

ARTICLE TYPE

Online Supplementary Materials: A multivariate parametric empirical Bayes screening approach for early detection of hepatocellular carcinoma using multiple longitudinal biomarkers

Nabihah Tayob*¹ | Anna S. F. Lok² | Ziding Feng³

¹Department of Data Science, Dana-Farber Cancer Institute, Massachusetts, United States of America

²Department of Internal Medicine, University of Michigan, Michigan, United States of America

³Biostatistics Program, Fred Hutchinson Cancer Research Center, Washington, United States of America

Correspondence

*Corresponding author: Nabihah Tayob.
Email: ntayob@ds.dfci.harvard.edu

Summary

The early detection of hepatocellular carcinoma (HCC) is critical to improving outcomes since advanced HCC has limited treatment options. Current guidelines recommend HCC ultrasound surveillance every six months in high-risk patients however the sensitivity for detecting early stage HCC in clinical practice is poor. Blood-based biomarkers are a promising direction since they are more easily standardized and less resource intensive. Combining of multiple biomarkers is more likely to achieve the sensitivity required for a clinically useful screening algorithm and the longitudinal trajectory of biomarkers contains valuable information that should be utilized. We propose a multivariate parametric empirical Bayes (mPEB) screening approach that defines personalized thresholds for each patient at each screening visit to identify significant deviations that trigger additional testing with more sensitive imaging. The Hepatitis C Antiviral Long-term Treatment against Cirrhosis (HALT-C) trial provides a valuable source of data to study HCC screening algorithms. We study the performance of the mPEB algorithm applied to serum α -fetoprotein, a widely used HCC surveillance biomarker, and des- γ carboxy prothrombin, an HCC risk biomarker that is FDA approved but not used in practice in the United States. Using cross-validation, we found that the mPEB algorithm demonstrated moderate but improved sensitivity compared to alternative screening approaches. Future research will validate the clinical utility of the approach in larger cohort studies with additional biomarkers.

KEYWORDS:

Biomarkers, Early detection, Empirical Bayes, Longitudinal screening history, Numeric optimization.

Web Appendix A | SEQUENTIAL QUADRATIC PROGRAMMING (SQP) ALGORITHM

The sequential quadratic programming (SQP) is a powerful, efficient and accurate optimization algorithm where at each iteration, an approximate subproblem with a quadratic objective function and linear constraints defines the search direction¹. We use the `fmincon` solver within the Matlab Optimization Toolbox to implement the SQP algorithm. The SQP is one possible algorithm that can be used to find a vector x that is a local minimum to a scalar function $f(x)$ where one (or more) of the following constraints hold:

$$g(x) \leq 0$$

$$g^*(x) = 0$$

$$A \cdot x \leq 0$$

$$A^* \cdot x = 0$$

$$x_{LB} \leq x \leq x_{UB}$$

The basic idea is that instead of solving the nonlinear problem, at each iteration x_t , you instead solve a quadratic subproblem to define the new iterate x_{t+1} that is ideally a good step for the nonlinear problem. The SPQ designs the quadratic subproblem as an application of Newton's method to the Karush-Kuhn-Tucker conditions, the necessary conditions for optimality for a constrained optimization problem. The details of this algorithm are beyond the scope of this manuscript but both Nocedal and Wright (2000)¹ and the Matlab Help Center (<https://www.mathworks.com/help/optim/ug/constrained-nonlinear-optimization-algorithms.html>) provide useful details. In our multivariate PEB algorithm (mPEB), the nonlinear optimization problem we are trying to solve is

$$\begin{aligned} \max_{\mathbf{c}_{i(n+1)}} Pr(\mathbf{Y}_{i(n+1)} \in \mathbf{A}_{i(n+1)} | \bar{\mathbf{Y}}_{in}, \mu^*, \Sigma_\theta^*, \Sigma^*, D_{i(n+1)} = 1), \text{ such that} \\ Pr(\mathbf{Y}_{i(n+1)} \in \mathbf{A}_{i(n+1)} | \bar{\mathbf{Y}}_{in}, \mu, \Sigma_\theta, \Sigma, D_{i(n+1)} = 0) \leq f_0. \end{aligned}$$

