Adjustment for the measurement error in evaluating biomarker performances at baseline for future survival outcomes: Time-dependent receiver operating characteristic curve within a joint modelling framework

Ruwanthi Kolamunnage-Dona and Adina Najwa Kamarudin

Abstract
The performance of a biomarker is defined by how well the biomarker is capable to distinguish between healthy and diseased individuals. This assessment is usually based on the baseline value of the biomarker; the value at the earliest time point of the patient follow-up, and quantified by ROC (receiver operating characteristic) curve analysis. However, the observed baseline value is often subjected to measurement error due to imperfect laboratory conditions and limited machine precision. Failing to adjust for measurement error may underestimate the true performance of the biomarker, and in a direct comparison, useful biomarkers could be overlooked. We develop a novel approach to account for measurement error when calculating the performance of the baseline biomarker value for future survival outcomes. We adopt a joint longitudinal and survival data modelling formulation and use the available longitudinally repeated values of the biomarker to make adjustment of the measurement error in time-dependent ROC curve analysis. Our simulation study shows that the proposed measurement error-adjusted estimator is more efficient for evaluating the performance of the biomarker than estimators ignoring the measurement error. The proposed method is illustrated using Mayo Clinic primary biliary cirrhosis (PBC) study.

Keywords
Time-dependent ROC curve, joint modelling, baseline biomarker, measurement error, primary biliary cirrhosis

Introduction
Due to current trends in medical practice towards personalised medicine, biomarkers have grown in importance in clinical studies. More and more studies are conducted to discover biomarkers that can accurately signal a clinical endpoint, e.g. measures of liver function such as prothrombin index as indicators of liver fibrosis, and in clinical practice, rapid tests of biomarkers hold the promise of prompt diagnosis of diseases for an improved outcome, e.g. sepsis. In this article, we refer the term “biomarker” to a single biomarker such as prothrombin index or to a composite risk score. A good biomarker can help identify patients who will have an early clinical benefit from a treatment or effectively guide the choice of therapeutic decisions, improving patients survival. However, due to imperfect laboratory conditions such as operator error, contamination, variable storage conditions, and limited machine precision, biomarkers are often subjected to substantial error in studies. Failing to adjust for such measurement error may hinder the explanatory power of the biomarker, and in a direct comparison,
useful biomarkers could be overlooked due to measurement error.\textsuperscript{4,5}

The performance of a biomarker is based on how well the biomarker is capable of discriminating between individuals who experience the disease onset (cases) from individuals who do not (controls). It is usually quantified by receiver operating characteristics (ROC) curve analysis, a well-established methodology in medical diagnostic research.\textsuperscript{6} The area under the ROC curve (AUC) is an effective way to summarise the discriminative capability of the biomarker. AUC takes values from 0 to 1, and a biomarker with high AUC is considered better. A single biomarker value at baseline is mainly used in this assessment. Baseline time is an important time horizon in practice, as it is considered as the earliest time point of the patient follow-up time and provides the time base to assess the disease progression. However, individuals who are free of disease at baseline may develop the disease later in the follow up, and therefore, the assumption of fixed disease status over time may not be appropriate when evaluating the biomarker performance. Hence, incorporating the time dimension in ROC curve analysis has recently been actively researched, enabling better clinical guideline in medical decision based on biomarkers. The time-dependent ROC curve is usually derived from risk regression models such as Cox proportional hazards model as they naturally account for censored failure times. This ROC curve estimates the performance of baseline biomarker at future time points. For example, in a breast cancer study, time-dependent ROC curve was used to assess whether the patients are free from subclinical disease if the clinical disease does not emerge by two years of screening.\textsuperscript{7} It has also been used to assess the predictive ability of the gene expression signatures in detecting early tumour response among metastatic colorectal cancer patients.\textsuperscript{8} Lu et al.\textsuperscript{9} identified a robust prognostic biomarker for tumour recurrence among lung cancer patients using time-dependent ROC curve analysis by estimating the AUC of the 51-gene expression signature at 60 and 100 months of follow-up. Using time-dependent AUC, Chen et al.\textsuperscript{10} made direct comparison of five recently recognised serum biomarkers and identified those can be recommended for use in clinical practice to surveillance of cirrhosis for hepatocellular carcinoma patients. A comprehensive review of current time-dependent ROC curve analysis approaches is provided by Kamarudin et al.\textsuperscript{11} However, Faraggi,\textsuperscript{12} Reiser\textsuperscript{13} and others concerned about ignoring the measurement error of biomarker values in ROC methodology, and showed that the effect can be substantial on the decision as to the diagnostic effectiveness of the biomarker.

