BIOMETRICS 71, 102–113
March 2015

DOI: 10.1111/biom.12232

Quantifying and Comparing Dynamic Predictive Accuracy of Joint Models for Longitudinal Marker and Time-to-Event in Presence of Censoring and Competing Risks

Paul Blanche,[1,2,3,*] Cécile Proust-Lima,[1,2] Lucie Loubère,[1,2] Claudine Berr,[4] Jean-François Dartigues,[1,2] and Hélène Jacqmin-Gadda[1,2]

[1]Université Bordeaux Segalen, ISPED, Inserm Research Center U897, F33076 Bordeaux, France
[2]INSERM, ISPED, Centre INSERM U897-Epidémiologie-Biostatistique, F-33000 Bordeaux, France
[3]Department of Biostatistics, University of Copenhagen, DK-1014 Copenhagen K, Denmark
[4]INSERM, Centre INSERM U1061, Université Montpellier 1, Montpellier, France
*email: pabl@sund.ku.dk

SUMMARY. Thanks to the growing interest in personalized medicine, joint modeling of longitudinal marker and time-to-event data has recently started to be used to derive dynamic individual risk predictions. Individual predictions are called dynamic because they are updated when information on the subject's health profile grows with time. We focus in this work on statistical methods for quantifying and comparing dynamic predictive accuracy of this kind of prognostic models, accounting for right censoring and possibly competing events. Dynamic area under the ROC curve (AUC) and Brier Score (BS) are used to quantify predictive accuracy. Nonparametric inverse probability of censoring weighting is used to estimate dynamic curves of AUC and BS as functions of the time at which predictions are made. Asymptotic results are established and both pointwise confidence intervals and simultaneous confidence bands are derived. Tests are also proposed to compare the dynamic prediction accuracy curves of two prognostic models. The finite sample behavior of the inference procedures is assessed via simulations. We apply the proposed methodology to compare various prediction models using repeated measures of two psychometric tests to predict dementia in the elderly, accounting for the competing risk of death. Models are estimated on the French Paquid cohort and predictive accuracies are evaluated and compared on the French Three-City cohort.

KEY WORDS: Competing risks; Dynamic prediction; Joint model; Longitudinal data; Prediction accuracy.

1. Introduction

For patient counseling or targeting early detection of disease, risk prediction models constitute a basis of personalized medicine. In neurology for instance, current research focuses on preventive treatments against Alzheimer's disease that could be administered in the pre-clinical phase to subjects at high risk of dementia (Aisen et al., 2011). Accurate prediction models for dementia onset are thus required. Since the decline in cognitive functions could begin long before a dementia diagnosis (Amieva et al., 2008), repeated measures of cognitive tests over time could be helpful for the prediction of Alzheimer's disease.

Whatever the clinical setting, predictions should ideally be as personalized as possible. For instance, in Alzheimer's disease they should make the best use of the information about individual cognitive decline available at the landmark time s at which predictions are made. We thus use the terminology of individual *dynamic predictions* to define such predictions that are expected to be updated when information on the subject's clinical profile grows with time. In Section 2, we recall how the joint modeling of a longitudinal marker and of a time-to-event can be used for building such dynamic predictions (Rizopoulos, 2011, 2012; Proust-Lima et al., 2014).

To be useful in clinical practice and health care, dynamic predictions need to have a good prediction accuracy. Appropriate measures for quantifying prediction accuracy and proposals for corresponding estimators are thus essential. Both the dynamic nature of the prediction and the potential presence of competing events should impact the definitions of predictive accuracy since they are real phenomena. For example, in the context of dementia prognosis in the elderly, predictive accuracy definitions should depend on both (i) the competing risk of dementia-free death and (ii) the landmark time at which predictions are made and on which the amount of available information depends. By contrast, censoring induced by loss to follow-up is a nuisance that should not impact the definition of predictive accuracy but that has to be handled by the estimators. Thus, in Section 3 we propose adaptations of the definitions of Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) and expected Brier Score (BS) in the setting of dynamic prediction with competing risks. The proposed definitions combine previously proposed definitions of AUC and BS in the setting of dynamic prediction (Schoop, Graf, and Schumacher, 2008; Parast, Cheng, and Cai, 2012) and competing risk (Schoop et al., 2011; Zheng et al., 2012; Blanche, Dartigues, and Jacqmin-Gadda, 2013). We then discuss their meaning and propose non-parametric Inverse Probability of Censoring Weighting (IPCW) estimators.

Thanks to the breakthroughs in medical and statistical knowledge and the new computation opportunities, more and more prediction models are becoming available. Before using one of them in clinical practice, it would be useful to compare rigorously their predictive abilities. Following on from the prediction of Alzheimer's disease using aging cohort data, the main objective of this work is therefore to provide inference procedures to quantify and compare several dynamic prediction strategies. Thus, in Section 3 we provide large sample results. They rigorously justify and enable computation of pointwise confidence intervals, simultaneous confidence bands as well as tests for comparing two dynamic prediction accuracy curves of two dynamic prediction strategies. Within the context of dynamic predictions based on joint modeling, simultaneous confidence bands are useful for comparing dynamic predictions simultaneously over a set of landmark times. The finite sample behavior of the asymptotic inference procedures is investigated through a simulation study in Section 4.

Finally, Section 5 illustrates the value of the proposed methodology. Two joint models for dementia prediction using repeated measures of two different cognitive tests are estimated on the French Paquid cohort (training set). The 5-year predictive accuracy of the two models is then compared on a validation data set coming from the independent French Three-City cohort.

2. Building Dynamic Prediction Tools by Joint Modeling

Let T denote a time-to-event and η the cause of the event. For the sake of simplicity, we assume only two competing events and denote $\eta = 1$ the main event (e.g., dementia) and $\eta = 2$ the competing event (e.g., dementia-free death). Owing to censoring C induced by loss to follow-up, we observe $\widetilde{T} = \min(T, C)$ and $\widetilde{\eta} = \Delta\eta$, where $\Delta = \mathbb{1}_{(T \le C)}$ and $\mathbb{1}_{(\cdot)}$ denotes the indicator function. Let $Y(t)$ be a marker measurement at time t. We assume that we observe the independent and identically distributed (i.i.d.) sample of n subjects $\left\{(\widetilde{T}_i, \Delta_i, \widetilde{\eta}_i, \mathbf{Y}_i, \mathbf{X}_i), i = 1, \ldots, n\right\}$, where $\mathbf{X}_i$ denotes a vector of baseline covariates (e.g., gender and education level) and $\mathbf{Y}_i$ denotes the vector of n_i observed repeated marker measurements $Y_{ij} \equiv Y_i(t_{ij})$ for subject i at time t_{ij}, $j = 1, \ldots, n_i$ (e.g., repeated measurements of a cognitive test).

2.1. *Joint Modeling of Longitudinal Marker and Time-to-Event with Competing Risks*

Usually, a fully parametric approach is used and the joint probability distribution of $\big(T, \eta, Y(\cdot)\big)$ is thus parametrized by a vector of parameters $\boldsymbol{\xi}$. Modeling strategies are diverse in the literature, but in most cases, the full likelihood of the data $\left\{(\widetilde{T}_i, \widetilde{\eta}_i, \mathbf{Y}_i), i = 1, \ldots, n\right\}$ given baseline covariates $\mathbf{X}_i, i = 1, \ldots, n$, denoted by $\mathcal{L}_{(\widetilde{T},\widetilde{\eta},\mathbf{Y})}$, can be written as

$$\mathcal{L}_{(\widetilde{T},\widetilde{\eta},\mathbf{Y})} = \prod_{i=1}^{n} \int_{\boldsymbol{\gamma}_i} \mathcal{L}_{\mathbf{Y}}\left\{\mathbf{Y}_i|\boldsymbol{\gamma}_i\right\} \mathcal{L}_{(\widetilde{T},\widetilde{\eta})}\left\{(\widetilde{T}_i, \widetilde{\eta}_i)|\boldsymbol{\gamma}_i\right\} f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}_i)\mathrm{d}\boldsymbol{\gamma}_i.$$

The term $\mathcal{L}_{\mathbf{Y}}\left\{\mathbf{Y}_i|\boldsymbol{\gamma}_i\right\}$ denotes the contribution to the likelihood of the repeated marker measurements $\mathbf{Y}_i$ of subject i given a shared latent variable $\boldsymbol{\gamma}_i$ (discrete or continuous) of distribution $f_{\boldsymbol{\gamma}}(\cdot)$. The contribution to the likelihood of the survival data $(\widetilde{T}_i, \widetilde{\eta}_i)$ of subject i given $\boldsymbol{\gamma}_i$ is defined as

$$\mathcal{L}_{(\widetilde{T},\widetilde{\eta})}\left\{(\widetilde{T}_i, \widetilde{\eta}_i)|\boldsymbol{\gamma}_i\right\} = \lambda_1(\widetilde{T}_i|\boldsymbol{\gamma}_i)^{\mathbb{1}_{(\widetilde{\eta}_i=1)}} \lambda_2(\widetilde{T}_i|\boldsymbol{\gamma}_i)^{\mathbb{1}_{(\widetilde{\eta}_i=2)}} S(\widetilde{T}_i|\boldsymbol{\gamma}_i),$$

where $S(\widetilde{T}_i|\boldsymbol{\gamma}_i) = \exp\left(-\int_0^{\widetilde{T}_i}\left\{\lambda_1(u|\boldsymbol{\gamma}_i) + \lambda_2(u|\boldsymbol{\gamma}_i)\right\}\mathrm{d}u\right)$.

