Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

Author Manuscript

ARTICLE TYPE

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

Nabihah Tayob*[1] | Anna S. F. Lok[2] | Ziding Feng[3]

[1]Department of Data Science, Dana-Farber Cancer Institute, Massachusetts, United States of America
[2]Department of Internal Medicine, University of Michigan, Michigan, United States of America
[3]Biostatistics 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.

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/sim.9358

This article is protected by copyright. All rights reserved.

Author Manuscript

1 | INTRODUCTION

The early detection of hepatocellular carcinoma (HCC) is currently the best available strategy towards potentially improving the mortality rates associated with HCC in the United States (US) since possibly curative treatments are only recommended at early or very early stages[1]. Patients who receive these treatments have greatly improved five-year survival – transplantation (84%), radiofrequency ablation (53%) and resection (47%). By comparison, those diagnosed with advanced disease have limited treatment options and five-year survival <10%. In the US, the overall age-adjusted incidence rates for liver cancer have risen, on average, 2.7% each year between 2005 and 2014. Approximately 60% of HCC cases are diagnosed with advanced stage disease indicating significant room for improvement in the current HCC surveillance strategy[2].

The 2017 updated guidelines of the American Association for the Study of Liver Diseases (AASLD) recommends ultrasonography surveillance with or without serum α-fetoprotein (AFP) every six months in cirrhosis patients at high risk for HCC[3]. However, in clinical practice, the sensitivity of ultrasound for detecting early stage HCC is only 32%[4]. The poor performance of ultrasound is multifactorial including operator dependency, poor sensitivity for early lesions and difficulties performing ultrasound in obese patients. More sensitive imaging modalities, such as computed tomography (CT) or magnetic resonance imaging (MRI), are not recommended for surveillance for several reasons including possible harm, high cost and unknown sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV)[5]. Blood-based surveillance tests are a promising screening modality for HCC since they are non-invasive, more standardized and more easily applied in limited resource settings. Advances in treatments for hepatitis C virus (HCV), resulting in cure rates >90%, and the increasing incidence of nonalcoholic fatty liver disease (NAFLD) and nonalcoholic steatohepatitis (NASH)[6] may shift the predominant etiology from HCV to NAFLD/NASH in the future. Ultrasound performs poorly in these patients and may be impractical given the sheer number of NAFLD/NASH patients in the United States, further motivating the utility of blood-based surveillance tests.

Serum AFP is widely used in the US to complement ultrasonography, despite it being optional in current guidelines. While there is evidence that screening with ultrasound and AFP leads to increased earlier detection versus no surveillance, there have been no randomized controlled studies evaluating the additive benefit over ultrasound alone[3]. A meta-analysis found the pooled sensitivity of ultrasound with and without AFP was 63% and 45% respectively[7]. In most cancer settings, including HCC, a single biomarker will not cover the heterogeneous subtypes in the target surveillance population. Des-γ carboxy prothrombin (DCP) and lens culinaris agglutinin-reactive alpha-fetoprotein (AFP-L3) are serum biomarkers that have been evaluated in Phase-2 biomarker studies[8] and are FDA-approved for HCC risk.

The Hepatitis C Antiviral Long-term Treatment against Cirrhosis (HALT-C) trial is a rich source of data to better understand HCC screening. The trial enrolled patients with cirrhosis or advanced fibrosis and active hepatitis to evaluate if long-term low-dose pegylated interferon would prevent fibrosis progression and other clinical outcomes, including HCC. The study found

Author Manuscript

no reduction in the incidence of HCC compared with placebo[9]. Patients underwent extensive follow-up with visits scheduled every three months post-randomization for the first 42 months and every six months thereafter. At each visit patients had local laboratory tests including AFP. Patients had scheduled ultrasounds at 6, 18, 30 and 42 months post-randomization and every six months thereafter. DCP was measured at a central laboratory using stored samples collected during the first 42 months within an ancillary study. An earlier generation of the AFP-L3 assay was used in the HALT-C Trial and hence it was not included in our algorithms at this stage of the development but will be evaluated in future studies.

Until recently, most studies evaluating AFP and other biomarkers have focused on comparing current levels to a fixed threshold however more recent studies have found that trends in AFP have prognostic value[10]. The univariate parametric empirical Bayes (PEB) algorithm was initially proposed by McIntosh and Urban (2003)[11] as a computationally straightforward approach to incorporate longitudinal biomarker observations into screening with fewer assumptions than the alternative longitudinal screening algorithms available to date. The PEB algorithm uses a personalized threshold that combines the sample average of prior biomarker observations in the patient with a model for the expected behavior of biomarker in the disease-free population to evaluate whether the current biomarker level represents a significant elevation. Tayob et al (2015)[12] evaluated the univariate PEB algorithm applied to AFP in the HALT-C trial and found that at 10% screening-level false positive rate (FPR), the PEB algorithm improved patient-level sensitivity from 60.4% to 77.1% (p-value < 0.0005) compared to a fixed threshold in those with cirrhosis at baseline and from 72.5% to 87.5% (p-value=0.0015) in those with advanced fibrosis at baseline.

We have previously proposed a fully Bayesian screening algorithm that incorporates the longitudinal trajectory of multiple biomarkers into the calculations of the posterior risk of having cancer[13]. In the HALT-C Trial, we demonstrated that a fully Bayesian algorithm with AFP and DCP further improved patient-level sensitivity at 10% screening-level FPR from 77.1% to 89.5% in those with cirrhosis at baseline. Other approaches that have been proposed for early detection of cancers include shared random effects models and pattern mixture models (PMM)[14,15]. The PMM and our fully Bayesian algorithm both estimate the posterior risk of having cancer but differ in the estimation approach and the model assumed for the biomarker trajectories. The PPM model assumes a linear mixed model for biomarker trajectories that is anchored from the time patient enters screening in both cancer free patients and those that develop cancer during screening. Our fully Bayesian approach anchors time from clinical diagnosis in those that develop cancer during screening. The advantage of this approach is that we do not require patients to enter screening at similar risk levels or to have risk factors that are sufficiently predictive so that conditional on those factors, we can specify a model for the biomarker trajectories. The fully Bayesian algorithm then uses a changepoint model, where we assume that the changepoint in the biomarkers that signals onset of cancer occurs in a period prior to clinical diagnosis and is a key parameter in the model. Prior to the changepoint, we include patient specific means biomarker levels that capture the heterogeneity in the screening population. A disadvantage of this approach is that prospective implementation of the fully Bayesian algorithm requires an estimate for the clinical diagnosis time. A univariate PEB algorithm, which only looks for

Author Manuscript

deviations from expected behavior, does not depend on clinical diagnosis time in any way during prospective application of the algorithm. A multivariate PEB algorithm that is able to retain this feature would be a useful tool in our early detection toolbox.

Here we generalize the univariate PEB algorithm to enable screening with multiple correlated longitudinal biomarkers. This is non-trivial. The univariate PEB algorithm requires only a model for the biomarker in disease-free patients and a target population-level FPR to identify a unique personalized threshold. The generalization to multiple biomarkers will result in many thresholds that achieve the target population-level FPR and we require a rule for selecting optimal thresholds. We propose a fundamentally different PEB algorithm, where a minimal model for the biomarker levels in diseased patients is used to select the personalized thresholds to maximize sensitivity. In Section 2, we describe the details of our proposed multivariate PEB algorithm, including the model assumptions, implementation and evaluation measures. The operational characteristics of the screening algorithm are studied in simulations (Section 3). In Section 4 we present the results from applying the screening approach to the data from the HALT-C Trial. A discussion follows in Section 5.

2 | METHODS

An important feature of the univariate PEB approach is that it only requires a model for the biomarker behavior in disease-free patients. The algorithm then uses deviations from expected disease-free biomarker-behavior to identify patients that should receive additional screening or diagnostic work-ups (i.e a rule-in screening program where controlling the observed FPR is critical). The univariate PEB screening threshold at each screening occasion for each patient is selected to achieve a target population-level FPR f_0, based on the model assumed for the biomarker in disease-free patients and conditional on the patient's screening history to date. There is a unique threshold that satisfies these conditions when screening with a single biomarker.

When we generalize the framework to screening with multiple biomarkers, we no longer have a unique solution. There are multiple combinations of thresholds for the biomarkers that achieve a target f_0, conditional on the patient's screening history. We have chosen to make minimal assumptions about the biomarker levels in patients that develop the disease and select the combination of thresholds that also maximizes the probability of a positive screen if the patient is diseased. The multivariate PEB algorithm is described for the most general scenario since screening with multiple biomarkers is an area of active research in many cancer screening settings.

We assume that there are K biomarkers under consideration. The k^{th} marker level for i^{th} patient at j^{th} screening visit is denoted by Y_{ijk}. The subscript i indexes the N patients in the study and j indexes the J_i screening times for the i^{th} patient. The true disease status of the i^{th} patient at j^{th} screening visit is a time-dependent indicator variable, where $D_{ij} = 0$ if the patient is disease-free and $D_{ij} = 1$ otherwise.

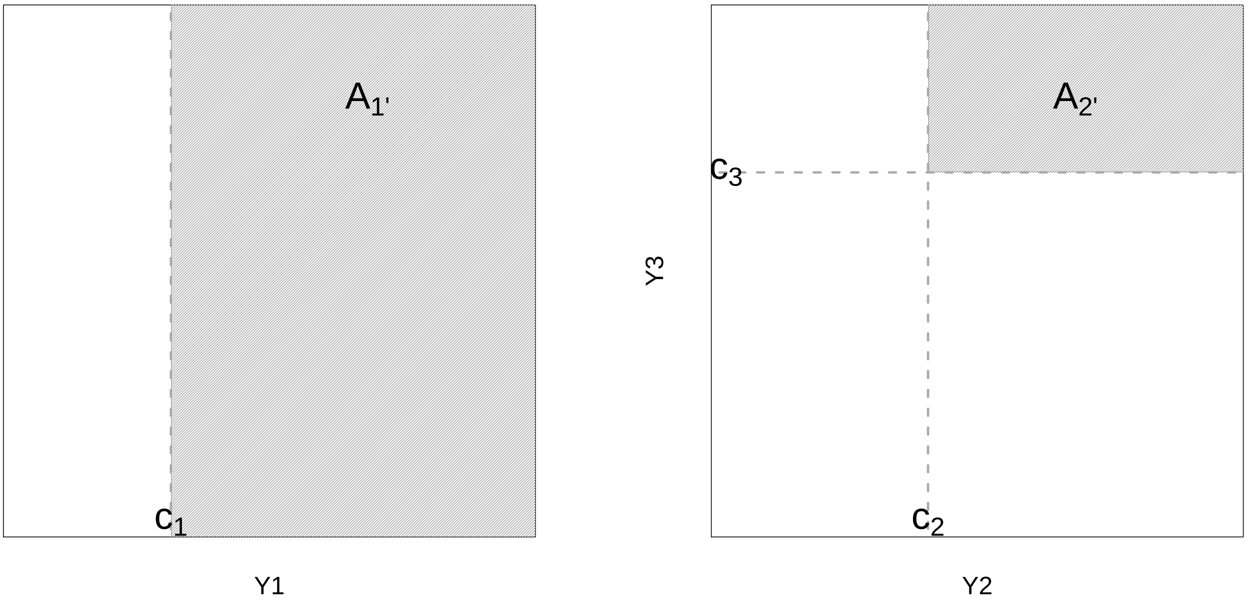
Author Manuscript

We define general positive and negative regions based on multiple biomarkers, where Boolean operators OR and AND are used to stratify the multidimensional biomarker support into regions corresponding to a positive or negative screening results. For example, with three biomarkers a possible rule could be $Y_1 > c_1$ OR ($Y_2 > c_2$ AND $Y_3 > c_3$). We assume that we have prior knowledge of the structure of this rule based on known properties of the biomarkers from case-control studies. The example rule describes a scenario where Y_1 identifies one subgroup of diseased patients while Y_2 and Y_3 are both necessary to identify a second subgroup. Alternatively, logic regression could be used to search for the best rules for multiple biomarkers using prior phase 2 studies[16] when considering the inclusion of more than three biomarkers. The results from these logic regression models, in addition to prior clinical and biological knowledge, would then be used to define the positivity region of the multivariate PEB algorithm. Note that the goal at this stage is not to select the thresholds (e.g. c_1, c_2 and c_3) but to define the shape of the positivity region using the Boolean expressions identified. Based on these guidelines we can construct flexible positivity regions that are able to cover a wide range of possible joint behaviors for biomarkers. As a bonus, the OR and AND rules are intuitive and familiar to clinical practitioners.

We use set notation to define the region of positivity in the most general setting. An OR rule defines the union of two sets while the AND rule defines the intersection of two sets. Note that in this setting, each set is a subset of $\mathbf{R}^K$. For the above example when K=3, the rule $Y_2 > c_2$ AND $Y_3 > c_3$ defines the intersection of the following two sets: $\{(y_1, y_2, y_3) : y_1 \in \mathbf{R}, y_2 > c_2, y_3 \in \mathbf{R}\}$ and $\{(y_1, y_2, y_3) : y_1 \in \mathbf{R}, y_2 \in \mathbf{R}, y_3 > c_3\}$.

