Published in final edited form as:

Stat Methods Med Res. 2016 December ; 25(6): 2972–2991. doi:10.1177/0962280214535763.

Individualized dynamic prediction of prostate cancer recurrence with and without the initiation of a second treatment: development and validation

Mbéry Sène[1,2,*], **Jeremy M. G. Taylor**[3], **James J. Dignam**[4], **Hélène Jacqmin-Gadda**[1,2], and **Cécile Proust-Lima**[1,2]

[1]INSERM, U897, Epidemiology and Biostatistics Research Center, 33076 Bordeaux, France

[2]Université de Bordeaux, 33076 Bordeaux, France

[3]Department of Biostatistics, department of Radiation Oncology, University of Michigan, Ann Arbor, MI

[4]Department of Health Studies, University of Chicago and Radiation Therapy Oncology Group, American College of Radiology, Philadelphia

Abstract

With the emergence of rich information on biomarkers after treatments, new types of prognostic tools are being developed: dynamic prognostic tools that can be updated at each new biomarker measurement. Such predictions are of interest in oncology where after an initial treatment patients are monitored with repeated biomarker data. However, in such setting, patients may receive second treatments to slow down the progression of the disease. This paper aims to develop and validate dynamic individual predictions that allow the possibility of a new treatment in order to help understand the benefit of initiating new treatments during the monitoring period. The prediction of the event in the next x years is done under two scenarios: (1) the patient initiates immediately a second treatment, (2) the patient does not initiate any treatment in the next x years. Predictions are derived from shared random-effect models. Applied to prostate cancer data, different specifications for the dependence between the PSA repeated measures, the initiation of a second treatment (hormonal therapy) and the risk of clinical recurrence are investigated and compared. The predictive accuracy of the dynamic predictions is evaluated with two measures (Brier score and prognostic cross-entropy) for which approximated cross-validated estimators are proposed.

Keywords

Brier score; Dynamic predictions; Hormonal treatment; Joint model; Prognostic models; Prostate cancer; Shared random-effect models

Correspondence to: Cécile Proust-Lima, INSERM U897, ISPED, Université de Bordeaux, 146 rue Léo Saignat, 33076 Bordeaux Cedex, FRANCE, cecile.proust-lima@inserm.fr.

Declaration of Conflicting Interests: none.

1 Introduction

The increase of rich longitudinal information in health studies[1–6] has motivated the development of joint models of longitudinal and time-to-event data. In addition to providing an efficient framework to correctly model correlated longitudinal markers and clinical events and to better understand their interrelationship, these models have more recently offered a new approach for monitoring the patients after the diagnosis of a chronic disease.

Indeed, dynamic individual predictions of an event of interest can be easily derived from joint models[7–9]. They consist of the predicted probability of having the event in a certain window of time conditional on the longitudinal marker history. As such, in contrast with standard predictive tools that only use information at diagnosis, these predictions can dynamically adapt the risk of the event of interest to the individual trajectory of the biomarker of progression.

Such predictions are of interest in oncology where after a first treatment, patients are monitored to detect early recurrence of the cancer. In particular, after prostate cancer diagnosis and initial treatment by radiation therapy, some dynamic tools that are based on repeated measures of PSA (prostate specific antigen), in addition to the standard prognostic factors were found to more accurately predict the risk of clinical recurrence than standard prognostic tools[7,10].

In practice, after a first treatment of the cancer, patients may receive a second treatment (ST) to slow down the progression of the disease, and prevent or delay the recurrence of the cancer. In particular for prostate cancer, after a first treatment by radiation therapy, the patient may initiate hormonal therapy (HT) when he has a high risk of clinical recurrence. The optimal time to initiate such HT is unknown. The timing is mainly determined by the clinician based on experience, the knowledge of the disease and the observation of the PSA trajectory. Yet, HT has consequences for the patient's personal life, so that in accordance with the principle of personalized medicine[11], assessing the benefit for the patient in terms of reduction of his risk of recurrence of initiating a ST is of great importance.

Dynamic predictive tools can be used for this purpose as they provide an up-to-date quantification of the risk of recurrence. However most dynamic predictions currently developed do not take into account a possible change of treatment[7,8,12,13]: they are computed by assuming the patient will not initiate any ST in the window of prediction. As ST's usually change the dynamics of the biomarker and of the time-to-event, such a joint model would be more complex to define[13,14] and differential dynamic predictions would be required to distinguish the risk of an event according to the initiation of ST.

In this context, the objective of this paper is to compute and validate dynamic individual predictions of an event in the next x years based on different scenarios, in particular two conditions: (1) whether the patient initiates a ST today or (2) whether the patient does not initiate any ST in the window of prediction. The idea is to provide tools to help the clinician quantifying the potential benefit of starting immediately a new treatment compared to postponing the decision to the next monitoring visit. To reach this goal, we utilized the joint model methodology and focused on the prostate cancer example with initiation of hormonal therapy. We explored and compared different specifications of the interrelationships between the PSA repeated measures, the initiation of HT and the risk of clinical recurrence. Dynamic individual predictions were computed in situations (1) and (2), and several measures of predictive abilities were considered.

In order to validate the dynamic predictive tools, we considered the quadratic error of prediction (Brier Score - BS)[10,15] and developed a new estimator which is valid both on external data and on estimation data. The usual overoptimism due to the use of the same data for the estimation and the prediction was corrected by applying a general formula of approximated cross-validation recently developed[16]. In addition to the Brier Score, we also considered an information criterion for assessing the prognostic value of joint models (EPOCE)[10,17]. Specifically in situation (1) that focuses on the subsequent years after initiation of a ST, integrated versions of these two predictive accuracy measures over the times of initiation of ST were defined.

The paper is organized as follows. Section 2 describes the classical joint model methodology. Section 3 details the development of the differential individual dynamic predictions and their validation with BS and EPOCE estimators. Section 4 describes the application to Prostate Cancer with the evaluation of the risk of clinical recurrence based on the PSA trajectory after initial radiation therapy, and on the immediate initiation of hormonal therapy. Finally, the model and the results are discussed in section 5.

2 Joint models

The shared random-effect approach was chosen to jointly model the biomarker repeated measures and the time-to-event[1–3,5,6]. A general description is given below and more specific models will be described in the application section.

2.1 Notation

Let T_i^* be the time to the event of interest and C_i the censoring time for subject i, $i = 1, \ldots, N$. We observe the time $T_i = \min(T_i^*, C_i)$ and $E_i = {}_{\{T_i^* \leq C_i\}}$ is the indicator of event. Let τ_i be the time to ST (unobserved if $\tau_i > T_i$) and $\mathbb{1}_{\{t \geq \tau_i\}}$ the indicator of ST status at any time t. For each subject we also collect n_i repeated measures of the biomarker $Y_i = (Y_i(t_{i1}), \ldots, Y_i(t_{in_i}))$ at times $(t_{i1}, \ldots, t_{in_i})$.

2.2 Longitudinal submodel

The biomarker trajectory in the absence of second treatment initiation is described using a linear mixed model[18]. The biomarker pattern after a possible ST initiation is not modelled. We assume that for $j = 1, \ldots, n_i$, the repeated measures $Y_i(t_{ij})$ are noisy measures of $Y_i^*(t_{ij})$ the true unobserved biomarker value. The mean change over time of $Y_i^*(t_{ij})$ can depend on covariates and the within-subject correlation of the biomarker repeated measures is captured using subject-specific random effects:

$$\begin{aligned} Y_i(t_{ij}) &= Y_i^*(t_{ij}) + \varepsilon_i(t_{ij}) \\ &= X_{Li}(t_{ij})^T \beta + Z_i(t_{ij})^T b_i + \varepsilon_i(t_{ij}) \end{aligned} \quad (1)$$

where $X_{Li}(t_{ij})$ and $Z_i(t_{ij})$ are a p-vector and a q-vector of time-dependent covariates associated respectively with the p-vector of fixed effects β and the q-vector of random effects b_i, $b_i \sim \mathscr{N}(0, B)$. The vector of independent errors of measurement $\varepsilon_i = (\varepsilon_i(t_{i1}), \ldots, \varepsilon_i(t_{in_i})) \sim \mathscr{N}(0, \Sigma_i = \sigma^2 I_{n_i})$; ε_i and b_i are independent.

2.3 Survival submodel

To model the risk of the event and to quantify the effects of the biomarker dynamics and the initiation of ST on this risk adjusted for other covariates, we define a proportional hazard model as follows:

$$\lambda_i(t|X_{Si}, b_i, \tau_i) = \lambda_0(t)\exp\left[X_{Si}^T \gamma + W_i(b_i, \tau_i, t)^T \phi\right] \quad (2)$$

where $\lambda_0(t)$ is the baseline hazard function, γ is the r-vector of coefficients associated with the r-vector of time-independent covariates X_{Si}, and ϕ is the vector of parameters associated with $W_i(b_i, \tau_i, t)$, the multivariate function of the random effects b_i from model (1) and the initiation of ST when ST is considered. In the standard framework without any ST, $W_i(b_i, \tau_i, t)$ defines the nature of the dependence between the longitudinal and the survival processes and ϕ measures the corresponding strength of association. The most common example is $W_i(b_i, \tau_i, t) = Y_i^*(t)$ that assumes an association with the risk of event through true current level of the biomarker. In the presence of a ST, $W_i(b_i, \tau_i, t)$ captures both the dependence between the two processes and the effect of ST on the hazard. For example, $W_i(b_i, \tau_i, t) = (Y_i^*(t);_{\{t \geq \tau_i\}})$ models independent effects of the true current level of the biomarker and a change of risk after ST initiation. Other examples of $W_i(b_i, \tau_i, t)$ will be described in the application (section 5.2).