In practice, parametric proportional hazards models are usually chosen for modeling the cause-specific hazards given $\boldsymbol{\gamma}_i$, denoted by $\lambda_k(\cdot|\boldsymbol{\gamma}_i), k = 1, 2$, while a linear mixed model is often used for modeling the marker trajectory $Y(\cdot)$, leading $\mathcal{L}_{\mathbf{Y}}\left\{\mathbf{Y}_i|\boldsymbol{\gamma}_i\right\}$ to be the likelihood of a multivariate Gaussian variable.

By exploiting the link between the two contributions to the likelihoods $\mathcal{L}_{\mathbf{Y}}\left\{\mathbf{Y}_i|\boldsymbol{\gamma}_i\right\}$ and $\mathcal{L}_{(\widetilde{T},\widetilde{\eta})}\left\{(\widetilde{T}_i, \widetilde{\eta}_i)|\boldsymbol{\gamma}_i\right\}$ modeled by $\boldsymbol{\gamma}_i$, the maximization of the full likelihood $\mathcal{L}_{(\widetilde{T},\widetilde{\eta},\mathbf{Y})}$ make it possible to estimate the vector of model parameters $\boldsymbol{\xi}$ consistently.

Finally, note that two main approaches are often used for modeling the shared latent variable $\boldsymbol{\gamma}$. Either a continuous or a discrete distribution can be assumed for $\boldsymbol{\gamma}$, leading to the so-called "shared random effect" or "latent class" joint models, respectively. We refer to Proust-Lima et al. (2014) for a recent overview contrasting the two approaches. In this work, we make use of the joint latent class approach for our application, taking advantage of its computational assets and of our previous experiences, in particular for predictive purposes (Proust-Lima and Taylor, 2009; Proust-Lima et al., 2014). However, the following methodology can be applied whatever the joint modeling approach.

2.2. *Predictions from Joint Models*

There is currently an increasing interest in making individual predictions using the joint model framework (Proust-Lima and Taylor, 2009; Rizopoulos, 2011, 2012; Proust-Lima et al., 2014, among others). Based on the model specification and the vector of estimated parameters $\widehat{\boldsymbol{\xi}}$, various subject-specific probabilities can be computed. Of particular interest are the subject-specific probabilities of experiencing the main event within a time interval $(s, s + t]$ given the whole information available on the subject accumulated till the landmark time s. Here, t denotes a fixed window of prediction whereas the varying landmark time s denotes the time at which predictions are made conditionally to the subject-specific history. For $0 \le s < T_i$, we thus define the dynamic predictions of event 1 (main event) for subject i by

$$\pi_i(s, t) = \mathbb{P}_{\widehat{\boldsymbol{\xi}}}(s < T_i \le s + t, \eta_i = 1|T_i > s, \mathcal{Y}_i(s), \mathbf{X}_i),$$

where $\mathbb{P}_{\widehat{\boldsymbol{\xi}}}$ denotes the probability distribution parametrized by $\widehat{\boldsymbol{\xi}}$, $\mathbf{X}_i$ denotes the vector of baseline covariates and $\mathcal{Y}_i(s)$ denotes the entire information on the marker trajectory available for subject i by time s, that is, $\mathcal{Y}_i(s) = \left\{Y_{ij} : 0 \le t_{ij} \le s, j = 1, \ldots, n_i\right\}$. These predictions are dynamic in the sense that they change with increasing landmark time s and available information $\{\mathcal{Y}_i(s), T_i > s\}$. From a practical point of view, they can be computed using Bayesian formulae, Monte-Carlo simulations or numerical integrations (Proust-Lima et al., 2014; Rizopoulos, 2012, Sec. 7.1). For the joint latent class approach used in our application, additional details are provided in Section 5 and in the Web Appendix B.

3. Quantifying and Comparing Dynamic Predictive Accuracy

In the following, we assume an i.i.d. sample of n subjects $\{(\widetilde{T}_i, \Delta_i, \widetilde{\eta}_i, \pi_i(\cdot,\cdot)), i = 1, \ldots, n\}$, where $\pi_i(\cdot,\cdot)$ denotes a subject-i-specific prediction process computable for all landmark times s and prediction horizon t. Note that assuming an i.i.d. sample implies that the joint model used to compute $\pi_i(s,t)$ from the specific characteristics of subject i has been fitted on an independent learning dataset. Without loss of generality, we set $\pi_i(s,t) = 0$ for all subjects i that are no longer at risk at s, and we focus on prediction of event $\eta = 1$ (main event).

3.1. *Definitions and Meaning*

We propose an adaptation of the definitions of two well-established prediction accuracy measures, the Area Under the ROC Curve (AUC) and the expected Brier score (BS), to account simultaneously for both (i) the dynamic nature of the predictions and (ii) the competing risks setting.

3.1.1. Dynamic AUC with competing risks. In practice, one would like a prediction tool that gives higher predicted risks of event for subjects who are more likely to experience the event than for subjects who are less likely to experience it. This is the concept of predictive accuracy in terms of discrimination, for which AUC is a meaningful measure.

Formally, by combining the definition of ROC curve for competing risks (Zheng et al., 2012; Blanche et al., 2013) and the one for dynamic prediction (Parast et al., 2012), we propose the following definition of the dynamic AUC, at landmark time s for a prediction horizon t:

$$\begin{aligned}&\mathrm{AUC}(s,t)\\ &\quad = \mathbb{P}\Big(\pi_i(s,t) > \pi_j(s,t) \Big| D_i(s,t) = 1, D_j(s,t) = 0, T_i > s, T_j > s\Big),\end{aligned}$$

where $D_i(s,t) = \mathbb{1}_{(s<T_i\le s+t, \eta_i=1)}$. With this notation (where "D" is for "diseased"), for any subject i at risk at time s, $D_i(s,t) = 1$ when subject i experiences the main event within the time interval $(s, s+t]$, and $D_i(s,t) = 0$ when either subject i experiences a competing event within the time interval or is event-free at $s+t$. Within the terminology of the ROC methodology, at fixed landmark time s and prediction horizon t, subject i at risk at time s is defined as a case when $D_i(s,t) = 1$ and a control when $D_i(s,t) = 0$ (Blanche et al., 2013).

It is worth noting that by using a similar argument as in McIntosh and Pepe (2002), it can be shown that the "true underlying" risk of event has the best of all possible AUC (i.e., the highest), which is a desirable property for any measure of prediction accuracy.

3.1.2. Dynamic Brier score with competing risks. By combining definitions of the Expected Brier score for competing risks (Schoop et al., 2011) and for dynamic prediction (Schoop et al., 2008), we propose the following definition for the dynamic expected Brier score

$$\mathrm{BS}(s,t) = \mathbb{E}\Big[\Big(D(s,t) - \pi(s,t)\Big)^2 \Big| T > s\Big],$$

which is a mean squared error. Following Graf, Schmoor, and Schumacher (1999), it can be expressed as

$$\begin{aligned}\mathrm{BS}(s,t) = {}& \mathbb{E}\Big[\Big(\mathbb{E}\big[D(s,t)\big|\mathcal{H}(s)\big] - \pi(s,t)\Big)^2 \Big| T > s\Big]\\ &+ \mathbb{E}\Big[\Big(D(s,t) - \mathbb{E}\big[D(s,t)\big|\mathcal{H}(s)\big]\Big)^2 \Big| T > s\Big], \quad (1)\end{aligned}$$

where $\mathcal{H}(s) = \{\mathbf{X}, \mathcal{Y}(s), T > s\}$ denotes the history at time s used for computing the prediction $\pi(s,t)$. The second term in (1), named "inseparability," does not depend on the distribution of the predictions. The first term, named "imprecision" or "calibration," measures how close the predictions are to $\mathbb{E}\big[D(s,t)\big|\mathcal{H}(s)\big]$, that is, the "true underlying" risk of event in $(s, s+t]$ given $\mathcal{H}(s)$. Note that if the predictions were perfect, that is, if $\pi(s,t) = \mathbb{E}\big[D(s,t)\big|\mathcal{H}(s)\big]$, then the best (i.e., the lowest) of all possible BS for predictions using $\mathcal{H}(s)$ would be reached and the "calibration" term would be equal to zero.

3.1.3. Contrasts between AUC and BS. Both AUC and BS are interesting and complement each other for measuring how good a dynamic prediction tool $\pi(s,t)$ is, where $\pi(s,t)$ depends on some estimated parameters $\widehat{\boldsymbol{\xi}}$, baseline covariates $\mathbf{X}$ and past marker trajectory $\mathcal{Y}(s)$. AUC is particularly convenient for communication purposes, as it has a simple interpretation as a concordance index, does not depend on the main event cumulative incidence $\mathbb{P}(s < T \le s+t, \eta = 1 | T > s)$, and therefore has an easily understandable scaling. By contrast, BS has the advantage of being a more complete predictive accuracy measure as it quantifies both calibration and discrimination. Indeed, it can be shown that the "inseparability" term in (1) depends on the inherent discrimination ability of the information $\mathcal{H}(s)$. However, as BS depends on the main event cumulative incidence, note that the scaling of this predictive accuracy measure changes with s and has to be interpreted carefully.