As discussed by Henderson et al.\textsuperscript{14} and many others, a framework such as joint longitudinal and failure-time outcome modelling is capable of avoiding biases not only due to informative missingness in biomarker measurement schedule, but also due to measurement error. In a joint model, both longitudinally repeated biomarker and censored failure-time processes is modelled simultaneously. This novel modelling framework has rapidly been developed in the past decade (see Gould et al.\textsuperscript{15} and Tsiatis and Davidian\textsuperscript{16} for comprehensive reviews of the model). Many have adopted or extended this framework to investigate the association between the biomarker and the hazard of failure (e.g.\textsuperscript{17}), or to derive risk predictions (e.g. Proust-Lima and Taylor,\textsuperscript{18} Garre et al.\textsuperscript{19}). However, adopting the joint models for estimation of diagnostic effectiveness of a biomarker has been limited. Kolamunnage-Dona and Williamson\textsuperscript{20} used joint modelling framework to evaluate time-dependent discriminative capability of a biomarker within the ROC curve analysis. In other studies, ROC curve has been used to evaluate the accuracy of the predicted survival probabilities from the joint model (e.g. Rizopoulos\textsuperscript{21}). Henderson et al.\textsuperscript{22} has parameterised the underlying association between longitudinal biomarker and failure-time processes by individual-level deviation of the longitudinal profile from the population mean, but $R^2$ like statistic is used to quantify the predictive accuracy of a biomarker for failure rather than the ROC curve.

According to our review,\textsuperscript{11} measurement error of the biomarker has been ignored in all current time-dependent ROC curve approaches. And, to our knowledge, joint modelling framework has not been adopted to make adjustment for measurement error when evaluating the performance of the biomarker in ROC curve analysis. As its main contribution, this article provides a new development of time-dependent ROC curve to evaluate the performance of baseline biomarker correcting for the measurement error. We propose to utilise a joint model to link the baseline biomarker and failure-time process, and use the individual-level deviation of the biomarker from the population mean to develop an estimator to evaluate the time-dependent ROC curve. In health research, often biomarkers are recorded longitudinally as patients are followed up over time, and we use available longitudinal measurements of the biomarker to make adjustment of the measurement error in our proposed approach. By incorporating the longitudinally repeated biomarker measurements, we make the most efficient use of the data available.

**General notation**

Let $T_i$ be the true failure time (e.g. time to death or time to disease onset) for the $i$th individual. Let $\delta_i = I(T_i \leq C_i)$ be the indicator of the failure, taking values
1 if the failure is occurred at time $T_i$, and 0 if it is not occurred, so censored at time $C_i$. We observe the failure-time process $\{X_i, \delta_i\}$ where $X_i = \min(T_i, C_i)$ defines the observed failure-time for $i = 1, \ldots, n$ individuals in the study dataset. Let $y_i = \{y_{ij}(t_{ij}), j = 1, \ldots, m_i\}$ be a set of all available biomarker measurements recorded at times $t_{ij}, j = 1, \ldots, m_i$ for the $i$th individual. $y_{ij}$ is the biomarker measurement of the $i$th individual at baseline level (observed at baseline time $t_{ij} = 0$).

### Estimation of measurement error-adjusted estimator of baseline biomarker

Firstly, we formulate the joint model. A joint model is usually consisted of two submodels; a submodel for longitudinal measurements of the biomarker $y_i$, and a submodel for failure-time $\{X_i, \delta_i\}$. The two components are linked together through some shared parameters. Longitudinal data are typically modelled by linear mixed effect models, while the failure-times assume various choice of modelling approaches through shared latent effects. In general, a Gaussian linear model is assumed for longitudinal data, and proportion-al hazards is assumed for failure-times. Through the joint model, we can link the true (without measurement error) biomarker trajectory and the hazard of the failure for each individual. We follow joint modelling approach proposed by Henderson et al., but as we only consider that the baseline value of the biomarker is predictive of the failure, we formulate the joint model by

