Universiteit
Leiden
The Netherlands

Estimation of incident dynamic AUC in practice

Geloven, N. van; He, Y.; Zwinderman, A.H.; Putter, H.

Citation
Geloven, N. van, He, Y., Zwinderman, A. H., & Putter, H. (2021). Estimation of incident dynamic AUC in practice. *Computational Statistics & Data Analysis*, *154*. doi:10.1016/j.csda.2020.107095

Version:	Publisher's Version
License:	Creative Commons CC BY 4.0 license
Downloaded from:	https://hdl.handle.net/1887/3276346

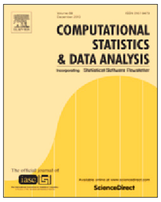
Note: To cite this publication please use the final published version (if applicable).

Computational Statistics and Data Analysis 154 (2021) 107095

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda



Estimation of incident dynamic AUC in practice[☆]

N. van Geloven [a,*], Y. He [a], A.H. Zwinderman [b], H. Putter [a]

[a] Department of Biomedical Data Sciences, Leiden University Medical Center, PO box 9600, zone S5-P, 2300 RC Leiden, The Netherlands

[b] Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Amsterdam UMC, University of Amsterdam, Meibergdreef 9, AZ, Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 13 December 2019
Received in revised form 8 September 2020
Accepted 9 September 2020
Available online 18 September 2020

Keywords:
Time-dependent AUC
Dynamic survival model
Discriminative ability
Concordance index
Prediction 1

ABSTRACT

The incident/dynamic time-dependent AUC (Area Under the ROC Curve) is an appealing measure to express the discriminative value of a dynamic survival model over time. However, estimation of this measure is not straightforward. Four recently proposed estimation approaches are studied. In an extensive simulation study, a head-to-head comparison between these four estimation methods is made. The estimation algorithms of some of the methods are extended. Results are illustrated with a motivating dynamic survival model from Reproductive Medicine.

© 2020 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Dynamic prediction models that can incorporate longitudinal covariate data and allow temporal updating of predictions are becoming increasingly popular. Such models allow repeated predictions over time using all information available up to that time point. Several methodologies have been developed for dynamic prediction of time-to-event outcomes, including landmark approaches and joint models (Van Houwelingen and Putter, 2011; Rizopoulos, 2012). As a logical result of the introduction of these dynamic survival models, attention has been given to the evaluation of the accuracy of their predictions. We focus here on the discriminative ability, which reflects the desired ability of the prediction model to assign higher predictions to patients with events ('cases') than to patients without events ('controls').

Due to the nature and intended repeated use over time of a dynamic survival model, global discrimination measures such as Harrell's c-index (Harrell et al., 1982) are not sufficient for their evaluation. In particular, the discriminative ability of a dynamic prediction model may vary over the course of time. Sole use of a global measure would obscure this fact and could falsely suggest acceptable discrimination over the whole time range. We thus need time-dependent measures of discrimination. The challenge in assessing time-dependent discrimination of a dynamic survival model, is that the sets of cases and controls change over time. Several time-dependent discrimination measures have been proposed in the literature, each one choosing a different definition of the case and control set. In this article, we consider the incident/dynamic time-dependent area under the Receiver Operator Characteristic curve (I/D AUC) which was first defined by Heagerty and Zheng (2005). This measure compares the predictions of incident cases, defined as patients who experience the event at the time point where we want to assess the discriminative ability, with dynamic controls,

[☆] The online version of this article contains supplementary material: appendix I contains the analysis code, appendix II contains the simulation results of all simulation scenarios.

[*] Corresponding author.
E-mail address: n.van_geloven@lumc.nl (N. van Geloven).

https://doi.org/10.1016/j.csda.2020.107095
0167-9473/© 2020 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

defined as patients who have not yet experienced the event of interest up to and including that time point. Though the cumulative definition of cases, i.e. defined as patients who experience their event within a certain time range, may be clinically relevant, most methodologists use incident cases instead (Pepe et al., 2008). One appealing property of this measure is that the evaluation of the model at a certain time point does not take into account patients who had their event earlier. As prediction models target the future (and not the past), this property is very useful when evaluating dynamic prediction models. A second appealing property is that the focus on incident cases is better suited for picking up trends of AUC values over time compared to cumulative AUC definitions. With cumulative case definition adjacent AUCs over time will be based on partly overlapping information and that may lever out time trends (Blanche et al., 2013b; Pepe et al., 2008). Thirdly, the I/D AUC allows a direct way to incorporate longitudinal markers (for example blood markers that change in value over time or risk scores from a dynamic prediction model) by plugging in the marker value at time t. The measures defined by cumulative cases may require joint modeling with an extra linear mixed model to evaluate the discriminative ability of a longitudinal marker (Rizopoulos, 2011). For simplicity we use a time fixed marker in our data application and simulations in this paper, but all I/D AUC methods we present are able to handle time-varying markers.

Estimation of the I/D AUC is not straightforward. In particular, because the number of events at a certain time point is generally low. If the event occurrence is measured in continuous time, at most one event per time point is present. Usually, event times are recorded with some degree of rounding, e.g. whole number of days since randomization, or whole number of years since diagnosis. In such situations there may be more than one case per time point, but still usually few. Estimating sensitivity, one of the components of the I/D AUC, for a small event number will in general lead to unstable results. In the most extreme situation, the AUC could jump from 1 to 0 and back between adjacent time points. Therefore, estimation of the I/D AUC requires some degree of smoothing (Saha-Chaudhuri and Heagerty, 2013).