3.2. *Estimation*

In the presence of censored data induced by loss to follow-up, for all subjects i censored within $(s, s+t]$, the indicator $D_i(s,t)$ cannot be computed and is thus unknown. To overcome this "missing data" issue, the Inverse Probability of Censoring Weighting (IPCW) technique has been recently applied in several closely related settings (Hung and Chiang, 2010; Blanche et al., 2013; Parast et al., 2012, among others). Thus, we propose the IPCW estimators:

$$\widehat{\mathrm{AUC}}(s,t) = \frac{\sum_{i=1}^n \sum_{j=1}^n \mathbb{1}_{(\pi_i(s,t)>\pi_j(s,t))} \widetilde{D}_i(s,t)\left(1 - \widetilde{D}_j(s,t)\right) \widehat{W}_i(s,t)\widehat{W}_j(s,t)}{\sum_{i=1}^n \sum_{j=1}^n \widetilde{D}_i(s,t)\left(1 - \widetilde{D}_j(s,t)\right) \widehat{W}_i(s,t)\widehat{W}_j(s,t)},$$

and

$$\widehat{\mathrm{BS}}(s,t) = \frac{1}{n\widehat{S}_{\widetilde{T}}(s)} \sum_{i=1}^{n} \widehat{W}_i(s,t)\left(\widetilde{D}_i(s,t) - \pi_i(s,t)\right)^2,$$

where $\widehat{S}_{\widetilde{T}}(s) = (1/n)\sum_{i=1}^{n} \mathbb{1}_{(\widetilde{T}_i>s)}$ estimates the probability of observing a subject at risk at s. The indicator $\widetilde{D}_i(s,t) = \mathbb{1}_{(s<\widetilde{T}_i\le s+t,\widetilde{\eta}_i=1)}$ equals 1 when subject i is known to have experienced the main event within $(s, s+t]$, and equals 0 otherwise. To account for censoring, the weights are defined as

$$\widehat{W}_i(s,t) = \frac{\mathbb{1}_{(\widetilde{T}_i>s+t)}}{\widehat{G}(s+t|s)} + \frac{\mathbb{1}_{(s<\widetilde{T}_i\le s+t)}\Delta_i}{\widehat{G}(\widetilde{T}_i|s)},$$

where $\widehat{G}(u)$ is the Kaplan–Meier estimator of survival function of the censoring time at u, that is, $\mathbb{P}(C > u)$, and $\forall u > s$, $\widehat{G}(u|s) = \widehat{G}(u)/\widehat{G}(s)$ estimates the conditional probability of not being censored at time u conditionally on being uncensored at time s.

Importantly, these estimators are model-free in the sense that there is no assumption about the correctness of the specification of the joint model used for computing $\pi_i(s,t), i = 1,\dots,n$. Indeed, it is frequently unrealistic to assume that a model is well-specified but it makes sense to consider that it could be useful enough for prediction purposes. For such misspecified models, the inference procedures for AUC and BS must be model-free to be unbiased. Secondly, for comparing prediction tools from two different joint models, the use of model-based estimators considered for example in Proust-Lima and Taylor (2009) and Henderson, Diggle, and Dobson (2002) would assume that the two models are simultaneously both well-specified, whereas it is often paradoxical. Consequently, model-based comparisons would likely lead to biased and potentially unfair comparisons.

3.3. *Large Sample Results*

For the sake of brevity, let $\theta(s,t)$ denote either $\mathrm{AUC}(s,t)$ or $\mathrm{BS}(s,t)$ and $\widehat{\theta}(s,t)$ the corresponding IPCW estimator. To define some usual identifiability constraints, let $\tau_0 < \sup\left\{u : \mathbb{P}(\widetilde{T} > u) > 0\right\}$, $\tau_1(s) > \inf\left\{u : \mathbb{P}(T \le s+u, \eta = 1|T > s) > 0\right\}$ and $\tau_2(s) < \sup\left\{u : \mathbb{P}(\widetilde{T} > s+u) > 0\right\}$. The key result that both justifies and enables the computation of inference procedures for $\theta(s,t)$ is the following.

Lemma 1. *Assume that the censoring time C is independent of $(T, \eta, \pi(\cdot,\cdot))$, then $\forall s \in [0,\tau_0]$, $\forall t \in [\tau_1(s), \tau_2(s)]$:*

$$\sqrt{n}\left(\widehat{\theta}(s,t) - \theta(s,t)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} IF_\theta(\widetilde{T}_i, \widetilde{\eta}_i, \pi_i(s,t), s, t) + o_p, \quad (1)$$

where $\mathbb{E}\left[IF_\theta(\widetilde{T}_i, \widetilde{\eta}_i, \pi_i(s,t), s, t)\right] = 0$ and $IF_\theta(\cdot)$ is the influence function of the estimator which is detailed in Web Appendix A, at formulae (14) *and* (27).

Proof. A proof is provided in Web Appendix A. □

From the decomposition of the estimator $\widehat{\theta}(s,t)$ in a sum of asymptotically i.i.d. terms in Lemma 1, the central limit theorem induces the asymptotic normality of the estimator:

$$\sqrt{n}\left(\widehat{\theta}(s,t) - \theta(s,t)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma^2_{s,t}\right). \quad (2)$$

Using a simple plug-in estimator $\widehat{\mathrm{IF}}_\theta(\widetilde{T}_i, \widetilde{\eta}_i, \pi_i(s,t), s, t)$ detailed in Web Appendix A.3, variance $\sigma^2_{s,t}$ can be consistently estimated by the empirical estimator

$$\widehat{\sigma}^2_{s,t} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathrm{IF}}_\theta(\widetilde{T}_i, \widetilde{\eta}_i, \pi_i(s,t), s, t)^2.$$

3.4. *Confidence Regions and Tests*

Let $\pi^{(1)}(\cdot,\cdot)$ and $\pi^{(2)}(\cdot,\cdot)$ be two rival prediction processes computable for all landmark times s and prediction horizon t and, respectively; $\theta^{(1)}(\cdot,\cdot)$ and $\theta^{(2)}(\cdot,\cdot)$ their predictive accuracy processes. For the sake of brevity, we display confidence regions and tests only for the difference $\Delta\theta(\cdot,t) = \theta^{(1)}(\cdot,t) - \theta^{(2)}(\cdot,t)$ at a given prediction horizon t.

3.4.1. Inference procedures at a specific landmark time s. Let $\Delta\widehat{\theta}(\cdot,t) = \widehat{\theta}^{(1)}(\cdot,t) - \widehat{\theta}^{(2)}(\cdot,t)$ define the IPCW estimator of $\Delta\theta(\cdot,t)$. As a consequence of Lemma 1, for any landmark time s, we directly obtain the asymptotic $(1-\alpha)$-level pointwise confidence interval

$$\left\{\Delta\widehat{\theta}(s,t) \pm z_{1-\alpha/2}\frac{\widehat{\sigma}_{\Delta,s,t}}{\sqrt{n}}\right\}, \quad (3)$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the univariate standard normal distribution and

$$\widehat{\sigma}^2_{\Delta,s,t} = \frac{1}{n}\sum_{i=1}^{n} \Delta\widehat{\mathrm{IF}}_\theta(\widetilde{T}_i, \widetilde{\eta}_i, \pi_i^{(1)}(s,t), \pi_i^{(2)}(s,t), s, t)^2,$$

with $\Delta\widehat{\mathrm{IF}}_\theta(\widetilde{T}_i, \widetilde{\eta}_i, \pi_i^{(1)}(s,t), \pi_i^{(2)}(s,t), s, t) = \widehat{\mathrm{IF}}_\theta(\widetilde{T}_i, \widetilde{\eta}_i, \pi_i^{(1)}(s,t), s, t) - \widehat{\mathrm{IF}}_\theta(\widetilde{T}_i, \widetilde{\eta}_i, \pi_i^{(2)}(s,t), s, t)$. A test for comparing the two prediction accuracy measures $\theta^{(1)}(s,t)$ and $\theta^{(2)}(s,t)$ can also be equivalently derived. Indeed, under $\mathcal{H}_0^{(s)} : \Delta\theta(s,t) = 0$, as $n \to \infty$ then $\sqrt{n}\left(\Delta\,\widehat{\theta}(s,t)/\widehat{\sigma}_{\Delta,s,t}\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$.