Therefore $x = \mathbf{c}_{i(n+1)}$, the the set of thresholds that characterize the positivity region $\mathbf{A}_{i(n+1)}$, and the functions $f(\mathbf{c}_{i(n+1)})$ and $g(\mathbf{c}_{i(n+1)})$ are defined to be

$$f(\mathbf{c}_{i(n+1)}) = -Pr(\mathbf{Y}_{i(n+1)} \in \mathbf{A}_{i(n+1)} | \bar{\mathbf{Y}}_{in}, \mu^*, \Sigma_\theta^*, \Sigma^*, D_{i(n+1)} = 1)$$

$$g(\mathbf{c}_{i(n+1)}) = Pr(\mathbf{Y}_{i(n+1)} \in \mathbf{A}_{i(n+1)} | \bar{\mathbf{Y}}_{in}, \mu, \Sigma_\theta, \Sigma, D_{i(n+1)} = 0) - f_0.$$

Therefore, we are solving for the optimal $\tilde{\mathbf{c}}_{i(n+1)}$ that minimizes $f(\mathbf{c}_{i(n+1)})$ such that $g(\mathbf{c}_{i(n+1)}) \leq 0$. The focus of this manuscript is the development of an early detection screening algorithm for multiple biomarkers and we take advantage of existing tools to develop this algorithm. It may be possible to further improve our algorithm by optimizing the solver but that is not our current goal and could be an avenue of future research.

Web Appendix B | HIERARCHICAL CHANGEPOINT MODEL USED FOR DATA GENERATION IN THE SIMULATIONS

In these simulations, we have generated the longitudinal biomarker data for the cohort from a biologically plausible hierarchical changepoint model². For each biomarker, we assumed that the levels vary randomly around a constant mean in the absence of disease. After the onset of disease (change point time), each biomarker may or may not increase linearly with time. The hierarchical model connects the multiple biomarkers using a Markov random field distribution for the parameters that reflect whether or not each biomarker increases after onset of disease. This distribution ensures the probability of observing a change point in one biomarker is conditional on the number of change points observed in the other biomarkers.

Definitions

- Y_{ijk} : k^{th} marker level for the i^{th} patient at the j^{th} screening time
- t_{ij} : j^{th} screening time for the i^{th} patient
- i indexes the N patients in the study
- j indexes the J_i screening times for the i^{th} patient
- k indexes the K biomarkers in the study.
- D_i : disease status of the i^{th} individual, $D_i = 0$ if the patient is disease-free during the study and $D_i = 1$ if patient develops the disease during the study.
- d_i : the last observation time if $D_i = 0$ and clinically diagnosis time if $D_i = 1$
- $\sim N(., .)$: normal distribution
- $\sim TN_{[.,.]}(., .)$: truncated normal distribution

Without loss of generality, we assume time is measured in years from entry into the cohort.

Model used for data generation

For each of the N patients in the cohort, we generate the variable D_i from the Bernoulli distribution with probability of disease based on the expected number of patients who develop the disease in the cohort using the annual incidence rate and study follow-up length. In addition, the time to last observation or clinical diagnosis time is generated from

$$d_i \sim U[0, 5]$$

and the screening interval is assumed to be

$$t_{ij} - t_{i(j-1)} \sim (6, 0.1).$$

For disease-free patients, with $D_i = 0$:

$$Y_{ijk} = \theta_{ik} + \varepsilon_{ijk}$$

$$\text{where } \varepsilon_{ijk} \sim N(0, \sigma_k^2)$$

$$\text{and } \theta_{ik} \sim N(\mu_{\theta k}, \sigma_{\theta k}^2)$$

For patients that develop the disease during the study, with $D_i = 1$, we generate an 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 increase after disease onset and follows the same model as control patients:

$$Y_{ijk} = \theta_{ik} + \varepsilon_{ijk}$$

$$\text{where } \varepsilon_{ijk} \sim N(0, \sigma_k^2)$$

$$\text{and } \theta_{ik} \sim N(\mu_{\theta k}, \sigma_{\theta k}^2)$$

If $I_{ik} = 1$, then we assume the k^{th} marker level increases after disease onset, under the following model:

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

$$\text{where } \varepsilon_{ijk} \sim N(0, \sigma_k^2),$$

$$\theta_{ik} \sim N(\mu_{\theta k}, \sigma_{\theta k}^2),$$

$$\log(\gamma_{ik}) \sim N(\mu_{\gamma k}, \sigma_{\gamma k}^2),$$

$$\tau_{ik} \sim TN_{d_i - \tau_k^*, d_i}(d_i - \mu_{\tau k}, \sigma_{\tau k}^2)$$

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

The parameter τ_k^* is fixed based on the known preclinical behavior of the disease. In the case of HCC, a fast growing cancer, the preclinical duration is assumed to be at most 2 years ($\tau_k^* = 2$).