Let $\mathbf{A} \subset \mathbf{R}^K$ be the positivity region based on our K biomarkers. Without loss of generality, we can define $\mathbf{A} = \bigcup_{k'=1}^{\tilde{K}} A_{k'}$, where $\tilde{K} \leq K$. Note that each set $A_{k'}$ can itself be the intersection of sets and hence the union could be over fewer than K sets. For each biomarker, indexed by k, we can define the set $B_k = \{(y_1, \ldots, y_K) : y_k > c_k, y_j \in \mathbf{R} \;\; \forall j \neq k\}$ to reflect each biomarkers contribution to the region of positivity. Note that we assume biomarker levels have a positive association with disease but we can easily define B_k for biomarkers that decrease post-onset. Then each set $A_{k'}$ is either defined by a single biomarker and hence $A_{k'} = B_k$ for some $k \in \{1, \ldots, K\}$, or $A_{k'}$ is defined by an AND rule combining multiple biomarkers, e.g. $A_{k'} = B_{k_1} \cap B_{k_2}$ for some $k_1, k_2 \in \{1, \ldots, K\}$. Figure 1 shows the two-dimension representation of the example rule with three biomarkers $Y_1 > c_1$ OR ($Y_2 > c_2$ AND $Y_3 > c_3$), and connects the graphical representation of the screening rule with the set notation that we have defined.

Let $\mathbf{Y}_{ij} = [Y_{ij1}, \ldots, Y_{ijK}]$ be the vector of K biomarker levels in the i^{th} patient at j^{th} screening visit. We require that Y_{ijk}, $k \in \{1, \ldots, K\}$, are non-missing and continuous random variables. Here "non-missing" means that all K biomarkers should be measured on the blood collected at the j^{th} screening visit. Note that the j^{th} screening visit for the i^{th} patient does not need to occur at the same time as the j^{th} screening visit for the i'^{th} patient. A patient is then defined to have a positive screen if $\mathbf{Y}_{ij} \in \mathbf{A}_{ij}$. We define $\mathbf{c}_{ij} = [c_{ij1}, \ldots, c_{ijK}]$ to be the set of thresholds that characterize $\mathbf{A}_{ij}$. The shape of $\mathbf{A}_{ij}$, as defined by the OR and AND rules, is fixed but the thresholds $\mathbf{c}_{ij}$ are unique to each patient and screening occasion in the mPEB algorithm implementation.



(a) Subgroup of diseased patients identified by rule $Y_1 > c_1$

(b) Subgroup of diseased patients identified by rule $Y_2 > c_2$ AND $Y_3 > c_3$

FIGURE 1 In the example, we define a positive screening rule $Y_1 > c_1$ OR ($Y_2 > c_2$ AND $Y_3 > c_3$). Then in set notation we define (a) $A_{1'} = B_1$, where $B_1 = \{(y_1, y_2, y_3) :, y_1 > c_1, y_2 \in \mathbf{R}, y_3 \in \mathbf{R}\}$ and (b) $A_{2'} = B_2 \cap B_3$, where $B_2 = \{(y_1, y_2, y_3) : y_1 \in \mathbf{R}, y_2 > c_2, y_3 \in \mathbf{R}\}$ and $B_3 = \{(y_1, y_2, y_3) : y_1 \in \mathbf{R}, y_2 \in \mathbf{R}, y_3 > c_3\}$. Then the positivity region $\mathbf{A} = A_{1'} \cup A_{2'}$

The biomarker levels in disease-free patients are assumed, after an appropriate monotone transformation, to follow a multivariate normal hierarchical model.

$$\mathbf{Y}_{ij} | \theta_i, \Sigma, D_{ij} = 0 \sim MVN(\theta_i, \Sigma)$$

$$\theta_i | \mu, \Sigma_\theta, D_{ij} = 0 \sim MVN(\mu, \Sigma_\theta)$$

In the those who develop the disease, we assume biomarker levels also follow a multivariate normal hierarchical model but with different mean and covariance parameters.

$$Y_{ij} | \theta_i^*, \Sigma^*, D_{ij} = 1 \sim MVN(\theta_i^*, \Sigma^*)$$

$$\theta_i^* | \mu^*, \Sigma_\theta^*, D_{ij} = 1 \sim MVN(\mu^*, \Sigma_\theta^*),$$

after disease onset. Multivariate normality of the K biomarkers is a stronger assumption than univariate normality, since even for continuous markers, a monotonic transformation to multivariate normality is not assured. However, it is a reasonable working assumption that allows us to jointly model the K biomarkers but ensure the computations of the algorithm are feasible. We

Author Manuscript

Author Manuscript

evaluate the robustness of our approach to this assumption in simulations. The mPEB algorithm cannot be used for discrete biomarkers since transformations to multivariate normality or other continuous distributions do not exist.

Baseline clinical risk factors could be incorporated into the algorithm by extending both models for biomarker levels to incorporate fixed covariates. In the above models, the mean parameters μ and μ^* could be replaced by linear predictors $\mu + \beta X_i$ and $\mu^* + \beta^* X_i$, respectively, to improve precision. Time-varying clinical covariates that can explain additional variability in the biomarker trajectories can be incorporated by redefining the models in terms of residuals after adjusting biomarker levels for time-varying covariates. This extension is a bit more complex since it requires the time-varying covariates to be measured at the same time as the biomarkers and the relationship between the two can be more complex. But a clinical time-varying covariate that has a substantial effect on the biomarkers would be important to incorporate to increase the likelihood that deviations in biomarkers that are not related to onset of cancer are explained.

There are many possible $\mathbf{c}_{ij}$ that ensure the population-level FPR is at mostf_0, conditional on the known screening history for the patient. We define the optimal $\mathbf{c}_{ij}$ to be the set of thresholds that also maximize the probability of a positive screen in patients that develop the disease. Suppose the i^{th} patient has completed n screenings and we know their observed sample means $\overline{\mathbf{Y}}_{in}$, then the optimal thresholds $\tilde{\mathbf{c}}_{i(n+1)}$ solve the following problem:

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

Note that the region $\mathbf{A}_{i(n+1)}$ characterized by optimal thresholds $\tilde{\mathbf{c}}_{i(n+1)}$ is both patient specific and visit specific since it conditions on the individual screening history of the patient to date but the shape of the region $\mathbf{A}_{i(n+1)}$ is fixed by the prespecified OR and AND rules.

The conditional distributions are straightforward to derive when we assume a multivariate normal hierarchical model for biomarker levels.

$$Y_{i(n+1)} | \overline{\mathbf{Y}}_{in}, \mu, \Sigma_\theta, \Sigma, D_{i(n+1)} = 0 \sim MVN(\hat{\theta}_n, V_n)$$
$$\text{where } \hat{\theta}_n = \left(\Sigma_\theta^{-1} + n\Sigma^{-1}\right)^{-1} \left(\Sigma_\theta^{-1}\mu + n\Sigma^{-1}\overline{\mathbf{Y}}_{in}\right)$$
$$\text{and } V_n = \left(\Sigma_\theta^{-1} + n\Sigma^{-1}\right)^{-1} + \Sigma,$$

and

Author Manuscript

$$Y_{i(n+1)}|\overline{\mathbf{Y}}_{in}, \mu^*, \Sigma_\theta^*, \Sigma^*, D_{i(n+1)} = 1 \sim MVN(\hat{\theta}_n^*, V_n^*)$$

$$\text{where } \hat{\theta}_n^* = \left(\Sigma_\theta^{*-1} + n\Sigma^{*-1}\right)^{-1}\left(\Sigma_\theta^{*-1}\mu^* + n\Sigma^{*-1}\overline{\mathbf{Y}}_{in}\right)$$

$$\text{and } V_n^* = \left(\Sigma_\theta^{*-1} + n\Sigma^{*-1}\right)^{-1} + \Sigma^*,$$

Implementation of mPEB algorithm

Before we can implement an mPEB algorithm, we require estimates for the fixed model parameters $\mu, \Sigma_\theta, \Sigma$ and $\mu^*, \Sigma_\theta^*, \Sigma^*$. We assume that we have an existing longitudinal Phase-3 study [17], where patients were recruited from the target screening population. In those who remain disease-free and we have sufficient follow-up, we can use restricted maximum likelihood estimation to obtain estimates of $\mu, \Sigma_\theta, \Sigma$. In those who develop the disease, we most likely do not know the specific screening visit where their true disease status changes from $D_{ij} = 0$ to $D_{ij} = 1$. In this setting, we require additional assumptions to obtain estimates for the parameters $\mu^*, \Sigma_\theta^*, \Sigma^*$.

Without loss of generality, let d_i be the time from first screening visit to the clinical detection of disease if the i^{th} patient is diseased and end of study time if the i^{th} patient is disease-free. We define t_{ij} to be the time from first screening visit to the j^{th} screening visit for the i^{th} patient, i.e. $t_{i1} = 0$. We assume that among those who develop the disease, for all visits where $t_{ij} > (d_i - \tau)$ (i.e. the screening visit occurs within a fixed pre-clinical interval of length τ) the patient likely has undetected disease. We use observations from this period to estimate parameters $\mu^*, \Sigma_\theta^*, \Sigma^*$ using restricted maximum likelihood estimation. In HCC, which is a fast growing cancer, we set $\tau = 12$ months. This time frame allows us to capture multiple screening visits within a patient since six-monthly screening is recommended for HCC and also aims to capture biomarker levels in an early pre-clinical phase where follow-up testing with MRI or CT is more likely to identify visible lesions. We can optimize the algorithm for the intended purpose by selecting τ to be the pre-clinical window of interest that is disease specific.

We also require a procedure to solve for the optimal patient-specific thresholds, $\tilde{\mathbf{c}}_{ij}$ at each screening visit. These thresholds $\tilde{\mathbf{c}}_{ij}$ are the solution to a constrained nonlinear optimization problem and we can therefore take advantage of existing numerical optimization algorithms that have been robustly implemented in widely available software. In this setting, both the objective function and the constraint are nonlinear and twice differentiable. Sequential quadratic programming (SQP) is a powerful, efficient and accurate algorithm where at each iteration, an approximate subproblem with a quadratic objective function and linear constraints defines the search direction [18]. Since the SQP approach, like most optimization routines, only guarantees iteration towards the local solution, we use multiple starting points defined by the demi-deciles of the biomarker distributions observed. The optimal thresholds $\tilde{\mathbf{c}}_{ij}$ are then the thresholds that maximize the probability of a positive screen in patients that develop the disease across the different starting values. At each screening visit, we use the patient screening history, the fixed model parameters and our pre-specified shape of positivity region $\mathbf{A}_{ij}$ to then solve for the $\tilde{\mathbf{c}}_{ij}$. Then if $\mathbf{Y}_{ij} \in \mathbf{A}_{ij}$, the patient has a positive screen at the j^{th} visit. See Web Appendix A for more details.

Evaluation of mPEB algorithm

The standard measures to evaluate screening are based on a single test: sensitivity (proportion of diseased with a positive test) and specificity (proportion of disease-free with a negative test). We have extended these definitions to the longitudinal screening setting[12,13]. Patient-level sensitivity is defined as the proportion of patients that develop the disease with at least one positive screening test during the screening period. Screening-level specificity was defined as the proportion of negative screening tests among all the screenings conducted in the disease-free group. The specificity (1-false positive rate) was defined at the screening level because each false positive result leads to further testing that can be expensive and may lead to complications and anxiety. The receiver operating characteristic (ROC) curves for the mPEB algorithm are constructed by varying f_0. Note that when the model assumptions hold, f_0 is the population-level FPR but in most settings these are unlikely to hold and instead we can think of f_0 as a parameter of the screening algorithm that needs to be fixed prior to implementation. By adjusting this parameter, we can increase the robustness of the mPEB algorithm by ensuring the observed specificity of the algorithm achieves target levels.

Note that although we believe the use of patient-level sensitivity and screening-level specificity is likely the most useful one for retrospective evaluation of cancer screening from banked specimens, other settings could arise where alternative combinations would be of interest to a user and the R-code provided could easily be modified. For example, a clinical setting where both sensitivity and specificity should be evaluated at the screening-level could arise when the treatment for a true test positive patient will not eliminate the disease nor the future needs for surveillance. In this case, a true test positive patient after having received appropriate treatment will go back to surveillance and hence screening-level sensitivity would be a useful measure for evaluation of the screening algorithm.

We consider the performance of the early detection screening algorithms with respect to time-frames prior to HCC diagnosis (e.g. one year prior) since positive screening tests within these windows are more likely to lead to earlier detection of HCC that is visualized with follow-up testing with MRI or CT. Very early positive screening tests in patients that develop HCC are unlikely to lead to early detection and we exclude these positive screening results when estimating patient-level sensitivity within the pre-clinical windows of interest.