In recent years, several methods have been proposed to estimate I/D AUC. Each method makes a different choice for how and when it applies the necessary smoothing. Heagerty and Zheng impose the smoothing by assuming a Cox model for the relation between the predictor under evaluation and the event (Heagerty and Zheng, 2005). For each event time point they then estimate the true positive component of the area under the ROC curve based on this Cox model. The smoothing is thus introduced before the actual AUCs are estimated. Van Houwelingen and Putter propose a model-free alternative where the smoothing is imposed at a later stage. They first estimate unsmoothed intermediate AUCs at each event time point and then use lowess regression for the estimation of a smoothly varying curve (Van Houwelingen and Putter, 2011). Similar non-parametric two-step 'afterward' approaches have been proposed by Song et al. (2012) and later on by Saha-Chaudhuri and Heagerty (2013). Both use kernel functions for generating a smoothly varying curve from unsmoothed intermediate AUC estimates. The main difference between those two methods is that Song et al. additionally use inverse probability of censoring weighting in their algorithm, to derive estimates that are robust to marker-dependent censoring, i.e. censoring that depends on the value of the predictions under evaluation. Though this addition can be very useful (Blanche et al., 2013a,b), to enhance comparability with the other methods that do not actively model the censoring distribution, we focus hereafter on the unweighed version of the kernel based estimator as formulated by Saha-Chaudhuri and Heagerty. We thus assume throughout this paper that the censoring distribution is independent of the values of the predictions. A fourth way of introducing the smoothing has been proposed by Shen et al. (2015). They propose a single-step approach where the AUC curve is expressed as a combination of fractional polynomials and the weights of these polynomials are estimated directly through a pseudo-likelihood optimization. The estimation of the AUC and the smoothing is thus performed simultaneously.

Each of the aforementioned papers describes the advantages presented by the newly introduced method and some compare their own method to one of the other methods in a simulated data scenario. So far it remains unclear what criteria one should follow to choose the most appropriate estimation method in a particular applied setting. One study compared estimators of the global 'summary' discrimination index (c-index) that can be derived from the I/D AUC, but no comparison for estimation of the I/D AUC over time exists (Schmid and Potapov, 2012). Our study was set up on that perspective. In Section 2 we introduce a motivating dynamic prediction model from Reproductive Medicine. In Section 3 we introduce notation, formally define the I/D AUC, describe the four estimation methods and propose extensions for some. In Section 4 we present results of an extensive simulation study comparing the performance of the estimation methods in many different scenarios. In Section 5 we apply the four methods to the motivating dataset to illustrate results. In Section 6 we provide recommendations for analysis choices in new applications and we end with a discussion in Section 7.

2. Introduction to the motivating dataset

Van Eekelen et al. developed a dynamic prediction model for the chance of natural conception leading to ongoing (beyond 12 weeks) pregnancy in couples diagnosed with unexplained subfertility (Van Eekelen et al., 2017). These are couples who have been trying to become pregnant for longer than a year and for whom no medical cause for their fertility problem can be found. Data used for the development of this model were collected in a prospective cohort study including 4999 couples who, after being diagnosed with unexplained subfertility, continued their attempts to conceive without medical interventions (expectant management). The predictions from this model can be used in clinical practice to counsel couples about their prognosis of a natural pregnancy after various periods of unsuccessful expectant management. The outcome variable was measured in discrete time as the number of days until natural conception, measured from the

Table 1
Properties of the four estimation methods.

	Heagerty and Zheng (2005)	Van Houwelingen and Putter (2011)	Saha-Chaudhuri and Heagerty (2013)	Shen et al. (2015)
Model dependence	Semi-parametric	Non-parametric	Non-parametric	Semi-parametric
Required choices	Model and neighborhood span for Local Cox and Schoenfeld models	Lowess neighborhood span	Kernel function and neighborhood span	Degrees of the polynomials and link function
Ties handling	Method generalizes to tied event times	Algorithm can be generalized to deal with ties	Method generalizes to tied event times	Direct application to tied event times possible
Software	R package RisksetROC	R package dynpred (only untied data)	None	None

moment of diagnosis on. 1053 (21%) women reached a natural conception within a mean follow-up of 239 days (5th and 95th percentiles: 21–636 days). The dynamic model proposed by Van Eekelen predicts the chance of natural conception at various time points after diagnosis based on female age, the duration of subfertility, female subfertility being primary or secondary, sperm motility and referral status and the amount of time that has elapsed since diagnosis.

Currently, the point of interest is to assess the accuracy of the predictions over time of this dynamic model in terms of discrimination. In the original publication, discrimination was only calculated at four time points. Here we want to know the discriminative ability over the full time range, as the model is proposed for repeated use at multiple points in time. During model development authors used a beta-geometric model for time-to-pregnancy (Van Eekelen et al., 2017). Owing to this approach, the rank order of patients' predictions is the same at each time point. Since only the ordering of predictions influences discrimination, we could treat the predictions as a time fixed covariate. Due to small numbers in follow up in the later years, Van Eekelen et al. propose the model for the first 2.5 years after diagnosis. The discriminative ability in these first 2.5 years is thus of primary clinical interest. Here we will evaluate the model on the full 5 year time range to also be able to study the behavior of the estimators with sparse data.

3. Methods

3.1. Notation and definition of I/D AUC

The unit of analysis is a couple trying to become pregnant. Each couple i has an event time T_i (number of days between diagnosis and conception), and a censoring time C_i (number of days between diagnosis and last follow up). The minimum of T_i and C_i is observed: $\tilde{T}_i = min(T_i, C_i) \in \{1, 2, \dots, t_k, \dots, t_{max}\}$, with $\delta_i = I(T_i < C_i)$ the event indicator. We denote the prediction for the ith couple as X_i. Note that we do not put any restriction on the model that generated the predictions, i.e., X may be a single marker, or maybe a summary prognosis (risk score) generated by any (non-)parametric dynamic regression or predictive method. We assume throughout that the higher X_i, the more likely couple i will experience the event.

