SUPPLEMENTARY MATERIAL

Simulation of time-to-event data

Survival times were simulated in a similar way to that described by Bender et al. (2005). Two classes, 1 and 2, and two treatment groups, A and B, were simulated in each scenario. The reference survival curve used was based on the Kaplan-Meier estimate of the gemcitabine arm in the ESPAC2 trial (Neoptolemos et al., 2010), Figure S1(a) and (b). Let S_0 represent this reference survival curve for subjects belonging to *true* latent class 2 and treatment group B (i.e. $c_i = 0$ and $z_i = 0$), so that survival probabilities corresponding to proportional hazard effects can be obtained using

$$S_i = S_0^{\exp(\beta z_i + \gamma_j c_i)}, \tag{8}$$

in this case for z_i and $c_i \in \{0, 1\}$. As described previously, the hazard ratio for the effect of Treatment A relative to Treatment B, $\exp(\beta)$, was fixed at 0.75 and the hazard ratio for the effect of latent class 1 relative to latent class 2 was varied, $\exp(\gamma_1) \in \{1, 1.5, 2, 3\}$. 'True' survival probabilities were obtained for each of the four permutations of class and treatment over a sequence of 0 to 60 months in steps of 0.1 months. High-dimensional spline fits were used to approximate these survival curves, as shown for the reference survival curve in Figure S1(a). The splines were fitted separately to each of the four survival curves by regressing the time sequence on polynomials of the survival probabilities. A survival probability was then simulated for each subject from $Uniform(0, 1)$ and a corresponding survival time is obtained from the relevant spline fit. Administrative censoring was applied at 60 months and uniform censoring was added by generating censoring times from an exponential distribution such that overall approximately 50%

of survival times were right-censored.

A label switching algorithm

Latent class models are only identifiable up to a permutation of class labels (McLachlan and Peel, 2004). Whilst this is not an issue in standalone applications, it is a problem for simulation studies since it is not always straight-forward to establish, for a particular simulated data set, the class label that corresponds to the true class. A useful discussion of this issue in latent variable models is given in Tueller et al. (2011), and the same labelling problem arises in Bayesian Monte Carlo Markov Chain simulations (Celeux et al., 2000; Grün and Leisch, 2009; Sperrin et al., 2010).

A number of solutions have been proposed (e.g. Tueller et al., 2011; Yao, 2015; Celeux et al., 2000). In this study, we used a clustering and relabelling strategy based on Euclidean distances, as in Celeux et al. (2000), where the distances between the true parameter values and their estimates is calculated for each simulated data set.

Assume that data are simulated according to a particular latent class model with P 'true' parameter values $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_P)$. There are $J!$ possible permutations of the class labels, $l = 1, \ldots, J!$, and we let the last permutation represent the correct labelling. We simulate $d = 1, \ldots, D$ data sets according to the true model. For each data set we fit a latent class model of the same form as the true model. Let $\hat{\boldsymbol{\theta}}_d = \left(\hat{\theta}_{d1}, \ldots, \hat{\theta}_{dP}\right)$ be a vector of parameter estimates from the latent class model fitted to the d^{th} data set. We assume that $\hat{\boldsymbol{\theta}}_d$ are unbiased estimates of the true values but possibly labelled incorrectly. If $\hat{\boldsymbol{\theta}}_d$ are labelled 'correctly', then

$$\frac{1}{se(\hat{\theta}_{dp})}\left(\hat{\theta}_{dp} - \theta_p\right) \sim N(0, 1)$$

and

$$\left[\frac{1}{se(\hat{\theta}_{dp})}\left(\hat{\theta}_{dp}-\theta_p\right)\right]^2 \sim \chi^2_{(1)},$$

for $d = 1, \ldots, D$ and where $1 \leq p \leq P$. We then assume that parameter estimates are independent, which in practice will be determined by the form of the model fitted; this issue is discussed further below. Summing over P,

$$\sum_{p=1}^{P}\left[\frac{1}{se(\hat{\theta}_{dp})}\left(\hat{\theta}_{dp}-\theta_p\right)\right]^2 \sim \chi^2_{(P)},$$