2.4 Maximum likelihood estimation

Maximum likelihood estimates were obtained using the JM R package[19], with modifications of the JM source code when necessary. The maximum likelihood estimates, denoted $\hat{\theta}$, were obtained by a quasi-Newton algorithm with a convergence criterion on the log-likelihood[19]. The two integrals involved in the log-likelihood computation were approximated using Gaussian quadrature[19]. Estimates of the variance-covariance matrix of the estimated parameters $\widehat{\mathrm{V}(\hat{\theta})}$ were provided by the inverse of the Hessian matrix.

3 Individual dynamic predictions

Individual dynamic predictions derived from joint models[7–10] consist in the individual predicted probability of event, $p_i(s, \mathscr{T}; \theta)$, between times s and $s + \mathscr{T}$ computed for a new subject given his biomarker history $Y_i^{(s)} = \{Y_i(t_{ij}), j = 1, \ldots, n_i, \text{such as}\, t_{ij} \leq s\}$ and his covariates history $X_i^{(s)}=\{X_{Li}(t_{ij}),Z_i(t_{ij}),j=1,\dots,n_i,\text{such as}\,t_{ij}\le s\}$ until time *s* as well as time-independent covariates X_{Si}. It is defined as:

$$p_i(s,\mathscr{T};\theta)=\mathbb{P}\left(T_i\le s+\mathscr{T}|T_i\ge s,Y_i^{(s)},X_i^{(s)},X_{Si};\theta\right) \qquad (3)$$

With a change in risk due to the initiation of a ST, different individual dynamic predictions can be distinguished from a time of prediction *s* in patients free of ST:

- The patient initiates ST at time *s*:

$$\begin{aligned}p_i(s,\mathscr{T};\theta)&=\mathbb{P}\left(T_i\le s+\mathscr{T}|T_i\ge s,\tau_i=s,Y_i^{(s)},X_i^{(s)},X_{Si};\theta\right)\\&=1-\frac{\int_{b_i}f_{Y^{(s)}}\left(Y_i^{(s)}|X_i^{(s)},b_i;\theta\right)S_i(s+\mathscr{T}|X_{Si},\tau_i=s,b_i;\theta)\,f_b(b_i;\theta)\,db_i}{\int_{b_i}f_{Y^{(s)}}\left(Y_i^{(s)}|X_i^{(s)},b_i;\theta\right)S_i\,(s|X_{Si},\tau_i=s,b_i;\theta)\,f_b(b_i;\theta)\,db_i}\end{aligned} \qquad (4)$$

- The patient does not initiate ST in *s*. In that case, many alternative scenarios can be considered. For example, ST could be initiated after a certain amount of time t_1 with $t_1<\mathscr{T}$ or $t_1>\mathscr{T}$ or ST could be initiated when the biomarker reaches a certain threshold *x*. For validating the dynamic prognostic tools, we focused on scenario "no initiation of ST in the window of prediction $\mathscr{T}$" (but we illustrate alternatives in section 5.4.3):

$$\begin{aligned}p_i(s,\mathscr{T};\theta)&=\mathbb{P}\left(T_i\le s+\mathscr{T}|T_i\ge s,\tau_i>\min\left(T_i,s+\mathscr{T}\right),Y_i^{(s)},X_i^{(s)},X_{Si};\theta\right)\\&=1-\frac{\int_{b_i}f_{Y^{(s)}}\left(Y_i^{(s)}|X_i^{(s)},b_i;\theta\right)S_i(s+\mathscr{T}|X_{Si},\tau_i>\min\left(T_i,s+\mathscr{T}\right),b_i;\theta)\,f_b(b_i;\theta)\,db_i}{\int_{b_i}f_{Y^{(s)}}\left(Y_i^{(s)}|X_i^{(s)},b_i;\theta\right)S_i\,(s|X_{Si},\tau_i>\min\left(T_i,s+\mathscr{T}\right),b_i;\theta)\,f_b(b_i;\theta)\,db_i}\end{aligned}$$

(5)

In (4) and (5), $f_{Y^{(s)}}$ and f_b are multivariate Gaussian density functions with respectively means $X_{Li}^{(s)}\beta+Z_i^{(s)}b_i$ and 0, and variance-covariance matrices $\sum_i^{(s)}$ and *B; S_i* is the survival function. $X_{Li}^{(s)}$ and $Z^{(s)}$ are design matrices with respectively $X_{Li}^{(s)}(t_{ij})^T$ and $Z^{(s)}(t_{ij})^T$ as row vectors with $t_{ij}\le s$, and $\sum_i^{(s)}$ is the submatrix of Σ_i with $t_{ij}\le s$.

A point estimate of these individual dynamic predictions can be obtained with $p_i(s,\mathscr{T};\hat{\theta})$. Alternatively, the posterior distribution of $p_i(s,\mathscr{T};\theta)$ can be also approximated by a Monte Carlo method[7]: a large set of $\theta^{(d)}$ ($d=1,\dots,D$) is generated from the asymptotic distribution of the estimates $\mathscr{N}\left(\hat{\theta},\widehat{\mathrm{V}(\hat{\theta})}\right)$ and used to compute $p_i(s,\mathscr{T};\theta^{(d)})$ for ($d=1,\dots,D$). The median value of $p_i(s,\mathscr{T};\theta^{(d)})$ provides the point estimate and the 2.5% and 97.5%

percentiles give a 95% confidence band. Instead of computing the probabilities using (4) and (5), $b_i^{(d)}$ can be sampled from its posterior distribution, and the probabilities computed given $b_i^{(d)}$8.

4 Evaluation of predictive accuracy

Validation of the prognostic tools is done in terms of predictive accuracy. Two measures adapted to the dynamic setting were considered: the prognostic cross-entropy and the Brier score. The measures and simple estimators of them are described. Then, the formula for the approximate cross-validation[16] is applied to provide estimators of the measures that can be used on the training data to correct overoptimism. Finally, integrated versions over the times of initiation of ST are proposed and confidence intervals are given.

In the following, N_s is the number of subjects at risk at time of prediction s. However, as predictive performances are evaluated differently in the absence of ST and after immediate initiation of ST, in the first case $N_s{=}\sum_{i=1\{T_i>s\&\tau_i>\min(T_i,s+\mathscr{T})\}}^{N}$ while in the second case $N_s{=}\sum_{i=1\{T_i>s\&\tau_i=s\}}^{N}$.

4.1 Measures of predictive accuracy

4.1.1 Expected Prognostic Observed Cross-Entropy—The expected prognostic observed cross-entropy (EPOCE) is a criterion that quantifies the prognostic information of a joint model from a time of prediction s[10,17]. It is formally defined as the expectation of the log of the conditional density $f_{T\,|\,Y^{(s)},T^*\geq s}$ of the time to event given the history of the biomarker $Y^{(s)}$ until the time of prediction s, E[−log ($f_{T|Y^{(s)},T^*\geq s}|T^* \geq s$)]. A simple estimator of EPOCE is given by the prognostic observed log-likelihood (POL) which is the log-likelihood to observe the event in the time window $[s, s+\mathscr{T}]$ conditional on the observed marker data until time s:

$$\mathrm{POL}(\hat{\theta}, s, \mathscr{T}){=}\frac{1}{N_s}\sum_{i=1}^{N_s} F_i\left(\hat{\theta}, S, \mathscr{T}\right) \qquad (6)$$

where $F_i(\hat{\theta}, s, \mathscr{T})$ is minus the observed individual contribution to the conditional log-likelihood defined below in (7) given that the subject is still at risk at time s and given his covariates history until s. For $i = 1, \ldots, N_s$, F_i is defined as :

$$F_i\left(\hat{\theta}, s, \mathscr{T}\right){=}-_{\{T_1\geq s\}}\log\left(\frac{\int_{b_i} f_Y\left(Y_i^{(s)}|X_i^{(s)}, b_i;\theta\right)\lambda\left(\tilde{T}_i|X_{Si}, b_i;\theta\right)^{\tilde{E}_i} S_i\left(\tilde{T}_i|X_{Si}, b_i;\theta\right) f_b\left(b_i;\theta\right) db_i}{\int_{b_i} f_Y\left(Y_i^{(s)}|X_i^{(s)}, b_i;\theta\right) S_i\left(s|X_{Si}, b_i;\theta\right) f_b(b_i;\theta)\, db_i}\right)$$

(7)

where $\tilde{T}_i = \min(T_i, s + \mathscr{T})$ and $\tilde{E}_i = \mathbb{1}_{\{T_i \le s+\mathscr{T}\}}$ which means that subjects are artificially censored at $s + \mathscr{T}$.