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + U_{0i} + U_{1i} t_{ij} + e_{ij}$$

where $U_{0i}$ and $U_{1i}$ are individual-level random intercept and random slope respectively, and they reflect the true difference between longitudinal profile of each individual from the population mean. In particular, $U_{0i}$ reflect the true deviation of the biomarker value from the population at baseline time. Therefore, through $U_{0i}$, the proposed submodel for failure-time links the risk of failure directly on the true scale of the biomarker at baseline for each individual. We assume $(U_{0i}, U_{1i})$ follows a bivariate normal distribution with mean 0 and variance $\Sigma_u = \begin{pmatrix} \sigma^2_{u0} & \rho \sigma_{u0} \sigma_{u1} \\ \rho \sigma_{u0} \sigma_{u1} & \sigma^2_{u1} \end{pmatrix}$. In the joint model, the measurement error process is accounted for by $e_{ij}$ in longitudinal data submodel. The measurement error process is non-differential and can be defined by a classical additive measurement error model $y_{ij} = y^*_{ij} + e_{ij}$ where $y^*_ij$ is the error-prone measure of $y_{ij}$. We assume $e_{ij}$ follows a Gaussian distribution with mean zero and variance $\sigma^2_e$. The above longitudinal data submodel assumes that in the absence of measurement error, the biomarker follows a perfectly linear trajectory. In failure-time submodel, $\lambda_0(t)$ is an unspecified baseline hazard, and $\gamma$ estimates the level of association between baseline biomarker and hazard for the failure.

We can estimate the model by maximising the joint likelihood of the observed data via the Expectation-Maximisation (EM) algorithm. The EM algorithm involves taking expectations with respect to the unobserved random effects $U_{0i}$ and $U_{1i}$, and it iterates between two steps (E and M) until convergence is achieved. For the proposed joint model, E-step determines expected values $E[U_{0i}]$ conditional on observed joint outcome $\{y_{ij}, X_i, \delta_i\}$. M-step maximises the complete data log-likelihood by $U_{0i}$ replaced by corresponding expectation. The EM algorithm provides the best linear unbiased estimates of the individual-specific deviations $U_{0i}$.

Secondly, based on the estimated values, we can compute measurement error-adjusted estimator based on the linear predictor of the failure-time submodel by

$$\hat{M}_i = \gamma \hat{U}_{0i}$$

where $\hat{U}_{0i}$ is the estimated true deviation of the biomarker value from the population mean at baseline for the $i$th individual, and $\gamma$ is the estimated association parameter between baseline biomarker and hazard for failure. Note that, $\exp(\gamma)$ is the hazard ratio associated with a unit increase in the value of biomarker at baseline with respect to the population mean. In our simulation study (Simulation investigations section), we will extensively explore the validity of $\hat{M}_i$ as the measurement error-adjusted estimator within time-dependent ROC curve analysis.

### Estimation of the time-dependent ROC curve at future time horizons

We need to define the cases and controls at time future time points $t_b(> 0)$. Let $R(t_b) = I(X_i \geq t_b)$ be at-risk indicator for each individual defining the riskset at a $t_b$. Then, dichotomise the riskset at time $t_b$ into two mutually exclusive groups: cases (experienced failure at $t_b$) and controls (survived failure beyond time $t_b$). At any $t_b$, each diseased individual (i.e. $\delta_i = 1$) plays a role as control for an early time $t_b < T_i$ but then play the role of case when $t_b = T_i$. In this case, the failure-time is represented through the counting process $N(t_b) = I(T_i \leq t_b)$, and the corresponding increment is defined by $dN(t_b) = N(t_b) - N(t_b^-)$ in terms of the failure time $T_i$ alone. This is the incident/dynamic
failure-times proposed by Heagerty and Zheng\(^2\) for the estimation of time-dependent ROC curve, and is the version adopted by most methodologists.

Finally, we can assess the discriminatory potential of the measurement error-adjusted estimator \(M_i\) at time \(t_h\) conditional on a threshold value \(c\). Following the standard ROC curve methodology, \(M_i \geq c\) determines the test positive (disease presence) and test negative (disease absent) if \(M_i < c\). The sensitivity and specificity of the error-adjusted baseline estimator at \(t_h\) can then be defined by

\[
\text{sensitivity}(c, t_h) = \Pr(M_i \geq c | dN(t_h) = 1)
\]

\[
\text{specificity}(c, t_h) = \Pr(M_i < c | N(t_h) = 0)
\]