with mean P, for $d = 1, \ldots, D$. The standardised Euclidean distance, δ_d, between the estimates from a model fitted to the dth data set and the vector of true parameter values is

$$\delta_d = \left\{\sum_{p=1}^{P}\left[\frac{1}{se(\hat{\theta}_{dp})}(\hat{\theta}_{dp}-\theta_p)\right]^2\right\}^{1/2} \sim \chi_{(P)},$$

i.e. a central χ distribution. If $\hat{\boldsymbol{\theta}}_d$ are labelled 'incorrectly', then

$$\frac{1}{se(\hat{\theta}_{dp})}\left(\hat{\theta}_p-\theta_p\right) \sim N\left(\mu_p, 1\right),$$

$$\sum_{p=1}^{P}\left[\frac{1}{se(\hat{\theta}_{dp})}\left(\hat{\theta}_{dp}-\theta_p\right)\right]^2 \sim \chi^2_{(P)}(\lambda),$$

i.e. a non-central χ^2 distribution with non-centrality parameter $\lambda = \sum_{p=1}^{P}\mu_p^2$ and mean $P + \lambda$. It therefore follows that

$$\delta_d \sim \chi_{(P)}\left(\lambda\right).$$

Letting $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_D)$, and assuming that the random starting values for the param-

eter estimates do not favour one label permutation over another,

$$\boldsymbol{\delta} \sim \frac{1}{J!} \sum_{l=1}^{J!} f_l(\boldsymbol{\delta}),$$

i.e. a $J!$ component mixture distribution with one central χ distribution, $f_{J!} \sim \chi_{(P)}$, and $J!-1$ non-central χ distributions, $f_l \sim \chi_{(P)}(\lambda_l)$, for $l = 1, \ldots, J!-1$. A histogram of $\boldsymbol{\delta}$ should therefore yield a mixture distribution of $J!$ (hopefully distinct) probability distributions for which the component with the lowest mean is labelled correctly. Larger differences in the true parameter values for the latent classes and greater numbers of class distinct parameters to estimate will result in clearer separation of the mixture components, making clustering and relabelling easier. An example of such a histogram for 2000 simulations from a latent class model with $J = 2$ and $P = 21$ parameters is depicted in Figure S2. Assuming sufficient separation between components, either by introducing some threshold or by clustering $\boldsymbol{\delta}$ (e.g. K-means clustering), estimates that have been labelled incorrectly can be easily identified and relabelled accordingly.

A label switching algorithm is therefore as follows

1. Fit latent class models to each of $d = 1, \ldots, D$ data sets.

2. For each of the $d = 1, \ldots, D$ data sets calculate the standardised Euclidean distances, δ_d, between each set of parameter estimates, $\hat{\boldsymbol{\theta}}_d$, and the true parameter values, $\boldsymbol{\theta}$.

3. Inspect a histogram of $\boldsymbol{\delta}$ for distinct component densities.

4. Use K-means clustering to assign each δ_d (and hence $\hat{\boldsymbol{\theta}}_d$) to a cluster (/component density). The cluster with the lowest mean corresponds to the cluster of correctly labelled parameter estimates.

5. Relabel those $\hat{\boldsymbol{\theta}}_d$ which do not belong to the correctly labelled component density.

As a check, the first three steps can be repeated using the relabelled estimates and the new histogram should reveal a unimodal (and central) χ distribution. If $J > 2$ it may be necessary to repeat this process using a permutation of the true parameter values in the place of the true values in order to distinguish between two or more incorrectly labelled clusters.

The histogram of $\boldsymbol{\delta}$ also serves as a useful diagnostic tool, since any outlying values, perhaps exceeding a selected critical threshold, can be identified and investigated further. These may represent local maximum and/or boundary solutions.

In practice, whilst we have provided theoretical justification for the distribution of the standardised Euclidean distances in the case of independent parameters, if dependencies are included in the model, then the algorithm can still be used. In this case, these parameters can be included or excluded in the algorithm, as long as the histogram of the standardised Euclidean distances reveals distinct clusters.