The incident/dynamic time-dependent concordance index as proposed by Heagerty and Zheng (2005) assesses the concordance of the predictions at time point t_k among *incident* cases, i.e. subjects with $T_i = t_k$, and *dynamic* controls, i.e. subjects with $T_i > t_k$ and is defined as:

$$\text{I/D AUC}(t_k) = P(X_i > X_j | T_i = t_k, T_j > t_k).$$

In words: for any two subjects for whom one subject gets the event at t_k, while the other is still event free, what is the probability that the model correctly assigns the subject with the event a higher prediction at that time point, i.e. a higher X, than the other subject? As noted in Heagerty and Zheng (2005), this definition allows using time-dependent predictions, so X_i may be written as $X_i(t_k)$. For brevity we use the shorter notation X_i here.

Furthermore, let $R(t_k)$ denote the risk set at time t_k, i.e. the set of indices of all patients with $T_i \geq t_k$ and $C_i > t_k$, with size $n(t_k)$. The risk set is split into two: $R(t_k) = \{R_0(t_k) \cup R_1(t_k)\}$, with $R_1(t_k)$ containing the indices of cases who have an event at t_k and $R_0(t_k)$ containing the indices of the controls who are still event free by t_k. The corresponding sizes are denoted by $n_1(t_k)$ and $n_0(t_k)$. The total number of couples included in the study is $n(t_0)$.

3.2. Estimators

Below we describe the four estimation methods briefly, for more details we refer to the original papers (Heagerty and Zheng, 2005; Van Houwelingen and Putter, 2011; Saha-Chaudhuri and Heagerty, 2013; Shen et al., 2015). A schematic description of the properties of the four methods is depicted in Table 1.

(a) *Heagerty and Zheng 2005*. Heagerty and Zheng first estimate ROC curves at each time point t_k where an event occurs and then calculate areas under these curves by numerical integration. The ROC curves consist of pairs of incident true positive fraction TPF and dynamic false positive fraction FPF at each cutpoint $X = c$. FPF can be estimated empirically because we usually have a sufficient number of control patients:

$$\hat{\text{FPF}}(t_k, c) = \hat{P}(X_i > c | T_i > t_k) = \sum_{i \in R_0(t_k)} 1\{X_i > c\}/n_0(t_k).$$

The true positive rate is harder to estimate since we only have one or few incident cases at a particular time point t_k. To assess this quantity, Heagerty and Zheng propose to first use a Cox proportional hazards model to estimate the relation between X and the event: $\lambda(t_k|X_i) = \lambda_0(t_k)\exp(X_i\gamma(t_k))$, with λ_0 the baseline hazard rate and $\gamma(t_k)$ a possibly time-varying regression coefficient. If $\gamma(t_k)$ is assumed time-varying, it can be estimated using either a “LocalCox” or a “Schoenfeld” smoothing algorithm. The Cox model can estimate the distribution of a covariate X given an event occurred at t_k (Xu and O’Quigley, 2000), which is needed for estimation of TPF:

$$\hat{\text{TPF}}(t_k, c) = \hat{P}(X_i > c | T_i = t_k) = \frac{\sum_l 1\{X_l > c\} 1\{\tilde{T}_l \geq t_k\} \exp(X_l \gamma(t_k))}{\sum_m 1\{\tilde{T}_m \geq t_k\} \exp(X_m \gamma(t_k))}.$$

This provides a semi-parametric estimator for TPF which can directly be applied to discrete survival times (Heagerty and Zheng, 2005).

An implementation of these estimators is available in the R package `RisksetROC`. For the LocalCox and Schoenfeld algorithms a choice has to be made for the size (span) of the local neighborhood. The suggested span is $n^{-0.2}$, with n the number of observed event times.

(b) *Van Houwelingen and Putter 2011*. Van Houwelingen and Putter (2011) present a model-free alternative for the estimation of the I/D AUC. Assuming no tied event times, they first compute an intermediate area A at each time point t_k where a subject i experiences an event as:

$$A(t_k) = \frac{\sum_{j \in R(t_k), j \neq i} \mathbf{1}\{X_j < X_i\} + 0.5 \times \mathbf{1}\{X_j = X_i\}}{n(t_k) - 1}.$$

After plotting $A(t_k)$ against t_k, a lowess smoothing curve is used to provide a non-parametric summary of the I/D AUC over time.

We here extend the I/D AUC definition of Van Houwelingen and Putter to accommodate the case of discrete event times with ties and compute the area A as:

$$A(t_k) = \frac{\sum_{j \in R_1(t_k)} \sum_{i \in R_0(t_k)} \mathbf{1}\{X_j < X_i\} + 0.5 \times \mathbf{1}\{X_j = X_i\}}{n_0(t_k) \times n_1(t_k)}. \tag{1}$$

For the lowess smoothing curve a choice needs to be made on how to weigh these extended intermediate estimates $A(t_k)$ as they are based on different numbers of observations (cases and controls) at risk. We explore three different weighing schemes: using equal weights for each $A(t_k)$ (w1), using the number of pairs (i.e. $n_0 \times n_1$) (w2), and using the size of the case set n_1 (w3).

The original implementation of the I/D AUC for untied data is available in the R package `dynpred`. We modified the code of that package to allow for the extension to tied event times.

