1. Introduction

In many studies the outcome of primary interest is the time passing until a certain event occurs, for example time to death for patients with malignant melanoma (skin cancer; Drzewiecki and Andersen, 1982). The focus of such studies may be to investigate whether patients with different background characteristics (e.g. patients of different age or gender) have different survival prognosis. Typically the data is analyzed using survival analysis (Andersen et al., 1993; Fleming and Harrington, 1991), and the most commonly used model is Cox’s Proportional Hazards Model (Cox, 1972), which is also a part of the framework for the present paper. The proportional hazards model has been extended in different ways. These extensions include situations with random effects (frailties; Vaupel et al., 1979), multivariate survival times (Hougaard, 2000), measurement error in covariates (Prentice, 1982), time-varying covariates (Wulfsohn and Tsiatis, 1997), longitudinal modeling with informative drop-out (Henderson et al., 2000), and others.

In this paper the Cox model is extended to encompass an explanatory class variable that is not directly observable, so that information about the class membership is only available indirectly through a set of variables, whose distribution depend
on the latent classes. The model is motivated by the following example, which is taken from the field of Public Health.

Quantifying human physical functioning is difficult. In the Women’s Health and Aging Study (WHAS; REF) the domain of mobility/exercise tolerance is sought measured by five questions. These are: "Without help, do you have any difficulty 1. walking $\frac{1}{4}$ mile, 2. climbing 10 steps, 3. getting in and out of bed or chair, 4. doing heavy housework, and 5. lifting up 10 pounds?". The possible answers to the five questions are "Yes" and "No". It is of interest to know to which extend mobility/exercise tolerance is associated with survival. One way to investigate this would be to analyze the women’s survival times with each or some of the five binary variables as covariates, for instance using a Cox model. This approach is unsatisfactory for at least three reasons, (i) none of the variables exactly measure mobility/exercise tolerance, but rather do each of the variables measure special aspects of the domain, (ii) some individuals may have misclassified themselves, and (iii) not all information (all five questions) can be used in the model simultaneously because of collinarity. In the present paper these problems are avoided by modeling the binary indicators using a latent class regression model (Goodman, 1974; Clogg, 1995).

The latent class regression model is a model for multiple indicators of latent classes. It is specified in two parts: first, a regression model for the relationship between covariates and the latent class membership. This part is modeled by a multinomial logit model (Agresti, 1984). Second, a model for the relationship between the latent classes and the observed indicators. These indicators are assumed to be conditionally independent given the latent class variable.

A model for joint analysis of multiple binary indicators and survival is presented. The distribution of the binary indicators are specified as a latent class regression model, and the distribution of the survival times are modeled by a proportional hazards model conditional on latent class membership. The model may be seen as an extension of the latent class regression model allowing survival to be predicted by latent class membership. Alternatively, one can view it as a proportional hazards model tailored towards situations with collinarity among covariates as in the WHAS example. The scope of this paper is - in addition to proposing a new model - to
suggest methods for drawing inference in the model, and to show how to investigate model fit.

Bandeen-Roche et al. (1997) analyzed the five Bernoulli variables from the domain of mobility/exercise tolerance from the WHAS data using the latent class regression model. They found that the domain may be described well by three latent classes. To motivate and illustrate the model presented here the WHAS data is revisited, and the relationship between mobility/exercise tolerance and later survival is investigated for these women. Their survival times are modeled by a proportional hazards model conditional on the latent classes, which are modeled by a latent class regression model.

The rest of this paper is organized as follows. The model is presented in Section 2. It is a proportional hazards model that is extended to encompass a latent class variable as predictor of survival. The class membership is measured indirectly through a set of binary indicators. In Section 3 the likelihood function for the joint modeling of survival and latent class indicators are maximized using the EM algorithm. In Section 4 different graphical checks and formal tests are suggested for detection of specific violations of model assumptions related to the survival part of the model. In Section 5 an extensive analysis of the association between mobility/exercise tolerance and survival is carried out using the methodology presented in Sections 2-4 on data from the WHAS study. The paper concludes with a discussion in Section 6.

2. Model

The model consists of two parts, a latent class regression model for the multiple indicators, and a proportional hazards model for the failure times.

2.1. Latent class regression model.

The latent class regression model is a model for the relationship between multiple binary indicators and some covariates. It may be thought of as an analogue to a factor analysis with covariates; with binary instead of normal outcomes, and with a latent class variable instead of a normal latent variable (Bartholomew, 1987).
Let $Y_i = (Y_{i1}, \ldots, Y_{iJ})^t$ denote a random vector of $J$ binary indicators for the $i$th individual in a population with $N$ individuals. Assume that the population consists of $K$ subpopulations, but that the group membership, $C_i$, of the $i$th individual is not observed. That is, $C_i$ is a latent class variable taking one of the values $1, \ldots, K$.