3 | SIMULATION STUDY

We use the same simulation study design proposed by Tayob et al (2018)[13]; assuming a hierarchical changepoint model for the biomarker trajectory. This decision is reasonable since this model reflects a biologically plausible description of biomarker

Author Manuscript

Author Manuscript

trajectory. For each biomarker, we assumed that the levels vary randomly around a constant mean in the absence of disease. After the onset of disease (changepoint 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 changepoint in one biomarker is conditional on the number of changepoints observed in the other biomarkers.The details of this hierarchical model are included in Web Appendix B. It is important to note that we do not simulate the longitudinal biomarkers from the multivariate normal hierarchical models assumed in the mPEB algorithm but instead use a biologically plausible hierarchical joint model. Hence, these simulations allow us to evaluate our proposed mPEB algorithm in settings where multivariate normality is not guaranteed.

Our goal is to compare the mPEB algorithm to existing approaches under different scenarios and compare which has the greater potential to increase early detection of HCC. The existing approaches we consider are the multivariate fully Bayesian (mFB) screening algorithm[13], the univariate fully Bayesian (uFB) screening[19], univariate parametric empirical Bayes (uPEB) screening[11], and a standard threshold (ST) approach that ignores screening history . Additional details of the uFB and uPEB algorithm are provided in the Web Appendices of Tayob et al (2018)[13]. We also included the results from a two-step approach that used generalized estimating equations to model the risk of developing cancer in the next six months and identified an optimal combination of the biomarkers in the training cohort and then applied the uPEB algorithm to the linear combination (*l*PEB). This approach is computationally simpler but does assume that the relative contribution of each biomarker at the different time points is fixed. This could be a restrictive assumption, particularly when applying it to a different population, and hence our mPEB approach utilized Boolean operators instead of a linear rule to combine the biomarkers in the panel. Web Appendix C includes additional details for each of the existing approaches.

For each approach, we compare the ROC(0.1): patient-level sensitivity corresponding to 90% screening-level specificity (reported specificity for AFP in clinical practice[8]) on the ROC curve. The patient-level sensitivity estimate included only positive screening tests within one year prior to clinical diagnosis of HCC, positive screening tests between one and two years prior to clinical diagnosis and greater than two years prior to clinical diagnosis of HCC. These time-frames were selected to reflect the short pre-clinical window of HCC, a fast growing cancer. Patients were simulated to have an average pre-clinical window of one year, with a maximum pre-clinical window of two years.

The simulation study design assumes we have three biomarkers measured on average every six months for five years in two cohorts, with additional variability incorporated since in practice patients do not undergo surveillance exactly every six months. The fixed model parameters are estimated using the 400 patients included in the training data. The screening algorithm was then implemented in the 400 patients included in the validation data. The same data generation model was used for both the training and validation cohorts, unless otherwise stated. The probability of developing HCC was assumed to be 50/400, reflecting approximately the number of patients with cirrhosis at baseline biopsy and the number of patients that developed HCC among

Author Manuscript

those with cirrhosis in the HALT-C Trial. We assume that the diagnosis time or end of follow-up time was uniformly distributed during the five-year study period. The parameter values of the hierarchical model used for data generation are included in Web Table 1.

In scenario A, we selected the parameters of the hierarchical models for markers (1) and (2) to reflect the behavior of AFP and DCP, respectively, in the HALT-C Trial. The parameters of marker (3) were selected to reflect a biomarker whose mean level prior to the onset of HCC is greater than that of biomarker (1) but less than that of biomarker (2), with a shallower slope after onset than either (1) or (2). Next, we assume all three biomarkers had lower rates of increase after onset (scenario B). This scenario was selected to examine our hypothesis that when biomarkers have flatter trajectories, a fully Bayesian approach that directly models the biomarker trajectory would be more powerful for detecting changes in the biomarker levels compared to an mPEB approach that aimed to identify deviations from expected behavior.

In the mFB approach, the parameter d_i, the time to clinical diagnosis, was a component of the model that required estimation. We are not aware of any approaches that will avoid this estimation step. We used a Bayesian imputation approach and incorporated random draws from the empirical distribution of d_i (estimated from the training data). This required an assumption that the distribution of d_i in the training data was reflective of the distribution of d_i in the validation dataset. This would be violated when the training data was from a study with a relatively short time frame, since most cohort studies have restricted specimen collection periods, while the validation data was from a cohort being followed for an extended time frame that is more reflective of a more stable screening population under long term follow-up. In HCC screening, patients with compensated cirrhosis have a median surveillance period of 8-10 years. The empirical distribution of d_i will then have an artificial truncation that does not exist in the validation cohort. The mPEB algorithm, which does not have any dependence on future clinical diagnosis time, could have an intrinsic advantage.

In scenario C, we generated the data to reflect this possible study design by modifying Scenario A. In both the training and validation cohorts, we assume we had 400 patients included. In the training cohorts, the patients were followed for up to 3 years, while the patients in the validation cohort were followed for up to 9 years. If we assume an incidence rate of 2% per year, we expect approximately 24 patients to develop HCC in the training cohort and 72 patients to develop HCC in the validation cohort. The biomarker trajectories were unchanged from Scenario A.

We then evaluated a few additional scenarios to explore the robustness of our proposed mPEB algorithm. In Scenario D, we examined the performance of the algorithms in patients that underwent annual surveillance. In Scenario A, the median number of follow-up visits was 5.5 under the assumption of semi-annual surveillance used in the data generation. When we reduce to annual surveillance, the median number of visits is 3, keeping all other aspects of the data generation constant.

In the last scenario, we further explore the robustness of the mPEB algorithm to distributional assumptions of multivariate normality. The hierarchical changepoint model assumed each patient's pre-clinical mean biomarker levels were normally

Author Manuscript

distributed, the changepoint times followed a truncated normal distribution and the biomarker trajectory post-onset was log-normally distributed. We further perturb the distributional assumptions by modifying the data generating mechanism so that these parameters follow bi-modal distributions instead. The bi-modal distributions are generated using mixtures of each distribution with location shifts to create bi-modality. This reflects a clinical setting where unmeasured covariates create subpopulations with different biomarker distributions but where the mPEB and mFB algorithms currently assume all patients are from the same population. It is our hypothesis that the mPEB algorithm that has fewer distributional assumptions would be more robust in this setting.

Results

In each of 200 simulation studies, the parameters of the mPEB screening algorithm (as well as the parameters of each of the existing approaches) were estimated from the training data. All screening algorithms were then implemented in the validation data. The empirical mean ROC(0.1) and standard error of the mean are presented in Tables 1-3 corresponding to patient-level sensitivity within one year prior to diagnosis, patient-level sensitivity between one and two years prior to diagnosis and patient-level sensitivity greater than two years prior to diagnosis, respectively. A well performing screening algorithm would have higher patient-level sensitivity within one year prior to diagnosis or between one and two years prior to diagnosis but lower patient-level sensitivity greater than two years prior to diagnosis when we do not expect the patient to have HCC lesions that can be imaged via CT/MRI.

We observe that under scenario A the mPEB algorithm has slightly greater patient-level sensitivity than the mFB approach in the one year prior to clinical diagnosis, with an empirical mean ROC(0.1) of 72.23% compared to 68.63% (Table 1). This indicates the utility of the mPEB algorithm since the mFB was optimized for this scenario, i.e. the data was simulated from the same hierarchical model fitted, but the mPEB approach has almost equivalent performance with fewer model assumptions. The *l*PEB approach had patient-level sensitivity (70.09%) that was greater than the mFB approach but less than the mPEB algorithm. In addition, we observe that the univariate biomarker algorithms and the fixed threshold approaches all have substantially reduced performance compared to the multivariate approaches. In Table 2, we observed reduced patient-level sensitivity between one to two years prior to clinical diagnosis for all methods but the mPEB algorithm retained slightly greater patient-level sensitivity than the mFB approach and the *l*PEB approach. If we focus on patient-level sensitivity greater than two years prior to diagnosis (Table 3), we observe that with the mPEB algorithm 29.26% of patients that develop HCC have at least one positive screen before HCC is considered detectable while with the mFB and *l*PEB algorithms, 44.92% and 30.44% of patients have at least one positive screen in this undetectable time-period, respectively.

In scenario B, we are attempting to study a setting where the mPEB algorithm may not be the preferred approach. If we examine the patient-level sensitivity within one year prior to clinical diagnosis in Table 1, we observe that the mPEB has an

Author Manuscript

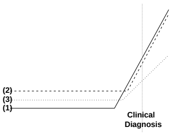
empirical mean ROC(0.1) of 58.76% compared to 52.95% for the mFB algorithm while the *l*PEB algorithm is comparable at 58.91%. In the one to two years prior to clinical diagnosis (Table 2), the mPEB , mFB and *l*PEB algorithms have an empirical mean ROC(0.1) of 26.84%, 21.91% and 26.16%, respectively. Therefore, the performance of algorithms are reduced but the mPEB and *l*PEB would be preferred compared to the mFB algorithm.

In scenario C, we examine a study design that violates the assumptions of the mFB approach. In this setting, the empirical distribution of d_i from the training cohort, with support of $[0, 3]$ is no longer a reasonable estimate of the distribution of d_i in the validation cohort where follow-up is up to 9 years. When we examine the patient-level sensitivity within one year prior to clinical diagnosis in Table 1, the mPEB empirical mean ROC(0.1) was 73.42%, the *l*PEB algorithm was comparable at 71.59% while the mFB approach had much drastically lower patient-level sensitivity of 45.33%, the largest difference in patient-level sensitivity observed across the scenarios. We observe a similar pattern of results when we compare the mPEB and *l*PEB algorithms to the mFB algorithm within one to two years prior to clinical diagnosis (Table 2). Of interest is that when we implement the mFB algorithm in this setting, 80.60% of patients that develop HCC have at least one positive screen before HCC is considered detectable and when we implement the mPEB or *l*PEB algorithms in this setting, 49.60% and 49.70% of HCC patients had at least one positive screen (Table 3). Therefore, while the study design increased the rate of too early positive screens for all algorithms, the impact of it was greater for the mFB algorithm. This pattern is observed for the univariate algorithms as well.

When we reduce the number of surveillance visits to every 12 months in Scenario D, we observe $\sim 13 - 14$ percentage point decrease in patient-level sensitivity within one year prior to clinical diagnosis for all the multivariate longitudinal biomarker algorithms studied (Web Table 2) and $\sim 10 - 13$ percentage point decrease in the patient-level sensitivity between one to two years prior to clinical diagnosis (Web Table 3). Similarly, we observe declines in the sensitivity of the univariate biomarker algorithms and a fixed single threshold approach. Therefore, increasing the surveillance interval to annual when the pre-clinical period is on average one year and the maximum pre-clinical window is two years, impacts the early detection performance of all surveillance approaches.

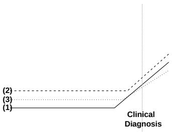
Lastly, in Scenario E we are generating biomarker trajectories from bimodal parameter distributions resulting from unmeasured covariates. Once again, in Web Table 2 we observe reduced patient-level sensitivity within one year prior to clinical diagnosis for all the screening algorithms compared to Scenario A but the mPEB algorithm is still slightly greater than the mFB and *l*PEB algorithms. We observe minimal impact of unmeasured confounders on either the patient-level sensitivity between one and two years prior to diagnosis (Web Table 3) or patient-level sensitivity greater than two years prior to diagnosis (Web Table 4).

Author Manuscript



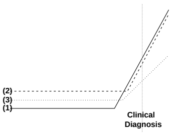
Scenario A

Biomarker	mPEB	mFB	*l*PEB	uPEB	uFB	ST
(1)				53.11 (0.48)	46.67 (0.56)	47.50 (0.55)
(2)	72.23 (0.47)	68.63 (0.53)	70.09 (0.50)	47.91 (0.51)	41.30 (0.54)	45.70 (0.51)
(3)				42.51 (0.51)	34.98 (0.48)	38.49 (0.50)



Scenario B

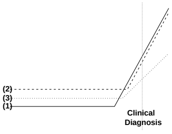
Biomarker	mPEB	mFB	*l*PEB	uPEB	uFB	ST
(1)				50.13 (0.49)	43.15 (0.52)	41.23 (0.51)
(2)	58.76 (0.53)	52.95 (0.55)	58.91(0.50)	35.29 (0.47)	26.49 (0.43)	33.28 (0.48)
(3)				37.06 (0.49)	27.99 (0.45)	32.81 (0.53)



Scenario C

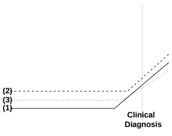
Biomarker	mPEB	mFB	*l*PEB	uPEB	uFB	ST
(1)				55.06 (0.42)	34.94 (0.48)	48.45 (0.43)
(2)	73.42 (0.39)	45.33 (0.50)	71.59 (0.45)	48.85 (0.43)	23.28 (0.36)	45.88 (0.43)
(3)				42.88 (0.40)	18.47 (0.34)	37.47 (0.41)

TABLE 1 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, *l*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.