4.1.2 Brier Score and Integrated Brier Score—The Brier Score (BS) developed in survival models[20,21] was extended to joint models[7,10,15,22]. It consists in $E[(\eta_i(s + \mathscr{T}) - \hat{S}(s + \mathscr{T} \mid s; \hat{\theta}))^2]$, the expectation of the squared difference between the observed survival status η and the predicted survival at a specific time: $\hat{S}(s + \mathscr{T} \mid s; \hat{\theta}) = 1 - p_i(s, \mathscr{T}; \hat{\theta})$ with $p_i(s, \mathscr{T}; \hat{\theta})$ defined in (5) in the absence of HT and in (4) after immediate initiation of HT. The simple estimator of BS at time $s + \mathscr{T}$ is:

$$\mathrm{BS}(\hat{\theta}, s, \mathscr{T}) = \frac{1}{N_s} \sum_{i=1}^{N_s} w_i \left(\eta_i(s+\mathscr{T}) - \hat{S}\left(s+\mathscr{T} \mid s; \hat{\theta}\right)\right)^2 \qquad (8)$$

where w_i is a weight that compensates for the loss of information due to censoring. The weights are defined according to the inverse probability of censoring[10,23] as

$w_i = \frac{{}_{\{T_i > s+\mathscr{T}\}}}{\hat{G}\,(s+\mathscr{T})/\hat{G}(s)} + \frac{E_i{}_{\{T_i \le s+\mathscr{T}\}}}{\hat{G}\,(T_i)/\hat{G}(s)}$ with $\hat{G}$ the Kaplan-Meier estimate of the survival function for the censoring.

An average prediction accuracy is derived from BS in a window $[s, s + \mathscr{T}]$ by integrating the quantity over the horizon times[10,22]. Again, to account for the loss of events due to censoring, a weighted mean can be used[22] and estimated by :

$$\mathrm{IBS}(\hat{\theta}, s, \mathscr{T}) = \left[\sum_{k}^{n_s^{\mathscr{T}}} d_k^{(s)} \left(\hat{G}(s)/\hat{G}(t_k)\right) \mathrm{BS}(\hat{\theta}, s, t_k - s)\right] / \left[\sum_{k}^{n_s^{\mathscr{T}}} d_k^{(s)} \left(\hat{G}(s)/\hat{G}(t_k)\right)\right] \qquad (9)$$

where $t_k (k=1, \cdots, n_s^{\mathscr{T}})$ are the distinct $n_s^{\mathscr{T}}$ times of events in the window $[s, s + \mathscr{T}]$ and $d_k^{(s)}$ is the number of events at each time t_k among subjects at risk at time s.

4.2 Approximated cross-validated estimators

To provide a valid assessment of predictive accuracy, these measures should be computed for independent data. On the model estimation data, a cross-validation technique is required to correct for overoptimism. With complex models, cross-validation is numerically too expensive to be used, so we applied the approximate leave-one-out cross-validation formula for regular problems proposed by Commenges et al.[16]. It is defined as:

$$\mathrm{CV}\,\mathscr{M}_{\mathrm{a}}(\hat{\theta}, s, \mathscr{T}) = \mathscr{M}(\hat{\theta}, s, \mathscr{T}) + N\,\mathrm{Trace}\,(H^{-1}K_{s,\mathscr{T}}) \qquad (10)$$

where $\mathfrak{M}(\hat{\theta}, s, \mathscr{T})$ is the simple estimator (POL, BS or IBS) and the second term is the correction term added by the approximated leave-one-out cross-validation. This is a penalty for the statistical complexity of the model that captures the overfit and the corresponding overoptimism of predictions. H is the Hessian matrix of the joint log-likelihood, and $K_{s,\mathscr{T}}=\frac{1}{N_s(N-1)}\sum_{i=1}^{N}{}_{\{T_1\geq s\}}\hat{v}_i(s,\mathscr{T})\hat{d}_i^T$ is the product of the gradients $\hat{v}_i(s,\mathscr{T})$ and $\hat{d}_i$ of the individual contributions respectively to the simple estimator $\mathfrak{M}(\hat{\theta}, s, \mathscr{T})$ and the maximized joint log-likelihood. The gradients are computed using finite differences. Applied to POL, BS and IBS respectively, the approximate cross-validation estimators are respectively called CVPOL_a, CVBS_a and CVIBS_a; H and $\hat{d}_i$ are the same in the three approximate cross-validation measures and $\hat{v}_i(s,\mathscr{T})=\frac{\partial F_i(\theta,s,\mathscr{T})}{\partial\theta}|_{\hat{\theta}}$ for CVPOL_a, $\hat{v}_i(s,\mathscr{T})=-2\,w_i\frac{\partial S(s+\mathscr{T}|s;\theta)}{\partial\theta}|_{\hat{\theta}}\left(\eta_i(s+\mathscr{T})-\hat{S}\left(s+\mathscr{T}|s;\hat{\theta}\right)\right)$ for CVBS_a and $\hat{v}_i(s,\mathscr{T})=\sum_{k}^{n_s^{\mathscr{T}}}d_k^{(s)}\frac{\left(\hat{G}(s)/\hat{G}(t_k)\right)}{\sum_{k}^{n_s^{\mathscr{T}}}d_k^{(s)}\left(\hat{G}(s)/\hat{G}(t_k)\right)}\left[-2\,w_i\frac{\partial S\,(t_k|s;\theta)}{\partial\theta}|_{\hat{\theta}}\left(\eta_i(t_k)-\hat{S}\left(t_k|s;\hat{\theta}\right)\right)\right]$ for CVIBS_a.

4.3 Averaged predictive accuracy

When predicted probabilities are computed for the case of immediate initiation of a ST, the predictive accuracy is evaluated from the time of ST initiation. In practice, this time is different for each subject. So instead of an evaluation at fixed times of prediction s, we focused on the average predictive accuracy over the times of ST initiation computed using the inverse probability of censoring weighting technique[22]:

$$\overline{\mathscr{M}(\hat{\theta},\mathscr{T})}=\left[\sum_{i=1}^{n^{ST}}\frac{d_i^{ST}}{\hat{G}_\tau(\tau_i)}\mathscr{M}(\hat{\theta},\tau_i,\mathscr{T})\right]/\left[\sum_{i=1}^{n^{ST}}\frac{d_i^{ST}}{\hat{G}_\tau(\tau_i)}\right]\qquad(11)$$

where $\mathfrak{M}$ can be BS or POL (simple or cross-validated) computed in (6), (8) and (10); n^{ST} is the number of distinct times of ST initiations, d_i^{ST} is the number of ST initiations at time τ_i and $\hat{G}_\tau$ is the Kaplan-Meier estimate of the survival function of censoring related to times of ST initiation τ_i.

4.4 Confidence interval

Predictive accuracy measures between two models can be compared by computing the difference of their approximated cross-validation estimators with a 95% confidence interval (CI)[16]. It is computed from the asymptotic distribution of the predictive accuracy difference and its empirical variance. Let $\hat{\theta}_A$ and $\hat{\theta}_B$ be the vectors of parameter estimates for the two models A and B. Let $\Delta(A(\hat{\theta}_A), B(\hat{\theta}_B))$ be the difference of predictive accuracy between the two models and $\mathscr{D}(A(\hat{\theta}_A), B(\hat{\theta}_B))$ its estimator. Let m be the number of subjects on which the predictive accuracy is computed. It is shown[16] that the difference between $\mathscr{D}(A(\hat{\theta}_A), B(\hat{\theta}_B))$ and $\Delta(A(\hat{\theta}_A), B(\hat{\theta}_B))$ is asymptotically normal:

$$m^{1/2}\left[\mathscr{D}\left(A(\hat{\theta}_A),B(\hat{\theta}_B)\right)-\Delta\left(A(\hat{\theta}_A),B(\hat{\theta}_B)\right)\right]\to\mathscr{N}\left(0,w_*^2\right) \qquad (12)$$

where w_*^2 can be estimated by the empirical variance $\hat{w}^2$ of the difference of the simple estimators. With z_u the u^{th} quantile of a standard normal variable, the confidence interval is then $[\mathscr{D}\,(A(\hat{\theta}_A), B(\hat{\theta}_B)) - z_{\alpha/2}m^{-1/2}\ \hat{w}; \mathscr{D}\,(A(\hat{\theta}_A), B(\hat{\theta}_B)) + z_{\alpha/2}m^{-1/2}\ \hat{w}]$.

In absence of ST, the predictive accuracy is evaluated at different times s with $m = N_s$ and $\mathscr{D}$ is the approximated cross-validation estimate of POL, BS or IBS at time s. After initiation of HT, predictive accuracy is evaluated once with $m=\sum_{i=1}^{n^{ST}} d_i^{ST}$ and $\mathscr{D}$ is the approximated cross-validation estimate of the average measures defined in (11).