3.4.2. Simultaneous inference procedures over a set of landmark times s. Pointwise confidence intervals and tests are of interest when aiming to quantify and compare predictive accuracy at a specific landmark time s. However, within the joint modeling framework that enables computations of dynamic predictions for a set (or an interval) of landmark times s, denoted $\mathcal{S}$, it is also desirable to quantify the overall predictive accuracy over $\mathcal{S}$ for a fixed prediction horizon t. A simultaneous confidence band is then useful to estimate the variability of the estimation of the curve $\left\{\left(s, \Delta\theta(s,t)\right), s \in \mathcal{S}\right\}$. A $(1-\alpha)$-simultaneous confidence band for this curve is defined as a region containing this curve with probability level $1-\alpha$. By definition, a simultaneous confidence band is larger that the band of pointwise confidence intervals. We propose to compute an asymptotic $(1-\alpha)$-simultaneous confidence band by:

$$\left\{ \Delta\widehat{\theta}(s,t) \pm \widehat{q}^{(\mathcal{S},t)}_{1-\alpha} \frac{\widehat{\sigma}_{\Delta,s,t}}{\sqrt{n}} \right\}, \quad s \in \mathcal{S}, \tag{4}$$

where $\widehat{q}^{(\mathcal{S},t)}_{1-\alpha}$ is a $(1-\alpha)$-quantile that depends on both the landmark time set $\mathcal{S}$ and the prediction horizon t. By using Lemma 1, the following simulation-based technique is applied to estimate this quantile by accounting properly for the autocorrelation of the process $\Delta\widehat{\theta}(\cdot, t)$:

(1) For $b = 1, \ldots, B$, with B large enough, say $B = 4000$:
   (a) Generate a random sample $(\omega_1^b, \ldots, \omega_n^b)$ from n independent standard normal distributions.
   (b) Using the estimator $\widehat{\mathrm{IF}}_\theta(\cdot)$ detailed in Web Appendix A.3, compute:

$$\Upsilon^b = \sup_{s\in\mathcal{S}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \omega_i^b \frac{\Delta\widehat{\mathrm{IF}}_\theta(\widetilde{T}_i, \widetilde{\eta}_i, \pi_i^{(1)}(s,t), \pi_i^{(2)}(s,t), s, t)}{\widehat{\sigma}_{\Delta,s,t}} \right|,$$

(2) Compute $\widehat{q}^{(\mathcal{S},t)}_{1-\alpha}$ as the $100(1-\alpha)$th percentile of $\left\{\Upsilon^1, \ldots, \Upsilon^B\right\}$.

Note that when $\mathcal{S}$ reduces to any singleton $\{s\}$ then $\widehat{q}^{(\mathcal{S},t)}_{1-\alpha}$ is close enough to $z_{1-\alpha/2}$ as soon as B is large enough. By contrast, the larger the set $\mathcal{S}$, and the weaker is the autocorrelation of the process $\widehat{\theta}(\cdot, t)$ and the larger $\widehat{q}^{(\mathcal{S},t)}_{1-\alpha}$ will be compared to $z_{1-\alpha/2}$. Based on the conditional multiplier central limit theorem, this kind of simulation approach has already been successfully applied in several related settings (Martinussen and Scheike, 2006).

The simultaneous confidence band of the curve $\left\{ \left(s, \Delta\theta(s,t)\right), s \in \mathcal{S} \right\}$ can either be used:

- to test whether or not the two dynamic prediction accuracy curves are different, that is, $\mathcal{H}_0 : \forall s \in \mathcal{S} \; \Delta\,\theta(s,t) = 0$, by observing whether or not the zero line is contained within the band;
- or to assert that one dynamic prediction accuracy is significantly uniformly higher than another in the set of landmark times, that is, $\forall s \in \mathcal{S} \; \Delta\,\theta(s,t) > 0$, by observing whether or not the confidence band overlaps the zero line.

3.4.3. Remarks. Similarly, confidence regions can be derived for $\theta(s,t)$ instead of $\Delta\theta(s,t)$. The previous methodology applies by replacing $\Delta\widehat{\theta}(s,t)$ by $\widehat{\theta}(s,t)$, $\Delta\widehat{\mathrm{IF}}_\theta(\cdot)$ by $\widehat{\mathrm{IF}}_\theta(\cdot)$ and $\widehat{\sigma}_{\Delta,s,t}$ by $\widehat{\sigma}_{s,t}$ in formulae (3) and (4) and in the algorithm to compute $\widehat{q}^{(\mathcal{S},t)}_{1-\alpha}$. Note also that there is no equivalence between the confidence regions for predictive accuracy overlapping and the significance of the differences, in particular because of paired data. Confidence regions for $\theta(s,t)$ and $\Delta\theta(s,t)$ are thus complementary to each other.

4. Simulation Study

The objective of the simulation study was to check the finite sample behavior of the inference procedures, in particular those for simultaneous inference procedures over a set of landmark times. Scenarii were partly inspired by the cognitive aging cohort data described in Section 5.

4.1. *Simulation Scenarii*

We considered a setting with two competing risks and two longitudinal markers $Y^{(l)}(\cdot)$, $l = 1, 2$, both associated with the two competing events. For simplicity, we generated the data from simple joint latent class models that consider a discrete shared latent variable γ. We generated only two classes for γ from a Bernoulli distribution with parameter $p = 0.5$. Class-specific constant hazard functions were chosen for cause-specific hazards of the two events. The two class-specific maker trajectories $Y^{(l)}(\cdot)$, $l = 1, 2$, were modeled by

$$Y_i^{(l)}(t_{ij})|_{\gamma_i = g} = (\beta_{0g}^{(l)} + b_{i0}^{(l)}) + (\beta_{1g}^{(l)} + b_{i1}^{(l)}) \times t_{ij} + \varepsilon_i^{(l)}(t_{ij}),$$

for class $g = 1, 2$, with $(b_{i0}^{(l)}, b_{i1}^{(l)}) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{b}}^{(l)})$ independent of γ_i, $\mathbf{\Sigma}_b^{(l)} = \mathrm{diag}\left({\sigma_{b_0}^{(l)}}^2, {\sigma_{b_1}^{(l)}}^2 \right)$ and an independent noise $\varepsilon_i^{(l)}(t_{ij}) \sim \mathcal{N}(0, {\sigma_\varepsilon^{(l)}}^2)$.

We generated a 10-year follow-up for the longitudinal data with repeated measurements every year. Independent censoring was generated with an exponential distribution. The landmark times were chosen to be equal to the measurement times, that is, $\mathcal{S} = \{0, 1, \ldots, 10\}$ and the prediction horizon was set to $t = 5$ years. Parameters for survival data generation were chosen such that the proportions of observed main event, competing event and censored observations within the time interval $(s, s+t]$ and observed event-free at time $s + t$ were, respectively, around 20%, 10%, 20%, and 50% for all landmark times s (see Web Figure I).

For each time s, two rival dynamic predictions $\pi_i^{(l)}(s,t)$, $l = 1, 2$, were computed from two joint latent class models, using the formulae of Web Appendices B and C, each using one of the two sets of repeated marker measurements $\mathcal{Y}_i^{(l)}(s)$, $l = 1, 2$. Eight scenarii illustrated in Figure 1 were investigated with sample sizes $n = 1000$, 2000, and 3000. Changes from one scenario to the other consisted only in changes in the marker trajectories. Two scenarii were generated under the null hypothesis asserting that the dynamic predictive accuracy curves were equal, by generating $Y_i^{(1)}(\cdot)$ and $Y_i^{(2)}(\cdot)$ from the same distribution. The two scenarii differed by the distance between the two class-specific slopes $\beta_{11}^{(l)}$ and $\beta_{12}^{(l)}$, $l = 1, 2$ (H0.1 and H0.2). Six scenarii were generated under the alternative hypothesis by breaking the symmetry between the distribution of the two marker trajectories $Y_i^{(1)}(\cdot)$ and $Y_i^{(2)}(\cdot)$. Three of them were generated by increasing the difference $\beta_{11}^{(1)} - \beta_{12}^{(1)}$ between the class-specific slopes for marker 1 (H1.1.1, H1.1.2, H1.1.3). Two scenarii were generated by making the standard deviations $\sigma_{b_0}^{(l)}$, $\sigma_{b_1}^{(l)}$ and $\sigma_\varepsilon^{(l)}$ 1.5 or 2 times larger for $l = 1$ compared to $l = 2$ (H1.2.1, H1.2.2). Contrary to the previous scenarii, one additional scenario was generated such that the two class-specific trajectories of one of the two markers crossed (H1.3). Parameter values for data generation are given in Web Appendix C.

