Supplementary Materials: Personalized statistical learning algorithms to improve the early detection of cancer using longitudinal biomarkers

Nabihah Tayob[1] and Ziding Feng[2]

[1]Department of Data Science, Dana-Farber Cancer Institute, MA, U.S.A.

[2]Biostatistics Program, Fred Hutchinson Cancer Research Center, WA, U.S.A.

We have provided additional details of the multivariate fully Bayesian screening algorithm including model specification, assumed priors and posterior risk calculation utilized in the decision rule.

1 Biomarker models

Let Y_{ijk} be the k^{th} marker level for the i^{th} patient at the j^{th} screening time t_{ijk}, where

- $i = 1, \ldots, N$
- $j = 1, \ldots, J_{ik}$
- $k = 1, \ldots, K$.

$D_i = 0$ if the i^{th} patient is cancer-free at the last observation time d_i and $D_i = 1$ if the patient is clinically diagnosed at time d_i.

For cancer-free patients, with $D_i = 0$, the k^{th} marker level is assumed to randomly fluctuate around a constant mean θ_{ik} and follows the model

$$Y_{ijk} = \theta_{ik} + \varepsilon_{ijk}$$
$$\text{where } \varepsilon_{ijk} \sim N(0, \sigma_k^2).$$

For those that develop cancer, with $D_i = 1$, we define an unobserved indicator I_{ik} to distinguish between the two possible models for the k^{th} marker. If $I_{ik} = 0$, then we assume that the k^{th} marker level does not change after cancer onset and follows the same model as control patients.

If $I_{ik} = 1$, then we assume the k^{th} marker levels randomly fluctuates around a constant mean θ_{ik} until an unobserved changepoint time τ_{ik}, after which the k^{th} marker level changes linearly at a rate of γ_{ik} and follows the model

$$Y_{ijk} = \theta_{ik} + \gamma_{ik}(t_{ijk} - \tau_{ik})^+ + \varepsilon_{ijk}$$

where $(.)^+$ indicates the positive part of the expression.

2 Priors for model parameters

The priors assumed in the Bayesian hierarchical model structure are as follows. For parameters common to the biomarker models in both those that develop cancer and that remain cancer free:

- $\theta_{ik} \sim \text{Normal}(\mu_{\theta k}, \sigma^2_{\theta k})$.

We assume uninformative Jeffreys' priors, $1/\sigma_k^2$ where $k = 1, \ldots, K$ for the variability of each biomarker since we have large numbers of patients for paramter estimation.

For the changepoint time and slope parameters in the biomarker models assuming a change in trajectory in those that develop cancer:

- $\log(\gamma_{ik}) \sim \text{Normal}(\mu_{\gamma k}, \sigma^2_{\gamma k})$
- $\tau_{ik} \sim \text{Truncated Normal}_{[d_i - \tau^*_k, d_i]}(d_i - \mu_{\tau k}, \sigma^2_{\tau k})$.

Note that the parameter γ_{ik} is positive reflecting our assumption that biomarker levels increase after cancer onset but appropriate transformations can be accommodated for biomarkers that decrease. The parameter τ^*_k is fixed based on the known preclinical behavior of the cancer. In the case of hepatocellular cancer, a fast growing cancer, the preclinical duration is assumed to be at most 2 years ($\tau^*_k = 2$).

The K biomarkers are connected via the Markov Random Field (MRF) prior assumed for the binary indicators, $\mathbf{I}_i = (I_{i1}, \dots, I_{iK})$.

$$P(\mathbf{I}_i) \propto \exp\left\{\mu_I \left(\sum_{k=1}^{K} I_{ik}\right) + \eta_I \left(\mathbf{I}_i^T R \mathbf{I}_i\right)\right\},$$

where R is a strictly upper triangular matrix (entries above the diagonal are 1, entries in and below the diagonal are 0) reflecting the assumption that all K markers are correlated. Not all biomarkers are expected to increase in all the cases and the parameter μ_I controls the sparsity of the model while η_I regulates the smoothness of the distribution of $\mathbf{I}_i$. These properties are clearer upon examination of the conditional distribution of I_{ik} given all other elements of $\mathbf{I}_i$:

$$P\{I_{ik} | (I_{ik'} : k' \neq k)\} = \frac{\exp\{I_{ik} F(I_{ik})\}}{1 + \exp\{F(I_{ik})\}}$$
$$\text{where } F(I_{ik}) = \mu_I + \eta_I \sum_{k' \neq k} I_{ik'}.$$

The probability of observing a change-point in the k^{th} marker of the i^{th} patient depends on both μ_I and the number of change-points observed in the other $K - 1$ markers, where η_I moderates this dependency. The MRF defines a dependence structure helpful for detecting borderline change-points when there are only a moderate numbers of patinets that develop cancer.

3 Screening rule: Posterior risk calculation

The decision rule for a new $(N+1)^{th}$ patient at screening time t_{ijk} is based on the posterior risk of cancer, given the longitudinal history of each biomarker up to time t_{ijk}. Specifically,

$$\frac{P(D_{N+1}=1|\mathbf{Y}_{N+1})}{P(D_{N+1}=0|\mathbf{Y}_{N+1})} = \frac{P(\mathbf{Y}_{N+1}|D_{N+1}=1)}{P(\mathbf{Y}_{N+1}|D_{N+1}=0)} \times \frac{P(D_{N+1}=1)}{1-P(D_{N+1}=1)},$$

where $\mathbf{Y}_{N+1} = \{Y_{(N+1)j'k},\, j' = 1,\ldots,j \text{ and } k = 1,\ldots,K\}$.