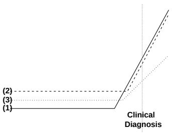
Scenario A

Biomarker	mPEB	mFB	*l*PEB	uPEB	uFB	ST
(1)				29.87 (0.52)	24.87 (0.51)	23.18 (0.45)
(2)	32.74 (0.56)	28.28 (0.51)	30.77 (0.59)	23.25 (0.49)	19.00 (0.42)	21.31 (0.45)
(3)				21.99 (0.49)	18.00 (0.43)	19.16 (0.46)



Scenario B

Biomarker	mPEB	mFB	*l*PEB	uPEB	uFB	ST
(1)				26.51 (0.48)	21.74 (0.46)	19.17 (0.44)
(2)	26.84 (0.46)	21.91 (0.47)	26.16 (0.59)	20.44 (0.46)	16.10 (0.40)	19.03 (0.45)
(3)				20.67 (0.48)	16.69 (0.44)	17.94 (0.43)

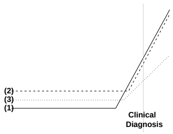


Scenario C

Biomarker	mPEB	mFB	*l*PEB	uPEB	uFB	ST
(1)				29.68 (0.40)	17.44 (0.34)	23.52 (0.38)
(2)	33.36 (0.44)	18.60 (0.35)	31.74 (0.41)	23.23 (0.37)	12.72 (0.31)	21.37 (0.39)
(3)				22.11 (0.38)	12.43 (0.31)	19.30 (0.40)

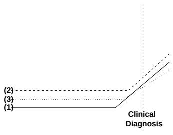
TABLE 2 Summary of simulation results in 200 studies: empirical mean ROC(0.1) (empirical standard error of the mean) within 1-2 years prior to clinical diagnosis. mEB: joint multivariate parametric empirical Bayes, mFB: joint multivariate fully Bayesian, *l*PEB: regression plus univariate PEB algorithm 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.

Author Manuscript



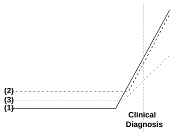
Scenario A

Biomarker	mPEB	mFB	*l*PEB	uPEB	uFB	ST
(1)				29.88 (0.61)	45.31 (0.74)	17.50 (0.51)
(2)	29.26 (0.60)	44.92 (0.75)	30.44 (0.59)	30.16 (0.60)	43.00 (0.71)	25.08 (0.57)
(3)				30.27 (0.62)	41.44 (0.64)	23.01 (0.62)



Scenario B

Biomarker	mPEB	mFB	*l*PEB	uPEB	uFB	ST
(1)				29.82 (0.58)	42.45 (0.73)	17.90 (0.47)
(2)	29.19 (0.58)	43.91 (0.65)	30.85 (0.63)	29.69 (0.62)	41.72 (0.68)	24.37 (0.57)
(3)				30.53 (0.61)	43.68 (0.71)	23.11 (0.59)



Scenario C

Biomarker	mPEB	mFB	*l*PEB	uPEB	uFB	ST
(1)				50.43 (0.49)	82.79 (0.45)	22.56 (0.39)
(2)	49.60 (0.47)	80.60 (0.45)	49.70 (0.49)	50.24 (0.50)	76.32 (0.48)	36.85 (0.47)
(3)				50.37 (0.50)	77.38 (0.55)	32.87 (0.53)

TABLE 3 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, *l*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.

4 | RESULTS FROM THE HALT-C TRIAL

While the HALT-C Trial included extensive follow-up (median follow-up time was 83 months), serum specimens were only collected and stored during the first 42 months post-randomization. Hence, we only have DCP measured concurrently with AFP during the initial phase of the study. Therefore in this analysis, we restrict our attention to the first 48 months post-randomization during which time 24 patients with cirrhosis at baseline biopsy and 18 patients with advanced fibrosis at baseline biopsy developed HCC within six months (the recommended HCC surveillance interval) from last concurrent measurement of AFP and DCP. We excluded 46 patients that developed HCC during the extended phase of the study from the analysis with more than six months between biomarker testing and the development of HCC. See Web Figure 1 for more details on the study cohort. HCC diagnosis was based on histology and in its absence, by imaging with or without AFP. We evaluated HCC screening in all patients, regardless of assigned treatment, since there was no evidence the incidence of HCC differed between the two treatment groups[9].

The proposed screening methodology performance was evaluated in the HALT-C Trial via 10-fold cross-validation. 918 disease-free patients were randomly divided into eight subsets of 92 patients and two subsets of 91 patients. 42 HCC patients were randomly divided into eight subsets of 4 patients and two subsets of 5 patients. At each iteration of the cross-validation,

Author Manuscript

the validation data consists of one subset of HCC patients and one subset of disease-free patients. The remaining nine subsets form the training data.

AFP and DCP are potentially complementary biomarkers for HCC early detection and among patients in the HALT-C trial who develop HCC, we observe elevated levels in either AFP or DCP prior to diagnosis. Hence, the implementation of the mPEB algorithm uses an OR rule to construct the positivity region, **A**. In the training cohort, we estimated the parameters $\mu, \Sigma_\theta, \Sigma$ via restricted maximum likelihood estimation using all study visits among disease free patients and we estimated $\mu^*, \Sigma_\theta^*, \Sigma^*$ using visits within $\tau = 12$ months prior to HCC diagnosis among those that develop HCC. The details for implementing the mFB algorithm have been previously described[13].

In Table 4 we present the cross-validated ROC estimates at 90% screening-level specificity, calculated by averaging the patient-level sensitivity at 90% screening-level specificity across each iteration within one year prior and between one and two years prior to clinical diagnosis. These were the periods during which a positive screen was more likely to lead to confirmation of HCC diagnosis using more sensitive imaging (CT or MRI). In addition, we estimated patient-level sensitivity at 90% screening-level specificity across each iteration greater than two years prior to diagnosis to evaluate the positive screens observed outside the pre-clinical window of HCC.

At 90% screening-level specificity, we observed that the mPEB algorithm and mFB algorithms had similar patient-level sensitivity within one year prior to clinical diagnosis (63.67% vs 63.17%, respectively) and the *l*PEB approach was slightly lower at 61.17%. The multivariate biomarker algorithms had greater patient-level sensitivity within one year compared to either of the univariate algorithms or a fixed threshold approach. Within one to two years prior to clinical diagnosis, we observed that the *l*PEB had greater patient-level sensitivity than either the mPEB algorithm (37.67%) or the mFB algorithm (25.33%) but the univariate PEB algorithm applied to DCP alone had the largest patient-level sensitivity (44.00%). In addition, the fixed threshold approach with DCP alone (37.00%) was comparable to the mPEB algorithm.

When we examined patient-level sensitivity greater than two years prior to diagnosis, screening approaches with the lowest patient-level sensitivity were preferred for early detection of HCC. Here we observed that with the mPEB algorithm 39.00% of patients that developed HCC had at least one positive screen before HCC was considered detectable, while with the mFB algorithm in this setting, 75.67% of HCC patients had at least one positive screen and 47.33% of HCC patients had a positive screen with the *l*PEB approach. More than two-years prior to HCC diagnosis, the univariate fully Bayesian algorithms had the highest patient-level sensitivity while the fixed threshold approaches had the lowest patient-level sensitivity.

We also compared the timing of the first positive screening result for the mPEB algorithm to the other approaches evaluated. For each approach, we compared the proportion of patients that had a positive screen first for the mPEB algorithm and the proportion of patients that had a positive screen first for the other approach. In Figure 2, we observed that in the year prior to clinical diagnosis, the mPEB algorithm is more likely to have a positive screen first compared to all other methods considered.

Author Manuscript

The improvement is largest when compared to the fully Bayesian approaches (17.33% vs 7.00% for the mFB; 32.33% vs 9.00% for the uFB AFP; 48.67% vs 12.00% for the uFB DCP) and more moderate when compared to the univariate PEB algorithms (20.33% vs 17.00% for the *l*PEB; 7.50% vs 2.50% for the uPEB AFP; 38.17% vs 24.50% for the uPEB DCP). In the one to two years prior to clinical diagnosis, the mPEB algorithm is more likely to have a positive screen first compared to the mFB (19.83% vs 7.50%), uFB AFP (19.00% vs 7.50%) and uFB DCP (30.17% vs 12.50%) algorithms; but less likely to have a positive screen first compared to the *l*PEB (12.83% vs 19.50%), uPEB AFP (2.50% vs 7.00%), uPEB DCP (23.17% vs 32.00%) and fixed threshold approaches with either AFP (13.67% vs 27.00%) or DCP (25.17% vs 27.00%), though these differences are mostly moderate. When we compare the timing of positive screens that are greater than two years prior to diagnosis, the mPEB algorithm is less likely to have a positive screen first compared to all the algorithms considered, except the univariate PEB algorithm applied to DCP or the fixed threshold approach with DCP.

While the mPEB algorithm was not shown to be preferred algorithm across all the measures used to evaluate the early detection screening algorithms, we do observe an increase in patient-level sensitivity within one year prior to diagnosis, the lead time that hepatologists believe is most likely to result in a stage shift of the cancer towards a curable stage. This demonstrated the potential clinical utility of the algorithm and has motivated its further study and validation in larger HCC surveillance cohorts. As noted in the simulation study, the mFB and univariate FB algorithms are more likely to have early positive screens in patients that develop HCC— indicating they are potentially more appropriate for risk prediction and less optimized for early detection. The boundary of risk prediction and early detection is not clear but we always use an operational criterion that the early detection of cancer using a blood test means that the cancer lesion would be seen if CT/MRI imaging was done on the detected patients. For this reason, patient-level sensitivity beyond two years prior to clinical diagnosis was considered to be most likely the "risk" for cancer rather than "early detection" of cancer. The univariate PEB from its conceptualization was to detect cancer rather than predicting future cancer risk and this property was carried forward in the development of the multivariate PEB algorithm. Note that it may be of interest to researchers to combine the mFB algorithm for risk prediction and the mPEB algorithm for early detection in the surveillance population but understanding the clinical utility of these comprehensive screening approaches is beyond the scope of this manuscript.

5 | DISCUSSION

When we extend screening to multiple biomarkers, an important question is how to optimally combine the multiple biomarkers. For a single screening occasion, the likelihood ratio has been shown to be the optimal manner to combine biomarkers[20]. In this case, a binary regression model is sufficient to estimate the optimal combination and the combination is optimal in that it provides the highest sensitivity among all possible combination rules. Our previously proposed fully Bayesian hierarchical

Time Period	Biomarker	mPEB	mFB	*l*PEB	uPEB	uFB	ST
1 year prior to clinical diagnosis	log(AFP)	63.67%	63.17%	61.17%	56.17%	46.17%	50.67%
	log(DCP+1)				57.83%	36.83%	51.33%
1-2 years prior to clinical diagnosis	log(AFP)	37.67%	25.33%	39.83%	37.67%	21.17%	36.50%
	log(DCP+1)				44.00%	20.00%	37.00%
> 2 years prior to clinical diagnosis	log(AFP)	39.00%	75.67%	47.33%	39.83%	84.00%	36.50%
	log(DCP+1)				38.33%	90.17%	31.67%