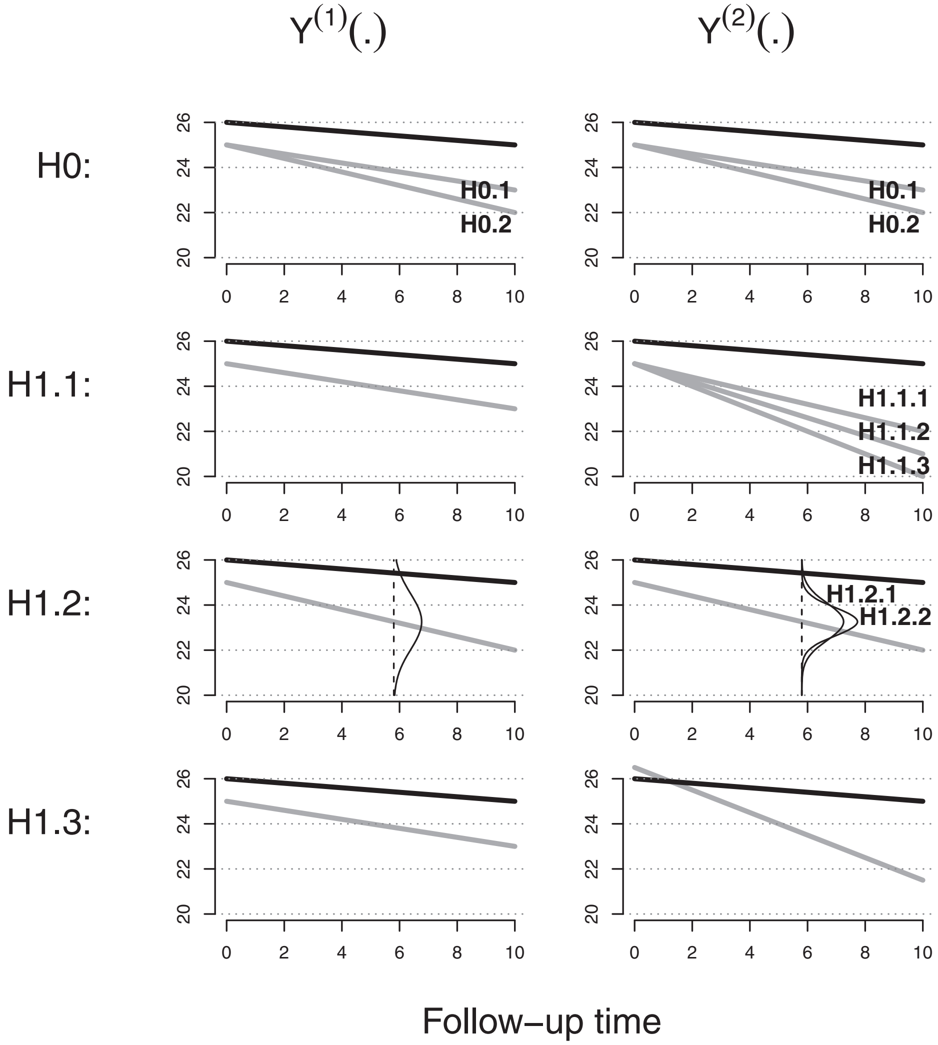

Figure 1. Comparison of marker trajectories according to each simulation scenario. Marker trajectories $Y_i^{(1)}(\cdot)$ and $Y_i^{(2)}(\cdot)$ conditionally to latent class $\gamma = 1$ (black) and $\gamma = 2$ (gray) are displayed as solid line. In Scenario H1.2, Gaussian densities of $Y^{(l)}(t = 6)|_{\gamma=2}$, $l = 1, 2$, display the impact of dividing the standard errors by 1.5 or 2.

To illustrate the impact of a potential violation of the independent censoring assumption, we additionally simulated the two scenarii under the null (H0.1 and H0.2) with marker-dependent censoring. To that end, we generated class-specific censoring distributions as detailed in Web Appendix C.

4.2. *Simulation Results*

For each scenario, we ran 1000 simulations and estimated curves of $\text{AUC}(s, t)$ and $\text{BS}(s, t)$ versus $s \in \mathcal{S}$ for the two dynamic predictions $\pi^{(l)}(s, t)$, $l = 1, 2$, and the curves of differences $\triangle\text{AUC}(s, t)$ and $\triangle\text{BS}(s, t)$ versus $s \in \mathcal{S}$. For all curves, pointwise confidence intervals, simultaneous confidence bands and the corresponding tests were estimated at α-level equal to 5% (using $B = 4000$). For the sake of brevity, we mainly present the empirical coverage probabilities of simultaneous confidence bands for the two dynamic predictions and the type I error and power of the simultaneous tests of $\mathcal{H}_0 : \forall s \in \mathcal{S} \triangle \theta(s, t) = 0$ (Table 1). Figure 2 illustrates the power of the pointwise tests of $\mathcal{H}_0^{(s)} : \triangle\theta(s, t) = 0$ for each landmark time $s \in \mathcal{S}$. Figure 3 displays the mean curves of $\text{AUC}(s, t)$ versus s for $\pi^{(2)}(s, t)$ and the coverage probabilities of the pointwise confidence intervals. Additional details about simulation results for pointwise inference are presented in the Web Appendix D.

For both BS and AUC, empirical coverages of confidence regions and type I error are close to the chosen 95% and 5%-levels when sample sizes are large enough (Table 1). These results show that the asymptotic inference procedures behave quite well under our data generated scenarii, with an increased power when sample sizes increase. However, with a sample size of $n = 1000$, the type I errors are correct, but the coverage rates of the confidence bands are slightly underestimated. Indeed, sample sizes that do not ensure enough

Table 1

Simulation results. Empirical coverage probabilities of simultaneous confidence bands and type I error and power of tests with $\mathcal{H}_0 : \forall s \in \mathcal{S}\ \triangle\,\theta(s,t) = 0$, for comparing dynamic predictions $\pi^{(1)}(\cdot,t)$ and $\pi^{(1)}(\cdot,t)$ ($t = 5$, $s \in \{0, 1, \ldots, 10\}$, 1000 runs).

	Coverage rates for confidence bands				Type I error or power	
	$\mathrm{AUC}_{\pi^{(1)}}$	$\mathrm{BS}_{\pi^{(1)}}$	$\mathrm{AUC}_{\pi^{(2)}}$	$\mathrm{BS}_{\pi^{(2)}}$	$\triangle$AUC	$\triangle$BS
$n = 1000$						
H0.1	92.6	91.2	92.4	91.6	5.2	5.9
H0.2	92.9	93.2	92.0	92.8	4.1	4.1
H1.1.1	93.4	93.4	90.9	93.2	9.2	14.0
H1.1.2	92.8	93.2	90.9	93.2	14.7	31.1
H1.1.3	93.5	92.1	93.3	92.3	21.4	47.7
H1.2.1	91.3	92.7	94.1	91.8	39.9	53.3
H1.2.2	91.3	92.7	94.2	92.0	82.3	92.0
H1.3	91.2	93.4	93.1	93.1	87.4	78.0
$n = 2000$						
H0.1	94.2	93.5	93.9	93.9	5.7	5.7
H0.2	92.8	93.1	93.6	94.0	5.0	5.2
H1.1.1	94.7	95.5	93.6	95.9	10.5	21.6
H1.1.2	95.3	95.2	93.6	95.9	21.9	56.9
H1.1.3	93.9	93.4	92.3	93.4	39.5	82.7
H1.2.1	93.7	94.0	93.3	93.4	68.4	85.1
H1.2.2	93.7	94.0	94.2	92.9	98.5	99.9
H1.3	94.1	93.2	93.6	94.7	99.3	97.7
$n = 3000$						
H0.1	94.5	93.9	94.0	94.2	4.2	3.7
H0.2	93.0	95.4	93.7	94.3	4.7	4.7
H1.1.1	94.3	93.9	94.4	95.4	14.5	34.4
H1.1.2	94.1	94.3	94.4	95.4	34.0	80.7
H1.1.3	94.5	94.8	93.6	94.4	59.3	95.8
H1.2.1	94.7	94.9	93.7	93.9	85.6	95.6
H1.2.2	94.7	94.9	93.4	93.9	99.8	100.0
H1.3	92.9	94.4	93.2	94.2	100.0	99.9

subjects at risk at the late landmark times may lead to erroneous asymptotic normal approximations (as illustrated in Web Figure I).

The BS comparison tests seem more powerful than the AUC ones in these simulations. However, this difference in power may largely depend on the simulation setting and is not confirmed in our application displayed in Section 5. It could be due to the fact that, for the sake of simplicity, we have generated well-calibrated dynamic predictions.

Note that Figure 3 displays an increase in $\mathrm{AUC}(s,t)$ for an early increasing s that precedes a decrease. Even if the models accumulate more and more information when s increases and thus become more and more accurate, this decrease in discrimination is realistic and rational. Indeed, the decrease essentially reflects the fact that the selection mechanism for the at-risk population makes it naturally more and more homogeneous. Figure 2 also displays an increase in power with increasing s that precedes a decrease under the alternatives. This decrease may also be due to the selection mechanism by decreasing the difference between the two dynamic prediction models with increasing s, or to the decreasing number of subjects at risk.

As expected, the additional simulations including marker-dependent censoring illustrate the decreasing performances of the inference procedures with increasing departure from the independent censoring assumption (Web Table I). Interestingly and unlike for BS, inference procedures for AUC appeared rather robust, in accordance with previous results (Blanche et al., 2013).

5. Application to Prediction of Dementia Onset in the Elderly

The objective of this analysis is the quantification and the comparison of 5-year dynamic predictions of dementia obtained from two different joint latent class models.

The first one is based on repeated measurements of the Mini Mental Score Examination (MMSE). The second one is based on repeated measurements of the Isaacs Set Test shortened at 15 seconds (IST). The MMSE is a sum score evaluating various dimensions of cognition such as memory, calculation and language, and is a widely used test for dementia screening or for evaluating cognitive impairment in the elderly. The IST mainly evaluates speed of verbal production and has been shown to be particularly sensitive to small changes in cognition in the elderly. Roughly, the IST consists in asking a subject to give the longest list of words belonging to four specific semantic categories in 15 seconds (truncated to 10 words per category). Proust-Lima et al. (2007) provide more insights into these two psychometric tests.

The choice of a 5-year prediction horizon was guided by the clinical perspective of targeting subjects at high risk of dementia, who could benefit from prevention programs or treatments.

5.1. *Data from Paquid Cohort (Training Cohort) and 3C Cohort (Validation Cohort)*

Paquid and Three-City (3C) are two French population-based studies that investigate cognitive aging and dementia onset. They included 3777 and 9294 subjects aged 65 years and older and living at home at enrollment in 1988 and 1999, respectively. In both studies, subjects were seen approximately every 2 or 3 years. Each visit included evaluations of their cognitive abilities through a battery of cognitive tests and a standardized diagnosis of dementia (Dartigues et al., 1992; The 3C Study Group, 2003).

Subjects demented, blind, deaf or confined to bed at the initial visit were excluded. At the time of the analyses, 10-year follow-up data were not available for subjects from the city of Dijon. Only subjects from Bordeaux and Montpellier were thus included in our 3C sample. Final samples from Paquid and 3C data included $n = 2970$ and $n = 3880$ subjects with follow-up of 20 and 10 years, respectively.