where \(c \in (-\infty, +\infty)\). Sensitivity\((c, t_h)\) estimates the expected fraction of individuals with \(M_i \geq c\) among those who experience the failure at \(t_h\), while specificity\((c, t_h)\) estimates the expected fraction of individuals with \(M_i < c\) among those who survived failure beyond \(t_h\). To estimate the two conditional probabilities (conditional on incident/dynamic failure-times), we can use the proportional hazards properties of the joint likelihood function related to the failure-time submodel in (1). Xu and O’Quigley\(^24\) proposed estimating the proportion of variation in a covariate that is explained by failure times. They estimated the distribution of the covariate conditional on failure at a time \(t\) based on the weights \(\pi_k(t)\) from the Cox proportional hazards model. The same approach was later used by Heagerty and Zheng\(^23\) to estimate the time-dependent sensitivities and specificities as defined as (3). Following them, for a given threshold value \(c\), we estimate the sensitivity (or true positive fraction, TPF) at \(t_h\) by

\[
\text{sensitivity}(c, t_h) = \Pr(M_i \geq c | T_i = t_h) = \sum_k I(M_k \geq c) \pi_k(t_h)
\]

where \(\pi_k(t_h) = R_k(t_h) \exp(\hat{M}_k) / W(t_h)\) are the weights under proportional hazards and \(W(t_h) = \sum_k R_k(t_h) \exp(\hat{M}_k)\) is the total weight for the riskset individuals, and \(I(\cdot)\) is an indicator. We can calculate the specificity (or 1—false positive fraction, FPF) empirically by

\[
\text{specificity}(c, t_h) = \Pr(M_i < c | T_i > t_h) = \frac{\sum_k I(M_k < c) R_k^0(t_h)}{\sum_k R_k^0(t_h)}
\]

where \(R_k^0(t_h)\) is the set of failure-free individuals in the riskset at time \(t_h\) and \(\sum_k R_k^0(t_h)\) is the size of that control-set.

Bansal and Heagerty\(^25\) have also used the same incident/dynamic failure-times definition when there exists time-specific cases of interest at a particular time \(t_h\). However, we apply this definition to estimate the discriminatory capability at any time \(t_h > 0\) with no such prior information. Note that the proportional hazard assumption does not require any case to exist at \(t_h\) to estimate the above sensitivity; it will force the FPF equal to zero and specificity equal to one. Thus, although there is no case (had a failure) exists at \(t_h\) (which usually happens in practice), sensitivity can still be estimated at \(t_h\). Once the above sensitivity and specificity are computed at \(t_h\), the corresponding time-dependent ROC curve and AUC at time \(t_h\) for all \(c \in (-\infty, +\infty)\) can be computed by kernel (density) smoothing which follows closely the details of the original data.\(^26\) When there is no specific time \(t_h\) of interest, but restricted to a fixed follow-up period \((0, \tau)\), a global summary of the AUC can be provided by a survival concordance index (C-index).\(^27\)

In the proposed approach, \(M_i\) is computed from joint model estimates, which is then used as the input to ROC curve analysis; hence, the 95% confidence intervals (CIs) of the AUC, sensitivity and specificity are estimated by the bootstrap sampling with replacement\(^28\) to account for uncertainty due to the two estimation processes. The previously suggested time-dependent ROC models for censored failure-times also used bootstrap approaches to estimate the corresponding CIs.\(^11,20,23,27\) The software to implement the proposed joint model has been developed in R language, and will be available as part of the current joineR package.\(^29\) The risksetROC package in R can be used to estimate the corresponding incident/dynamic ROC curve,\(^23\) and we have modified corresponding R functions to implement the proposed ROC curve. The R codes are available from the authors on request.