(c) *Saha-Chaudhuri and Heagerty 2013*. The method proposed by Saha-Chaudhuri and Heagerty starts in the same way as the method by Van Houwelingen and Putter in the sense that it first estimates $A(t_k)$ at each event time using Eq. (1) and then uses these intermediate estimates as input for the construction of a smooth estimator. In the original formulation in Saha-Chaudhuri and Heagerty (2013) $A(t_k)$ does not include the addition of 0.5 for cases and controls with tied predictions, but we extend the definition here to allow situations with tied predictions. Where Van Houwelingen and Putter use the lowess smoother, Saha-Chaudhuri and Heagerty propose a locally weighted mean rank (WMR) estimator for combining the intermediate estimates:

$$\text{WMR}(t) = \frac{1}{|N_t(h_n)|} \sum_{t_j \in N_t(h_n)} A(t_j), \tag{2}$$

with $N_t(h_n) = \{t_j : |t - t_j| < h_n\}$. In words: to estimate the AUC at a certain time point t, the $A(t_k)$ at the event time points in a neighborhood of t are averaged. Note that in this formulation each unique event time point receives equal weight, independent of the number of events occurring, i.e., equal weight for each $A(t_k)$. A more general definition using a kernel weighted local average is also provided:

$$AUC^{\text{SH}}(t) = \sum_j K_{h_n}(t - t_j) A(t_j), \tag{3}$$

where K_{h_n} is a standardized kernel function such that $\sum_j K_{h_n}(t - t_j) = 1$. The WMR estimator (2) can be deduced from (3) by using the uniform (rectangular) kernel function.

To our knowledge no software implementation of the proposed methods is available, so we wrote our own implementation (available in Appendix I).

(d) *Shen et al. 2015.* Shen et al. (2015) propose to model the I/D AUC, after transformation with link function η, as a parametric function of time using fractional polynomials:

$$\eta(AUC(t,\beta)) = \sum_{l=1}^{L} \beta_l t^{(p_l)},$$

where for $l = 1, \ldots, L$,

$$t^{(p_l)} = \begin{cases} t^{p_l} & \text{if} \quad p_l \neq 0 \\ \ln(t) & \text{if} \quad p_l = 0, \end{cases}$$

with $p_1 \leq p_2 \leq, \ldots, \leq p_L$ real-valued powers. The regression parameters $\beta = \beta_1, \ldots, \beta_L$ are then estimated by optimizing a pseudo-likelihood function that is based on the observation that the number of concordant controls belonging to each event occurring at time point t_k can be viewed as a realization of a binomial distribution with success probability $AUC(t_k, \beta)$, thus by maximizing:

$$L(\beta) \propto \prod_{k=1}^{k=K} AUC(t_k,\beta)^{c(t_k)}\{1 - AUC(t_k,\beta)\}^{d(t_k)},$$

with K the total number of event time points,

$$c(t_k) = \#\{i, j \in R(t_k); X_j < X_i, \tilde{T}_i = t_k, \delta_i = 1, \tilde{T}_j > t_k\},$$

the number of concordant controls at event time point t_k and

$$d(t_k) = \#\{i, j \in R(t_k); X_j \geq X_i, \tilde{T}_i = t_k, \delta_i = 1, \tilde{T}_j > t_k\},$$

the number of non-concordant controls at event time point t_k.

Choices that need to be made during the estimation are: the link function η, with logit as obvious candidate; the degrees of the polynomials, recommended to at most 2 (quadratic) or 3 (cubic) (Royston and Altman, 1994); possibly a cut-off point τ that determines up to which time point there is sufficient data for accurate estimation of the $AUC(t)$, for instance the 90th or 95th percentiles of the observed event times. The way we defined $c(t_k)$ and $d(t_k)$ above directly allows for tied event times, avoiding the adding of a small amount of time to one of the tied cases as advised in Shen et al. (2015).

No dedicated software package exists for this algorithm. The supplementary material of Shen et al. (2015) provides code where the pseudo-likelihood function is maximized by the *optim* function of R. However, with this *optim* implementation we experienced an undesirable high dependence on the choice of initial values and optimizer method. Therefore we wrote an alternative implementation of the pseudo-likelihood optimization step exploiting the general *glm* function. A second extension we make to this model is the use of natural cubic spline base functions instead of fractional polynomials, as polynomials could show unstable behavior at the boundaries of the time axis. We used the *ns* function of the R package `splines` for this and again performed the optimization with *glm*. The parameters that must be chosen in this approach are the number and the location of the knots of the spline.

4. Simulation study

In this section, we compare the aforementioned methods in a series of simulation scenarios. We assume a simple setting where there is one baseline marker predictive of time to event. The aims of the simulation are: (i) to compare the accuracy of the methods, i.e., compare the bias, empirical standard deviation and root mean squared error (RMSE) of their estimates of the I/D AUC over time; (ii) to investigate the influence of several data aspects (strength and proportionality of the marker, tied data, amount of censoring, sample size) on the accuracy of the methods; (iii) to compare within each method the accuracy under different configuration choices of span/bandwidth, kernel shape, type of base functions, link function, etc.

4.1. Simulation setup

The parameters of the data generation model were chosen based on estimates from the motivating dataset presented in Section 2. Thirty-two simulation scenarios were constructed, in which we varied the following five aspects that might influence the performance of the estimation methods.

(a) Proportional hazards. We generated two sets of scenarios: one with proportional hazards and one without. Under the assumption of proportional hazards the event times were generated from a Cox model $\lambda(t|X) = \lambda_0(t) \cdot exp(X\beta)$, where $X \sim N(0, 1)$ was the marker and $\lambda_0(t)$ was a Weibull baseline hazard estimated from the empirical data and β was the coefficient of marker X. The second set of scenarios assumed a bivariate normal distribution for the log of event times and markers values: $N_2(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$, where μ_1 and σ_1 were the mean and standard deviation of log event times, μ_2 and σ_2 were the mean and standard deviation of the marker X, ρ was the correlation between log event time and marker. Here $log(t) \sim N(\mu_1 = 3.6, \sigma_1 = 1.75)$ (estimated from the empirical data), and marker $X \sim N(\mu_2 = 0, \sigma_2 = 1)$. Simulations using such a bivariate normal distribution were also performed in Heagerty and Zheng (2005), Saha-Chaudhuri and Heagerty (2013) and Shen et al. (2015).