The binary indicators, $\mathbf{I}_i = (I_{i1}, \dots, I_{iK})$ are generated from a Markov Random Field (MRF) distribution

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

where \mathbf{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.

Fixed parameter values used in simulation study to generate data.

For scenario C, we used the same parameter values as those listed for scenario A (Web Figure 1 a). The only difference was in study follow-up and hence the probability of disease specified for the Bernoulli distribution used to generate D_i . In scenarios A, B and D, the probability of disease used was 50/400. In scenario C, the probability of disease was 24/400 in the training cohort and 72/400 in the validation cohort.

In scenario D (Web Figure 1 b), we used mostly the same data generation mechanism as scenario A except we generated the parameters θ_{ik} , γ_{ik} and τ_{ik} from bi-modal distributions rather than the unimodal distributions specified in the above hierarchical model.

$$\theta_{ik} \sim \begin{cases} N(\mu_{\theta k}, \sigma_{\theta k}^2), & \text{if } \rho_{\theta} = 1 \\ N(0, \sigma_{\theta k}^2), & \text{if } \rho_{\theta} = 0, \end{cases} \quad (1)$$

where $\rho_\theta \sim \text{Bernoulli}(0.5)$.

$$\log(\gamma_{ik}) \sim \begin{cases} N(\mu_{\gamma k}, \sigma_{\gamma k}^2), & \text{if } \rho_\gamma = 1 \\ N(0, \sigma_{\gamma k}^2), & \text{if } \rho_\gamma = 0, \end{cases} \quad (2)$$

where $\rho_\gamma \sim \text{Bernoulli}(0.5)$.

$$\tau_{ik} \sim \begin{cases} TN_{d_i - \tau_k^*, d_i}(d_i - \mu_{\tau k}, \sigma_{\tau k}^2), & \text{if } \rho_\tau = 1 \\ TN_{d_i - \tau_k^*, d_i}(d_i - 0, \sigma_{\tau k}^2), & \text{if } \rho_\tau = 0, \end{cases} \quad (3)$$

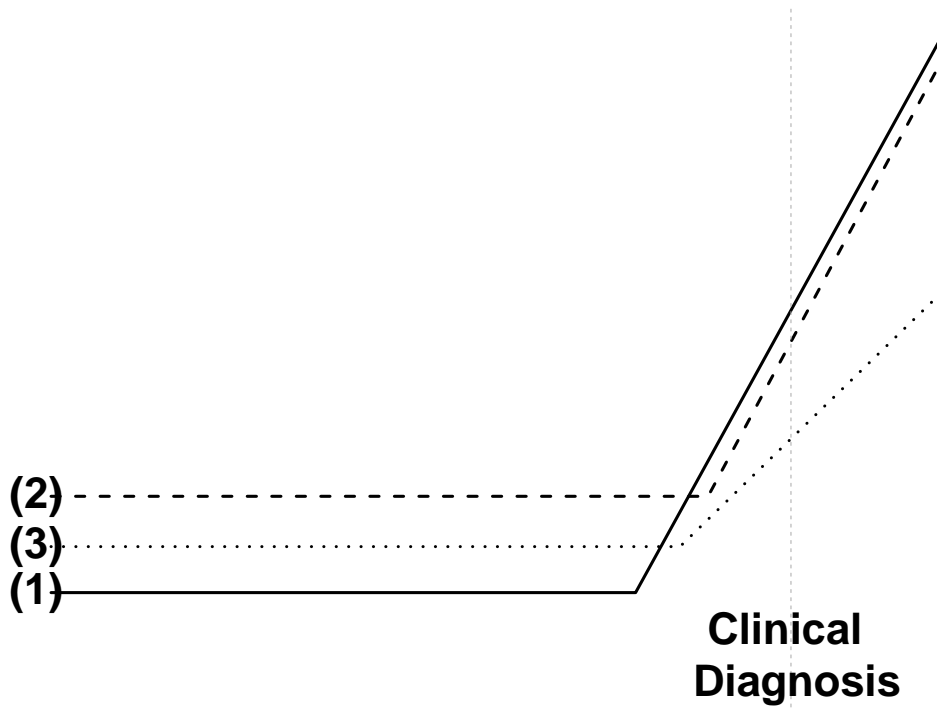
where $\rho_\tau \sim \text{Bernoulli}(0.5)$.