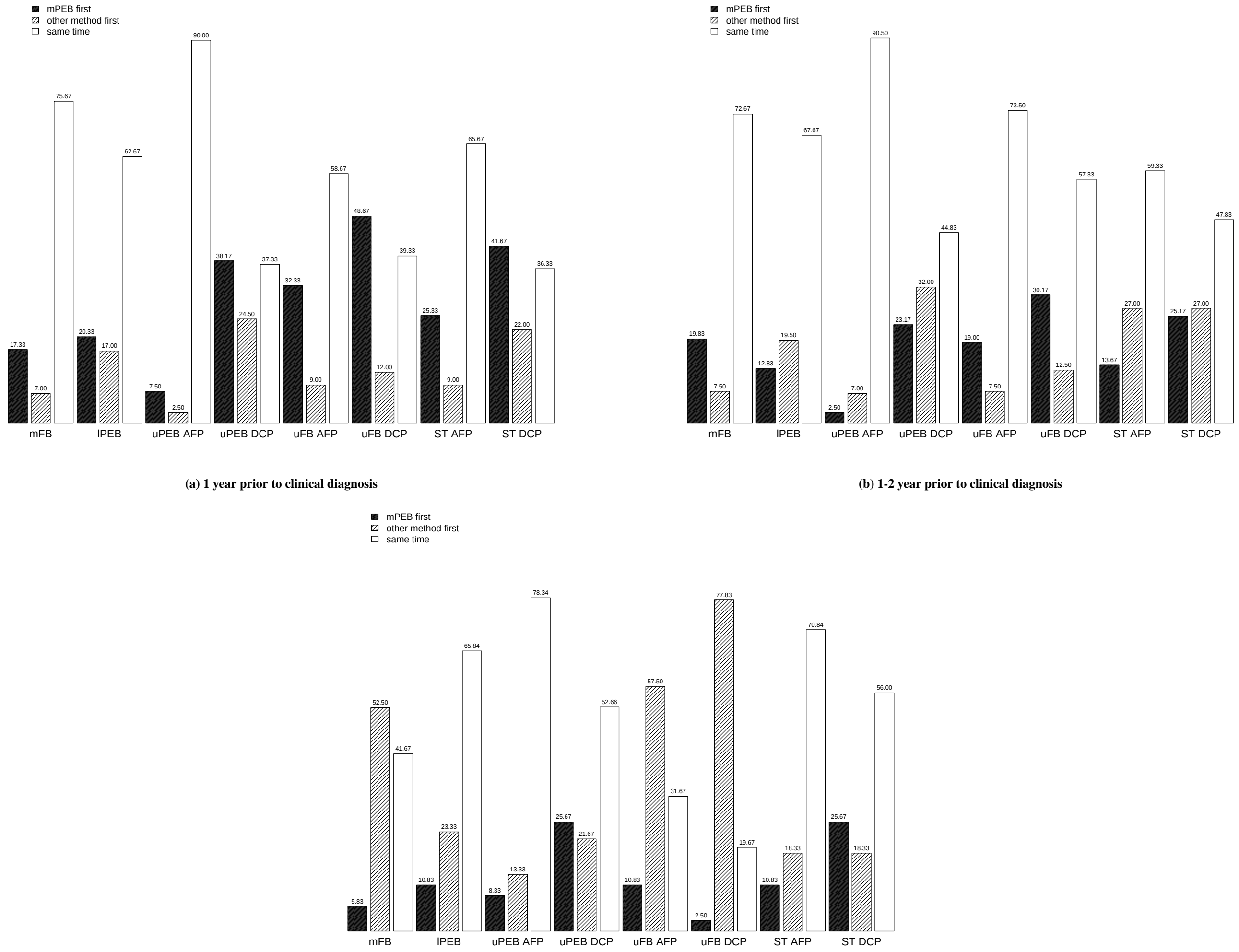
TABLE 4 Cross-valiated ROC(0.1) for mPEB: multivariate parametric empirical Bayes, mFB: multivariate fully Bayesian, *l*PEB: regression plus univariate PEB algorithm applied to the linear combination, uFB: univariate fully Bayesian and uPEB: parametric empirical Bayes and ST: single threshold in the three time periods.

Author Manuscript

changepoint model used a likelihood ratio approach to combine multiple longitudinal biomarkers and performs well in many settings. However, we have shown in our simulations that it is not necessary optimal for our longitudinal definitions of patient-level sensitivity and screening-level specificity, which are more appropriate in the HCC surveillance setting. Our mPEB algorithm defines regions of positivity using logic rules to combine the biomarkers. These can be intuitively understood by clinicians and hence improve the quality of prior information incorporated in developing these models. In addition, despite the simplicity, we can capture a wide range of joint biomarker behaviors and the approach is likely to approach optimality with minimal distributional assumptions. A theoretical derivation of optimality in this setting is an area of future research.

A computationally simpler approach that was also considered is to use regression to estimate a fixed linear combination of the biomarkers and then apply the univariate PEB algorithm to the derived biomarker score. While the approach has mostly comparable performance in our analyses, it does require assuming that the relative contribution of each biomarker in the panel is fixed over time. This assumption may not hold when it is applied to a different population. The proposed mPEB approach of using Boolean operators to combine the biomarkers was based on our goal of maintaining the flexibility and robustness of the univariate PEB algorithm when generalizing the method to multiple biomarkers, however we would encourage users to explore all the algorithms in their toolbox fully before identifying the optimal approach for their setting.

We have developed an mPEB algorithm that defines personalized screening thresholds for multiple screening biomarkers using the longitudinal history for each patient and a minimal model for the biomarker behavior. In simulation studies, we have demonstrated that the mPEB algorithm has superior performance for early detection of HCC with greater patient-level sensitivity within one year prior to diagnosis compared to both the multivariate fully Bayesian approach or any of the existing univariate algorithms. These simulation studies are particularly convincing since they are optimized for the multivariate fully Bayesian approach, since we used the changepoint model to generate the data. The simulations demonstrate that a minimal model for the biomarker behavior after onset of HCC is sufficient to develop an mPEB algorithm that is robust and has potential clinical utility in many settings. We advocate for a bigger toolbox that contains many different types of longitudinal algorithms. This would



(a) 1 year prior to clinical diagnosis

(b) 1-2 year prior to clinical diagnosis

(c) >2 year prior to clinical diagnosis

FIGURE 2 Cross-validated estimate of the percentage of times the proposed multivariate parametric empirical Bayes (mPEB) algorithm has a positive screen first, another method has a positive screen first, and the first positive screen for both is the same in the three time periods. mFB: multivariate fully Bayesian, *l*PEB: regression plus univariate PEB algorithm applied to the linear combination, uFB: univariate fully Bayesian and uPEB: parametric empirical Bayes and ST: single threshold.

allow a researcher to study and compare each approach in a retrospective phase 3 study before identifying the optimal algorithm to move forward with in a prospective phase 4 study.

A current limitation of the approach is that it can only be used for complete biomarker panel data and any screening occasions where only subset of the biomarkers is measured on the blood collected at the visit is excluded. For example, if one of the biomarkers in the panel is more expensive to measure, it may not be cost-effective to measure at every screening occasion. Future research would involve developing a more flexible approach that can accommodate incomplete information in the biomarker panel at a screening occasion. Another type of missingness that can occur in cohort studies is loss to follow-up that is missing not at random. In cancer screening cohort studies, it is important to ensure patients not diagnosed during the study period are cancer-free. In some contexts, we could apply a gold standard diagnostic test to all patients at the off-study visit but most often

Author Manuscript

Author Manuscript

cohort studies are designed to ensure a patient has sufficient follow-up to confirm they are cancer-free. If patients are lost to follow-up, we may need to censor the patients by a reasonable time frame to ensure all screening visits included were conducted when the patient most likely was cancer-free. For example, in our HALT-C analysis, we excluded visits within one year prior to last follow-up visit in those not diagnosed with HCC during the study. Informative missingness would then likely result in an underpowered study for evaluation of the biomarker screening algorithm because there are patients that dropped out of the study because they developed cancer and we are not able to capture the behavior the biomarker panel during this pre-clinical period. It is not likely to bias the algorithm unless the biomarker behavior after cancer onset is different in those that dropped out of the study compared to those that remain in the cohort and are diagnosed.

The HALT-C Trial analysis demonstrated a more nuanced message on the clinical utility of each of the algorithms. In the one year prior to clinical diagnosis, the mPEB algorithm had greater patient-level sensitivity compared to all other approaches. However, in the one to two years prior to clinical diagnosis, the univariate PEB algorithm applied to DCP alone had higher patient-level sensitivity than the mPEB algorithm. Therefore an argument could be made for the superiority of either approach. A bootstrap inferential procedure could be used to determine if the differences in the algorithms are statistically significant but comes with high computational cost (especially when combined with 10-fold cross-validation to adjust for overfitting). Research into developing a more feasible inferential procedure for comparing longitudinal algorithms based on our proposed measures is underway. There is also some difficulty in interpreting this result once we remember the context of the HALT-C Trial where patients were under HCC surveillance using AFP and ultrasound. A doubling of AFP from baseline would trigger additional follow-up leading to AFP-detected HCC patients in our cohort. This then presents difficulties when modeling the natural history of the joint behavior of AFP and DCP since this is terminated early in AFP-detected HCC patients. This is, in general, a problem with combining multiple screening biomarkers in a population that is currently under surveillance with a subset of the biomarkers and statistical methods to correct this are greatly needed to further improve our algorithms and are an avenue of future research.

A key question for researchers is which biomarkers to include in the multivariate biomarker screening algorithms. These algorithms have been developed in the context where we have validated cancer biomarkers that have shown ability to distinguish those with cancer at time of clinical diagnosis from cancer-free controls (phase 2 biomarker study) and have demonstrated utility prior to clinical diagnosis from longitudinal cohort studies (phase 3 biomarker studies)[17]. It is critical that the biomarkers have been measured in the same specimens from these retrospective studies so that their joint behavior can be explored. It is likely that highly correlated biomarkers would result in non-identifiable thresholds. The SQP algorithm requires both the objective function and the constraint to be twice differentiable and highly correlated biomarkers would likely violate this assumption. In addition, including additional biomarkers that are highly correlated with existing biomarkers in the panel is unlikely to increase the sensitivity of the algorithm for early detection and therefore would not be of interest for inclusion. Careful study of the

Author Manuscript

biomarkers being included in the algorithm is required and these algorithms should incorporate both prior information and sensitivity analyses to understand the optimal approach to move forward into prospective phase 4 studies.

Longitudinal biomarker algorithms gain power for improving early detection by incorporating the prior screening history for patients. A natural question is how the amount of screening history affects the performance of the algorithms. All the algorithms considered in this study can be implemented in both patients with no screening history as well as those with multiple prior visits. The recommended frequency of surveillance visits is cancer-specific and depends on the expected pre-clinical window and the cancer doubling time. Guidelines recommend semi-annual HCC surveillance since HCC is a fast-growing cancer with a short pre-clinical window that is not expected to be longer than two years. Therefore, surveillance visits are required to be more closely spaced to increase the opportunity to detect HCC at an early stage. For slower growing cancers, annual or even multiple years between surveillance visits may be more appropriate. From a technical perspective, the within-patient variability of the biomarker levels in the absence of cancer will also affect the ability of the algorithms to detect onset of cancer. A useful cancer biomarker either has lower within-patient variability compared to the between-patient variability, or we have clinical covariates that can explain the additional variability in biomarkers that can impact the performance of the algorithms.

Our multivariate screening algorithms have demonstrated potential in both simulation studies and detailed analysis of the HALT-C Trial. However, to truly understand their clinical utility we will need to validate them in two ongoing prospective cohorts. The first is the Early Detection Research Network's Hepatocellular Carcinoma cohort, where analysis can begin shortly since the cohort is sufficiently mature. The cohort includes 1,560 patients with cirrhosis of varying etiologies including HCV, HBV, alcoholic liver disease and NAFLD and 87 patient that developed HCC to date. This population is more reflective of the changing demographics of the HCC target surveillance population and it will be important to study the utility of the multivariate screening algorithms in this cohort. In addition, we will be able to refine our algorithms to enable screening with the FDA approved triplicate AFP, AFP-L3 and DCP and compare our algorithm to other approaches such as the GALAD score that combines the triplicate of biomarkers at the current screening occasion with age and gender[21]. It is important to demonstrate whether the additional complexity from using longitudinal biomarkers translates into clinically significant improvements in screening. These methods will be further validated in The Texas Hepatocellular Carcinoma Consortium (THCCC) cohort which is assembling the largest prospective cohort study of cirrhosis patients to study early detection of HCC in the United States to date and will be mature in the next few years.

The longitudinal screening algorithms for multiple biomarkers that we have developed are an important area of research in many settings though they remain a methodologically challenging problem. We have developed a general, robust methodology to address an salient clinical question in cancer surveillance and have provided code to implement these algorithms to promote the more widespread study and usage of these approaches.

Author Manuscript

ACKNOWLEDGMENTS

All authors are partially supported by NCI grant R01CA230503. Nabihah Tayob and Ziding Feng are partially supported by NCI grant U24CA230144. Ziding Feng is partially supported by NCI grant U24CA086368 and Anna S.F. Lok is partially supported by U01CA230669.

6 | SOFTWARE

R-code and Matlab code to implement the screening algorithms discussed here are available at https://github.com/ntayob.

7 | DATA SHARING

The data from the HALT-C Trial are available on request from https://repository.niddk.nih.gov/studies/halt-c/.

References

1. Bruix J, Sherman M. Management of hepatocellular carcinoma. *Hepatology* 2005; 42(5): 1208–1236.

2. Altekruse SF, McGlynn KA, Dickie LA, Kleiner DE. Hepatocellular Carcinoma Confirmation, Treatment, and Survival in Surveillance, Epidemiology, and End Results Registries, 1992-2008. *Hepatology* 2012; 55(2): 476-482. doi: 10.1002/hep.24710

3. Heimbach JK, Kulik LM, Finn RS, et al. AASLD guidelines for the treatment of hepatocellular carcinoma. *Hepatology* 2018; 67(1): 358-380. doi: 10.1002/hep.29086

4. Singal AG, Conjeevaram HS, Volk ML, et al. Effectiveness of Hepatocellular Carcinoma Surveillance in Patients with Cirrhosis. *Cancer Epidemiology and Prevention Biomarkers* 2012; 21(5): 793–799. doi: 10.1158/1055-9965.EPI-11-1005

5. El-Serag HB. Hepatocellular Carcinoma. *N Engl J Med* 2011; 365(12): 1118-1127.

6. Goldberg D, Ditah IC, Saeian K, et al. Changes in the Prevalence of Hepatitis C Virus Infection, Nonalcoholic Steatohepatitis, and Alcoholic Liver Disease Among Patients With Cirrhosis or Liver Failure on the Waitlist for Liver Transplantation. *Gastroenterology* 2017; 152(5): 1090–1099.e1.