Conditional on class membership the distribution of the $j$th indicator is

$$\Pr(Y_{ij} = y | C_i = c) = \pi_{cj}^y (1 - \pi_{cj})^{1-y}, \ y = 0, 1.$$  

The parameters $\pi_c = (\pi_{c1}, \ldots, \pi_{cJ})^t, c = 1, \ldots, K$ are the class specific probabilities of each of the indicators taking the value one. Further, conditional on class membership the $J$ indicators are assumed mutually independent, so that

$$\Pr(Y_i = y_i | C_i = c) = \prod_{j=1}^J \Pr(Y_{ij} = y_{ij} | C_i = c).$$  \hspace{1cm} (2.1)$$

Let $x_i = (x_{i1}, \ldots, x_{iP})$ be a $1 \times P$ row vector of covariates for the $i$th individual. We assume the generalized logit model (Agresti, 1984) for the relationship between the covariates, $x_i$, and the latent class, $C_i$:

$$\Pr(C_i = c | x_i) = \frac{\exp(x_i \kappa_c)}{\sum_{k=1}^K \exp(x_i \kappa_k)},$$

where $\kappa_k$ is a $P \times 1$ vector containing the parameters for the $k$th group. The covariates $x_i$ may be categorical or continuous.

The joint distribution of $(Y_i, C_i)$ is then

$$\Pr(Y_i = y_i, C_i = c | x_i) = \Pr(C_i = c | x_i) \Pr(Y_i = y_i | C_i = c, x_i)$$
$$= \frac{\exp(x_i \kappa_c)}{\sum_{k=1}^K \exp(x_i \kappa_k)} \prod_{j=1}^J \pi_{cj}^{y_{ij}} (1 - \pi_{cj})^{1-y_{ij}}.$$  

Note, that we have made the assumption of no differential item functioning, that is $\Pr(Y_i = y_i | C_i = c, x_i) = \Pr(Y_i = y_i | C_i = c)$. First, this assumption reflects how $C_i$ serves as a summary of the information in $Y_i$. Second, assuming a distribution on $X_i$ it implies that we have independence between $X_i$ and $Y_i$ conditional on $C_i = c$. 


2.2. Cox’s proportional hazards models.

For the survival times we consider proportional hazards models. First, assume that the covariates, \((z_i, c_i)\), are measured without error or misclassification. The covariates, \(z_i\), may be categorical or continuous, whereas \(c_i\) is a class variable taking the values \(1, \ldots, K\). In a proportional hazards model it is assumed that the hazard function for the failure time of the \(i\)th individual is on the form

\[
\lim_{h \to 0^+} \frac{1}{h} \Pr(T_i < t + h | T_i \geq t, z_i, c_i) = \alpha_i(t|z_i, c_i) = \alpha_0(t) \exp(z_i \beta + \nu c_i)
\]

where \(z_i = (z_{i1}, \ldots, z_{iQ})\) is a \(1 \times Q\) row vector, and \(\beta\) is the corresponding \(Q \times 1\) parameter vector. The \(K \times 1\) parameter vector \(\nu\) contains the effects of the class variable \(c_i\) on the hazard. The baseline hazard, \(\alpha_0(t)\), is left unspecified (non-parametric).

As a consequence of (2.2), the distribution of the failure time, \(T_i\), has survival function

\[
S_i(t|z_i, c_i) = \Pr(T_i > t|z_i, c_i) = \exp(-A_0(t) \exp(z_i \beta + \nu c_i)),
\]

where \(A_0(t) = \int_0^t \alpha_0(s) ds\) is the integrated baseline hazard, and the density is

\[
pr(t|z_i, c_i) = \alpha_0(t) \exp(z_i \beta + \nu c_i) \exp(-A_0(t) \exp(z_i \beta + \nu c_i)).
\]

Generally, the failure time, \(T_i\), is censored by a censoring time, \(V_i\), so that the observable variables are \(U_i = \min(T_i, V_i)\) and \(\Delta_i = I(T_i \leq V_i)\). That is, we observe which event comes first and at which time. The variable \(\Delta_i\) is one, if we observe a failure time, and it is zero, if we observe a censoring, and \(U_i\) is the time point for whichever comes first. In the case of non-informative censoring (Andersen et al, 1993) the probability distribution of \((U_i, \Delta_i)\) becomes

\[
pr(u_i, \delta_i|z_i, c_i) \propto [\alpha_0(u_i) \exp(z_i \beta + \nu c_i)]^{\delta_i} \exp(-A_0(u_i) \exp(z_i \beta + \nu c_i)).
\]

2.3. Cox’s proportional hazards model with a latent class variable as predictor.

We now extend the proportional hazards model described in Section 2.2 to the situation, when \(c_i\) is not measured directly but only indirectly via the information in \(Y_i\).
Using a latent class regression model for the regression of \((Y_i, C_i)\) on \(x_i\) and a proportional hazards model for the regression of \((U_i, \Delta_i)\) on \((z_i, c_i)\), the joint distribution of \((U_i, \Delta_i, Y_i, C_i)\) is

\[
\begin{align*}
\text{pr}(u_i, \delta_i, y_i, c_i | x_i, z_i) &= \text{pr}(c_i | x_i) \text{ pr}(y_i | c_i) \text{ pr}(u_i, \delta_i | c_i, z_i),
\end{align*}
\]

and integrating out the latent class variable, \(C_i\), the marginal distribution of the observable variables, \((U_i, \Delta_i, Y_i)\), becomes

\[
\begin{align*}
\text{pr}(u_i, \delta_i, y_i | x_i, z_i) &= \sum_{c=1}^{K} \text{pr}(c | x_i) \text{ pr}(y_i | c) \text{ pr}(u_i, \delta_i | c, z_i) \\
&= \sum_{c=1}^{K} \left\{ \frac{\exp(x_i \kappa_c)}{\sum_{k=1}^{K} \exp(x_i \kappa_k)} \left[ \prod_{j=1}^{J} \pi_{cj}^{y_{ij}} (1 - \pi_{cj})^{1-y_{ij}} \right] \right\} \left[ \alpha_0(u_i) \exp(z_i \beta + \nu_c) \right]^\delta_i \exp\left(-A_0(u_i) \exp(z_i \beta + \nu_c)\right) \exp\left(-A_0(u_i) \exp(z_i \beta + \nu_k)\right),
\end{align*}
\]