**Simulation investigations**

We have conducted three simulation investigations to demonstrate whether the proposed approach is an appropriate framework for estimating the time-dependent ROC curve. The details of data simulation and investigations are given in the supplementary file. Firstly, we explored the accuracy of estimation of association parameter \(\gamma\) from the joint model which is crucial for estimating the correct ROC curve from the proposed approach. For comparison with \(\gamma\), we also fitted Cox proportional hazard model including the observed baseline biomarker value as a covariate, and also the estimated random intercept terms from the linear mixed effect (LME) model \((\hat{U}_0)_{\text{LME}}\) (usually referred as two-stage model\(^11\)) see more details of these models in supplementary file. Figure 1 presents
the bias for estimated association from the proposed joint, observed and \( \hat{U}_0 \)\_lme models for 30% censoring. The corresponding numerical values together with mean square error (MSE) and 95% coverage probabilities are given in supplemental table S1, and S2 and S3 present these for 50% and 70% censoring respectively. All estimates were obtained from 500 bootstrap samples with replacement. We observe that the joint model provides the most accurate estimation of the association with smaller biases, lower MSE and coverage probabilities closer to 95% across all settings. Both observed and \( \hat{U}_0 \)\_lme approaches underestimate the level of association to a great extent (high bias) when the true association is fairly strong and measurement error is high. The underestimation of the association from the model including observed baseline biomarker value is anticipated, as this approach assumes the biomarker value is measured without error. Although the \( \hat{U}_0 \)\_lme approach is computationally simpler than the joint model, due to the two individual regressions it could lead to bias estimation for conditional effects such as the association between the two outcomes. Our observations are also consistent with the previously published simulation study results from various joint model specifications [e.g.17,20]. The proposed joint model estimates the association fairly close to the true value with lower bias even when the measurement error is high; indicating that the proposed model makes the proper adjustment of measurement error when estimating the underlying association at the baseline level, and strengthening the case of using the model for estimating association between biomarker and risk of failure at baseline.

Secondly, we evaluated how the proposed measurement error-adjusted estimator \( \hat{M} \) modifies the ROC curve on a given association \( \gamma \). The C-index of M for a fixed follow-up period of (0, 2) was computed for varying level of association and compared with the true C-index (based on the true biomarker value at baseline). To compare with the proposed estimator, we used the observed baseline value, as the observed baseline value is generally used in ROC curve analyses. And due to computational simplicity of \( \hat{U}_0 \)\_lme, we also considered it as a potential estimator for the time-dependent ROC curve analysis, but \( \hat{U}_0 \)\_lme has not been previously used as an estimator of its own for the time-dependent ROC curve.11 Figure 2 presents the estimated C-Indexes for 30% censoring for various strengths of association. The corresponding numerical values together with bias, MSE and 95% coverage probabilities are given in supplemental table S4, and S5 and S6 present these for 50% and 70% censoring respectively. When there is no association between the baseline biomarker and failure, the C-index is estimated from the proposed estimator \( \hat{M} \) fairly close to the null value of 0.5 (indicating that biomarker shows no discriminatory potential) across all settings of \( \gamma \) and censoring %s, see supplemental tables. As strength of the association is stronger (\( \gamma \) moves towards 1.0), the estimated C-index from M is also increased by acceptable margins. We can observe from Figure 2 that the point estimate of the C-Index from both M and \( \hat{U}_0 \)\_lme are fairly similar, especially when the association is
weak to moderate, and \( \hat{U}_0 \) point estimate is better as compared to the observed value. However, \( \hat{U}_0 \) substantially under-coverage the 95% confidence intervals as compared to that from M, especially when the association is moderate to strong; see Supplementary tables S4, S5, and S6. The coverage issue associated with \( \hat{U}_0 \) is also observed in association estimation (supplementary tables S1, S2, and S3). Therefore, we can expect the same extent of under-coverage for other ROC curve summaries, implying \( \hat{U}_0 \) is failed as an estimator for the time-dependent ROC curve analysis.

M provides the most accurate C-index estimation with smaller biases, and also with lower MSE and higher 95% coverage probabilities across all settings of \( \gamma \) and censoring %s.

Finally, the accuracy of the time-dependent ROC curve was further evaluated by comparing the estimated AUC(\( t_h \)) at future time points \( t_h \) with the true AUC(\( t_h \)). We also computed the sensitivity(\( t_h \)) and specificity(\( t_h \)) at optimal thresholds for the proposed M, and compared with the observed baseline value estimates. Table 1 presents the bias for estimated AUC(\( t_h \)) at \( t = t_h = 1, 2, 3, 4 \) for \( \gamma = 1 \) and 30% censoring and estimated sensitivity(\( t_h \)) and specificity(\( t_h \)). Supplemental table S7 to S21 present bias, MSE and 95% coverage probabilities of estimates for all values of \( \gamma \) and censoring rates. The estimated AUC(\( t_h \)) from the proposed measurement error-adjusted M is more accurate, with lower biases and MSE as compared to the observed baseline value. We observe that M is failed to archive the nominal coverage when the measurement error is considerably high at early time points; however, such high measurement error is rarely observed in current clinical data due to precision of latest machinery and better laboratory regulations. This investigation proves that the proposed methodology effectively corrected for a moderate measurement error when calculating performance of the baseline biomarker at future time points. As expected, AUC(\( t_h \)) decreases as \( t_h \) increases because discriminatory potential of the biomarker becomes weaker as departing from baseline.\(^{23}\)

