by the posterior class-membership probabilities.

More importantly, the simulations have shown that the score test is more powerful than three previously proposed tests even against two misspecified alternative hypotheses. In particular, it has a good power to detect when the number

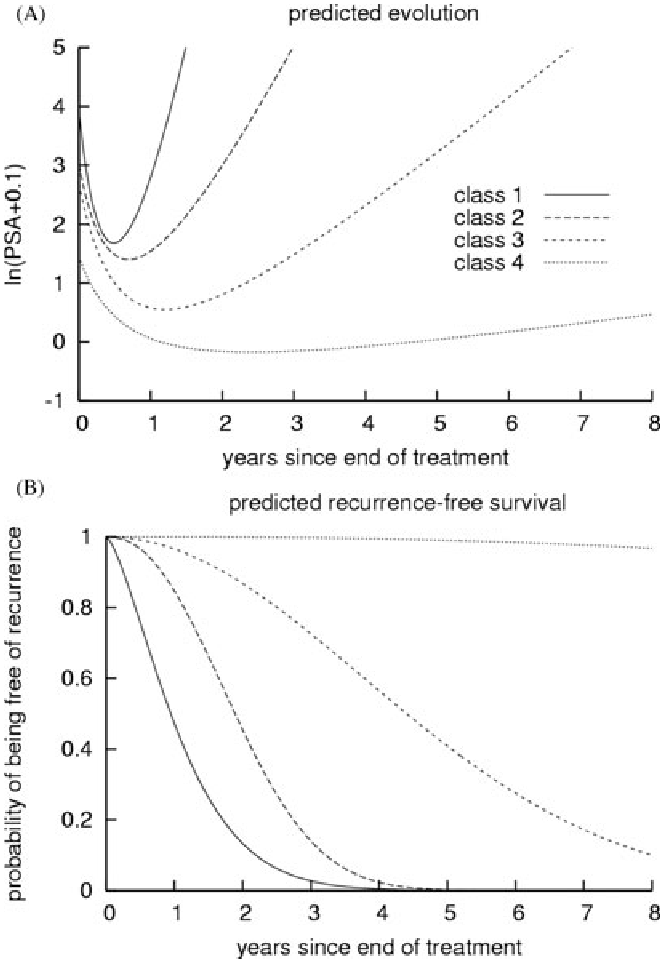

Figure 1. (A) Predicted mean evolution of PSA in the four latent classes and (B) associated probability of being free of recurrence.

of classes is too small and thus it can be used to validate the model selected using BIC. Note however that, for some applications, the number of classes may be driven by heterogeneity in the marker evolution without link with the risk of event. The simulations have also shown that the score tests using either the asymptotic variance or the empirical variance have similar behaviors.

Although the pattern mixture models may be preferred in this context (Dantan et al., 2008), joint LCMs are sometimes used for joint modeling time to dropout and longitudinal data to estimate without bias the change over time of the marker taking informative dropout into account (Lin et al., 2004). If we denote y_i^m the missing postdropout measures and y_i^o the observed predropout measures, the aim is to estimate $f(y_i^o, y_i^m)$. In this case, the likelihood (6) for the alternative model relies on the assumption that $(y_i^o, y_i^m) \perp T_i^* \mid c_i = g, b_{ig}$, which is uncheckable as y_i^m is not observed. As is often discussed, a test of the dropout mechanism is not possible without strong and uncheckable assumptions (Diggle and Kenward, 1994).

In the simulation study, the behavior of the test comparing the conditional residuals means was found to be highly dependent on the data structure. More thorough investigation of the distribution of these residuals through graphical analyses could probably be useful to highlight departures from model assumptions, but the mean comparison alone is too crude to identify some residual correlation structures. Tests in regression models adjusted for the posterior latent classes also exhibited mediocre power. The power could probably be improved by combining a test for the dependence of the marker on the time to event in a mixed model and a test for the dependence of the time to event on the marker in a survival model, but the score test is simpler and powerful against different alternatives. Moreover, the score test can be easily extended to other joint models using another time-to-event model or a logistic model for binary data.

7. Supplementary Materials

Web Appendices and Tables referenced in Section 3.3, 4.2, and 4.3 are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

The authors thank the associate editor and the referees for their helpful comments.

References

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., and Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association* **92,** 1375–1386.

Beunckens, C., Molenberghs, G., Verbeke, G., and Mallinckrodt, C. (2008). A latent-class mixture model for incomplete longitudinal Gaussian data. *Biometrics* **64,** 96–105.

Commenges, D., Letenneur, L., Jacqmin, H., Moreau, T., and Dartigues, J. F. (1994). Test of homogeneity of binary data with explanatory variables. *Biometrics* **50,** 613–620.

Dantan, E., Proust-Lima, C., Letenneur, L., and Jacqmin-Gadda, H. (2008). Pattern mixture models and latent class models for the analysis of multivariate longitudinal data with informative dropout. *International Journal of Biostatistics* **4**, art 14. Available at: `http://www.bepress.com/ijb/vol4/iss1/14`.

Diggle, P. and Kenward, M. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society Series C, Applied Statistics* **43,** 49–93.

Freedman, D. A. (2007). How can the score test be inconsistent? *The American Statistician* **61,** 291–295.

Guo, J., Wall, M., and Amemyia, Y. (2006). Latent class regression on latent factors. *Biostatistics* **7,** 145–163.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1,** 465–480.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38,** 963–974.

Lin, H., Turnbull, B. W., McCulloch, C. E., and Slate, E. H. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association* **97,** 53–65.

Lin, H., McCulloch, C. E., and Rosenheck, R. A. (2004). Latent pattern mixture models for informative intermittent missing data in longitudinal studies. *Biometrics* **60,** 295–305.

Proust-Lima, C. and Taylor, J. M. G. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures post-treatment PSA. To appear in *Biostatistics.*

Proust-Lima, C., Taylor, J. M. G., Williams, S. G., Ankerst, D. P., Liu, N., Kestin, L. L., Bae, K., and Sandler, H. M. (2008). Determinants of change of prostate-specific antigen over time and its association with recurrence following external beam radiation therapy of prostate cancer in 5 large cohorts. *International Journal of Radiation Oncology, Biology, Physics* **72**, 782–791.

Proust-Lima, C., Joly, P., and Jacqmin-Gadda, H. (2009). Joint modelling of multivariate longitudinal outcomes and a time-to-event: A nonlinear latent class approach. *Computational Statistics and Data Analysis* **53**, 1142–1154.

Roy, J. (2003). Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. *Biometrics* **59,** 829–836.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6,** 461–464.

Taylor, J. M. G., Yu, M., and Sandler, H. M. (2005). Individualized predictions of disease progression following radiation therapy for prostate cancer. *Journal of Clinical Oncology* **23,** 816–825.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association* **91,** 217–221.

Received June 2008. *Revised December* 2008.
Accepted December 2008.