Time-to-dementia was computed as the time elapsed between study entry and the mid-point between the time when the diagnosis was established and the time at the last visit without dementia. Subjects who died without a dementia diagnosis were considered as free of dementia at death if the last visit was less than 2 years before the death, and were considered as censored at the last visit if the duration

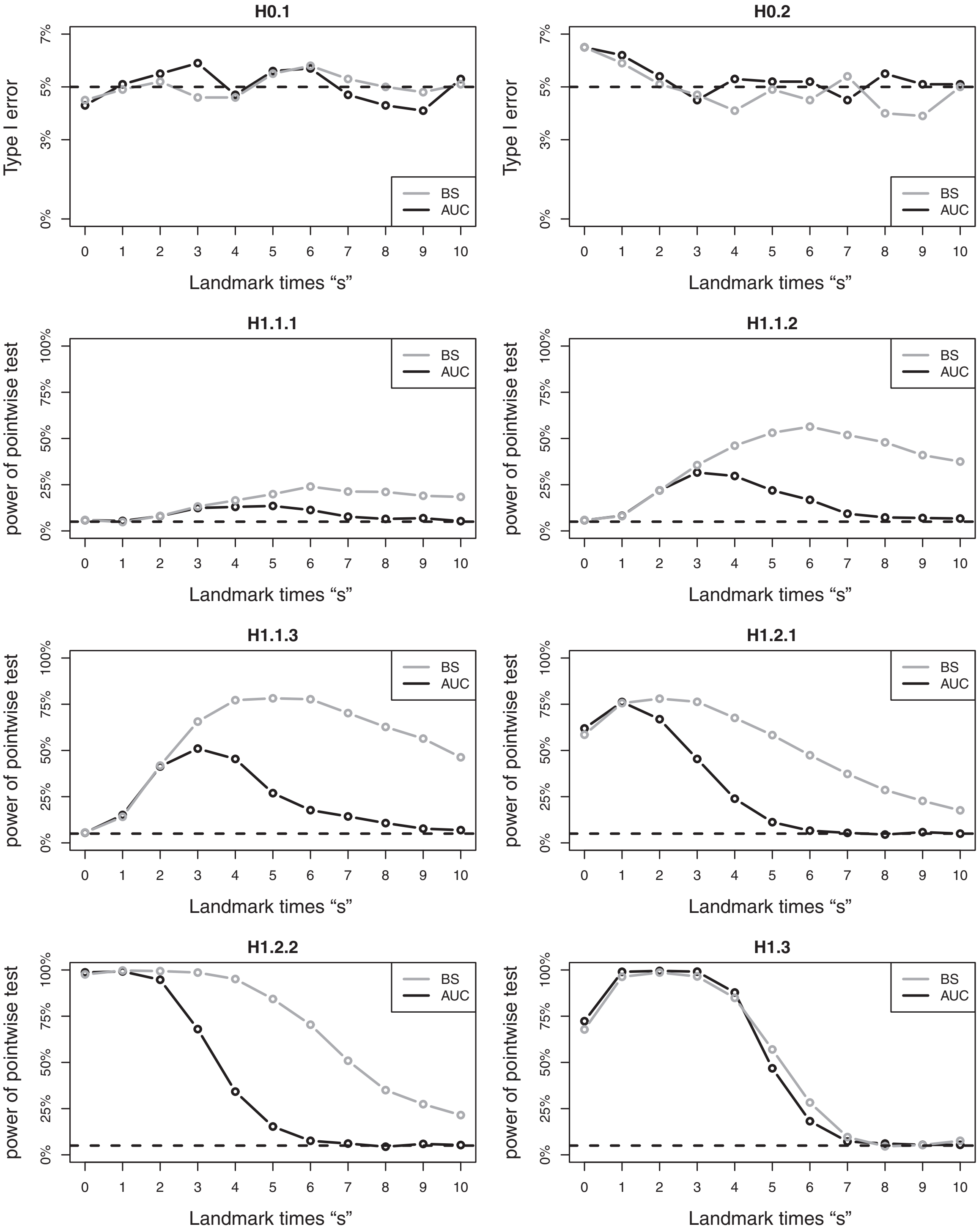

Figure 2. Type I error and power of pointwise tests for the difference in AUC$(s, t = 5)$ or BS$(s, t = 5)$ between the dynamic predictions $\pi^{(l)}(s, t = 5)$, $l = 1, 2$, when landmark time s changes, according to each simulation scenario. $n = 2000$ subjects, 1000 runs.

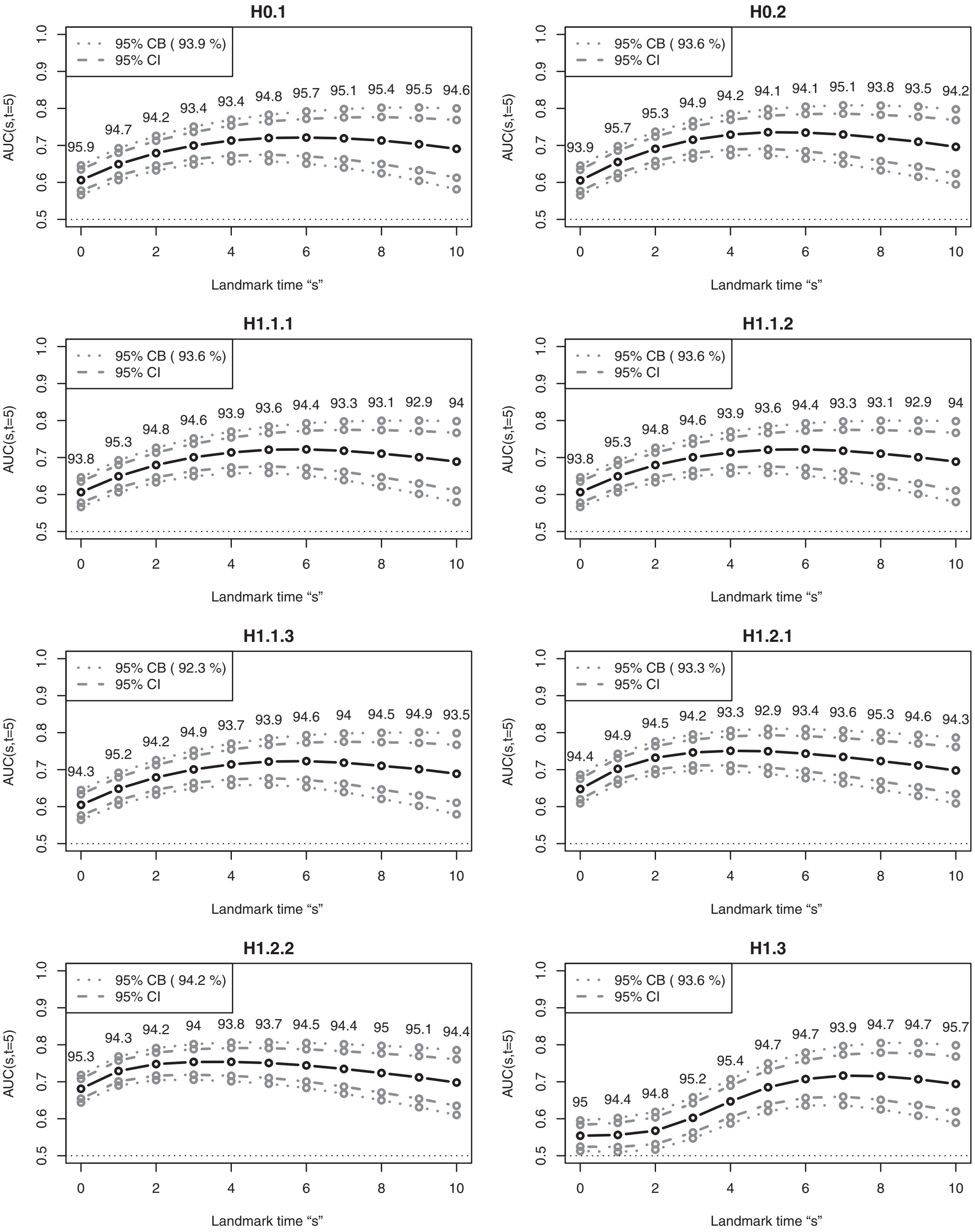

Figure 3. Average estimates of curve AUC$(s, t = 5)$ versus s for $\pi^{(2)}(s, t)$, $n = 2000$. Average 95% pointwise confidence intervals (CI) and 95% simultaneous confidence band (CB) are plotted from average standard errors and quantiles over 1000 runs. Empirical coverage probabilities of 95% CB are given in legends and 95% CI for each time s are displayed in plot.

between the last negative dementia diagnosis and death was longer.

The two samples are described in Web Table II. Age at inclusion (about 74 years) and gender frequencies (about 40% of male) were similar in the two cohorts. Level of primary education was higher in the 3C cohort (32.5% of high education) than in the Paquid cohort (8.6% of high education) and baseline cognitive scores were slightly higher in 3C than in Paquid. These differences were mainly due to the fact that 3C is a more recent cohort than Paquid and only includes subjects living in large cities.