(b) Effect size of the marker. In the set of scenarios based on the bivariate normal distribution, $\rho = -0.3$ and $\rho = -0.7$ were used. The negative correlations between the log event times and the marker indicated a positive predictive effect of the marker, meaning that subjects with higher marker values are expected to experience the event earlier. For the scenarios based on the Cox model, a small effect size $\beta = 0.448$ and a large effect size $\beta = 1.25$ were considered. The strength of these effect sizes was comparable with the ρ values.

(c) Percentage of censoring. Data with higher censoring percentages might have less events at later time points, which could lead to more challenging estimation. We considered independent censoring times generated from an exponential distribution. The empirical data contained 80% censoring. Besides this percentage, we also considered 20% censoring for contrast.

(d) Presence of ties. Tied event times are common in real-life data. The motivating dataset also contained ties (patients who conceived at the same number of days after diagnosis). We therefore generated scenarios both with and without tied event times. The tied event times were generated by rounding the simulated event times to the nearest lower integer value. This resulted in the % of tied event times ranging between 65 and 98% for the different scenarios (Table 1 in Appendix II).

(e) Sample size. Two sample sizes (n = 500 and n = 1000) were considered. We used 500 simulation runs for each simulation scenario.

The following estimation approaches were considered. For Heagerty and Zheng (2005), we evaluated the standard Cox model and the Cox model with time varying coefficient estimated based on a Schoenfeld smoothing algorithm with suggested span ($n^{-0.2}$). For the method of Van Houwelingen and Putter (2011), two different spans (0.33 and 0.66) for the lowess smoothing step were considered. In the tied data scenarios we evaluated the three different weighting approaches (w1, w2 and w3). For the method of Saha-Chaudhuri and Heagerty (2013), three different kernels (uniform, triangular and elliptic (Epanechnikov)) and two different bandwidths (6 months and 13 months) were adopted. For Shen et al. (2015), we used 2 and 3 as maximum degree of the polynomials. We also evaluated the use of natural cubic splines as base functions instead of fractional polynomials. We chose to place the knots at the 5th, 50th and 95th percentiles of the observed unique events times (knot = 1) and, in a second analysis, at the 5th, 50th and 95th percentiles of all (possibly duplicated) observed event times (knot = 2). For the untied data these two procedures result in the same knot locations.

The simulations were conducted in R (version 3.3.3). Codes for data generation and estimation methods can be found in Appendix I.

4.2. Simulation results

In this section, we first compare the accuracy of the methods on the two main scenarios that resemble the motivating data the most, namely the Cox model based scenario with $\beta = 0.448$ and the bivariate normal based scenario with $\rho = -0.3$, both with 80% exponential censoring, including ties and $n = 1000$. Secondly, we discuss the results from the other scenarios focusing on how different data settings may lead to different performance of the methods. Thirdly we focus on the performance of different configuration choices within each method.

Comparison of methods in the two main data scenarios

Tables 2 and 3 show the results from the main Cox model based scenario and the main bivariate normal based scenario respectively. For each time point (row), the lowest RMSE is indicated in bold. The bias was negligible for most estimation methods in both scenarios. Exceptions were the ‘standard Cox’ method of Heagerty and Zheng that showed a slight bias in the bivariate normal scenario where the proportional hazards assumption did not hold and Shen’s method that led to biased results in both scenarios at later time points (from month 26 onwards) where there were few observed events, both when using fractional polynomials and when using splines as base functions. In terms of RMSE, the ‘standard Cox’ method of Heagerty and Zheng was the most accurate method. In the Cox based scenario it had the lowest RMSE at all time points. In the bivariate normal scenario, for which the standard Cox model is misspecified, the low standard deviation of this method compensated the bias leading to the lowest RMSE at later time points. At earlier time points where there are more events, Shen’s method with spline base functions had lowest RMSE, with the standard Cox model at roughly second place. However, since Shen’s method showed considerable bias at later time points, the standard Cox model showed best results over the full time range. The non-parametric methods of Van Houwelingen and Putter and Saha-Chaudhuri showed low bias in both scenarios. However, due to a larger empirical standard deviation, the RMSEs were less favorable than those from the standard Cox model, especially at later points.

How do data settings influence accuracy?

Results from the other simulation scenarios are presented in Appendix II.

Proportionality of the marker: As in the main scenarios, the ‘standard Cox’ model of Heagerty and Zheng performed best in nearly all Cox based scenarios and also in the bivariate normal scenarios under heavy censoring (80%). When the relation between the marker and the event times was not proportional and when there was sufficient data to estimate this non-proportionality, other methods outperformed the method of Heagerty and Zheng. While no clear single winner

Table 2
Simulation results: Cox model based scenario with ties, n = 1000, $\beta = 0.448$, 80% censoring.