Web Appendix C | ADDITIONAL SIMULATION RESULTS

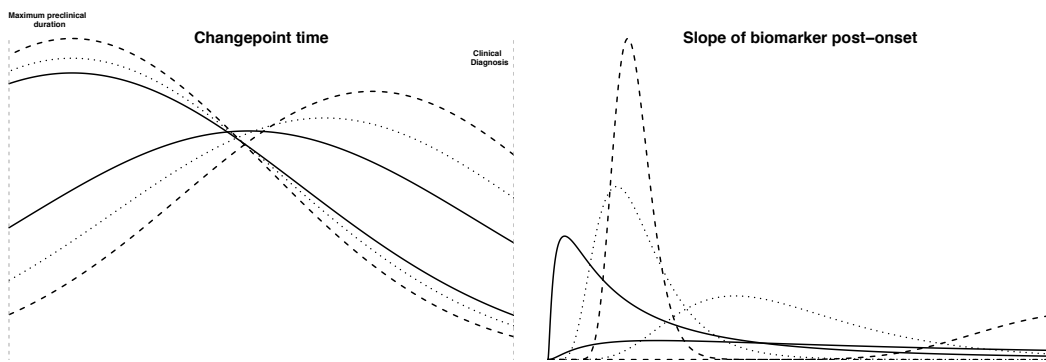
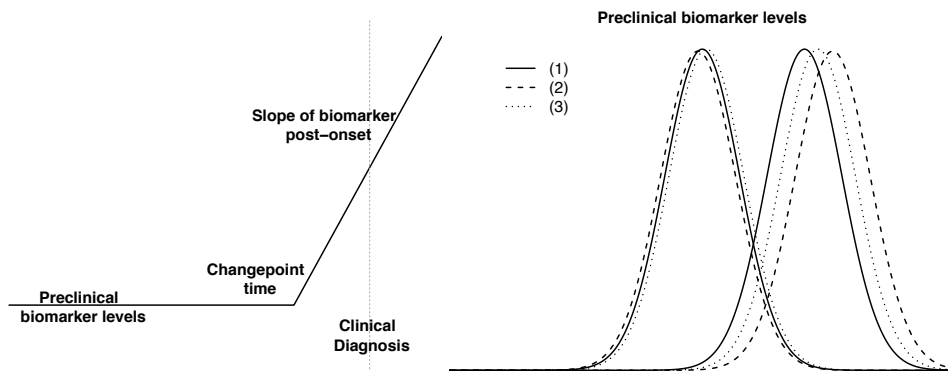
References

1. Nocedal J, Wright S. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering Springer New York . 2000.
2. Tayob N, Stingo F, Do KA, Lok AS, Feng Z. A Bayesian Screening Approach for Hepatocellular Carcinoma Using Multiple Longitudinal Biomarkers. *Biometrics* 2018; 74(1): 249-259.





(a) Scenario D

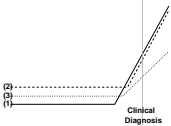
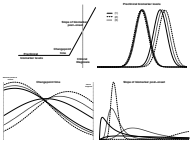


(b) Scenario E

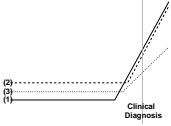
Web Figure 1 Simulation settings: Scenario D and E.

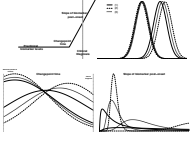
Parameter	Scenario A	Scenario B
σ_1^2	0.23	0.23
$\mu_{\theta 1}$	2.43	2.43
$\sigma_{\theta 1}^2$	0.79	0.79
$\mu_{\gamma 1}$	1.87	0.87
$\sigma_{\gamma 1}^2$	1.61	0.3
$\mu_{\tau 1}$	1.05	1.05
$\sigma_{\tau 1}^2$	0.82	0.82
σ_2^2	1.35	1.35
$\mu_{\theta 2}$	3.10	3.10
$\sigma_{\theta 2}^2$	0.80	0.80
$\mu_{\gamma 2}$	1.92	0.92
$\sigma_{\gamma 2}^2$	0.05	0.05
$\mu_{\tau 2}$	0.56	0.56
$\sigma_{\tau 2}^2$	0.58	0.58
σ_3^2	0.80	0.80
$\mu_{\theta 3}$	2.75	2.75
$\sigma_{\theta 3}^2$	0.79	0.79
$\mu_{\gamma 3}$	1.00	0.65
$\sigma_{\gamma 3}^2$	0.20	0.10
$\mu_{\tau 3}$	0.75	0.75
$\sigma_{\tau 3}^2$	0.70	0.70
μ_I	0.15	0.15
η_I	0.1	0.1