5.2. *Fitting a Joint Latent Class Model Using the Paquid (Training) Sample*

After preliminary analyses, we chose the same model specification for the two joint latent class models based either on IST or on MMSE. Paquid data were used to fit the two corresponding models. Based on BIC criteria, we retained models with three classes for both IST and MMSE. Conditionally on each latent class $\gamma_i = g$, $g = 1, 2, 3$, we modeled the cognitive test trajectory of subject i by

$$\begin{aligned} Y_i(t_{ij})|_{\gamma_i=g} &= \beta_0 + \beta_{0,\mathrm{age}}\mathtt{AGE}_i + \beta_{0,\mathrm{educ}}\mathtt{EDUC_i} + \beta_{0,\mathrm{learn}}\mathbb{1}_{(\mathtt{t_{ij}=0})} \\ &\quad + b_{i0|\gamma_i=g} + \left(\beta_{1g} + \beta_{1,\mathrm{age}}\mathtt{AGE_i} + \mathtt{b_{i1|\gamma_i=g}}\right) \times \mathtt{t_{ij}} \\ &\quad + \left(\beta_{2g} + \beta_{2,\mathrm{age}}\mathtt{AGE_i} + \mathtt{b_{i2|\gamma_i=g}}\right) \times \mathtt{t_{ij}^2} + \varepsilon_{\mathtt{i}}(\mathtt{t_{ij}}), \end{aligned}$$

with $(b_{i0|\gamma_i=g}, b_{i1|\gamma_i=g}, b_{i2|\gamma_i=g}) \sim \mathcal{N}(\mathbf{0}, \sigma_g^2\mathbf{B})$, where $\mathbf{B}$ is unstructured, and an independent noise $\varepsilon_i(t_{ij}) \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. The binary variable `EDUC` and the continuous variable `AGE` represent primary education level and age at baseline. Note that the term $\beta_{0,\mathrm{learn}}\mathbb{1}_{(t_{ij}=0)}$ was added to account for the learning effect after the first interview (Jacqmin-Gadda et al., 1997) and that the variables `EDUC` and `AGE` have a common effect over classes.

For cause-specific hazards of dementia (denoted by event type 1) and dementia-free death (denoted by event type 2), we, respectively, modeled

$$\begin{aligned} \lambda_{i,1}(t|\gamma_i = g) &= \lambda_{01,g}(t)\exp\left(\alpha_{11,g}\mathtt{AGE}_i + \alpha_{21,g}\mathtt{EDUC}_i\right) \\ \text{and} \quad \lambda_{i,2}(t|\gamma_i = g) &= \lambda_{02,g}(t)\exp\left(\alpha_{12,g}\mathtt{AGE}_i + \alpha_{22,g}\mathtt{EDUC}_i \right. \\ &\quad \left. + \alpha_{32,g}\mathtt{SEX}_i\right). \end{aligned}$$

Class-specific baseline hazards of both event types $\lambda_{0k,g}(t)$, $k = 1, 2$, $g = 1, 2, 3$, were parametrized by Weibull hazard functions. We directly used IST repeated measurements as input in the first joint model, that is, as the $Y_i(t_{ij})$. However, in order to account for ceiling effects and the curvilinearity of the MMSE, we used a recently proposed monotonic transformation of the MMSE (Philipps et al., in press) as input of the second joint model. Parameters of the two models $\boldsymbol{\xi}^{(l)}$, $l = 1, 2$, were estimated by maximizing the corresponding log-likelihoods.

5.3. *Building Predictions for the Subjects in the 3C (Validation) Sample*

Using the joint latent class models fitted on Paquid data, we computed subject-specific predictions at landmark times $s = 0$, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, and 4 years with a prediction window of $t = 5$ years for all subjects in the 3C sample. According to 3C study design, dynamic predictions use one cognitive test measurement at $s = 0$ and up to three repeated measurements when s increases till $s = 4$. For $l = 1, 2$, using estimated parameters $\widehat{\boldsymbol{\xi}}^{(l)}$, baseline covariates $\mathbf{X} = (\mathtt{AGE}, \mathtt{SEX}, \mathtt{EDUC})$, and MMSE ($l = 1$) or IST ($l = 2$) repeated measurements collected by time s, denoted by $\mathcal{Y}_i^{(l)}(s)$, dynamic subject-specific predictions for subjects in the 3C sample were computed as:

$$\begin{aligned} \pi_i^{(l)}(s,t) &= \mathbb{P}_{\widehat{\boldsymbol{\xi}}^{(l)}}\left(s < T_i \le s+t, \eta_i = 1 | T > s, \mathcal{Y}_i^{(l)}(s), \mathbf{X}_i\right) \\ &= \frac{\sum_{g=1}^{3} \mathbb{P}_{\widehat{\boldsymbol{\xi}}^{(l)}}(\gamma_i = g)\mathcal{L}_{g,\widehat{\boldsymbol{\xi}}^{(l)}}(\mathcal{Y}_i^{(l)}(s)|\mathbf{X}_i)\left\{F_{1g,\widehat{\boldsymbol{\xi}}^{(l)}}(s+t|\mathbf{X}_i) - F_{1g,\widehat{\boldsymbol{\xi}}^{(l)}}(s|\mathbf{X}_i)\right\}}{\sum_{g=1}^{3} \mathbb{P}_{\widehat{\boldsymbol{\xi}}^{(l)}}(\gamma_i = g)\mathcal{L}_{g,\widehat{\boldsymbol{\xi}}^{(l)}}(\mathcal{Y}_i^{(l)}(s)|\mathbf{X}_i)S_{g,\widehat{\boldsymbol{\xi}}^{(l)}}(s|\mathbf{X}_i)}, \end{aligned}$$

where $\mathcal{L}_{g,\widehat{\boldsymbol{\xi}}^{(l)}}(\mathcal{Y}_i^{(l)}(s)|\mathbf{X}_i)$, $F_{1g,\widehat{\boldsymbol{\xi}}^{(l)}}(s|\mathbf{X}_i)$ and $S_{g,\widehat{\boldsymbol{\xi}}^{(l)}}(s|\mathbf{X}_i)$ are, respectively, the density of $\mathcal{Y}_i^{(l)}(s)$, the cumulative incidence of dementia at time s and the all-event-free survival function at time s conditionally on $\mathbf{X}_i$ and on the latent class $\gamma_i = g$ (see details in Web Appendix B).

5.4. *Quantifying and Comparing Predictive Accuracy on the (Validation) 3C Cohort*

Figure 4 shows that estimated AUCs for both curves corresponding to the two prediction models are high (range from 0.79 to 0.90). This suggests that the proposed dynamic prediction tools, based on joint latent class models using repeated cognitive test measurements, have a good predictive accuracy in terms of discrimination.

Although the proportion of observed subjects developing dementia within each time window $(s, s+t]$ is low (range from 4.7% to 7.6%), the numbers are large enough (range from 182 to 227) owing to the large sample size $n = 3880$ (see also Web Figure II). This makes the confidence regions relatively narrow. Pointwise confidence intervals of the differences in AUCs do not overlap the zero line. Consequently, for each landmark time $s \in \mathcal{S} = \{0, 0.5, 1, 1.5, \ldots, 4\}$ the null hypotheses $\mathcal{H}_0^{(s)} : \triangle\mathrm{AUC}(s,t) = 0$ are rejected with a significance level of $\alpha = 5\%$. As the simultaneous confidence bands do not contain the zero line, then the simultaneous null hypothesis $\mathcal{H}_0 : \forall s \in \mathcal{S}\ \triangle\ \mathrm{AUC}(s,t) = 0$ is also significantly rejected. More importantly, as the simultaneous confidence band of the differences in AUCs does not even overlap the zero line, we can also assert with a confidence level of 95% that predictions from the IST have a uniformly better prediction accuracy in terms of AUC over all landmark times $s \in \mathcal{S}$ than those from the MMSE, that is, $\mathbb{P}\left(\forall s \in \mathcal{S}\ \mathrm{AUC}^{(\mathrm{IST})}(s,t) > \mathrm{AUC}^{(\mathrm{MMSE})}(s,t)\right) \ge 95\%$. Although not significant, BS for predictions from the IST are also estimated to be lower than those from the MMSE (the lower the better).

Finally, as similarly discussed for simulation results in Section 4.2, the decreasing trends for $\mathrm{AUC}(s,t)$ with increasing s is probably the consequence of a selection process that makes the at-risk population more and more homogeneous as s increases.