5 Application to the prediction of prostate cancer recurrence

5.1 Datasets

Data used in this application consist of 2386 men treated for localized prostate cancer by external beam radiation therapy (EBRT) in three different studies: 503 patients come from the cohort of the University of Michigan (UM) with a period of recruitment from 1988 to 2004; 1268 patients come from the cohort of Beaumont Hospital, in Michigan (BM), recruited between 1987 and 2003; 615 patients come from the multicenter clinical trial RTOG9406 recruited from 1994 to 2001. Among them, 261 (10.9%) received a ST that is hormonal therapy (HT) during their follow-up. The definition of clinical recurrence was any kind of recurrence (local, regional, distant) or death from prostate cancer and only the first clinical recurrence was considered. 312 (13.1%) patients had a clinical recurrence among which 53 received HT. The four baseline prognostic factors considered in this application are the pre-radiation therapy level of PSA, the T-stage which indicates how large the tumor is and how far it has spread (in three categories: 1;2;3–4), the Gleason score which quantifies the aggressiveness of the cancer (in three categories: 2–6; 7; 8–10) and the corrected total dose of radiation therapy[24]; full description of these covariates has been previously given[7,25].

Figure 1 shows 8 random individual observed trajectories of PSA after the end of EBRT for patients who recurred or were censored and patients who did or did not receive HT. After EBRT a drop in PSA is observed in the first year. Then a subsequent rise of PSA indicates a higher risk of recurrence and in some cases HT is initiated to reduce the risk and postpone the recurrence. Initiation of HT induces an immediate change in the PSA dynamics. Since the objective was to predict the risk of recurrence based on PSA dynamics prior to HT we chose to censor the post-HT PSA data. We call the PSA dynamics as if the person were not treated with HT the base PSA dynamics. A median of 9 (Inter-quartile Range =5,12) PSA repeated measures per subject were analyzed.

5.2 Specification of the joint models

PSA repeated measures were analyzed in the logarithm scale. As previously proposed[25] we used the two phases trajectory of PSA defined as:

$$\begin{aligned} Y_i(t) &= \log\left(\mathrm{PSA}_i(t)+0.1\right) \\ &= (X_{0i}^T\beta_0+b_{0i})+(X_{1i}^T\beta_1+b_{1i})\,f(t)+(X_{2i}^T\beta_2+b_{2i})\,t+\varepsilon_i(t), \forall t \in \mathbb{R}^+ \end{aligned} \quad (13)$$

where $f(t) = ((1 + t)^{-1.5} - 1)$ and t captured respectively the short-term decline and the long-term trend of PSA[25]; X_{0i} included 1, the pre-EBRT PSA and the cohort indicators; X_{1i} included X_{0i} plus 2 binary indicators for T-stage (2 *vs* 1, and 3–4 *vs* 1); and X_{2i} included X_{1i} plus 2 binary indicators for the Gleason score (7 *vs* 2–6, and 8–10 *vs* 2–6).

In the survival model, the baseline hazard function was approximated by splines and the four baseline prognostic factors were included in X_{Si}. In addition, different specifications of $W_i(b_i, \tau_i, t)$ were explored. $W_i(b_i, \tau_i, t)$ includes two components: the multivariate function $h(b_i, t)$ of the random effects derived from (1) that models the dependency between the PSA dynamics and the time-to-clinical-recurrence, and information about initiation of HT. In the following, we propose five specifications of $W_i(b_i, \tau_i, t)$ that differ in the way the initiation of HT enters into the model, and different variants that correspond to different functions of the PSA dynamics $h(b_i, t)$. The five specifications are:

1. $W_i(b_i, \tau_i, t)^T \phi = \phi_1 \mathbb{1}_{\{t \geq \tau_i\}}$. This is the standard survival model assuming there is no association between the PSA dynamics and the risk of event, but considering a change of risk of recurrence after initiation of HT.

2. $W_i(b_i, \tau_i, t)^T \phi = \phi_1 \mathbb{1}_{\{t \geq \tau_i\}} + h(b_i, t)^T \phi_2$. This is the standard joint model for describing PSA dynamics and risk of recurrence[13,14,25]**?**. This model assumes that the characteristics of the PSA dynamics have the same role before and after HT. After HT, these characteristics are extrapolated as if the patient did not initiate HT (see Figure 1) so that the change in risk after HT captured by parameter ϕ_1 summarizes the effect of HT adjusted for base PSA dynamics.

3. $W_i(b_i, \tau_i, t)^T \phi = \phi_1 \mathbb{1}_{\{t \geq \tau_i\}} + h(b_i, t)^T \phi_2 \mathbb{1}_{\{t<\tau_i\}} + h(b_i, t)^T \phi_3 \mathbb{1}_{\{t \geq \tau_i\}}$. This model considers an interaction between the initiation of HT and the PSA dynamics by including a different effect of the base PSA dynamics before and after HT. The assumption is that the effect of HT on the risk of recurrence depends on the shape of PSA trajectory preceding the initiation of HT.

4. $W_i(b_i, \tau_i, t)^T \phi = \phi_1 \mathbb{1}_{\{t \geq \tau_i\}} + h(b_i, t)^T \phi_2 \mathbb{1}_{\{t<\tau_i\}} + h(b_i, \tau_i)^T \phi_3 \mathbb{1}_{\{t \geq \tau_i\}}$. This model is a variant of model 3. As the extrapolated PSA dynamics after HT initiation no longer represents the actual PSA dynamics of the patient, the current extrapolated PSA values after HT are replaced by the PSA value at the time τ_i of HT initiation at which PSA measurements were censored.

5. $W_i(b_i, \tau_i, t)^T \phi = \mathbb{1}_{\{t \geq \tau_i\}} (\phi_1 + \alpha g(t - \tau_i)) + h(b_i, \mathrm{t})^T \phi_2 \mathbb{1}_{\{t<\tau_i\}} + h(b_i, \tau_i)^T \phi_3 \mathbb{1}_{\{t \geq \tau_i\}}$. This is a more flexible version of model 4, in which the baseline risk of recurrence after HT may change with time according to a function $g(t - \tau_i)$. We considered both $g(t - \tau_i) = \log(t - \tau_i)$ and $g(t - \tau_i) = t - \tau_i$.

In specifications 2 to 5, up to three variants of $h(b_i, t)$ were considered[9].

a. $h_{\mathrm{a}}(b_i,t)=(Y_i^*(t), \partial Y_i^*(t)/\partial t)^T$: the level and slope of PSA at time t are independent predictors of the time to clinical recurrence.

b. $h_{\mathrm{b}}(b_i,t)=(\Gamma\ (Y_i^*(t), \partial Y_i^*(t)/\partial t)^T$: instead of the crude PSA level, a transformed PSA level $\Gamma(Y_i^*(t))$ and the slope at time t are independent predictors of the time to clinical recurrence[25], with $\Gamma\ (Y_i^*(t))=\mathrm{logit}^{-1}\ ((Y_i^*(t)-0.71)/0.44)$.

c. $h_{\mathrm{c}}(b_i, t) = (b_{0i}, b_{1i}, b_{2i})^T$: the individual deviations from the mean PSA dynamics, that are the random effects, are independent predictors of the time to clinical recurrence. This variant was only considered with specification 2.

5.3 Estimation and goodness-of-fit of the joint models

Estimation of the joint models is summarized in Table 1 and parameter estimates that measure the effect of HT and the association between the PSA dynamics and the risk of clinical recurrence are shown in Table 2. Whatever the assumed nature of the dependence between the PSA dynamics and the time-to-clinical recurrence, the joint models provided a substantial gain in fit compared to model M1 which assumes independence between the two processes (minimum gain of 435.9 points of AIC for the joint model $\mathrm{M2_a}$).

Among the different joint models, considering a logistic transformation of the current level of PSA (models b) rather than the crude current level (models a) improved the fit. In previous work, a residual analysis[9] had noted a departure of the log-linearity assumption when considering the crude PSA level in the survival model, and the correction of this departure when considering the logistic transformation. This transformation that makes the effect of the PSA level increase in the range 0 to 4ng/ml and become maximal around 4ng/ml is particularly of importance in M2 where after initiation of HT, very high levels of PSA can be extrapolated from the longitudinal model (as illustrated in Figure 1), which may artificially increase the subsequent risk of recurrence.

Assuming that the effects of the crude current PSA level and the current slope differed before and after HT in $\mathrm{M3_a}$ greatly improved the fit (121.1 points of AIC) compared to $\mathrm{M2_a}$. In contrast, when considering the logistic transformation instead of the crude PSA value, assuming different effects of PSA dynamics before and after HT in $\mathrm{M3_b}$ provided only a small gain in fit (5.4 points of the AIC) compared to $\mathrm{M2_b}$. Indeed, after HT, most of the extrapolated PSA levels are very high so that they drive the estimate to a smaller overall impact of the current PSA level. When separating the effects pre and post-HT in $\mathrm{M3_a}$, the pre-HT crude effect (defined from relatively standard PSA levels) was four times bigger than the overall crude effect estimated in $M2_a$, and the effect post-HT was no longer significant. In contrast, when assuming a transformation of the PSA level, the overall effect in $M2_b$ was similar to the effect pre-HT in $M3_b$. We noted the same things for models M4 and M5 compared to the models M2.

We observed from Table 2 that only the slope of PSA was significantly predictive of the risk of recurrence after HT with relatively stable estimates ranging from 0.94 to 1.29 in models M3 through M5. Neither the extrapolated current level in models M3 (with $p = 0.38$ for $M3_a$ and $p = 0.42$ for $M3_b$) or the level reached at the time of initiation of HT in models M4 and M5 (with $p = 0.31$ and $p = 0.09$ for $M4_a$ and $M4_b$; $p = 0.35$ and $p = 0.11$ for $M5.1_a$ and $M5.1_b$; $p = 0.40$ and $p = 0.12$ for $M5.2_a$ and $M5.2_b$) were associated with the risk of recurrence post-HT after adjustment for the slope of PSA.