For the \(i\)th individual, the posterior class membership probabilities are defined by

\[
\omega_{ic} = \Pr(C_i = c | U_i = u_i, \Delta_i = \delta_i, Y_i = y_i, x_i, z_i) = \frac{\exp(x_i \kappa_c) \left[ \prod_{j=1}^{J} \pi_{cj}^{y_{ij}} (1 - \pi_{cj})^{1-y_{ij}} \right]}{\sum_{k=1}^{K} \exp(x_i \kappa_k) \left[ \prod_{j=1}^{J} \pi_{kj}^{y_{ij}} (1 - \pi_{kj})^{1-y_{ij}} \right] \left[ \exp(\nu_c)^\delta_i \exp(-A_0(u_i) \exp(z_i \beta + \nu_c)) \right] \left[ \exp(\nu_k)^\delta_i \exp(-A_0(u_i) \exp(z_i \beta + \nu_k)) \right]},
\]

for \(c = 1, \ldots, K\).

### 3. Inference

We employ the expectation-maximization (EM) algorithm (Dempster et al., 1977) to maximize the likelihood for the observed data, \((U, \Delta, Y)\). This is done by iterating between an E-step, where we compute the expected log-likelihood of the complete data, \((U, \Delta, Y, C)\), conditional on the observed data and the current estimate of the parameters, and an M-step, where new parameter estimates are computed by maximizing the expected log-likelihood.

Let \(\theta = (\pi, \kappa, \alpha_0(t), \beta, \nu)\). The complete data log-likelihood is

\[
\ell_{\text{com}}(\theta; u, \delta, y, c) = \sum_{i=1}^{N} \ell_{i,\text{com}}(\theta; u_i, \delta_i, y_i, c_i),
\]
where

\[ l_{i,\text{com}}(\theta; u_i, \delta_i, y_i, c_i) = x_i \kappa_{c_i} - \log \left( \sum_{k=1}^{K} \exp(x_i \kappa_k) \right) \]

\[ + \left( \sum_{j=1}^{J} y_{ij} \log \pi_{c_i j} + (1 - y_{ij}) \log(1 - \pi_{c_i j}) \right) \]

\[ + \delta_i [\log(\alpha_0(u_i)) + \mathbf{z}_i \beta + \nu_{c_i}] - A_0(u_i) \exp(\mathbf{z}_i \beta + \nu_{c_i}), \]

and the observed data log-likelihood is

\[ l(\theta; u, \delta, y) = \log \left( \sum_{c \in \{1, \ldots, K\}} \exp(l_{\text{com}}(\theta; \mathbf{U}, \mathbf{D}, \mathbf{Y}, \mathbf{C})) \right). \]

In the E-step we calculate

\[ Q(\theta; \theta^{(r)}) = E(l_{\text{com}}(\theta; \mathbf{U}, \mathbf{D}, \mathbf{Y}, \mathbf{C}) | \mathbf{U}, \mathbf{D}, \mathbf{Y}; \theta^{(r)}), \]

where \( \theta^{(r)} \) is the parameter value of \( \theta \) from the \( r \)th step. That is, we take the expectation of the complete data log-likelihood with respect to the conditional distribution of \( (C|U, D, Y) \).

In the M-step we maximize \( Q \) as a function of \( \theta \). The update of \( \pi \) is simple, whereas both \( \kappa \) and \( (\alpha_0(t), \beta, \nu) \) in principle need an iterative procedure within the E-step. Let \( \omega_{ic}^{(r)} = \Pr(C_i = c|U_i, D_i, Y_i; \theta^{(r)}) \), \( c = 1, \ldots, K \), denote the posterior distribution of \( C_i \) given data and \( \theta = \theta^{(r)} \).

The score equations for \( \pi \) are \( S\pi = 0 \), where

\[ S_{\pi_{c_j}} = \sum_{i=1}^{N} \omega_{ic}^{(r)} \frac{y_{ij} - \pi_{c_j}}{\pi_{c_j}(1 - \pi_{c_j})}, \]

\[ c = 1, \ldots, K, \quad j = 1, \ldots, J. \] These equations have closed form solutions

\[ \pi_{c_j}^{(r+1)} = \frac{\sum_{i=1}^{N} \omega_{ic}^{(r)} y_{ij}}{\sum_{i=1}^{N} \omega_{ic}^{(r)}}. \]

The update of \( \kappa \) is given as the solution to the score equations \( S\kappa = 0 \), where

\[ S_{\kappa_{c_p}} = \sum_{i=1}^{N} x_{ip} [\omega_{ic}^{(r)} - \frac{\exp(x_i \kappa_c)}{\sum_{k=1}^{K} \exp(x_i \kappa_k)}] \]

\[ c = 1, \ldots, K, \quad p = 1, \ldots, P. \] However, these equations do not have a closed form solution, so in principle it is necessary to employ an iterative scheme within the M-step to obtain \( \kappa^{(r+1)} \). We take just one step in a Newton-Raphson algorithm starting at \( \kappa^{(r)} \). That is,

\[ \kappa^{(r+1)} = \kappa^{(r)} + \mathbf{I}_{\kappa^{(r)}}^{-1} S\kappa^{(r)}, \]
where the matrix $I^{\kappa(r)}$ is the negative of the derivative of the score evaluated in $\kappa^{(r)}$ (see Appendix A for details).