**Application: Mayo Clinic primary biliary cirrhosis (PBC) study**

We apply the proposed approach to the data from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. PBC is a fatal, but rare liver disease. If PBC is not treated, and reaches an advanced stage, it can lead to several major complications, including death. The trial randomised 312 patients between D-penicillamine (n = 158) for the treatment of PBC and placebo (n = 154).\(^{30}\) Among the 312 patients randomised, 125 died during the follow-up. Although the study established that D-penicillamine is not effective for the treatment of PBC, the data have been used to develop clinical prediction models, and has been widely analysed using joint modelling methods.\(^{31–34}\) Patients with PBC typically have abnormalities in several blood tests; hence, during the study follow-up several biomarkers associated with liver function were serially recorded for these patients. In this

---

**Figure 2.** Estimated C-Index for 30% censoring. Square indicates the estimated value from the proposed measurement adjusted model, circle the Cox model with observed baseline value and triangle the estimated random intercept term from the LME model. The horizontal dashed line is the true C-index for corresponding association parameter.
In this article, we considered three biomarkers: serum bilirubin (measured in units of mg/dl), serum albumin (mg/dl), and prothrombin time (seconds) with the aim of assessing the performance of each biomarker at the baseline level for patient survival. The available longitudinal measurements of each biomarker were used to correct for measurement error. As the proposed modelling framework assumes Gaussian random effects and errors, the bilirubin measurements were log-transformed and the prothrombin time were transformed by $(0.1 / \text{prothrombin time})^{-4}$ as suggested by Box-Cox transformation. Albumin did not require transformation. A linear trajectory is assumed for each biomarker and the residual plots did not indicate any deviations from the linear form; see supplementary file for diagnostic plots.

Table 2 shows the estimated time-dependent AUC, sensitivity and specificity at times $t_h = \text{Year 1, Year 5 and Year 10}$ together with estimated measurement error and association parameter for both the measurement error-adjusted and observed baseline biomarker. The 95% confidence intervals (CI) were computed from 500 bootstrap samples with replacement. The measurement error-adjusted estimator is associated with a considerably high time-dependent AUC($t_h$) than the observed baseline biomarker at all $t_h$, and this is observed across all three biomarkers. The level of association between the baseline biomarker and risk of death is also substantially underestimated as the measurement error is ignored. Among the three biomarkers, once corrected for the measurement error, the highest time-dependent AUC($t_h$) was achieved for serum bilirubin, which means that among the 3 biomarkers, serum bilirubin is best for the earliest diagnosis of PBC. The ROC($t_h$) curves at three discrete time points and time-dependant AUC($t_h$) over continuous time for serum bilirubin are shown in Figure 3.

| $t_h$ | Measurement error adjusted | Observed baseline value |
|-------|-----------------------------|-------------------------|
|       | AUC (SE) | Bias | Sensitivity (SE) | Specificity (SE) | AUC (SE) | Bias | Sensitivity (SE) | Specificity (SE) |
|       |         |      |                |                |         |      |                |                |
| 1     | 0.73(0.02) | -0.02 | 0.67 (0.01) | 0.66 (0.01) | 0.70(0.02) | -0.04 | 0.65 (0.01) | 0.65 (0.01) |
| 2     | 0.70(0.01) | -0.02 | 0.65 (0.01) | 0.64 (0.01) | 0.69(0.01) | -0.03 | 0.64 (0.01) | 0.64 (0.01) |
| 3     | 0.68(0.01) | -0.02 | 0.64 (0.01) | 0.63 (0.01) | 0.68(0.01) | -0.02 | 0.63 (0.01) | 0.62 (0.01) |
| 4     | 0.66(0.02) | -0.02 | 0.62 (0.02) | 0.61 (0.02) | 0.66(0.02) | -0.02 | 0.62 (0.02) | 0.61 (0.02) |

**Table 1.** Time-dependent AUC (Standard Error, SE) and Bias at $t_h$ for the measurement-error adjusted and observed biomarker when $\gamma = 1$ and 30% censoring for varying measurement error variance. Sensitivity (SE) and specificity (SE) are estimated at the corresponding optimal threshold.
Discussion

The focus of this article was to develop a novel methodology for evaluating time dependent performance of the baseline biomarker correcting for measurement error. We proposed a novel utility of the joint modelling framework within the theory of time-dependent ROC curve analysis by developing a more efficient estimator that links the risk of failure and baseline biomarker. The baseline is an important time point as the biomarker value at baseline can serve as the earliest indicator of a potential future adverse clinical event (e.g. death). We have shown from our simulation investigations that measurement error could cause a severe bias in estimating the association between the baseline biomarker and risk of failure event. Although, this has been investigated in joint modelling literature in relation to various specifications of the model, this study was the first to show that observed baseline value could severely underestimate the true discriminative capability of the biomarker as estimated by AUC. Our simulation investigations proved that the proposed methodology effectively corrects for a moderate measurement error when calculating the performance of the baseline biomarker over time.