Assuming a dependence through the random effects ($M2_c$) rather than the PSA level or slope provided a fit in between models $M2_a$ and $M2_b$ even though the dependence was summarized by three parameters (all significant) instead of two. Finally, assuming a non constant change in the baseline risk function after HT (in models M5.1 and M5.2) did not improve substantially the fit of the models. In summary, the model $M4_b$ assuming an association with the transformed PSA level, separating effects of PSA prior and after HT, and focusing on characteristics at the time of HT after the initiation, provided the best fit of the data.

Regarding the specific effect of initiation of HT, the interpretation differs between models. Model $M2_a$ aims at capturing the actual protective effect of HT after adjustment for the base PSA trajectory ($\phi_1 = -1.89$, $p < 0.0001$). But as explained before, this model may suffer from the very high extrapolated PSA values after HT so that $M2_b$ may be more appropriate to accurately evaluate the effect of HT with an estimate $\phi_1 = -1.39$ ($p < 0.0001$). This corresponds to a relative reduction by 4 in the risk of recurrence when initiating HT and adjusted for the PSA characteristics.

In models M3 to M5, no single parameter represents the effect of HT, and particularly parameter ϕ_1 no longer represents the effect of HT and should not be interpreted as such. Indeed, distinct effects of PSA dynamics before and after HT are modeled so that (except for standard prognostic factors) the model is stratified on the initiation of HT and parameter ϕ_1 associated with the initiation of HT only represents a change in the baseline risk at HT initiation. This baseline risk appears to be substantially increased (e.g. ϕ_1=1.33 in $M3_a$ and ϕ_1=2.74 in $M3_b$) but this has to be put in balance with the different effects of PSA level and slope before and after HT, PSA level being highly significant before initiation of HT and no longer significant after initiation of HT.

5.4 Predictive accuracy of the joint models

For the comparison in terms of predictive accuracy, we focused on 6 joint models: the model assuming independence between the PSA dynamics and the risk of clinical recurrence (M1), the standard joint models in PSA studies ($M2_a$ and $M2_b$), the joint models in which the extrapolated PSA current level and slope after HT are replaced by the PSA level and slope at initiation of HT ($M4_a$ and $M4_b$) and the model with a dependence directly on the random effects ($M2_c$). The predictive accuracy was evaluated on the estimation data using the approximated cross-validated estimates. We assessed the ability of the joint models to predict the risk of clinical recurrence in a window of 3 years ($\mathscr{T}=3$) which was a clinically reasonable window. For all the measures, the lower the better.

5.4.1 Average predictive accuracy after immediate initiation of ST—

Among men who initiated a HT during the follow-up, the average POL and BS defined in section 4.3 are shown in Figures 2(a) and 2 (c). The differences between pairs of models and their 95% CI were also computed and shown in Figures 2(b) and 2 (d).

First, BS and POL measures mostly agreed even if a few differences were observed between the three or four most predictive models.

Whatever the nature of the dependency between the PSA dynamics and the risk of recurrence, the predictive accuracies of joint models were significantly better than those of model M1 which assumes independence between PSA dynamics and risk of recurrence.

In accordance with the goodness-of-fit measures, considering different effects prior to HT and after HT improved a lot more the predictive accuracy when the crude PSA level was considered (model $M4_a$ compared to $M2_a$) than when considering a transformed PSA level (model $M4_b$ compared to $M2_b$). The latter comparison is the only one with discordance between POL and BS results: BS concluded that predictive ability of $M4_b$ was significantly better than the one of $M2_b$ while no difference was found with POL.

Among models M2, considering a transformation of the PSA current level rather than the crude PSA level improved significantly the predictive accuracy ($M2_b$ compared to $M2_a$) while among models M4, this did not induce any significant difference for either measure between $M4_b$ and $M4_a$. Finally, assuming a dependence on the random effects ($M2_c$) rather than on the PSA transformed level and slope ($M2_b$) did not alter much the ability to predict the risk of recurrence.

In summary, BS tended to favor model $M4_b$ while POL tended to slightly favor model $M2_b$. As the difference in POL between $M4_b$ and $M2_b$ was not significant, we chose $M4_b$ as the final best model to predict clinical recurrence after immediate initiation of HT.

5.4.2 Predictive accuracy in absence of ST—

To evaluate the predictive accuracy of the joint models in absence of HT initiation in the next 3 years, predictive accuracy measures were computed at different times of prediction s (from 1 to 6 years after end of EBRT) among men who did not initiate any HT in the window $[s, s + 3]$ years. These curves are displayed in Figures 3(a), 3 (c) and 3 (e) for the approximated cross-validation estimates of POL, BS and IBS. The corresponding differences between pairs of models and their 95% confidence bands were computed at the same times of prediction and shown in Figures 3(b), 3 (d) and 3 (f).

Whatever the predictive accuracy measure and the nature of the dependency between the PSA dynamics and the risk of recurrence, the joint models provided globally a significantly better predictive accuracy compared to model M1 which assumes independence between the two processes (with the surprising exception for $M4_b$ in the first years according to the BS and IBS measures).

Whatever the predictive accuracy measure, models M4 and M2 had similar predictive performances (differences not shown) in the absence of HT. Indeed, the overall estimates in M2 are mostly driven by the high proportion of subjects who did not initiate HT.

Whatever the measure, considering in models M2 a logistic transformation of the PSA level ($M2_b$) instead of the crude PSA level ($M2_a$) did not really improve the predictive performances in the short-term for $s \in [1, 4]$. This was expected as the transformation of PSA level is supposed to mainly correct very high extrapolated PSA values not observed among subjects who will not initiate any HT. In the long-term ($s \geq 4$ years) joint models considering the crude PSA level ($M2_a$) provided even a significantly better predictive accuracy. This contrasted with conclusions in terms of goodness-of-fit or after HT initiation where specification b was systematically better.

When considering a dependence on the random effects ($M2_c$) rather than on the PSA crude level and slope ($M2_a$), conclusions based on BS, IBS and POL measures differed: $M2_c$ was found largely better than $M2_a$ at times of prediction greater than 1.5years with POL and was also found better with BS and IBS but only for shorter times of prediction. At longer times of prediction, model $M2_a$ was even slightly better with BS and IBS.

Although results differed substantially depending on the type of measure, the joint model with a dependence directly through the random effects ($M2_c$) provided a nice alternative to the more standard ($M2_a$) joint model among men who did not undergo any HT. These two models that are the most predictive in absence of HT are also the ones in which the effects of the PSA long-term slope are the highest. This was previously observed in **(author?)** [9], where among patients who did not initiate any HT, joint models having the largest effects of the slope of log PSA were also the ones having the best predictive ability suggesting that after a few years, the slope of log PSA would be the major predictor of the risk of recurrence in the absence of HT.

In summary, while the best model to predict the risk of clinical recurrence assuming an immediate initiation of HT was $M4_b$, the best model we chose to predict the risk of recurrence assuming the patient will not initiate any HT within 3 years was $M2_c$.

5.4.3 Example of differential dynamic prediction of prostate cancer recurrence

—We provide here an illustrative example of how these differential dynamic predictions can be used in practice. We consider a subject who had a T-stage of 2, a Gleason of 6, an initial PSA of 12.7 ng/ml a corrected dose of radiation of 65.7 Gy and who recurred at 2.7 years after the end of EBRT. After each PSA measurement, we computed his individual predicted probability of clinical recurrence in the next 3 years under the two extreme and validated assumptions: whether he initiates HT immediately (probabilities computed according to model $M4_b$) and whether he does not initiate any HT in the next 3 years (probabilities computed according to model $M2_c$), as well as two intermediate scenarios in which the patient initiates HT after 1 and 2 years respectively (probabilities computed according to model M4$_b$). Indeed, although for validation purposes, we chose to focus on the two first scenarios, in practice any clinically relevant scenario could be investigated, as for example a delayed initiation of the treatment.

Figure 4 provides for 4 times of prediction and according to the observed history of PSA (left side of the figure), the individual predictions of clinical recurrence in the next 3 years computed according to each of the four scenarios (right side of the figure).

This example illustrates that initiating HT early would have reduced largely the probability of having a recurrence for this patient. For example, at the 1.6-year visit, the man has a probability of having a clinical recurrence in the next 3 years of 25% which would reduce to 5% if he initiated immediately the hormonal therapy. Moreover, by reporting the predicted probabilities according to intermediate scenarios, we observe that the probability of recurrence for this patient increases with the delayed initiation of HT up to the largest predicted probability in case of no initiation in the window of time.

6 Discussion

Using the joint model methodology, we provided individualized dynamic prognostic tools depending on hypothetical clinical decisions, namely the initiation of new treatments. We focused mainly on two scenarios: the prediction of the event assuming no change in the treatment, and the prediction of the event assuming the immediate initiation of a new (second) treatment. Indeed, deciding whether to initiate a second treatment or not has become central in the individual monitoring of chronic diseases such as cancers. While the joint model development requires some subjects with at least three repeated measures, the derived individual predictions can be computed as soon as one measure is available even if in practice more information may be necessary to provide precise predictions[13].