The update of $(\beta, \nu)$ is more difficult because of $\alpha_0(t)$ being non-parametric. Given $\alpha_0(t)$, the score equations for $\beta$ and $\nu$ are $S_{\beta} = 0$ and $S_{\nu} = 0$, where

$$S_{\beta q} = \sum_{i=1}^{N} z_{iq} \{ \delta_i - A_0(u_i) \ E[\exp(z_i \beta + \nu_c_i) | U_i, \Delta_i, Y_i; \theta^{(r)}] \},$$

$$S_{\nu c} = \sum_{i=1}^{N} \omega_{ic} \{ \delta_i - A_0(u_i) \ \exp(z_i \beta + \nu_c) \},$$

$q = 1, \ldots, Q$, and $c = 1, \ldots, K$ for $S_{\beta q}$ and $S_{\nu c}$, respectively. The score equations do not have a closed form solution. As with $\kappa$ we take one step in a Newton-Raphson algorithm starting at $(\beta^{(r)}, \nu^{(r)})$. The step becomes

$$(\beta^{(r+1)}_t, \nu^{(r+1)}_t)^t = (\beta^{(r)}_t, \nu^{(r)}_t)^t + I^{-1}_{\beta^{(r)}, \nu^{(r)}} (S'_{\beta^{(r)}}, S'_{\nu^{(r)}})^t,$$

where $I_{\beta^{(r)}, \nu^{(r)}}$ is the negative of the derivative of the score (see Appendix A for details).

Given the new estimates for $(\beta, \nu)$ the baseline hazard is then updated in the following way. It is easily seen that $\alpha_0(t)$ must be zero almost everywhere outside the time points, where an event happens. The hazard at the event time for the $i$th individual is denoted $\alpha_i$. The score for $\alpha_i$ becomes

$$S_{\alpha_i} = \frac{\delta_i}{\alpha_i} - \sum_{j \in R_{\alpha_i}} E[\exp(z_j \beta + \nu_c_j) | U_j, \Delta_j, Y_j; \theta^{(r)}],$$

$\alpha_i > 0$, $i = 1, \ldots, N$, where $R_{\alpha_i}$ is the set of individuals under risk at time point $t$. If $\delta_i = 1$, then the score equation $S_{\alpha_i} = 0$ may be solved directly. If $\delta_i = 0$, then $Q(\theta; \theta^{(r)})$ is maximized when $\alpha_i = 0$. So, the update of $\alpha_0(t)$ may be written as

$$\alpha_0(t) = \sum_{i=1}^{N} \frac{\delta_i I(U_i = t)}{\sum_{j \in R_{\alpha_i}} E[\exp(z_j \beta + \nu_c_j) | U_j, \Delta_j, Y_j; \theta^{(r)}]}.$$

Thus, one way to update $(\alpha_0(t), \beta, \nu)$ in the M-step would be iteratively following the scheme

1. Update $(\beta, \nu)$, e.g. using a Newton-Raphson algorithm.
2. Calculate $\alpha_0(t)$ from (3.6).
3. Iterate between (1) and (2) until convergence.
However, we run through steps (1) and (2) only once. That is, we just plug-in $\beta^{(r+1)}$ and $\nu^{(r+1)}$ in the denominator of (3.6).

Standard errors are not straightforward to calculate. Louis’ formula is often useful, when maximum likelihood estimates are obtained using the EM algorithm. However, this is not feasible due to the non-parametric part of the model. Alternatively we use numerical differentiation treating $\alpha_0(t)$ as nuisance parameters. That is, we calculate the second derivative of the profile log-likelihood,

$$l_{\text{profile}}(\pi, \kappa, \beta, \nu) = l(\pi, \kappa, \psi[\pi, \kappa, \beta, \nu], \beta, \nu),$$

where $\psi[\pi, \kappa, \beta, \nu]$ maximizes $l$ with respect to $\alpha_0(t)$ for a given set of parameters, $(\pi, \kappa, \beta, \nu)$. Evaluated at the maximum likelihood estimator for $(\pi, \kappa, \beta, \nu)$, this gives the negative of the observed Fisher information for the parameters, and inverting the information, we have an estimator for the variance of $(\hat{\pi}_t, \hat{\kappa}_t, \hat{\beta}_t, \hat{\nu}_t)^t$.

4. Fit

The model consists of two parts, the latent class regression model and the proportional hazards model, and it is natural to investigate the fit of each part separately, though there may be model violations that are only detectable considering the joint distribution of $(U, \Delta, Y, C)$. Methods for checking the latent class model has already been discussed extensively in the literature (see Bandeen-Roche et al. (1997) for a review). Therefore here the focus is on the fit of the proportional hazards model.