Web Table 1 Fixed parameter values used in simulation study to generate data.

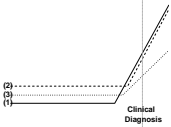
		Scenario D						
		Biomarker	mPEB	mFB	lPEB	uPEB	uFB	ST
(1)					43.24 (0.47)	37.45 (0.50)	40.37 (0.49)	
(2)		57.81 (0.55)	55.41 (0.50)	56.05 (0.52)	35.79 (0.50)	30.75 (0.49)	34.63 (0.47)	
(3)					30.16 (0.51)	22.47 (0.42)	27.71 (0.47)	
		Scenario E						
		Biomarker	mPEB	mFB	lPEB	uPEB	uFB	ST
(1)					48.37 (0.55)	39.92 (0.49)	36.63 (0.47)	
(2)		62.79 (0.57)	57.69 (0.53)	61.60 (0.55)	40.57 (0.59)	32.87 (0.50)	35.14 (0.50)	
(3)					36.42 (0.79)	28.24 (0.66)	28.57 (0.48)	

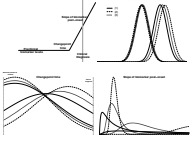
Web Table 2 Summary of simulation results in 200 studies: empirical mean ROC(0.1) (empirical standard error of the mean) within 1 year prior to clinical diagnosis. mEB: joint multivariate parametric empirical Bayes, mFB: joint multivariate fully Bayesian, lPEB: regression plus univariate PEB algorithm applied to the linear combination, uFB: univariate fully Bayesian, uPEB: univariate parametric empirical Bayes and ST: single threshold. The mean biomarker trajectories assumed for each scenario are shown in column 1.

	Scenario D						
	Biomarker	mPEB	mFB	/PEB	uPEB	uFB	ST
(1)					18.29 (0.45)	15.03 (0.45)	16.98 (0.45)
(2)	20.00 (0.48)	17.67 (0.43)	19.30 (0.53)	13.68 (0.41)	12.00 (0.36)	13.30 (0.43)	
(3)					13.97 (0.40)	10.82 (0.36)	12.95 (0.40)