7. Tzartzeva K, Obi J, Rich NE, et al. Surveillance Imaging and Alpha Fetoprotein for Early Detection of Hepatocellular Carcinoma in Patients With Cirrhosis: A Meta-analysis. *Gastroenterology* 2018; 154: 1706-1718.e1.

Author Manuscript

8. Marrero JA, Feng Z, Wang Y, et al. α-Fetoprotein, Des-γ Carboxyprothrombin, and Lectin-Bound α-Fetoprotein in Early Hepatocellular Carcinoma. *Gastroenterology* 2009; 137(1): 110-118.

9. Lok AS, Everhart JE, Wright EC, et al. Maintenance peginterferon therapy and other factors associated with hepatocellular carcinoma in patients with advanced hepatitis C. *Gastroenterology* 2011; 140(3): 840–849.

10. Lee E, Edward S, Singal AG, Lavieri MS, Volk M. Improving Screening for Hepatocellular Carcinoma by Incorporating Data on Levels of α-Fetoprotein, Over Time. *Clin Gastroenterol Hepatol* 2013; 11(4): 437 - 440.

11. McIntosh MW, Urban N. A parametric empirical Bayes method for cancer screening using longitudinal observations of a biomarker. *Biostatistics* 2003; 4(1): 27-40.

12. Tayob N, Lok AS, Do KA, Feng Z. Improved Detection of Hepatocellular Carcinoma by Using a Longitudinal Alpha-Fetoprotein Screening Algorithm. *Clinical Gastroenterology and Hepatology* 2015. epub ahead of print.

13. 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.

14. Han Y, Albert PS, Berg CD, Wentzensen N, Katki HA, Liu D. Statistical approaches using longitudinal biomarkers for disease early detection: A comparison of methodologies. *Statistics in Medicine* 2020; 39(29): 4405-4420. doi: https://doi.org/10.1002/sim.8731

15. Liu D, Albert PS. Combination of longitudinal biomarkers in predicting binary events. *Biostatistics* 2014; 15(4): 706-718. doi: 10.1093/biostatistics/kxu020

16. Ruczinski I, Kooperberg C, LeBlanc M. Logic Regression. *Journal of Computational and Graphical Statistics* 2003; 12(3): 475-511.

17. Pepe MS, Etzioni R, Feng Z, et al. Phases of Biomarker Development for Early Detection of Cancer. *Journal of the National Cancer Institute* 2001; 93(14): 1054–1061.

18. Nocedal J, Wright S. *Numerical Optimization*. Springer Series in Operations Research and Financial EngineeringSpringer New York . 2000.

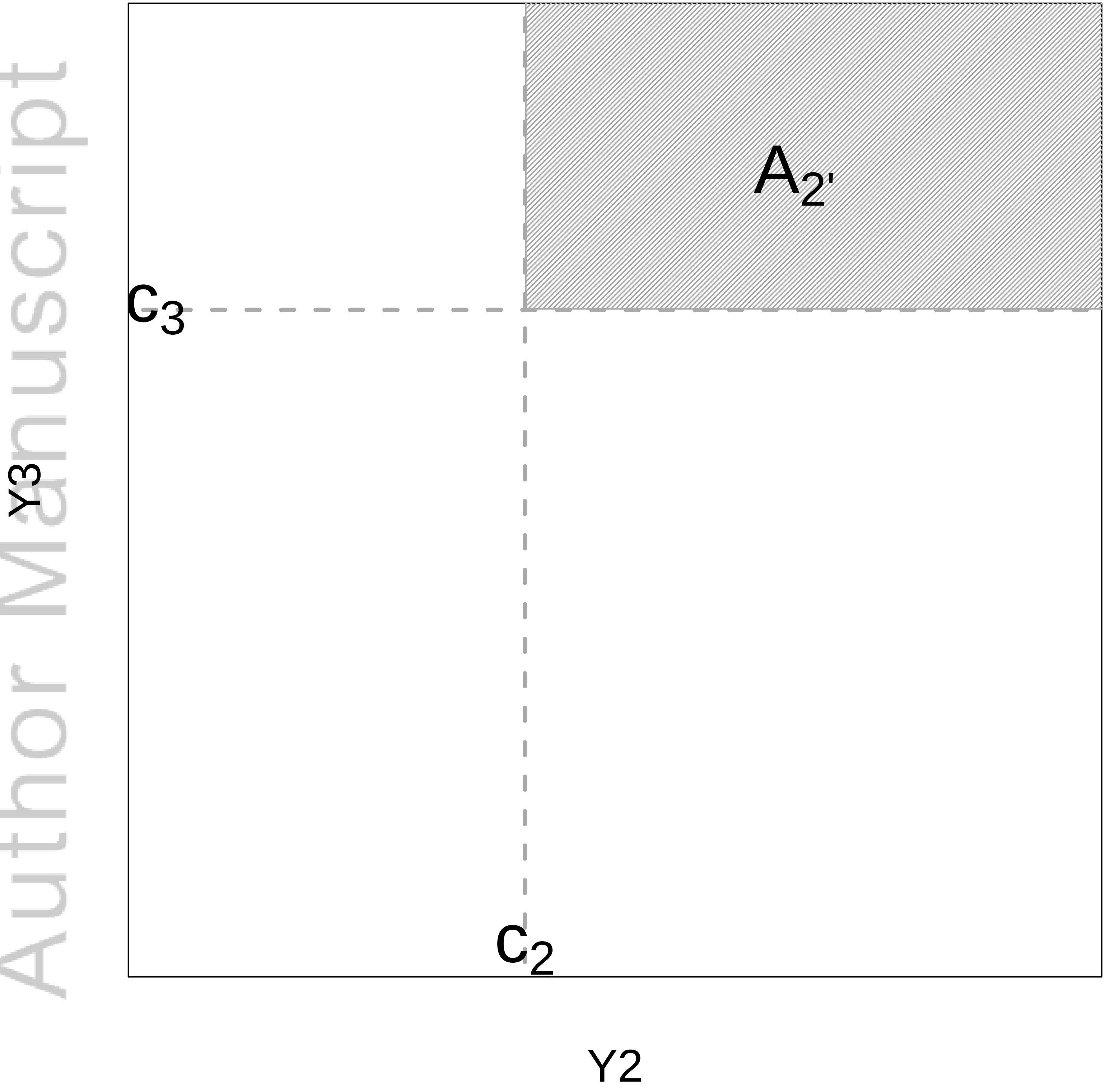
19. Skates SJ, Pauler DK, Jacobs IJ. Screening Based on the Risk of Cancer Calculation from Bayesian Hierarchical Changepoint and Mixture Models of Longitudinal Markers. *JASA* 2001; 96(454): 429-439.

20. McIntosh MW, Pepe MS. Combining several screening tests: Optimality of the risk score. *Biometrics* 2002; 58: 657-664.

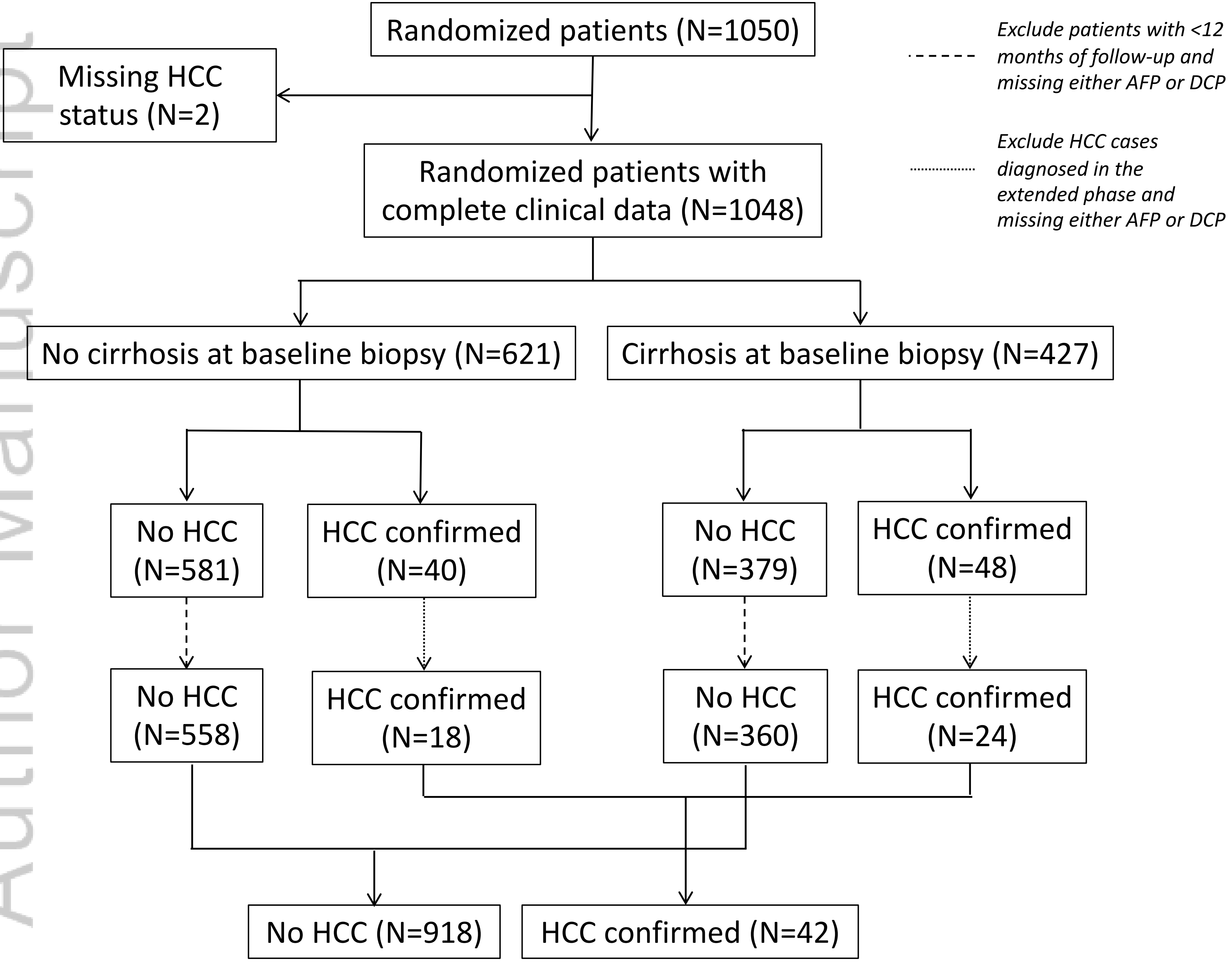
21. Johnson PJ, Pirrie SJ, Cox TF, et al. The Detection of Hepatocellular Carcinoma Using a Prospectively Developed and Validated Model Based on Serological Biomarkers. *Cancer Epidemiology and Prevention Biomarkers* 2014; 23(1): 144–153. doi: 10.1158/1055-9965.EPI-13-0870