Before bringing the proportional hazards model into play, the latent class regression model should be validated and established as a good model for the association between the multiple binary indicators and the covariates. Probably the most difficult problem when fitting a latent class regression is determining the number of latent classes. Classical test theory for nested models cannot easily be applied, because the likelihood ratio test statistic is not $\chi^2$ distributed, and no convincing alternative has to our knowledge been suggested in the literature. When the number of classes has been fixed, classical test theory may be applied, and it is possible to test hypotheses regarding effects of covariates on latent class membership, differential item functioning and local dependence between the items.
As mentioned above, the focus is here on checking the proportional hazards part of the model, assuming that the latent class regression model is a good model for the regression of the binary indicators on the covariates. We present two methods to check assumptions that are implied by the model. These are the assumptions of

(1) proportional hazards, and
(2) no additional effect of an item on survival given latent class.

The idea underlying both methods is the following. Obtain "complete data", \((U, \Delta, Y, C^*)\), by simulating \(C^*\) using the posterior class membership probabilities in equation (2.4). If the model holds, and the parameter for \(\theta\) is correct, then \((U, \Delta, Y, C^*)\) will be distributed as \((U, \Delta, Y, C)\). Therefore, finding a model violation based on \((U, \Delta, Y, C^*)\) would indicate that \((U, \Delta, Y, C)\) is not distributed according to the model. However, as \(\theta\) is unknown, we use \(\hat{\theta}\) instead.

4.1. Proportional hazards.

Violations of the proportional hazards assumption may be investigated graphically by plotting \(\hat{A}^*_c(t), c = 1, \ldots, K\) against \(t\) or some function \(f\) of \(t\), where

\[
\hat{A}^*_c(t) = \sum_{i=1}^{N} \frac{\delta_i I(u_i \leq t, C^*_i = c)}{\sum_{j \in R_{u_i}} \exp(z_j \hat{\beta} + \nu_c)}.
\]

That is, \(\hat{A}^*_c(t)\) is the estimated integrated hazard based on the individuals with \(C^*_i = c\). If the hazards are in fact proportional, then the estimated integrated hazards should show to be approximately proportional. Alternatively, one may plot \(\hat{A}^*_{c_1}(t)\) against \(\hat{A}^*_{c_2}(t)\) for two classes, \(c_1\) and \(c_2\) (Gill and Schumacher, 1987). If the proportional hazards model is correct, then \((\hat{A}^*_{c_1}(t), \hat{A}^*_{c_2}(t))\) should (approximately) be a straight line going through the origin. Either of the approaches suffer from the drawback that it is difficult to obtain meaningful confidence intervals/bands. Consequently, the graphs serve mainly as exploratory tools.

If the graphical check described above indicates non-proportionality of the hazards, then it may be of interest to perform a formal test. We suggest to specify an
alternative on the form

\begin{equation}
\alpha_i(t) = \alpha_0(t) \exp(z_i \beta + \nu_{ci} + \lambda_{ci} g(t)),
\end{equation}

where \( \lambda \) is a \( K \times 1 \) vector and \( g(t) \) is a suitable choice of function modeling the interaction. The test of the model against the alternative in (4.1) can be carried out either as a likelihood ratio test or as a \( \chi^2 \)-test fitting only (4.1).

### 4.2. No additional effect of an item on survival given latent class.

Violations of the assumption no additional effect of an item on survival given latent class may be investigated graphically somewhat similar to the proportional hazards assumption. Define

\[
\hat{A}_j^y(t) = \sum_{i=1}^{N} \frac{\delta_i I(u_i \leq t, Y_{ij} = y)}{\sum_{j \in \mathcal{R}_i} \exp(z_j \beta + \hat{\nu}_C)}, \quad y = 0, 1,
\]

for the items \( j = 1, \ldots, J \). If the is correct, then \( \hat{A}_j^0(t) \) and \( \hat{A}_j^1(t) \) should estimate the same function. This can be illustrated by plotting the two estimates against \( t \) (or \( f(t) \)), or by plotting them against each other. Doing the latter should show (approximately) a straight line with slope one going through the origin.

If the graphical check indicates association between an item and survival even with the latent class variable in the model, then we suggest to test the model against the alternative given by

\begin{equation}
\alpha_i(t) = \alpha_0(t) \exp(z_i \beta + \nu_{ci} + \mu I(Y_{ij} = 1)),
\end{equation}

where \( \mu \) is a scalar. The test of the model can be carried out either as a likelihood ratio test or as a t-test fitting only (4.2).

### 5. Example

WHAS is a prospective population-based study of disability in community-dwelling women of age 65 years and older. The study is carried out by Johns Hopkins, NIA, and it takes place in Baltimore, where 1002 women were selected after an initial screen, so that the group mainly consists of women with moderate to poor physical function.
The WHAS screening instrument measures aspects of disability and disease, and information about some demographic variables and death are also included. The demographic variables used here are age in years of the woman at the time of the interview, and education, which is the indicator of the woman having more than 10 years of school. The domain of mobility/exercise is sought measured by five binary items of difficulty doing specific tasks. These are: "With or without help, do you have any difficulty,…", walking a \( \frac{1}{4} \) of a mile (walk), climbing 10 steps (steps), getting in and out of bed or chairs (chair), doing heavy housework (hhw) and lifting up to 10 pounds (lift). Further, information about time of their death (from their death certificate), or last time they are known to be alive, are used in the analyses.