Time		H & Z 2005		S & H 2013				vH & P 2011				Sh 2015			
		Default		$h_n = 6$			$h_n = 13$	sp = 0.33			sp = 0.66	Frac.Poly(glm)		Spline	
		Cox	Residual	uni	elp	tri	uni	w1	w2	w3	w1	deg. = 2	deg. = 3	knot = 1	knot = 2
1	Bias($\times 10^2$)	0.2	0.1	−0.2	−0.1	−0.1	−0.2	0.0	−0.1	−0.1	−0.2	0.0	0.0	0.1	0.0
	SD	0.019	0.056	0.025	0.025	0.025	0.025	0.030	0.059	0.050	0.028	0.093	0.125	0.058	0.027
	RMSE	**0.019**	0.056	0.026	0.025	0.025	0.025	0.030	0.059	0.050	0.028	0.093	0.125	0.058	0.027
6	Bias($\times 10^2$)	0.1	−0.2	−0.1	−0.1	−0.1	0.0	0.0	−0.3	−0.2	0.1	−0.2	−0.1	−0.5	−0.1
	SD	0.018	0.046	0.025	0.027	0.029	0.029	0.033	0.042	0.043	0.021	0.042	0.058	0.069	0.033
	RMSE	**0.018**	0.046	0.025	0.027	0.029	0.029	0.033	0.042	0.043	0.021	0.042	0.058	0.069	0.033
13	Bias($\times 10^2$)	0.0	0.1	0.2	0.1	0.2	0.1	0.8	0.2	0.6	0.9	0.1	0.0	0.7	0.2
	SD	0.018	0.042	0.042	0.048	0.050	0.033	0.056	0.045	0.052	0.039	0.074	0.088	0.079	0.057
	RMSE	**0.018**	0.042	0.042	0.048	0.050	0.033	0.056	0.045	0.052	0.040	0.074	0.088	0.079	0.057
26	Bias($\times 10^2$)	−0.2	0.1	0.1	−0.1	−0.1	0.2	3.4	0.9	1.9	3.4	−1.5	−3.1	−1.9	−1.0
	SD	0.020	0.056	0.087	0.102	0.107	0.056	0.129	0.088	0.104	0.096	0.245	0.265	0.237	0.211
	RMSE	**0.020**	0.056	0.087	0.102	0.107	0.056	0.133	0.088	0.105	0.102	0.246	0.267	0.238	0.212
39	Bias($\times 10^2$)	−0.4	0.3	−0.2	−0.5	−0.5	0.2	3.0	0.6	2.5	4.0	−4.4	−5.8	−5.5	−4.9
	SD	0.044	0.080	0.160	0.182	0.190	0.105	0.194	0.133	0.176	0.173	0.355	0.376	0.344	0.322
	RMSE	**0.044**	0.080	0.160	0.182	0.190	0.105	0.196	0.133	0.177	0.178	0.358	0.381	0.348	0.326
52	Bias($\times 10^2$)	−6.1	−4.2	−0.1	−0.2	0.0	−0.4	2.6	0.5	5.4	6.1	−8.5	−5.2	−7.8	−7.2
	SD	0.159	0.178	0.283	0.285	0.286	0.243	0.347	0.191	0.297	0.297	0.395	0.386	0.381	0.360
	RMSE	**0.171**	0.183	0.283	0.285	0.286	0.243	0.348	0.192	0.302	0.303	0.404	0.389	0.389	0.367

H & Z: Heagerty and Zheng, S & H: Saha-Chaudhuri and Heagerty, vH & P: Van Houwelingen and Putter, Sh: Shen et al. residual: Cox model with time varying coefficient estimated based on a Schoenfeld smoothing algorithm, h_n bandwidth, sp = span, frac.poly(glm) = fractional polynomials estimated by the glm function, uni = uniform, elp = elliptic, tri = triangular, w1/w2/w3 = different weighting schemes, deg = maximum degree of the polynomials, SD = standard deviation, RMSE = root mean squared error.

Table 3
Simulation results: bivariate normal based scenario with ties, n = 1000, $\rho = -0.3$, 80% censoring.

Time		H & Z 2005		S & H 2013				vH & P 2011				Sh 2015			
		Default		$h_n = 6$			$h_n = 13$	sp = 0.33			sp = 0.66	Frac.poly(glm)		Spline	
		Cox	Residual	uni	elp	tri	uni	w1	w2	w3	w1	deg. = 2	deg. = 3	knot = 1	knot = 2
1	Bias($\times 10^2$)	−4.8	−1.5	−3.9	−2.9	−2.5	−5.9	−0.6	−1.5	−1.1	−1.9	−2.7	−3.4	−2.1	−1.6
	SD	0.019	0.054	0.026	0.025	0.026	0.028	0.032	0.055	0.042	0.030	0.087	0.115	0.065	0.001
	RMSE	0.052	0.056	0.047	0.039	0.036	0.065	0.032	0.057	0.044	0.035	0.091	0.120	0.068	**0.016**
6	Bias($\times 10^2$)	1.2	−0.1	0.9	0.2	0.1	−0.3	0.0	−0.2	0.0	1.1	−1.3	−1.0	−1.9	−2.1
	SD	0.018	0.045	0.026	0.028	0.029	0.029	0.034	0.042	0.047	0.023	0.045	0.057	0.073	0.002
	RMSE	0.022	0.045	0.027	0.028	0.029	0.029	0.034	0.042	0.047	0.025	0.046	0.058	0.075	**0.021**
13	Bias($\times 10^2$)	3.1	0.1	0.1	−0.1	−0.1	1.2	0.4	0.1	0.3	0.6	−1.1	−1.3	−0.6	0.4
	SD	0.018	0.042	0.043	0.050	0.053	0.033	0.064	0.043	0.056	0.040	0.108	0.123	0.115	0.008
	RMSE	0.036	0.042	0.043	0.050	0.053	0.035	0.064	0.043	0.056	0.041	0.108	0.124	0.115	**0.009**
26	Bias($\times 10^2$)	4.4	0.9	0.0	0.0	0.0	0.3	1.4	−0.3	0.7	1.8	−3.5	−2.8	0.8	−9.6
	SD	0.020	0.054	0.095	0.111	0.118	0.059	0.149	0.085	0.109	0.113	0.319	0.335	0.292	0.087
	RMSE	**0.048**	0.055	0.095	0.111	0.118	0.059	0.150	0.085	0.109	0.115	0.321	0.336	0.292	0.129
39	Bias($\times 10^2$)	4.3	3.5	−0.2	−0.1	0.1	0.6	0.6	−0.8	1.8	1.1	−7.2	−2.4	6.5	−17.1
	SD	0.041	0.082	0.183	0.197	0.204	0.130	0.214	0.144	0.195	0.204	0.392	0.397	0.347	0.144
	RMSE	**0.059**	0.089	0.183	0.197	0.204	0.130	0.214	0.144	0.196	0.204	0.399	0.398	0.353	0.224
52	Bias($\times 10^2$)	−2.1	0.3	−2.7	−2.7	−2.9	−3.0	−2.4	−5.1	−1.3	−1.8	−0.7	−3.7	6.5	−19.3
	SD	0.145	0.171	0.265	0.261	0.261	0.264	0.300	0.204	0.298	0.302	0.386	0.379	0.367	0.154
	RMSE	**0.146**	0.171	0.267	0.262	0.262	0.266	0.301	0.210	0.299	0.303	0.386	0.381	0.372	0.246