Although promising for clinical practice, such differential dynamic predictive tools were never developed or validated in the literature. A website calculator associated with a publication[13] does include dynamic predictions under two scenarios, but it was not described and validated in that publication. Until now, dynamic predictive tools in the literature only predicted the risk of event based on a biomarker value[8] or a biomarker trajectory[7,13] by assuming that there was no change in treatment or patient characteristics that might impact the subsequent risk of event. Indeed, developing and validating dynamic prognostic tools that can be conditioned on scenarios of initiation is challenging.

First, it requires a very precise specification of the dependency between the biomarker dynamics, the treatment initiation and the risk of the event. This was accomplished here by using as series of sophisticated joint shared random-effect models. However other approaches like joint latent class models or landmark analyses[7] could also be considered.

Second, the predictive performances have to be validated specifically for each scenario. Indeed, it may be unrealistic to expect that the same model provides the best predictions in different situations. This required the development of integrated measures in the “immediate initiation of second treatment” scenario to focus on the predictive performances following the initiation. In the application, even if all the joint models had a relatively good predictive accuracy in both situations, we did find that the best predictive tool in each scenario did not come from the same models. This illustrates that prognostic tools should be strictly validated for what they are aimed to quantify in practice.

Third, even in the absence of HT, the prognostic tool validation is still not straightforward. We chose to focus here on patients who did not initiate any HT in the window of prediction. However these patients may not be a representative sample of the patients free of HT at the time of prediction. **(author?)** [13] proposed instead to validate the prognostic tools by focusing on the sample of subjects free of HT at the time of prediction and by considering either all the HT initiations during the window of prediction as recurrences or as censoring. As shown in Web supplementary materials, the results concerning the relative performance of the models did not change when using this technique.

Fourth, due to the models complexity and the differential validation procedure, the use of the whole sample was preferred to a data splitting approach so that estimation and validation of the predictive tools were done on the same data. This motivated the development of a new estimator of the Brier Score by approximated leave-one-out crossvalidation[16] which is valid and easy to compute on the estimation data.

Predictive performances were assessed using two different measures that do not tackle the predictive accuracy in the same way. The Brier Score directly measures the Mean Square Error between the event process and the prediction of the model while the EPOCE assesses the prognostic value of the joint models by measuring the distance between the conditional density of the time to event assumed in the model and the true one. This may explain why we found differences between the conclusions given by the two measures in the application. We still chose to select the best models as a balance between the results given by these two measures. Moreover, the models providing the best goodness-of-fit did not necessarily have the best predictive ability. This illustrates the difference between these two types of assessment. While the goodness-of-fit measures use all the information, the predictive accuracy measures focus only on a part of the sample and use only the history of the biomarker up to the time of prediction. This is why when interested in dynamic predictions, the predictive ability of the models should be assessed[9].

AUC derived measures[26] were not considered here for assessing the predictive ability of the joint models. Indeed, first they focus on discrimination while our focus was really on predictiveness since we wanted to quantify individual probabilities of recurrence. Second, their use in dynamic settings has been rather limited[8,27]. Third, providing an approximated cross-validation estimate was not straightforward.

Finally, in prostate cancer, the initiation of a second treatment, namely hormonal therapy, has raised many questions about how to take it into account in the model for the risk of clinical recurrence or how to evaluate its causal effect in the presence of indication bias[?]. In the present paper, our focus was only on the dynamic individual predictions. As such, we chose to compare descriptive joint models that treated HT intuitively as a time-dependent covariate, possibly in interaction with other characteristics. However, using the same strategy, more causal or mechanistic models could also be investigated. Alternatively, time to

HT initiation could be treated as a censored time-to-event along with other clinical recurrences by defining a multistate model or a multivariate survival model jointly with the biomarker longitudinal model.

Acknowledgments

Funding: This work was supported by the French National Institute of Cancer INCa [grant PREDYC number 2010-059] and by the US National Cancer Institute [grants CA110518, U10-CA21661, U10-CA37422, and U10-CA180822].

References

1. Faucett CL, Thomas DC. Simultaneously modelling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. Statistics in medicine. Aug; 1996 15(15):1663–85. [PubMed: 8858789]
2. Wulfsohn MS, Tsiatis AA. A joint model of survival and longitudinal data measured with error. Biometrics. Mar.1997 53:330–339. [PubMed: 9147598]
3. Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event time data. Biostatistics (Oxford, England). Dec; 2000 1(4):465–80.
4. Lin H, Turnbull BW, McCulloch CE, Slate EH. Latent class models for joint analysis of longitudinal biomarker and event process data : Application to longitudinal prostate-specific antigen readings and prostate cancer. Journal of the American Statistical Association. Mar; 2002 97(457):53–65.
5. Tsiatis AA, Davidian M. Joint modeling of longitudinal and time-to-event data : An overview. Statistica Sinica. 2004; 14:809–834.
6. Rizopoulos, D. Joint models for longitudinal and time-to-event data: With applications in R. 2012.
7. Proust-Lima C, Taylor JMG. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach. Biostatistics (Oxford, England). Jul; 2009 10(3):535–49.
8. Rizopoulos D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. Biometrics. Sep; 2011 67(3):819–29. [PubMed: 21306352]
9. Sène M, Bellera CA, Proust-Lima C. Shared random-effect models for the joint analysis of longitudinal and time-to-event data: application to the prediction of prostate cancer recurrence. Journal de la Société Française de Statistique. In press.
10. Proust-Lima C, Sène M, Taylor JMG, Jacqmin-Gadda H. Joint latent class models for longitudinal and time-to-event data: A review. Statistical Methods in Medical Research. Apr.2012 doi: 10.1177/0962280212445839
11. Welsh, SJ.; Powis, G. Personalized cancer medicine. Springer-Verlag; Berlin Heidelberg: 2009.
12. Yu M, Taylor JMG, Sandler HM. Individual Prediction in Prostate Cancer Studies Using a Joint Longitudinal Survival-Cure Model. Journal of the American Statistical Association. Mar; 2008 103(481):178–187.
13. Taylor JMG, Park Y, Ankerst DP, Proust-Lima C, Williams S, Kestin L, Bae K, Pickles T, Sandler H. Real-time individual predictions of prostate cancer recurrence using joint models. Biometrics. 2013; 69(1):206–213. [PubMed: 23379600]
14. Kennedy EH, Taylor JMG, Schaubel DE, Williams S. The effect of salvage therapy on survival in a longitudinal study with treatment by indication. Statistics in medicine. Nov; 2010 29(25):2569–80. [PubMed: 20809480]
15. Schoop R, Schumacher M, Graf E. Measures of prediction error for survival data with longitudinal covariates. Biometrical journal. Mar; 2011 53(2):275–93. [PubMed: 21308724]
16. Commenges D, Proust-Lima C, Samieri C, Liquet B. A universal approximate cross-validation criterion and its asymptotic distribution. arXiv:1206.1753 [math.ST] Submitted;
17. Commenges D, Liquet B, Proust-Lima C. Choice of prognostic estimators in joint models by estimating differences of expected conditional kullback-leibler risks. Biometrics. Jun; 2012 68(2): 380–7. [PubMed: 22578147]

18. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics. 1982; 38:963–974. [PubMed: 7168798]
19. Rizopoulos DJM. An R package for the joint modelling of longitudinal and time-to-event data. Journal of Statistical Software. 2010; 35(9):1–33. [PubMed: 21603108]
20. Gerds TA, Schumacher M. Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times. Biometrical Journal. Dec; 2006 48(6):1029–1040. [PubMed: 17240660]
21. Gerds TA, Schumacher M. Efron-type measures of prediction error for survival analysis. Biometrics. Dec; 2007 63(4):1283–7. [PubMed: 17651459]
22. Henderson R, Diggle P, Dobson A. Identification and efficacy of longitudinal markers for survival. Biostatistics (Oxford, England). Mar; 2002 3(1):33–50.
23. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. Statistics in medicine. 1999; 18(17–18):2529–45. [PubMed: 10474158]
24. Proust-Lima C, Taylor JMG, Sécher S, Sandler H, Kestin L, Pickles T, Bae K, Allison R, Williams S. Confirmation of a low α/β ratio for prostate cancer treated by external beam radiation therapy alone using a post-treatment repeated-measures model for psa dynamics. International journal of radiation oncology, biology, physics. Jan; 2011 79(1):195–201.
25. Proust-Lima C, Taylor JMG, Scott W, Ankerst D, Liu N, Kestin L, KB, Howard S. Determinants of change in prostate-specific antigen over time and its association with recurrence after external beam radiation therapy for prostate cancer in five large cohorts. International Journal of Radiation Oncology Biology Physics. Aug; 2008 72(3):782–791.
26. Heagerty PJ, Zheng Y. Survival model predictive accuracy and roc curves. Biometrics. 2005; 61:92–105. [PubMed: 15737082]
27. Zheng Y, Heagerty PJ. Prospective accuracy for longitudinal markers. Biometrics. 2007; 63(2): 332–341. [PubMed: 17688486]