Comparison of the use of different hazard functions in three-step models

In this study one and two-step models used a piecewise exponential baseline hazard function whilst the three-step models used a Cox model in which the baseline hazard is left unspecified. The choice of a piecewise exponential model for the one and two-step models was primarily motivated by the fact that standard errors are easier to obtain when there are few baseline hazard parameters (see Discussion). To illustrate that the three-step methods are not disadvantaged by the choice of hazard function the table below contains results from a small simulation study for Scenario 17 (Low entropy, $N = 500$, HR=2). The results demonstrate that the results are practically unaffected by the choice of baseline hazard function.

[Table S1 about here]

SUPPLEMENTARY TABLE

Model	Estimate	Bias	CI Coverage (%)
MA using Cox model	-0.30	0.39	23.4
MA using PE model	-0.30	0.39	23.8
PA using Cox model	-0.57	0.13	90.4
PA using PE model	-0.57	0.13	90.7

Table S1: Comparison of simulation results for modal assignment and partial assignment when using unspecified (Cox) an piecewise exponential baseline hazard functions. MA Modal assignment, PA Partial assignment. The results demonstrate that the statistical properties of the latent class effect estimates are practically unaffected by the different hazard functions compared here.

SUPPLEMENTARY FIGURES

Figure S1: Kaplan-Meier estimate of overall survival for the gemcitabine arm from the ESPAC3v2 study and overlaid fitted models.

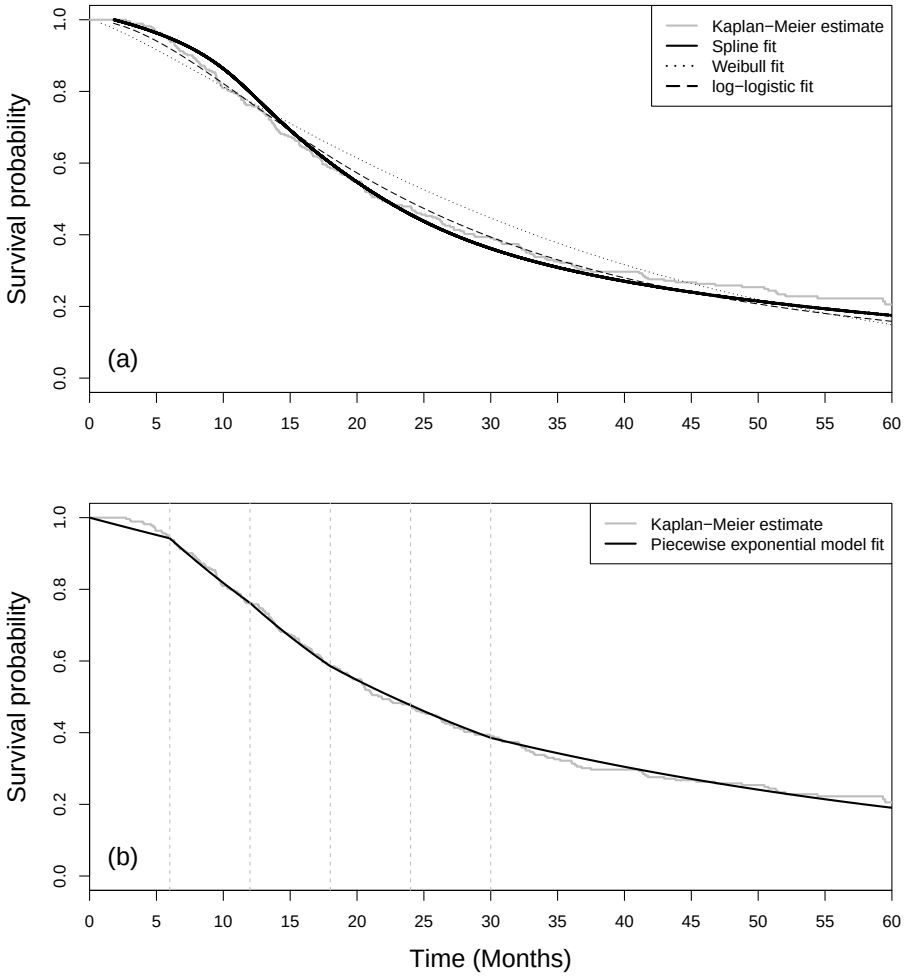

Figure S1: (a) Fitted polynomial spline, Weibull and log-logistic (parametric) models. (b) A piecewise exponential survival model with five partitions approximates the Kaplan-Meier estimate well.