H & Z: Heagerty and Zheng, S & H: Saha-Chaudhuri and Heagerty, vH & P: Van Houwelingen and Putter, Sh: Shen et al. residual: Cox model with time varying coefficient estimated based on a Schoenfeld smoothing algorithm, h_n bandwidth, sp = span, frac.poly(glm) = fractional polynomials estimated by the glm function, uni = uniform, elp = elliptic, tri = triangular, w1/w2/w3 = different weighting schemes, deg = maximum degree of the polynomials, SD = standard deviation, RMSE = root mean squared error.

could be pointed out from those scenarios, favorable performances were observed for the non-parametric methods of Van Houwelingen/Putter and Saha-Chaudhuri/Heagerty and Shen's method with spline based functions.

Effect size of the marker: Having a stronger or weaker predictive value of the marker did not influence the accuracy of the methods much. Most notably, with stronger predictive value of the marker the events occurred earlier in time, leading to very small risk sets at the later time points. In some of the simulation scenarios this led to biased or unestimable AUCs beyond three years.

Sample size and percentage of censoring: As expected, the RMSE increased with higher censoring percentage, at later time points and with lower sample size. In general, when there was a sufficient number of data points (n = 1000, 20% censoring), all methods were accurate (RMSE roughly below 0.05). When there were less data points (n = 500 with 20% censoring), the non-parametric models with smaller bandwidth (Saha-Chaudhuri/Heagerty with bandwidth = 6 months and to a lesser extent Van Houwelingen/Putter with span = 0.33) showed a modest increase in RMSE due to increase in empirical standard deviation. When data was scarce (scenarios with 80% censoring), the method of Saha-Chaudhuri and Heagerty showed undesirably high standard deviations at later time points when using the small bandwidth and Shen's method resulted in both unacceptable high standard deviation and high bias at the later time points. The methods of Heagerty and Zheng and of Van Houwelingen and Putter could be considered relatively most stable with regard to a decrease in effective sample size over time.

Tied data: Results from the scenarios with tied and untied data were highly comparable indicating that all methods and method extensions were capable of dealing with tied data.

Advice on configuration of methods

For the method of Heagerty and Zheng, allowing for a time varying coefficient in the Cox model did not result in better RMSE in most simulation scenarios. The standard Cox model has a clear advantage in terms of lower standard deviation.

For the method of Van Houwelingen and Putter, using a larger span for the lowess smoothing resulted in somewhat better RMSEs. The performance under the three different weighting schemes of tied pairs did not differ much. In the majority of scenarios, weighing by unique event time points showed a slight advantage in RMSE, but differences were small and not consistent across scenarios.

For the method of Saha-Chaudhuri and Heagerty, the larger kernel bandwidth resulted in somewhat better RMSE than the smaller bandwidth. Note that the large bandwidth averaged the AUC over a maximum time span of more than 2 years (13 months back and forward), which may seem a lot on the total time range of 52 months. But apparently, the slightly worse bias resulting from smoothing over a larger neighborhood was dominated by lower empirical standard deviation. Different kernels performed similarly under large bandwidth ($h_n = 13$); we only show the results of the uniform kernel in the tables for this bandwith. However, when the bandwidth was small ($h_n = 6$ months), the uniform kernel outperformed the other two kernels at later time points. The uniform kernel assigns the same weights to all event time points in the neighborhood, but triangular and Epanechnikov kernels assign higher weights to events in the middle of the moving window. For later time points where there are less events, triangular and Epanechnikov kernels could be more sensitive to the influence of single events, which might lead to the larger empirical standard deviations.

For Shen's method with fractional polynomial base function, the maximum polynomial degree of 2 resulted in somewhat lower RMSE compared to degree 3. The spline base functions outperformed the fractional polynomial base functions. With tied data, the knots were best placed at percentiles of all event times (knot = 2).

5. Applying the methods to the motivating dataset

Figs. 1 to 4 show the I/D AUC curves estimated from the motivating dataset. Since asymptotic variance formulas were only available for two out of the four methods (Saha-Chaudhuri and Heagerty, 2013; Shen et al., 2015), we constructed the confidence intervals by a bootstrapping procedure for all four methods. Shen et al. advise to only present the estimated curve up to the, e.g., 90th percentile of observed event time points to avoid boundary problems, which would be 18 months in the studied dataset. We here plotted the total time range to allow for comparison with the other methods on the full time range. Limiting the time range over which the AUC curve is calculated might be good practice in general applications (see discussion Section).