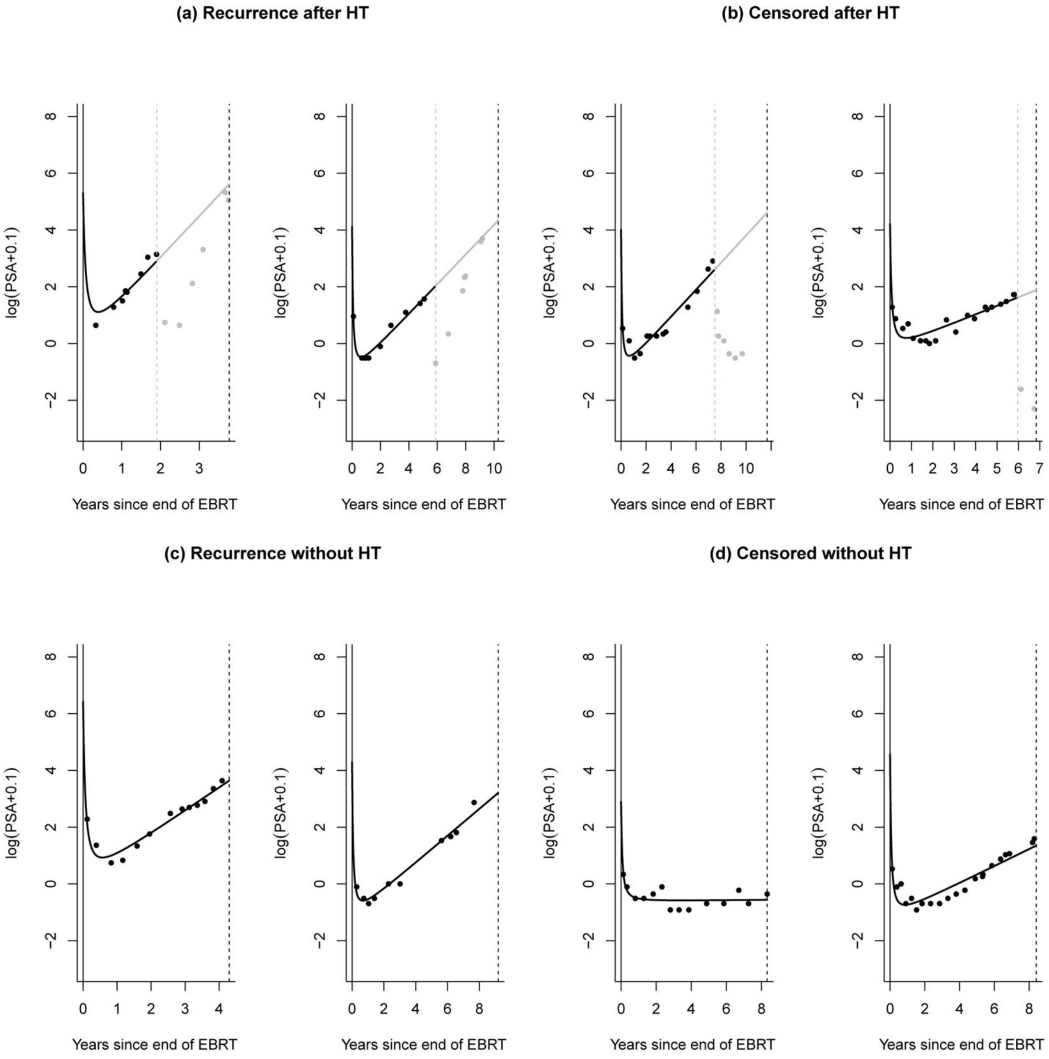

Figure 1.

Individual observed trajectories of log(PSA + 0.1) after the end of EBRT until the observed survival time (at the vertical black dash line): (a) for two patients who received HT (at vertical grey dashed line) and who subsenquently recurred, (b) for two patients who received HT and were subsequently censored, (c) for two patients who recurred without initiating any HT and (d) for two patients who were censored without initiating any HT. Black dots represent observed values of PSA before HT and the black curve represents the subject-specific PSA predictions from the linear mixed model. Grey dots are the observed PSA values after HT and the grey curve represents the extrapolated subject-specific PSA predictions from the mixed model based only on pre-HT data. It gives the expected PSA trajectory assuming the patient did not receive any HT.

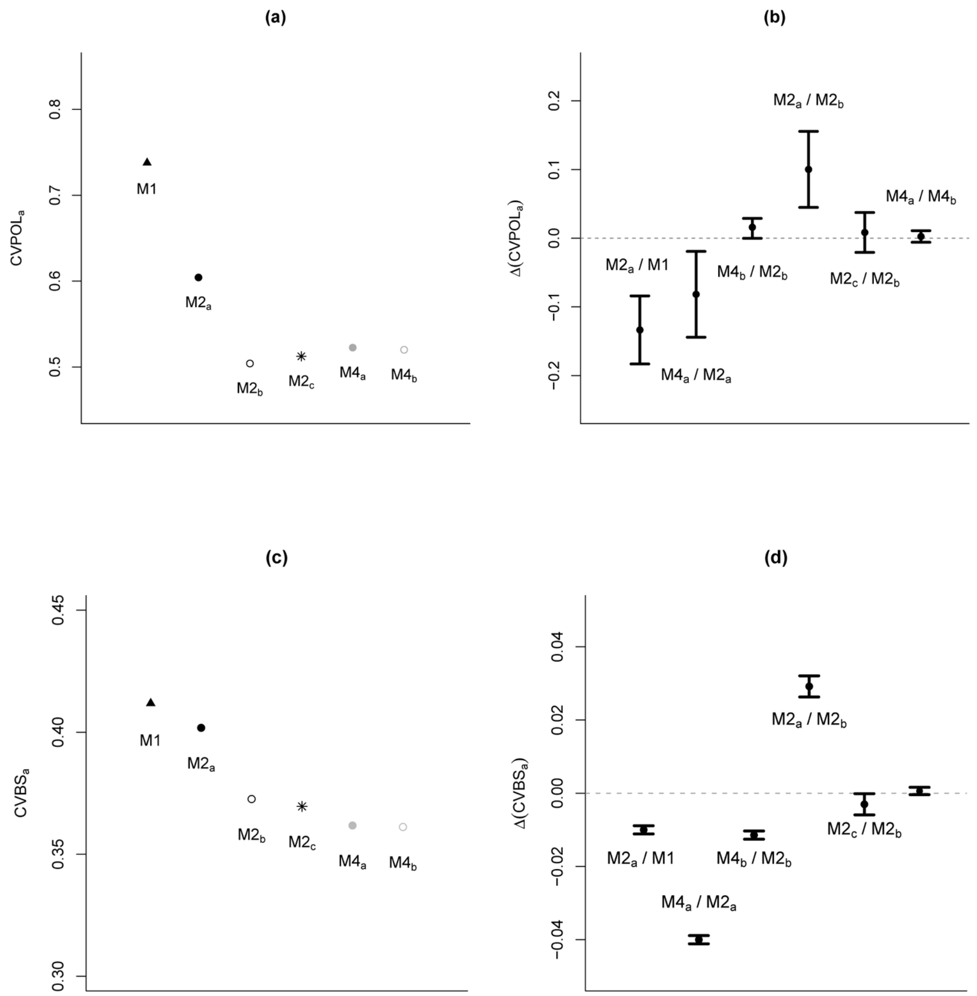

Figure 2.

Predictive accuracy measures after an immediate initiation of ST averaged over the times of ST initation for 6 joint models: with (a) POL estimate, (b) difference in POL and 95% CI, (c) BS estimate, (d) difference in BS and 95% CI. Negative (respectively positive) differences indicate the first model has a better (respectively worse) predictive ability.