The proposed methodology is motivated in Subsection 5.1, the latent class regression model is established in Subsection 5.2, and the full model is illustrated in Subsection 5.3.

All analyses are carried out using Splus.

5.1. **Naïve survival analyses.**

A conventional approach to the analyses of time to death is to use a proportional hazards model with the hazard on the form:

\[
\alpha_i(t|z_i) = \alpha_0(t) \exp(z_i\beta),
\]

where \( z_i \) is the vector of covariates, all measured without measurement error or misclassification, and \( \beta \) is the vector of corresponding parameters.

Estimates are obtained by maximizing the partial likelihood (Cox, 1972), and the results of the analyses are in Table 1. The analyses show a large and highly significant effect of age on survival: each additional year of age increase the death intensity about 6%. Further, the age effect seems to be linear or close to linear, and there does not seem to be an effect of having more than 10 years of school on survival.

In analyses with age and each of the five items at a time, excessive intensities for women with difficulties as measured by the items lie between 37% and 87%
compared to those with no difficulties. Whereas these effects are all highly significant, each of the analyses use only a fraction of the information available from the questionnaire. However, in a joint analysis with age and all five items, only one of the variables (walk) is significant on a 5% level. Including several items as covariates in this way introduces a number of problems: (i) collinarity among the items will make it difficult to put believable substantive meaning in the parameters, (ii) all the items are specific to doing a concrete task and do each measure special aspects of exercise/mobility, rather than the more general domain itself, and (iii) the model does not allow for misclassification.

An analysis using age and the sum scale of the five items shows a large effect of the scale. This model is equivalent to the model with the five items as covariates with the extra assumption of all items having the same effect on the relative hazard. This assumption is indeed hard to justify, and the estimates from the five separate analyses do not support it. This and the lack of ability to handle misclassification using the scale disqualify this approach as well.

In conclusion, the conventional approach using Cox’s model gives rise some difficulties when covariates are multiple items from a questionnaire. To accommodate for the shortcomings of the approach taken here, a latent class model is introduced for the items.

5.2. **Latent class modeling.**

The relationship between the five items and the covariates is modeled by a latent class model. Bandeen-Roche (1997) found that the exercise/mobility domain may be modeled well using three classes.

In Table 2 estimates and test statistics from three models are shown. The models are latent class models with (a) no covariates, (b) age as covariate, and (c) age and education as covariates including interaction between the two variables. Likelihood ratio tests show that both age and education are significant (p-values of < 0.001 and 0.03, respectively). Class 1 consists of individuals with no or few problems in mobility/exercise, class 2 contains those with moderate problems and class 3 those with multiple problems. The item characteristic parameters in model (b) and (c) are very similar, and somewhat different from those of model (a). The same is true
Table 1. Survival analyses without latent variables.

<table>
<thead>
<tr>
<th>Effect</th>
<th>Est.</th>
<th>Exp(est.)</th>
<th>Std.err.</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.056</td>
<td>1.06</td>
<td>0.007</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Age with each of the following variables:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Age − 78)^2</td>
<td>0.0005</td>
<td>1.00</td>
<td>0.0008</td>
<td>0.55</td>
</tr>
<tr>
<td>Education</td>
<td>0.005</td>
<td>1.01</td>
<td>0.056</td>
<td>0.92</td>
</tr>
<tr>
<td>Hhw</td>
<td>0.312</td>
<td>1.37</td>
<td>0.127</td>
<td>0.01</td>
</tr>
<tr>
<td>Walk</td>
<td>0.628</td>
<td>1.87</td>
<td>0.113</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Step</td>
<td>0.473</td>
<td>1.60</td>
<td>0.113</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Lift</td>
<td>0.361</td>
<td>1.44</td>
<td>0.112</td>
<td>0.001</td>
</tr>
<tr>
<td>Chair</td>
<td>0.423</td>
<td>1.53</td>
<td>0.156</td>
<td>0.007</td>
</tr>
<tr>
<td>Hhw+Walk+Step+Lift+Chair</td>
<td>0.200</td>
<td>1.22</td>
<td>0.036</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Age jointly with the following variables:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hhw</td>
<td>0.062</td>
<td>1.06</td>
<td>0.136</td>
<td>0.65</td>
</tr>
<tr>
<td>Walk</td>
<td>0.500</td>
<td>1.64</td>
<td>0.131</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Step</td>
<td>0.185</td>
<td>1.20</td>
<td>0.134</td>
<td>0.17</td>
</tr>
<tr>
<td>Lift</td>
<td>0.062</td>
<td>1.06</td>
<td>0.129</td>
<td>0.63</td>
</tr>
<tr>
<td>Chair</td>
<td>0.114</td>
<td>1.12</td>
<td>0.168</td>
<td>0.50</td>
</tr>
</tbody>
</table>

for the average posterior probabilities: model (a) estimates more individuals to belong to class 1 than model (b) and (c), whereas there is little difference between the two latter models.