Figure S2: Example of Euclidean distances between true and estimated parameters.

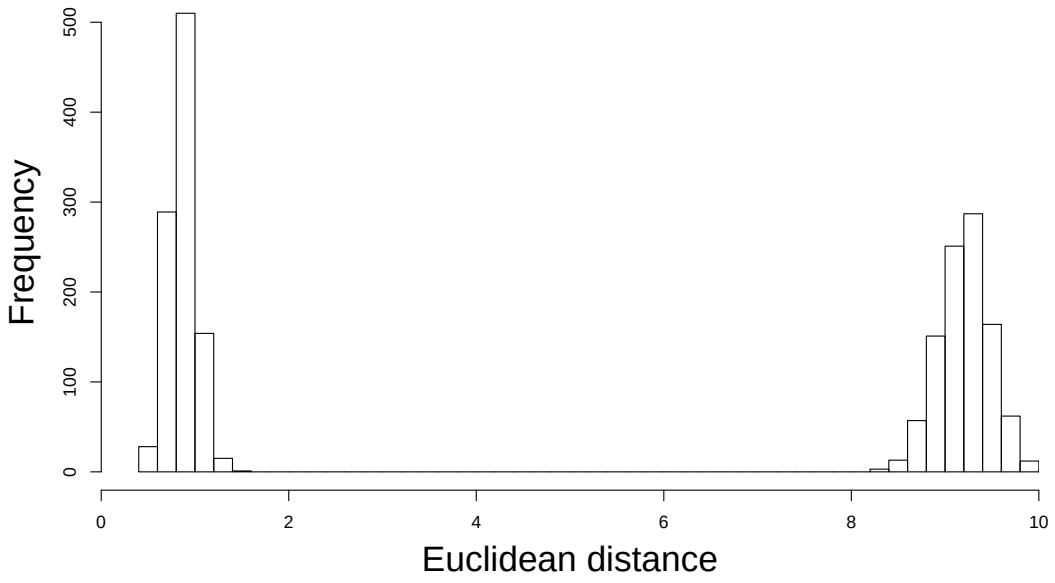

Figure S2: Example of Euclidean distances for 2000 simulations from a latent class model with 2 classes, before relabelling. The distribution on the left contains the models for which the class is correctly labelled.

Supplementary References

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723.

Celeux, G., Hurn, M., and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970.

Grün, B. and Leisch, F. (2009). Dealing with label switching in mixture models under genuine multimodality. *J. Multivar. Anal.*, 100(5):851–861.

McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

Neoptolemos, J. P., Stocken, D. D., Bassi, C., Ghaneh, P., Cunningham, D., Goldstein, D., Padbury, R., Moore, M. J., Gallinger, S., Mariette, C., et al. (2010). Adjuvant chemotherapy with fluorouracil plus folinic acid vs gemcitabine following pancreatic cancer resection: A randomized controlled trial. *JAMA*, 304(10):1073–1081.

Sperrin, M., Jaki, T., and Wit, E. (2010). Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models. *Statistics and Computing*, 20(3):357–366.

Tueller, S. J., Drotar, S., and Lubke, G. H. (2011). Addressing the problem of switched class labels in latent variable mixture model simulation studies. *Structural Equation Modeling: A Multidisciplinary Journal*, 18(1):110–131.

Yao, W. (2015). Label switching and its solutions for frequentist mixture models. *Journal of Statistical Computation and Simulation*, 85(5):1000–1012.