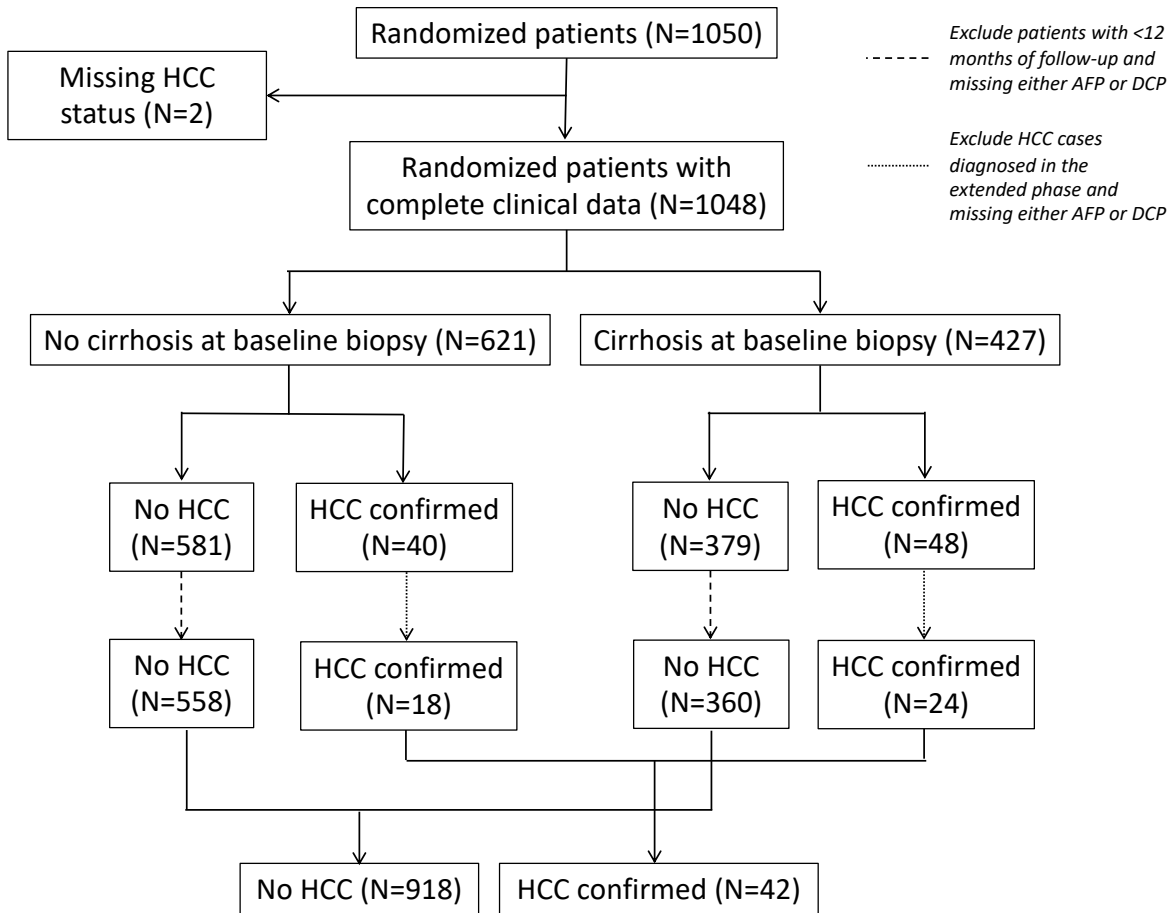
	Scenario D						
	Biomarker	mPEB	mFB	/PEB	uPEB	uFB	ST
(1)					27.55 (0.51)	23.34 (0.50)	20.09 (0.40)
(2)	31.31 (0.52)	27.25 (0.52)	30.79 (0.54)	23.91 (0.51)	19.22 (0.44)	20.10 (0.47)	
(3)					20.87 (0.59)	17.62 (0.51)	17.05 (0.39)

Web Table 3 Summary of simulation results in 200 studies: empirical mean ROC(0.1) (empirical standard error of the mean) within 1-2 years prior to clinical diagnosis. mEB: joint multivariate parametric empirical Bayes, mFB: joint multivariate fully Bayesian, /PEB: regression plus univariate PEB algorithm applied to the linear combination, uFB: univariate fully Bayesian, uPEB: univariate parametric empirical Bayes and ST: single threshold. The mean biomarker trajectories assumed for each scenario are shown in column 1.

	ScenarioD						
	Biomarker	mPEB	mFB	/PEB	uPEB	uFB	ST
(1)					19.13 (0.53)	28.68 (0.61)	14.35 (0.46)
(2)	18.64 (0.50)	29.62 (0.60)	19.18 (0.49)	18.87 (0.49)	26.48 (0.62)	17.21 (0.49)	
(3)					18.69 (0.53)	27.19 (0.55)	16.80 (0.50)

	Scenario E						
	Biomarker	mPEB	mFB	/PEB	uPEB	uFB	ST
(1)					30.27 (0.65)	44.69 (0.65)	16.41 (0.45)
(2)	28.41 (0.63)	44.18 (0.64)	30.66 (0.62)	30.27 (0.63)	42.24 (0.70)	21.83 (0.56)	
(3)					27.93 (0.75)	38.50 (0.90)	19.61 (0.51)

Web Table 4 Summary of simulation results in 200 studies: empirical mean ROC(0.1) (empirical standard error of the mean) greater than 2 years prior to clinical diagnosis. mEB: joint multivariate parametric empirical Bayes, mFB: joint multivariate fully Bayesian, /PEB: regression plus univariate PEB algorithm applied to the linear combination, uFB: univariate fully Bayesian, uPEB: univariate parametric empirical Bayes and ST: single threshold. The mean biomarker trajectories assumed for each scenario are shown in column 1.



Web Figure 2 HALT-C Trial: Standards for Reporting of Diagnostic accuracy (STARD) flow diagram.