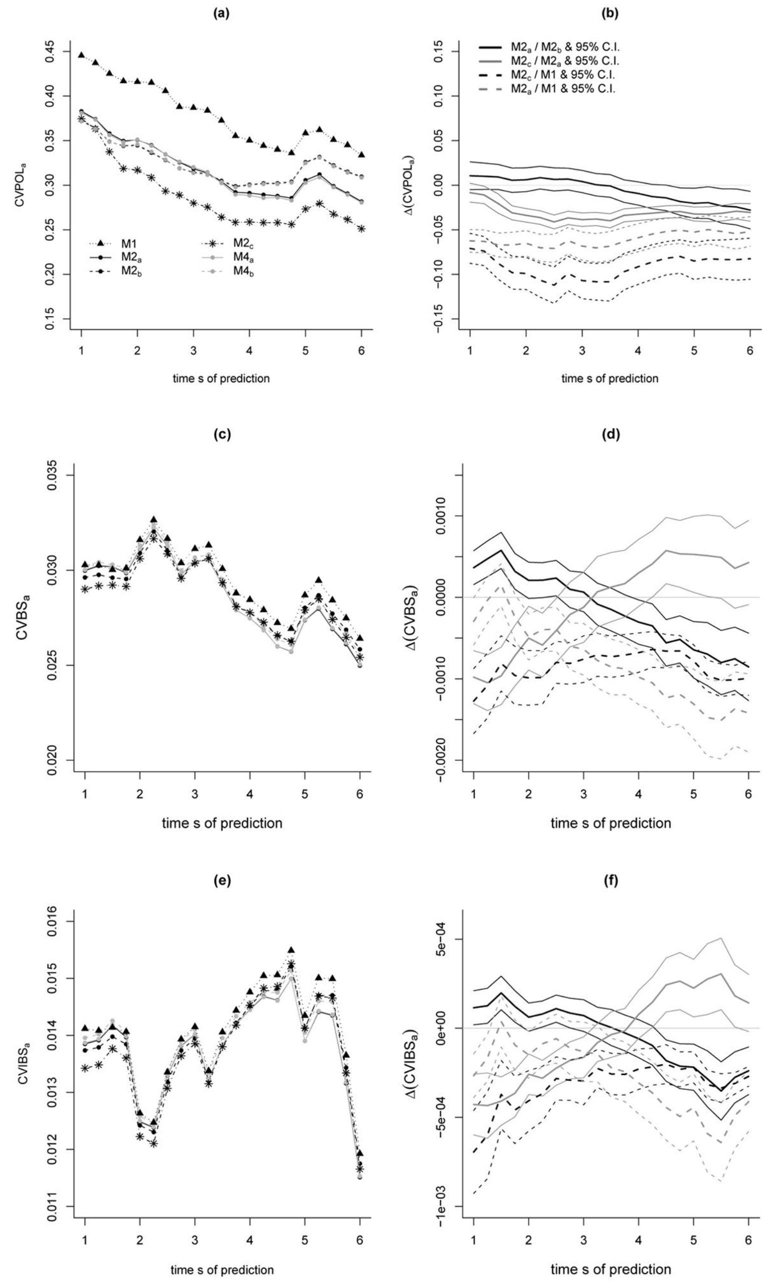

Figure 3.

Predictive accuracy measures in absence of ST for 6 joint models at times from 1 to 6 after EBRT with (a) EPOCE estimate, (b) difference in EPOCE and 95% CI, (c) BS estimate, (d) difference in BS and 95% CI, (e) IBS estimate and (f) difference in IBS and 95% CI. Negative (positive) differences indicate the first model has a better (worse) predictive ability.

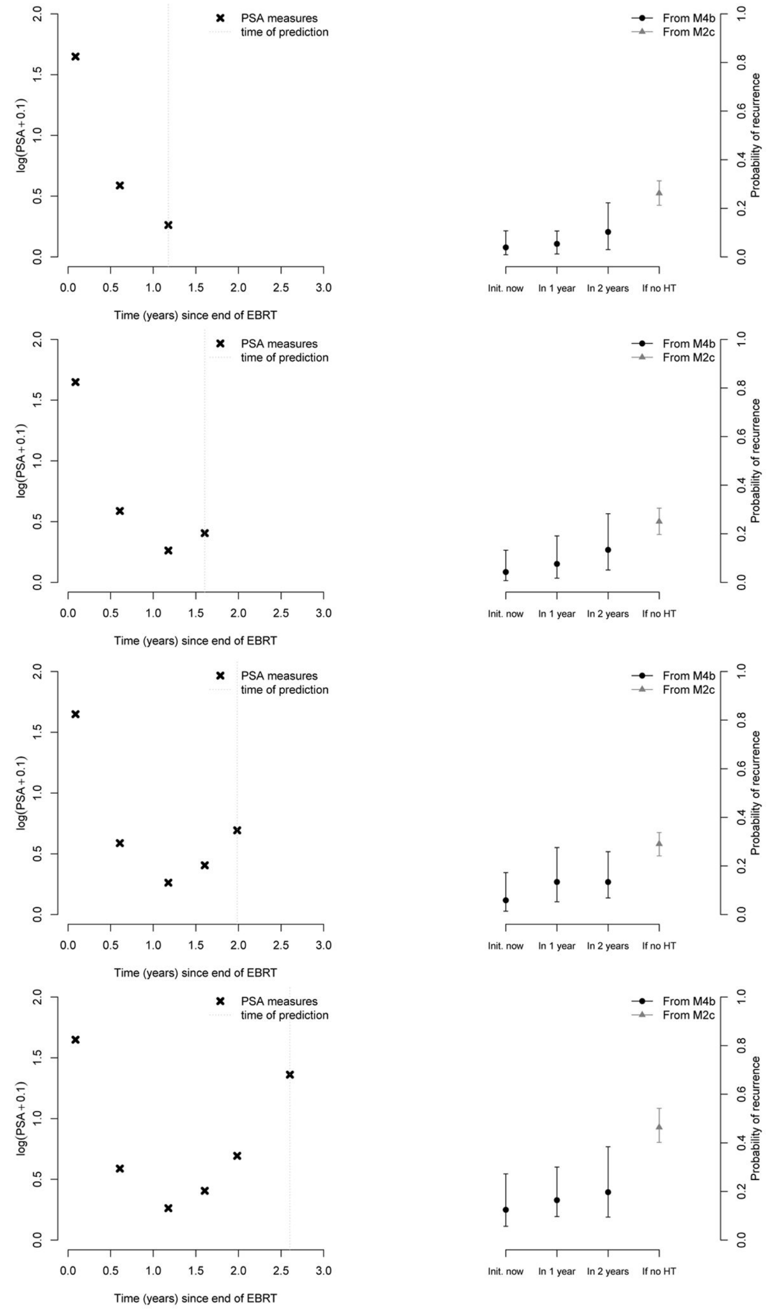

Figure 4.

Observed PSA history (denoted by × on the left) and individual predicted probabilities of clinical recurrence within 3 years according to four scenarios of treatment (on the right). The four scenarios are: immediate initiation of HT, initiation in 1 year, in 2 years or no initiation of HT in the next 3 years. After each new PSA measurement, the distribution of the prediction is approximated by a 2000-draw Monte Carlo method (solid black circle and solid grey triangle indicate the median and the intervals indicate the 95% bands)

Table 1

Goodness-of-fit statistics of the different joint models.

Model	L	AIC	# param.
1	−13549.4	27184.7	43
2.a	−13329.4	26748.8	45
2.b	−13222.9	26535.8	45
2.c	−13261.6	26615.1	46
3.a	−13266.8	26627.7	47
3.b	−13218.2	26530.4	47
4.a	−13265.5	26625.1	47
4.b	−13214.8	26523.5	47
5.1.a[†]	−13264.0	26624.0	48
5.1.b[†]	−13213.9	26523.7	48
5.2.a[‡]	−13263.2	26622.4	48
5.2.b[‡]	−13213.4	26522.7	48

[†] For this model, we assume a change in baseline risk after HT with the function $g(t-\tau_j) = t-\tau_j$ which corresponds to a Gompertz hazard function.

[‡] For this model, we assume that $g(t-\tau_j) = \log(t-\tau_j)$ which corresponds to a Weibull hazard function.

Table 2

Parameters estimates (and standard error (se)) of the HT and the association between the PSA dynamics and the risk of clinical recurrence adjusted on the prognostic factors.

	HT		**Before HT**										**After HT**			
			Level		**Slope**		**Random Effects**						**Level**		**Slope**	
Parameters Model	$\hat{\phi}_1$	**(se)**	$\hat{\phi}_{21}$	**(se)**	$\hat{\phi}_{22}$	**(se)**	$\hat{\phi}_{21}$	**(se)**	$\hat{\phi}_{22}$	**(se)**	$\hat{\phi}_{23}$	**(se)**	$\hat{\phi}_{31}$	**(se)**	$\hat{\phi}_{32}$	**(se)**
1.	0.16	(0.17)														
2.a	**−1.89**	**(0.25)**	**0.13**	**(0.05)**	**2.44**	**(0.18)**										
2.b	**−1.39**	**(0.17)**	**4.82**	**(0.39)**	**1.10**	**(0.14)**										
2.c	**−2.56**	**(0.22)**					**0.92**	**(0.14)**	**−0.31**	**(0.06)**	**3.70**	**(0.22)**				
3.a	**1.33**	**(0.45)**	**0.62**	**(0.06)**	**1.56**	**(0.19)**							−0.05	(0.06)	**1.29**	**(0.26)**
3.b	**2.74**	**(1.28)**	**4.77**	**(0.41)**	**1.19**	**(0.16)**							1.13	(1.39)	**0.95**	**(0.20)**
4.a	**1.20**	**(0.46)**	**0.64**	**(0.06)**	**1.50**	**(0.19)**							0.15	(0.14)	**1.10**	**(0.25)**
4.b	**2.17**	**(1.01)**	**4.77**	**(0.41)**	**1.18**	**(0.16)**							1.90	(1.13)	**0.94**	**(0.22)**
5.1.a	**1.33**	**(0.47)**	**0.62**	**(0.06)**	**1.55**	**(0.19)**							0.14	(0.15)	**1.20**	**(0.26)**
5.1.b	**2.31**	**(1.01)**	**4.77**	**(0.41)**	**1.20**	**(0.16)**							1.84	(1.14)	**0.96**	**(0.23)**
5.2.a	**1.35**	**(0.47)**	**0.62**	**(0.06)**	**1.56**	**(0.19)**							0.12	(0.15)	**1.23**	**(0.25)**
5.2.b	**2.36**	**(1.01)**	**4.71**	**(0.41)**	**1.22**	**(0.16)**							1.74	(1.13)	**1.00**	**(0.23)**

bold underlined: highly significant ($p < 0.001$); **bold**: significant ($0.001 \le p \le 0.05$); nonbold: not significant ($p > 0.05$).