A similar joint model specification was suggested by Crowther et al.\textsuperscript{35} to predict survival for new patients. In their model, association was defined on the current biomarker value rather than the individual-level...
deviation, and restricted cubic splines were used to define the longitudinal biomarker while failure-time assumes a parametric distribution. This level of complexity is necessary to model highly nonlinear biomarker trajectories over time, and to capture complex baseline hazards when predicting the future survival probabilities. However, our aim was to quantify the true discriminant capability of the baseline biomarker at future time points, and a more classical modelling and estimation framework has been proven sufficient from our thorough simulation study. To facilitate the use of the methods in practice, software is written in R language (which is a free software environment). The proposed approach can be implemented with a relatively low computational burden; for example, in our application dataset with 312 patients, the proposed joint model for each biomarker took under 1 minutes to converge on a standard desktop computer, and the time-dependent AUCs were derived in few seconds.

More recently, quantities such as proportion of information gain (PIG) have been proposed to measure the importance of a biomarker. Li and Qu adjusted for the measurement error in calculating the PIG for continuous, binary and failure-time outcomes. However, our focus in this article was to account for the measurement error of a more familiar and well established quantity among the medical research community. We proposed a computationally simple approach to estimate the true time-dependent ROC curve for a baseline biomarker subjected to measurement error. Although information from longitudinally repeated measurements is required for the proposed approach in addition to the single biomarker measurement at baseline, often in clinical studies, longitudinal measurements are recorded alongside the main study as secondary outcomes, e.g. to monitor the progression of a disease. Therefore, the prospects of utilising the proposed framework to detect the true performance of biomarkers is quite substantial.

The proposed ROC curve approach can be extended to incorporate multiple biomarkers by utilising multivariate joint models (e.g. Hickey et al.). In our application, we evaluated the measurement error-corrected performance of three biomarkers in separation for the survival of PBC patients. It may be of interest to assess the performance in a combination of biomarkers, as in many diseases it is unlikely that a single biomarker will ever be more effective due to complexity of the disease (e.g. Aerts et al.).

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: AK was supported by Malaysia government PhD studentship Majlis Amanah Rakyat Malaysia (MARA) during 2014-2018. This work was also partly supported by the Medical Research Council [grant number MR/M013227/1].

ORCID iD
Ruwanthi Kolamunnage-Dona https://orcid.org/0000-0003-3886-6208

Supplemental material
Supplemental material for this article is available online.