□ Author Manuscript

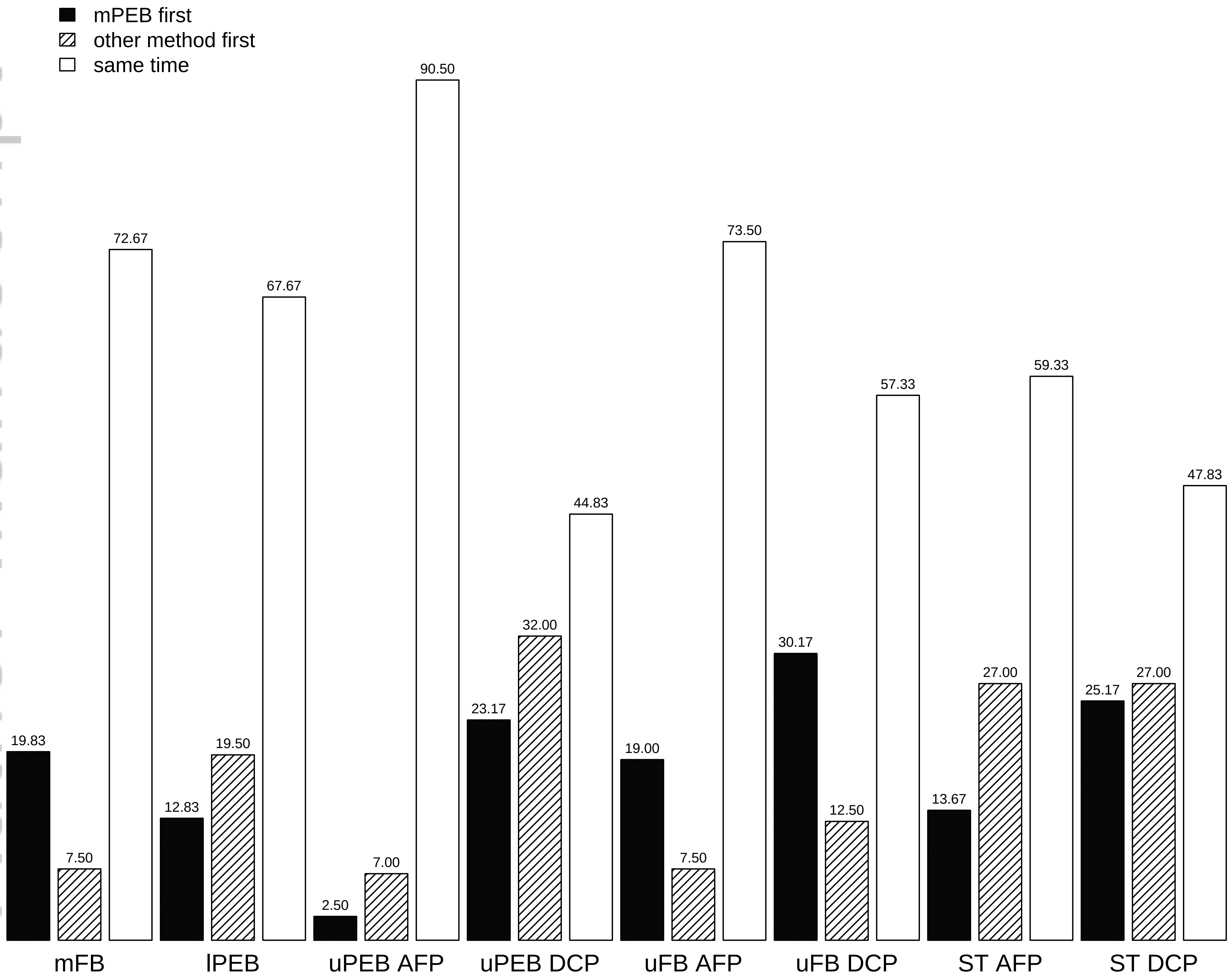
Author Manuscript



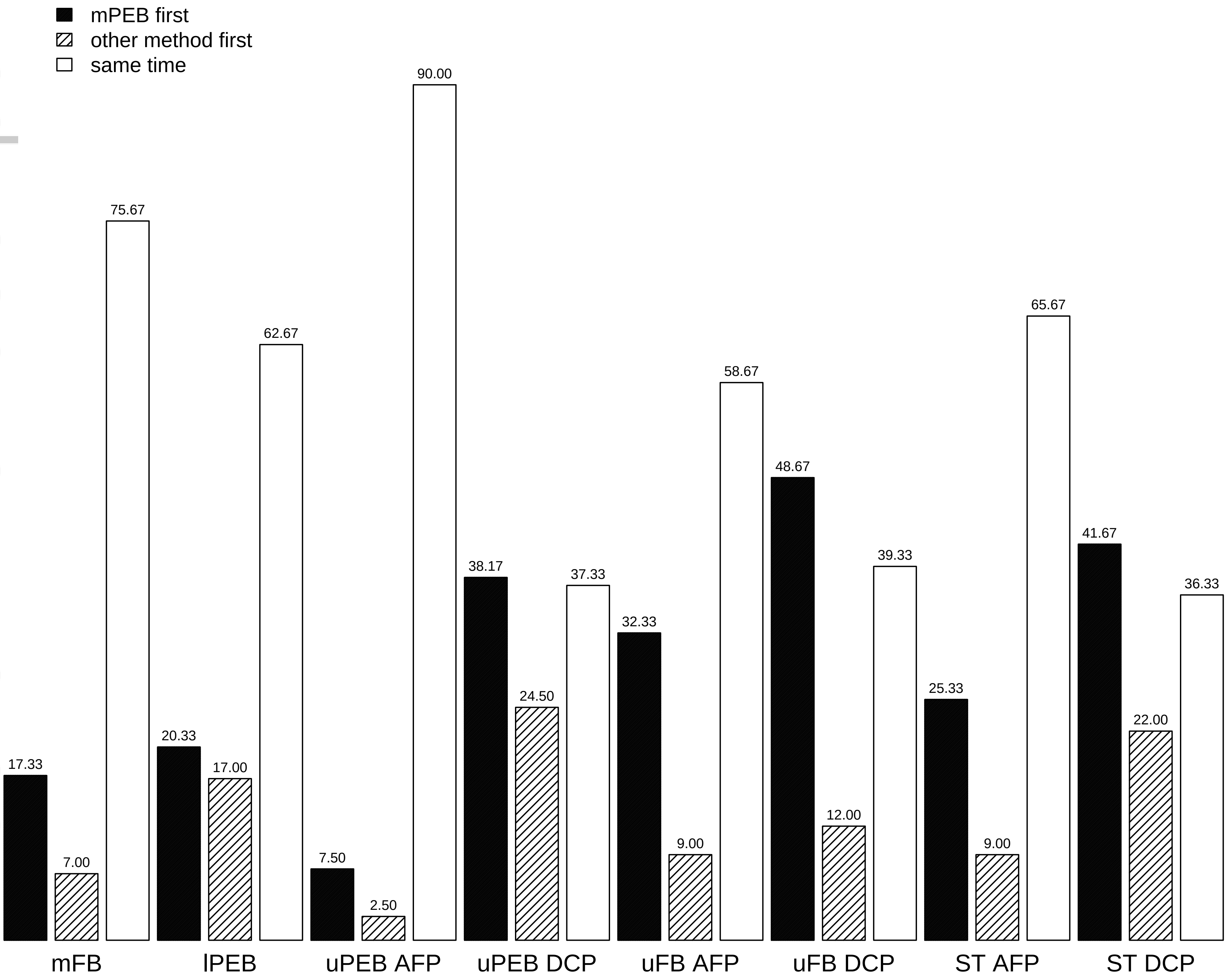
Author Manuscript


Randomized patients (N=1050)
Missing HCC status (N=2)
Randomized patients with complete clinical data (N=1048)
No cirrhosis at baseline biopsy (N=621)
Cirrhosis at baseline biopsy (N=427)
No HCC (N=581)
HCC confirmed (N=40)
No HCC (N=379)
HCC confirmed (N=48)
No HCC (N=558)
HCC confirmed (N=18)
No HCC (N=360)
HCC confirmed (N=24)
No HCC (N=918)
HCC confirmed (N=42)
Exclude patients with <12 months of follow-up and missing either AFP or DCP
Exclude HCC cases diagnosed in the extended phase and missing either AFP or DCP