Graphs for the class membership probabilities calculated from model (c) are shown in Figure 1 (the dotted curves). The fraction of women in class 3 is fairly constant around 20% independently of age and education, except for low educated, older women, where the fraction is higher. More women belong to class 1 in the group of higher educated than in the group of lower educated women, though in both groups the fraction shrinks with age.

In the following analysis (c) is preferred for the latent class part of the model.
Table 2. Estimated class characteristic parameters and posterior class probabilities in the latent class model (a) without covariates, (b) with time as covariate and (c) with time and education as covariates. Parameter estimates for the latent class regressions. Likelihood ratio test for the models.

<table>
<thead>
<tr>
<th>Mobility/Class: 1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>exercise domain</td>
<td>(a)</td>
<td>(b)</td>
</tr>
<tr>
<td>Hlw</td>
<td>0.44</td>
<td>0.35</td>
</tr>
<tr>
<td>Walk</td>
<td>0.12</td>
<td>0.04</td>
</tr>
<tr>
<td>Step</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>Lift</td>
<td>0.07</td>
<td>0.05</td>
</tr>
<tr>
<td>Chair</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Ave. class prob.:</td>
<td>0.45</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Class comparison

<table>
<thead>
<tr>
<th>Class 3 vs. class 1</th>
<th>Intercept</th>
<th>Age - 78</th>
<th>Education</th>
<th>(Age - 78)×Education</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>-0.93</td>
<td>0.050</td>
<td>-0.48</td>
<td>-0.001</td>
</tr>
<tr>
<td>(b)</td>
<td>-0.19</td>
<td>0.047</td>
<td>-0.48</td>
<td>-0.001</td>
</tr>
<tr>
<td>(c)</td>
<td>0.11</td>
<td>0.047</td>
<td>-0.48</td>
<td>-0.001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class 2 vs. class 1</th>
<th>Intercept</th>
<th>Age - 78</th>
<th>Education</th>
<th>(Age - 78)×Education</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>-0.20</td>
<td>0.047</td>
<td>-0.65</td>
<td>0.037</td>
</tr>
<tr>
<td>(b)</td>
<td>0.38</td>
<td>0.028</td>
<td>-0.65</td>
<td>0.037</td>
</tr>
<tr>
<td>(c)</td>
<td>0.77</td>
<td>0.028</td>
<td>-0.65</td>
<td>0.037</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>-2 log likelihood</th>
<th>LR</th>
<th>df</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) No covariates</td>
<td>5084.682</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(b) Age as covariate</td>
<td>5070.786</td>
<td>13.896</td>
<td>2</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>(c) Age and education as covariates</td>
<td>5060.040</td>
<td>10.746</td>
<td>4</td>
<td>0.03</td>
</tr>
</tbody>
</table>

5.3. Joint modeling of latent class and survival.
The models considered in this subsection consist of two parts, a latent class regression model, and a proportional hazards model. The analyses in subsection 5.2 show an interacting effect of age and education on mobility/exercise, and therefore all models considered here have age and education as covariates for the latent class regression. However, different covariates are included for the survival part of the model. Likelihood ratio tests for covariate effects are provided, and plots illustrating specific violations of model assumptions are presented along with tests against specific model violations.

Table 3. Test statistics for combined analyses. All models have interacting effect of age and education on mobility/exercise. Models with different covariates entering the hazard are compared.

<table>
<thead>
<tr>
<th>Model</th>
<th>-2 log likelihood</th>
<th>LR</th>
<th>df.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(d) Age and mobility/exercise</td>
<td>10094.67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Models that are nested in (d):</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(e) Age</td>
<td>10126.81</td>
<td>32.14</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>(f) Mobility/exercise</td>
<td>10147.03</td>
<td>52.36</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Model in which (d) is nested:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(g) Age, mobility/exercise and (Age – 78)^2</td>
<td>10094.43</td>
<td>0.24</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>(h) Age, mobility/exercise and education</td>
<td>10094.19</td>
<td>0.48</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>(i) Age, mobility/exercise and a linear effect of time within latent class</td>
<td>not yet ready</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(j1) Age, mobility/exercise and hhw</td>
<td>10093.89</td>
<td>0.78</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>(j2) Age, mobility/exercise and walk</td>
<td>10090.25</td>
<td>4.42</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>(j3) Age, mobility/exercise and step</td>
<td>10094.36</td>
<td>0.31</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>(j4) Age, mobility/exercise and lift</td>
<td>10094.12</td>
<td>0.55</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>(j5) Age, mobility/exercise and chair</td>
<td>10094.66</td>
<td>0.01</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Table 3 contains test statistics from fitting a number of different models. Model (d) includes effects of age and mobility/exercise as defined by the latent class model. Likelihood ratio tests of model (e) and (f) show that both age and mobility/exercise are highly significant. The assumption of the age effect being linear is tested by
adding a quadratic term (model (g)), providing no evidence against linearity. Including education (model (h)) does not improve the fit significantly either.

Estimates and standard errors from model (d) are found in Table 4. There is little difference between the class specific parameters from this model and those of model (c). The same is true for the latent class regression parameters. These are

**Figure 1.** Estimated class probabilities from model (c), the latent class regression without the time to event part (dotted) and (d), the combined analysis with age and latent class predicting time to event (solid).
Figure 2. Plots for checking model assumptions: proportional hazards between latent classes, and possible additional effects of hhw, walk, step, lift and chair on time to event.
Table 4. Parameter estimates and standard errors for model (d): the combined model with interacting effect of age and education on mobility/exercise, and effect of age and mobility/exercise on time to event.

