The Oncologist
The official journal of the Society for Translational Oncology

Supplemental Appendix for:
Heterogeneity in quality of life of long-term colon cancer survivors: a latent class analysis of the population-based PROFILES registry
Felix Johannes Clouth et al.

Supplemental Material

When using LCA, we want to estimate the probability that a patient belongs to an unobserved class of patients based on the response pattern on a set of categorical indicators. We used the three-step procedure proposed by Vermunt[1] to correctly estimate model parameters which we present below.

In step 1, a standard latent class model without covariates is estimated. Let Y_{ik} denote the response of patient *i* on indicator *k* and $\mathbf{Y}_\mathrm{i}$ the full response vector of patient *i*. The categorical latent class variable is given by *X* with *s* being a particular class of that vector with *S* classes. A latent class model for the probability of the response vector $P(Y_i)$ can be defined as:

$$P(\mathbf{Y}_i) = \sum_{s=1}^{S} P(X = s)P(\mathbf{Y}_i|X = s)\,\#(1)$$

With responses being assumed independent given class membership (conditional independence):

$$P(\mathbf{Y}_i|X = s) = \prod_{k=1}^{K} P(Y_{ik}|X = s)\,.\#(2)$$

Model parameters for the class proportion $P(X = s)$ and the class-specific response probabilities $P(\mathbf{Y}_i|X = s)$ are estimated by maximum likelihood.

In step 2, we obtain the posterior class membership probabilities. Additional to the estimates obtained in the first step, usually we are interested in the classification of the patients. The posterior class membership probability $P(X = s|\mathbf{Y}_i)$ (the probability that patient *i* belongs to the latent class *s* given the patient's response vector $\mathbf{Y}_\mathrm{i}$) can easily be obtained using Bayes rule:

$$P(X = s|\mathbf{Y}_i) = \frac{P(X = s)P(\mathbf{Y}_i|X = s)}{P(\mathbf{Y}_i)}\#(3)$$

with the terms on the right-hand side being defined above. In this second step, we use modal assignment. We estimate the assigned class W_i which is the value of *s* for which $P(X = s|\mathbf{Y}_i)$ is the largest.

In step 3, we obtain parameter estimates for the effect of covariates on the class membership using class assignment W_i and a maximum likelihood based correction which we present in detail here. Relating a set of covariates **Z** on class *X* can easily be done using a multinomial logistic (MNL) regression model:

$$P(X = s|\mathbf{Z}_i) = \frac{\exp\left(\gamma_{0s} + \sum_{q=1}^{Q} \gamma_{qs} Z_{iq}\right)}{\sum_{s=1}^{S} \exp\left(\gamma_{0s} + \sum_{q=1}^{Q} \gamma_{qs} Z_{iq}\right)} \#(4)$$

With Z_{iq} being one of Q covariates in the covariate vector $\mathbf{Z}_i$ for patient *i* and γ_{qs} being the regression coefficients for $0 \leq q \leq Q$.

Note that the MNL regression in Eq. 4 considers the effect of the covariates **Z** on the true class *X* and not on the class assignment *W*. However, as shown in Eq. 5, the classification error can be quantified as the probability of *r* being a realization of the class assignment *W* conditional on *s* being a realization of the true class *X* where $r \neq s$. Note that this classification error can be easily obtained by summing over all observed posterior class membership probabilities and class assignments in the data set:

$$P(W = r|X = s) = \frac{\sum_{i=1}^{N} P(X = s|Y_i) P(W_i = r|Y_i)}{P(X = s)} . \#(5)$$

It can be shown that $P(W = r|\mathbf{Z}_i)$ is related to $P(X = s|\mathbf{Z}_i)$ as follows:[2]

$$P(W = r|\mathbf{Z}_i) = \sum_{s=1}^{S} P(X = s|\mathbf{Z}_i) P(W = r|X = s) . \#(6)$$

Equation 6 represents a special type of latent class model with a single indicator *W* and known error probabilities $P(W = r|X = s)$. Since the error probabilities are known from Eq. 5 in step two, they do not need to be estimated in Eq. 6. The correct estimate for the effect of the covariates on the class assignment is obtained by maximizing the likelihood function given in Eq. 7:

$$\log L_{ML} = \sum_{i=1}^{N} \log \sum_{s=1}^{S} P(X = s|\mathbf{Z}_i) P(W = r|X = s) . \#(7)$$

Note that since $P(X = s|\mathbf{Z}_i)$ is defined in Eq. 4 as a MNL regression model, maximizing the likelihood function in Eq. 7 will yield estimates for the γ coefficients.

Additionally to the effect of the covariates on the class membership, we were interested in the stability of this membership over time. We therefore estimated transition probabilities between the classes using LTA. However, because of a considerable drop out of participants after the first measurement occasion (Fig. 1), we used data from only the first wave to identify our cluster solution using three-step LCA. Further, we fixed this class-solution or measurement model for the second and third wave instead of re-estimating it. That is, we applied the posterior class membership probabilities obtained from LCA on the first measurement occasion to the observed response patterns of wave two and three. For the drop out cases, the classifications were considered missing values in subsequent analyses since we did not apply these posterior class membership probabilities. Based on these classifications, we estimated the transition probabilities over the three measurement occasions.

In detail, we apply the three-step procedure as described above but substitute $P(X = s|\mathbf{Z}_i)$, the probability of X conditional on the set of covariates Z, by the Markov chain $P(X_0)\prod_{t=1}^{T} P(X_t|X_{t-1})$ with *t* indicating the time points *t* = 0,…,*T.*[3,4] As shown in Eq. 8, the set of class assignments **W** is given by the initial class probability $P(X_0)$, the probabilities of transitioning between classes $P(X_t|X_{t-1})$, and the known error probabilities $P(W = r|X = s)$, aggregated over all time points *t*; that is,

$$P(\mathbf{W} = \mathrm{r}) = \sum_{s_0,\dots,s_T} P(X_0 = s_0) \prod_{t=1}^{T} P(X_t = s_t|X_{t-1} = s_{t-1}) \prod_{t=0}^{T} P(W_t = s_t|X_t = s_t)\,.\#(8)$$

Note that we assume a first-order Markov process for the transition probabilities, i.e., each class at time point *t* only depends on the class at time point *t*-1. Parameter estimates are obtained by maximizing the likelihood function given in Eq. 9:

$$L_{step3} = \sum_{i=1}^{N} log \sum_{s_0,\dots,s_T} P(X_0 = s_0) \prod_{t=1}^{T} P(X_t = s_t|X_{t-1} = s_{t-1}) \prod_{t=0}^{T} P(W_{it} = r_{it}|X_t = s_t)\,.\#(9)$$

References

1 Vermunt JK. Latent class modeling with covariates: Two improved three-step approaches. Polit Anal 2010;18:450–469.

2 Bolck A, Croon M, Hagenaars J. Estimating Latent Structure Models with Categorical Variables: One-Step Versus Three-Step Estimators. Polit Anal 2004;12:3–27.

3 Bartolucci F, Montanari GE, Pandolfi S. Three-step estimation of latent Markov models with covariates. Comput Stat Data Anal 2015;83:287–301.

4 Di Mari R, Oberski DL, Vermunt JK. Bias-Adjusted Three-Step Latent Markov Modeling With Covariates. Struct Equ Model 2016;23:649–660.