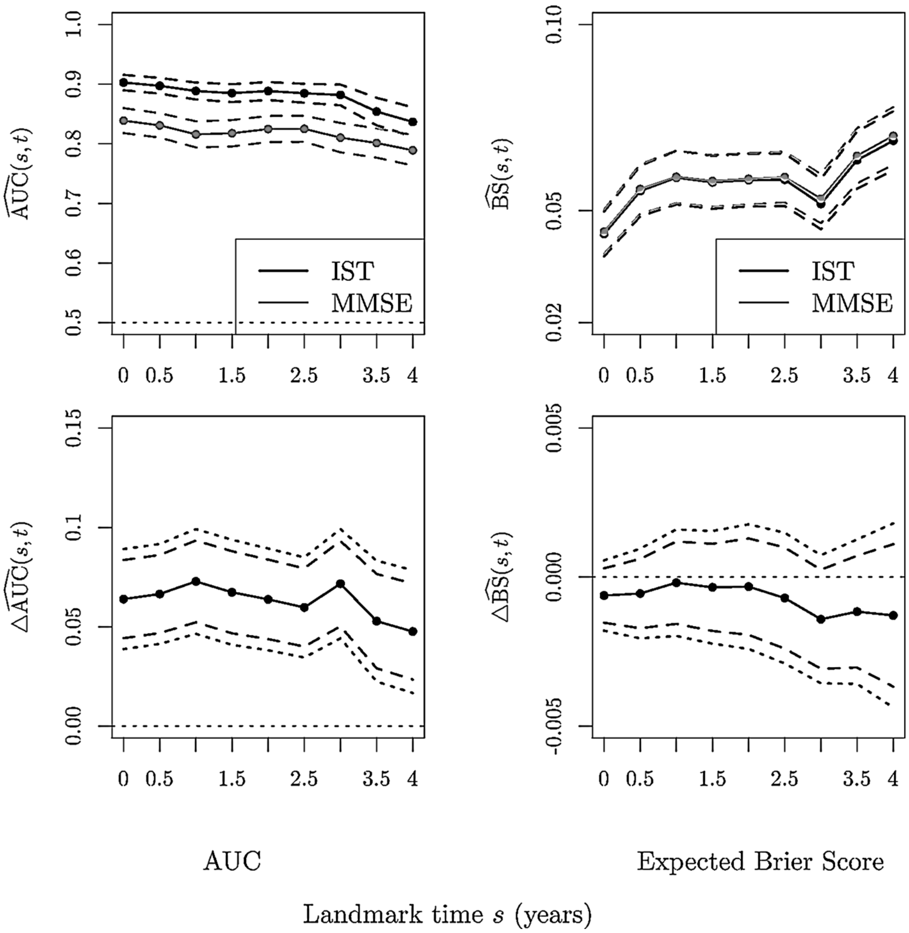

Figure 4. Comparison of predictive accuracy of the two predicted risks of dementia within time window $(s, s+t)$ when $s = \{0, 0.5, 1, 1.5, \ldots, 4\}$ and $t = 5$ years. 95% pointwise confidence intervals are displayed as dashed lines, 95% simultaneous confidence bands as dotted lines. 3C data, $n = 3880$ subjects.

6. Discussion

We propose two dynamic predictive accuracy curves to quantify and compare different dynamic risk prediction tools derived from the joint modeling framework. The dynamic AUC curve is easy to interpret and quantifies discrimination abilities. The dynamic BS curve also makes it possible to compare calibration. Estimators dealing with competing risk and censored data are proposed. Asymptotic results were established and we derived pointwise and simultaneous confidence regions as well as comparison tests that performed well in the simulation study. Furthermore, by applying the proposed methodology to cohort data, we show that: (i) dynamic predictions of dementia using repeated measurements of cognitive tests have a high predictive accuracy in terms of discrimination, and that (ii) the discrimination is uniformly better for the IST than for the MMSE.

The proposed methods assume independent censoring. This is reasonable for many applications. However, this assumption could also sometimes appear too restrictive. If only the baseline covariates are assumed to impact the censoring mechanism, the method could be extended by using a regression model instead of the Kaplan–Meier estimator for the weighting, as in Blanche et al. (2013) and Hung and Chiang (2010). The practical implementation could, however, become somehow challenging and require alternative bootstrap approaches. On the other hand, it seems difficult to provide a general methodology when the censoring distribution depends on a longitudinal biomarker without making unrealistic and uncheckable parametric assumptions.

A summary of the curve of AUC or BS for a range of landmark times s could appear appealing. However, due to the increasing selection of the subjects at risk with time s, it is not easy to define a meaningful summary measure and we think that curves will generally be more informative than a single value.

Although we focused on the competing risks setting, note that the proposed methodology is still relevant when there is only one type of event. In addition, in this more usual setting, the proportions of observed events in each prediction window is often higher and so the finite sample behaviors of the asymptotic procedures are better (results not shown).

Being model-free, our inference procedures do not assume any correct model specification. In addition, they allow the comparison of non-nested models and could also be applied beyond the joint modeling framework, for example, to rival semi-parametric landmarking approaches (van Houwelingen and Putter, 2012). This contrasts with some likelihood-based approaches, such as the comparison of prediction models using likelihood-ratio test or information criteria, such as AIC, BIC or other Kullback–Leibler divergence-based criteria (Commenges, Liquet, and Proust-Lima, 2012). Moreover, these likelihood approaches do not specifically aim to compare the specific t-year dynamic prediction abilities, but instead compare more globally the goodness-of-fit of the models or the prognostic abilities of the models, which potentially lead to different results and interpretations.

7. Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 2–5, as well as R codes and data used in the application are available at the *Biometrics* website on Wiley Online Library.

Acknowledgements

Computer time for this study was provided by the computing facilities MCIA (Mésocentre de Calcul Intensif Aquitain) of the Université de Bordeaux and of the Université de Pau et des Pays de l'Adour. This work was partly funded by a grant from France Alzheimer awarded to H. Jacqmin-Gadda in 2009. The Paquid study is funded by IPSEN and Novartis laboratories. The 3C Study supports are listed on the study website (www.three-city-study.com).

References

Aisen, P. S., Andrieu, S., Sampaio, C., Carrillo, M., Khachaturian, Z. S., Dubois, B., Feldman, H. H., Petersen, R. C., Siemers, E., Doody, R. S., Hendrix, S. B., Grundman, M., Schneider, L. S., Schindler, R. J., Salmon, E., Potter, W. Z., Thomas, R. G., Salmon, D., Donohue, M., Bednar, M. M., Touchon, J., and Vellas, B. (2011). Report of the task force on designing clinical trials in early (predementia) AD. *Neurology* **76**, 280–286.

Amieva, H., Le Goff, M., Millet, X., Orgogozo, J. M., Pérès, K., Barberger-Gateau, P., Jacqmin-Gadda, H., and Dartigues, J. F. (2008). Prodromal alzheimer's disease: Successive emergence of the clinical symptoms. *Annals of Neurology* **64**, 492–498.

Blanche, P., Dartigues, J.-F., and Jacqmin-Gadda, H. (2013). Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine* **32**, 5381–5397.

Commenges, D., Liquet, B., and Proust-Lima, C. (2012). Choice of prognostic estimators in joint models by estimating differences of expected conditional Kullback–Leibler risks. *Biometrics* **68**, 380–387.

Dartigues, J., Gagnon, M., Barberger-Gateau, P., Letenneur, L., Commenges, D., Sauvel, C., Michel, P., and Salamon, R. (1992). The Paquid epidemiological program on brain ageing. *Neuroepidemiology* **11**, 14–18.

Graf, E., Schmoor, C., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* **18**, 2529–2545.

Henderson, R., Diggle, P., and Dobson, A. (2002). Identification and efficacy of longitudinal markers for survival. *Biostatistics* **3**, 33–50.

Hung, H. and Chiang, C. (2010). Estimation methods for time-dependent AUC models with survival data. *Canadian Journal of Statistics* **38**, 8–26.

Jacqmin-Gadda, H., Fabrigoule, C., Commenges, D., and Dartigues, J.-F. (1997). A 5-year longitudinal study of the mini-mental state examination in normal aging. *American Journal of Epidemiology* **145**, 498–506.

Martinussen, T. and Scheike, T. (2006). *Dynamic Regression Models for Survival Data*. New York: Springer.

McIntosh, M. and Pepe, M. (2002). Combining several screening tests: Optimality of the risk score. *Biometrics* **58**, 657–664.

Parast, L., Cheng, S.-C., and Cai, T. (2012). Landmark prediction of long-term survival incorporating short-term event time information. *Journal of the American Statistical Association* **107**, 1492–1501.

Philipps, V., Amieva, H., Andrieu, S., Dufouil C.and Berr, C., Dartigues, J.-F., Jacqmin-Gadda, H., and Proust-Lima, C. (in press). Normalized mini-mental state examination for assessing cognitive change in population-based brain aging studies. *Neuroepidemiology.*

Proust-Lima, C., Amieva, H., Dartigues, J.-F., and Jacqmin-Gadda, H. (2007). Sensitivity of four psychometric tests to measure cognitive changes in brain aging-population-based studies. *American Journal of Epidemiology* **165**, 344–350.

Proust-Lima, C., Mbéry, S., Taylor, J. M. G., and Jacqmin-Gadda, H. (2014). Joint latent class models for longitudinal and time-to-event data: A review. *Statistical Methods in Medical Research* **23**, 74–90.

Proust-Lima, C. and Taylor, J. M. G. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment psa: A joint modeling approach. *Biostatistics* **10**, 535–549.

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* **67**, 819–829.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-event Data: With Applications in R*, Vol. 6. Boca Raton: Chapman & Hall.

Schoop, R., Beyersmann, J., Schumacher, M., and Binder, H. (2011). Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. *Biometrical Journal* **53**, 88–112.

Schoop, R., Graf, E., and Schumacher, M. (2008). Quantifying the predictive performance of prognostic models for censored survival data with time-dependent covariates. *Biometrics* **64**, 603–610.

The 3C Study Group (2003). Vascular factors and risk of dementia: Design of the three-city study and baseline characteristics of the study population. *Neuroepidemiology* **22**, 316–325.

van Houwelingen, H. and Putter, H. (2012). *Dynamic Prediction in Clinical Survival Analysis.* Boca Raton: CRC Press.

Zheng, Y., Cai, T., Jin, Y., and Feng, Z. (2012). Evaluating prognostic accuracy of biomarkers under competing risk. *Biometrics* **68**, 388–396.

Received September 2013. *Revised June* 2014.
Accepted August 2014.