<table>
<thead>
<tr>
<th>Mobility/ exercise domain</th>
<th>Class: 1</th>
<th>Class: 2</th>
<th>Class: 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hhw</td>
<td>0.30</td>
<td>0.10</td>
<td>0.74</td>
</tr>
<tr>
<td>Walk</td>
<td>0.03</td>
<td>0.07</td>
<td>0.44</td>
</tr>
<tr>
<td>Step</td>
<td>0.01</td>
<td>0.02</td>
<td>0.16</td>
</tr>
<tr>
<td>Lift</td>
<td>0.02</td>
<td>0.04</td>
<td>0.27</td>
</tr>
<tr>
<td>Chair</td>
<td>0.00</td>
<td>—</td>
<td>0.04</td>
</tr>
<tr>
<td>Ave. class prob.</td>
<td>0.26</td>
<td></td>
<td>0.48</td>
</tr>
</tbody>
</table>

Effect on class membership

<table>
<thead>
<tr>
<th>Class 3 vs. class 1</th>
<th>Intercept</th>
<th>Est.</th>
<th>Std. err.</th>
<th>Class 2 vs. class 1</th>
<th>Intercept</th>
<th>Est.</th>
<th>Std. err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age - 78</td>
<td>0.048</td>
<td>0.021</td>
<td></td>
<td>Age - 78</td>
<td>0.026</td>
<td>0.023</td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>-0.49</td>
<td>0.26</td>
<td></td>
<td>Education</td>
<td>-0.56</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>(Age - 78)×Education</td>
<td>0.004</td>
<td>0.030</td>
<td></td>
<td>(Age - 78)×Education</td>
<td>0.044</td>
<td>0.031</td>
<td></td>
</tr>
</tbody>
</table>

Effect on time to event

<table>
<thead>
<tr>
<th></th>
<th>Est.</th>
<th>Std. err.</th>
<th>Exp(est.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>0.051</td>
<td>0.007</td>
<td>1.05</td>
</tr>
<tr>
<td>Class 3 vs. class 1</td>
<td>1.080</td>
<td>0.259</td>
<td>2.95</td>
</tr>
<tr>
<td>Class 2 vs. class 1</td>
<td>0.552</td>
<td>0.285</td>
<td>1.74</td>
</tr>
</tbody>
</table>

shown for the two models (c) and (d) in Figure 1. The age effect for the time-to-event part of the model is similar to that of the simple analysis in Table 2. That is, each additional year adds about 5% to the hazard. The latent class effects are more difficult to compare with the results from the naive analyses. Reporting multiple problems (class 3) implies an excessive risk of 195% compared with those reporting
little or no problems (class 1) given age, and moderate problems (class 2) implies an excessive risk of 74% compared to the class with little or no problems.

To check the model, latent class variables are simulated from the posterior distribution given the observed variables in model (d) using the maximum likelihood estimates from Table 4. The simulated class variables and the observed data are used for graphical model checking of the assumption of proportional hazards and no additional effect of items on survival given latent class.

The upper left plot of Figure 2 shows the estimated integrated hazard for the three classes based on the simulation. The plot should reveal deviations from the proportional hazards assumption between the classes. The proportionality assumption does not seem to be violated, and this is also supported by the test of model (i) in Table 3.

The five latter plots in Figure 2 show the estimated integrated hazards, each stratified by the response on one of the five items. The solid curve is the estimated integrated hazard for those reporting no problems doing that specific task, and the dotted curve is the hazard for those having problems doing it. If there is no additional effect of the item on survival, then the two curves should be close to identical. None of the five plots suggests a serious violation of the model. The lihw and walk plots may suggest a small extra effect of the two variables on survival. Test for extra effect of the five variables are found in Table 3. Only walk shows a significant difference ($p = ????$) between the two groups.

In short, neither the graphical nor the parametric tests detected any serious violation of the model. However, both approaches indicate an extra effect of having problems walking on survival, though the evidence is not overwhelming.

6. Discussion

...
\[ c_1, c_2 = 1, \ldots, K, p_1, p_2 = 1, \ldots, P, \]
\[ I_{\beta_1, \beta_2} = \sum_{i=1}^{N} z_{iq_1} z_{iq_2} A_0(u_i) \exp(\mathbf{z}_i \mathbf{\beta} + \nu_{c_i}) [U_i, \Delta_i, \mathbf{Y}_i; \mathbf{\theta}^{(r)}], \]
\[ q_1, q_2 = 1, \ldots, Q, \]
\[ I_{\nu_1, \nu_2} = I(c_1 = c_2) \sum_{i=1}^{N} \omega_{ic_1}^{(r)} A_0(u_i) \exp(\mathbf{z}_i \mathbf{\beta} + \nu_{c_1}), \]
\[ c_1, c_2 = 1, \ldots, K, \]
\[ I_{\beta_1, \nu_c} = \sum_{i=1}^{N} \omega_{ic}^{(r)} z_{iq} A_0(u_i) \exp(\mathbf{z}_i \mathbf{\beta} + \nu_c), \]
\[ q = 1, \ldots, Q, c = 1, \ldots, K. \]

REFERENCES