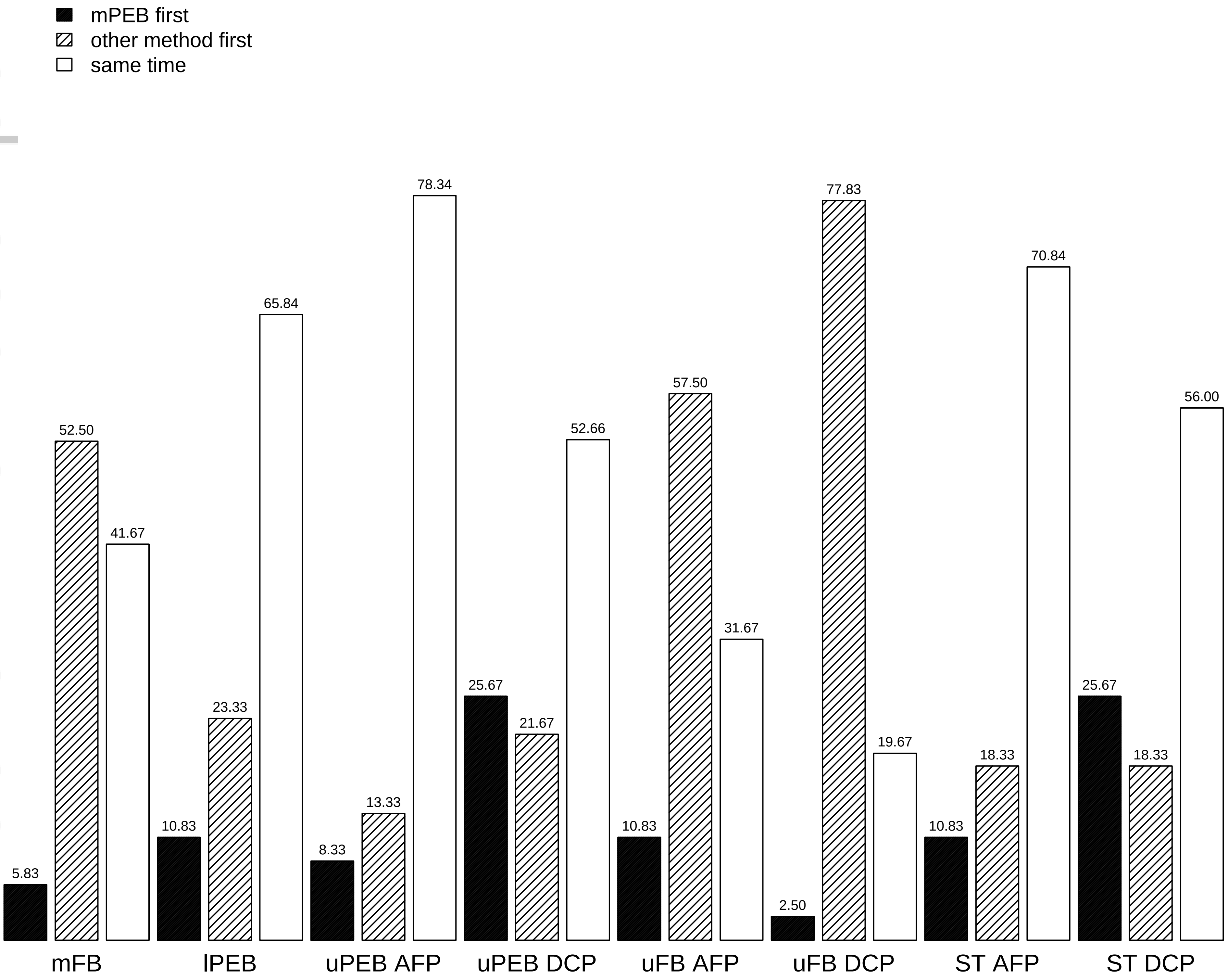
Author Manuscript



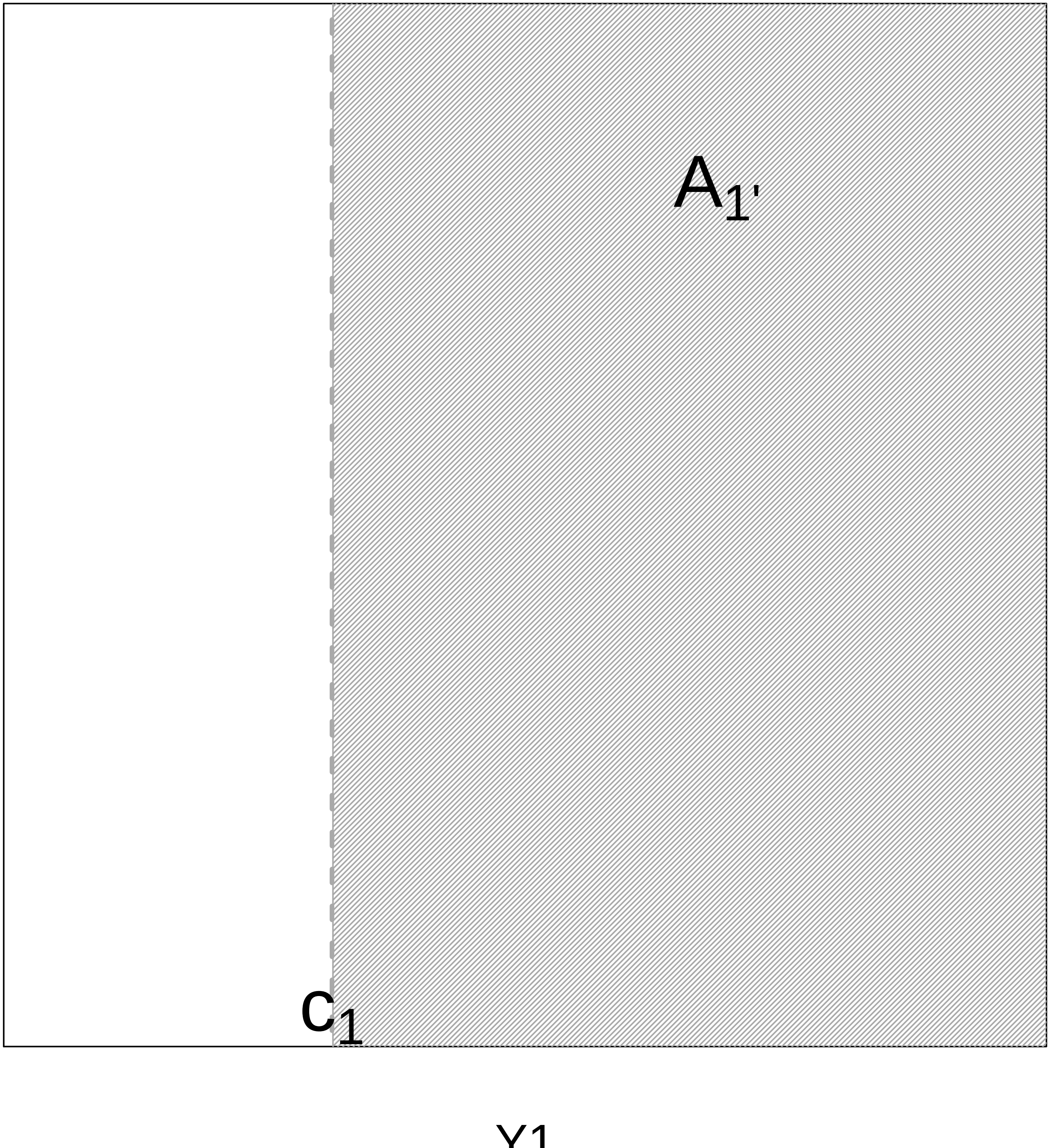
Author Manuscript



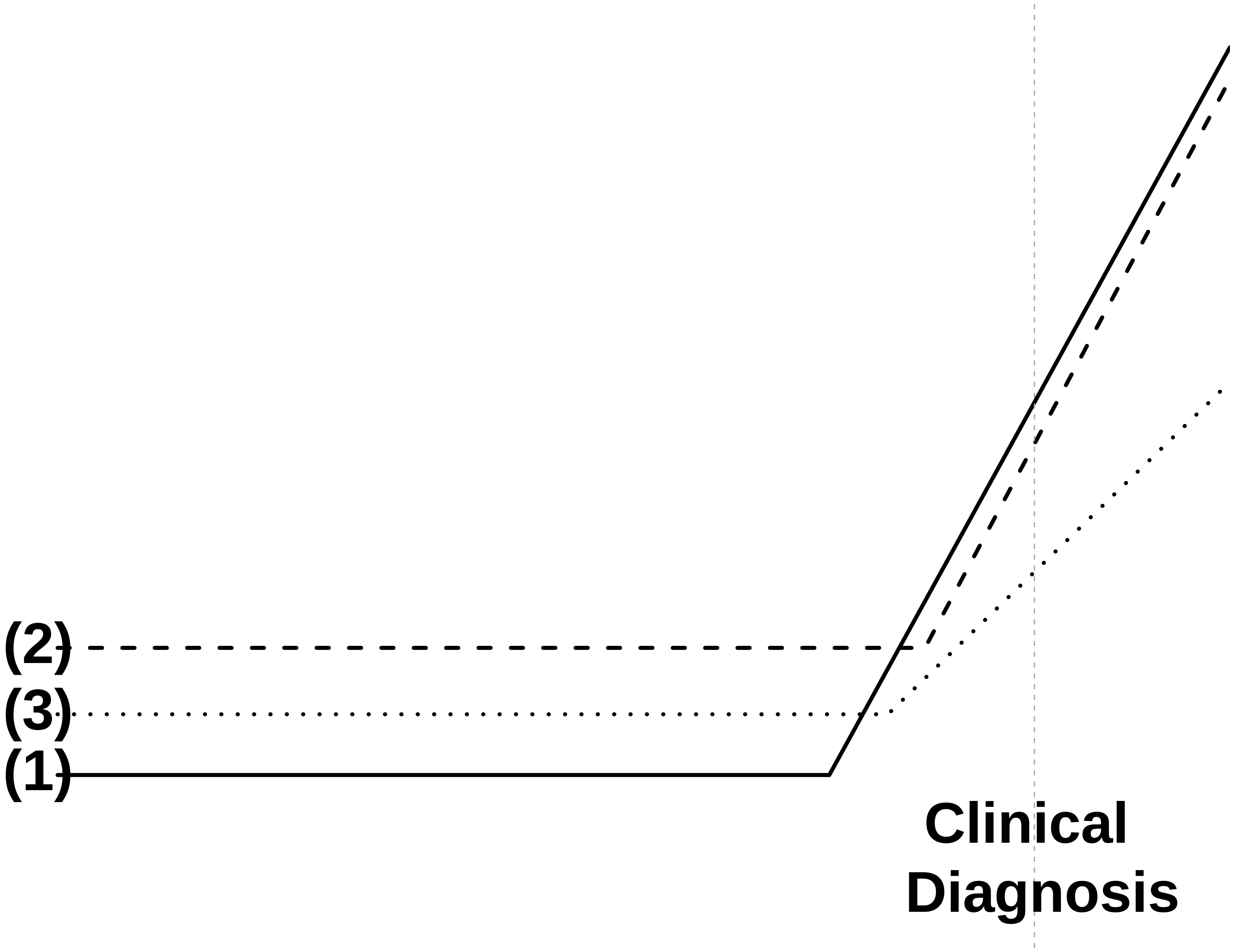
Author Manuscript



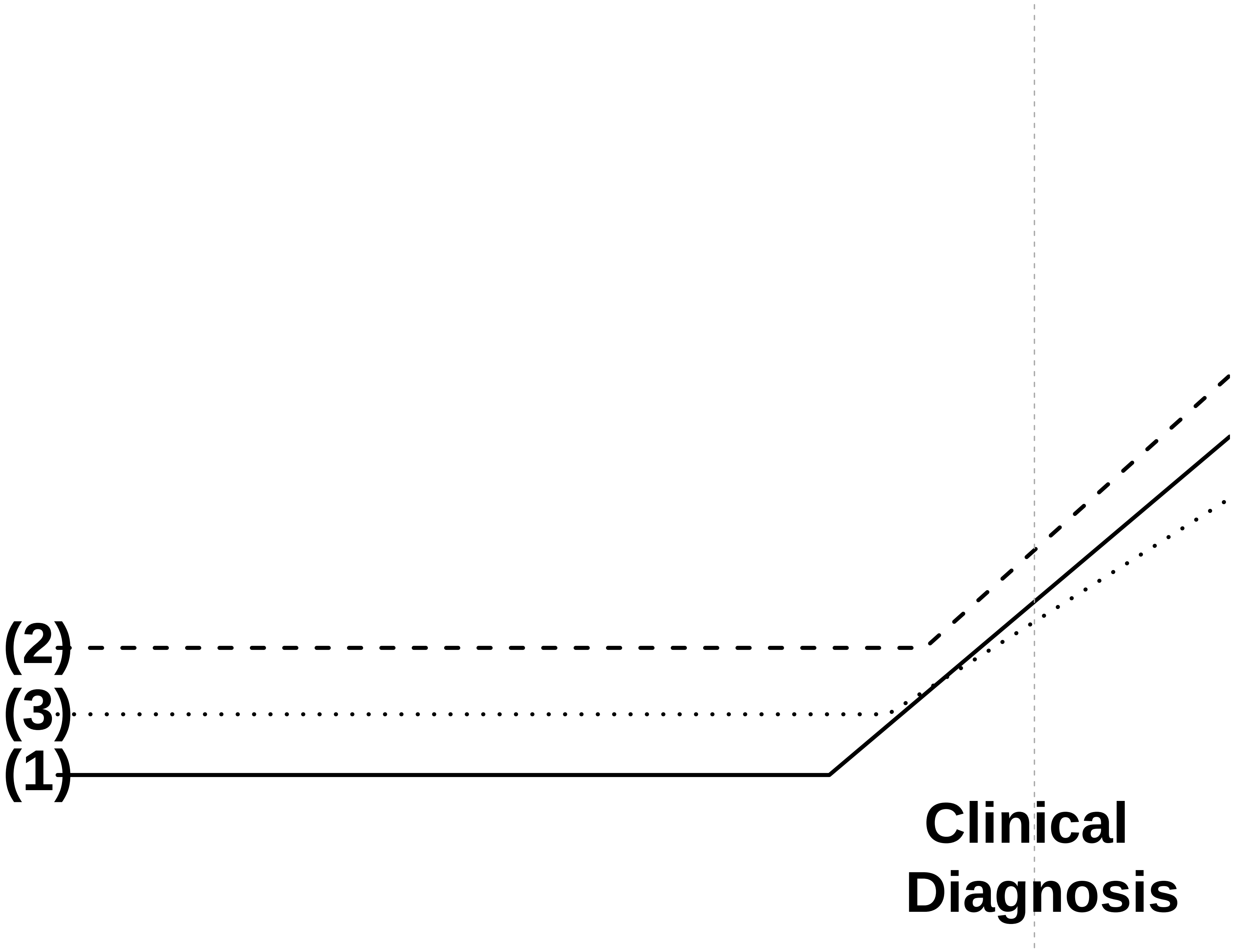
Author Manuscript



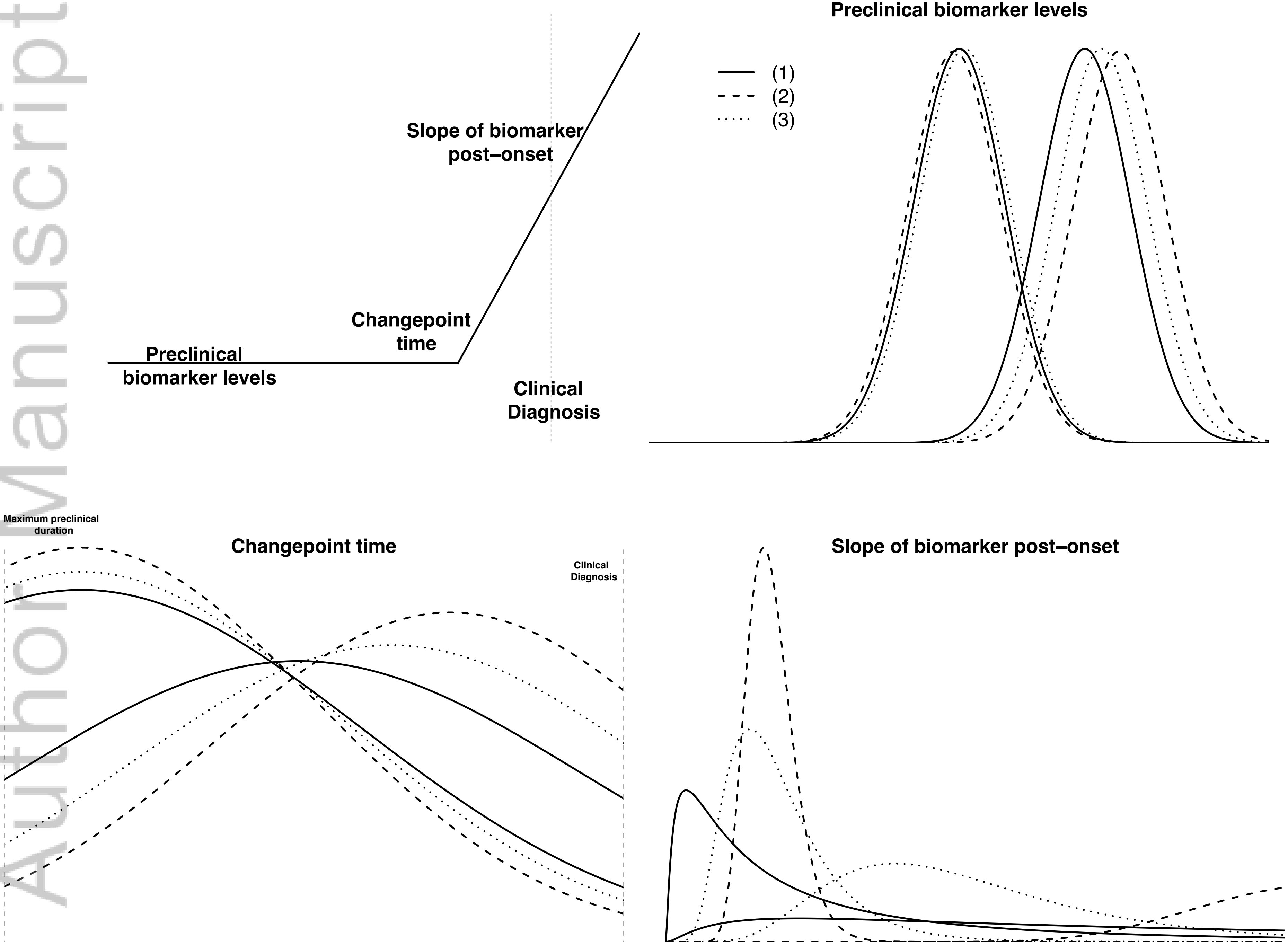
Author Manuscript



Author Manuscript



Author Manuscript


Slope of biomarker post–onset
Changepoint time
Preclinical biomarker levels
Clinical Diagnosis
Preclinical biomarker levels
(1)
(2)
(3)
Maximum preclinical duration
Changepoint time
Clinical Diagnosis
Slope of biomarker post–onset