References
1. Adams LA. Biomarkers of liver fibrosis. J. Gastroen Hepatol 2011; 26: 802–809.
2. Toh CH, Ticknor LO, Downey C, et al. Early identification of sepsis and mortality risks through simple, rapid clot-waveform analysis. Intensive Care Med 2003; 29: 55–61.
3. Keogh RH and White IR. A toolkit for measurement error correction, with a focus on nutritional epidemiology. Stat Med 2014; 33: 2137–2155.
4. Tsiatis AA, Degruttola V and Wulfsohn MS. Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. J Am Stat Assoc 1995; 90: 27–37.
5. Wulfsohn MS and Tsiatis AA. A joint model for survival and longitudinal data measured with error. Biometrics 1997; 53: 330–339.
6. Zweig MH and Campbell G. Receiver-operator characteristic plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993; 39: 561–577.
7. Cai T, Pepe MS, Zheng Y, et al. The sensitivity and specificity of markers for event times. Biostatistics 2006; 7: 182–197.
8. George J, Claes P, Vunckx K, et al. A textural feature based tumor therapy response prediction model for longitudinal evaluation with PET imaging. In: 2012 9th IEEE international symposium on biomedical imaging (ISBI), 2012, pp.1048–1051. Barcelona, Spain.
9. Lu Y, Wang L, Liu P, et al. Gene-expression signature predicts postoperative recurrence in stage I non-small cell lung cancer patients. PLoS One 2012; 7: e30880.
10. Chen H, Zhang Y, Li S, et al. Direct comparison of five serum biomarkers in early diagnosis of hepatocellular carcinoma. Cancer Manage Res 2018; 10: 1947–1958.
11. Kamarudin AN, Cox T and Kolamunnage-Dona R. Time-dependent ROC curve analysis in medical research: current methods and applications. BMC Med Res Methodol 2017; 17: 53.
12. Faraggi D. The effect of random measurement error on receiver operating characteristic (ROC) curves. Stat Med 2000; 19: 61–70.
13. Reiser B. Measuring the effectiveness of diagnostic markers in the presence of measurement error through the use of ROC curves. *Stat Med* 2000; 19: 2115–2129.
14. Henderson R, Diggle P and Dobson A. Joint modelling of longitudinal measurements and event time data. *Biomometrics* 2000; 1: 465–480.
15. Gould AL, Boye ME, Crowther MJ, et al. Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group. *Stat Med* 2015; 34: 2181–2195.
16. Tsiatis A and Davidian M. Joint modeling of longitudinal and time-to-event data: an overview. *Stat Sin* 2004; 14: 809–834.
17. Sweeting MJ and Thompson SG. Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biom J* 2011; 53: 750–763.
18. Proust-Lima C and Taylor JM. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach. *Biomometrics* 2009; 10: 535–549.
19. Garre F, Zwinderman A, Geskus R, et al. A joint latent class change point model to improve the prediction of time to graft failure. *J R Stat Soc Ser A* 2008; 171: 299–308.
20. Kolamunnage-Dona R and Williamson PR. Time-dependent efficacy of longitudinal biomarker for clinical endpoint. *Stat Methods Med Res* 2018; 27: 1909–1924.
21. Rizopoulos D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* 2011; 67: 819–829.
22. Henderson R, Diggle P and Dobson A. Identification and efficacy of longitudinal markers for survival. *Biomometrics* 2002; 3: 33–50.
23. Heagerty PJ and Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics* 2005; 61: 92–105.
24. Xu R and O’Quigley J. Proportional hazards estimate of the conditional survival function. *J R Stat Soc Ser B* 2000; 62: 667–680.
25. Bansal A and Heagerty PJ. A tutorial on evaluating time-varying discrimination accuracy for survival models used in dynamic decision-making. *Med Decis Making* 2018; 38: 904–916.
26. Zou KH, Hall WJ and Shapiro DE. Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Stat Med* 1997; 16: 2143–2156.
27. Heagerty PJ, Lumley T and Pepe MS. Time dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 2000; 56: 337–344.
28. Efron B and Tibshirani RJ. *An introduction to the bootstrap*. New York: Chapman and Hall, 1993.
29. Philipson P, Diggle P, Sousa I, et al. jointR: joint modeling of repeated measurements and time-to-event data. UK: Comprehensive R Archive Network, 2012.
30. Fleming TR and Harrington DP. *Counting processes and survival analysis*. New Jersey: John Wiley & Sons, 2011.
31. Andrinopoulou E-R and Rizopoulos D. Bayesian shrinkage approach for a joint model of longitudinal and survival outcomes assuming different association structures. *Stat Med* 2016; 35: 4813–4823.
32. Albert PS and Shih JH. An approach for jointly modeling multivariate longitudinal measurements and discrete time-to-event data. *Ann Appl Stat* 2010; 4: 1517–1532.
33. Crowther MJ, Abrams KR and Lambert PC. Joint modeling of longitudinal and survival data. *Stat J* 2013; 13: 165–184.
34. Hickey G, Philipson P, Jorgensen A, et al. jointRML: a joint model and software package for time-to-event and multivariate longitudinal outcomes. *BMC Med Res Methodol* 2018; 18: 50.
35. Crowther MJ, Lambert PC and Abrams KR. Adjusting for measurement error in baseline prognostic biomarkers included in a time-to-event analysis: a joint modelling approach. *BMC Med Res Methodol* 2013; 13: 146.
36. Li W and Qu Y. Adjustment for the measurement error in evaluating biomarkers. *Stat Med* 2010; 29: 2338–2346.
37. Aerts M, Benteyn D, Van Vlierberghe H, et al. Current status and perspectives of immune-based therapies for hepatocellular carcinoma. *World J Gastroenterol* 2016; 22: 253–261.