The four different methods estimated quite different I/D AUC curves. All curves started out between 0.6 and 0.65 during the first one and a half years, but showed quite different trends within this time frame. At the later time points (beyond 2.5 years) where there were few events, differences between methods were large. Visual assessment of the bootstrapped confidence intervals suggested there are probably too few events beyond 2.5 years for reliable estimation of the I/D AUC and none of the methods detected a significant increase or decrease in AUC over time.

The 'standard Cox' method by Heagerty and Zheng that was most accurate according to the simulation study resulted in a rather smooth AUC curve, more or less following a horizontal line around 0.63. The other methods either showed 'bumpy' curves moving up and down within small time intervals or resulted in smooth curves with unrealistic steep slopes at later time points. The bumpy curves resulted from the method of Heagerty and Zheng with time varying coefficient and from the non-parametric methods of Van Houwelingen/Putter and Saha-Chaudhuri/Heagerty when using smaller bandwidths or spans. The smooth curves with steep slopes at later time points resulted from using larger bandwidth and from Shen's method.

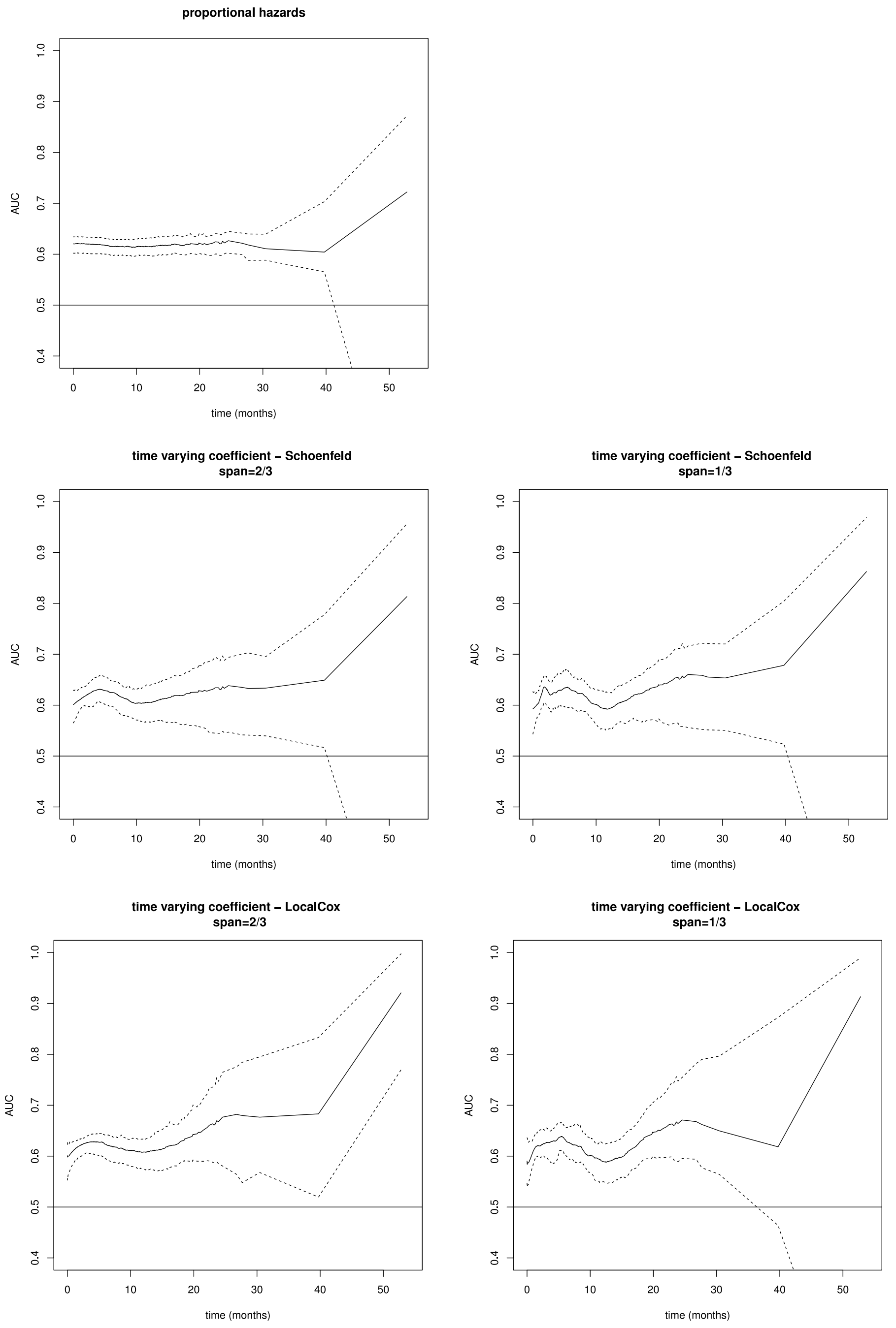


Fig. 1. I/D AUC according to different Heagerty and Zheng algorithms.

6. Recommendations

Our overall conclusion from the data application is that one should be aware that different estimation methods may give very different I/D AUC curves over time. A single curve should not be taken at face value and especially one should be cautious interpreting any increasing or decreasing trend in an I/D AUC curve. We recommend to always explore different estimation methods and also to vary the modeling choices and smoothing parameters in order to assess the robustness of the curve.

The main conclusion resulting from our simulation study is that when the amount of data is limited, imposing stronger assumptions during the estimation of the I/D AUC over time will lead to more accurate results. In particular, the 'standard

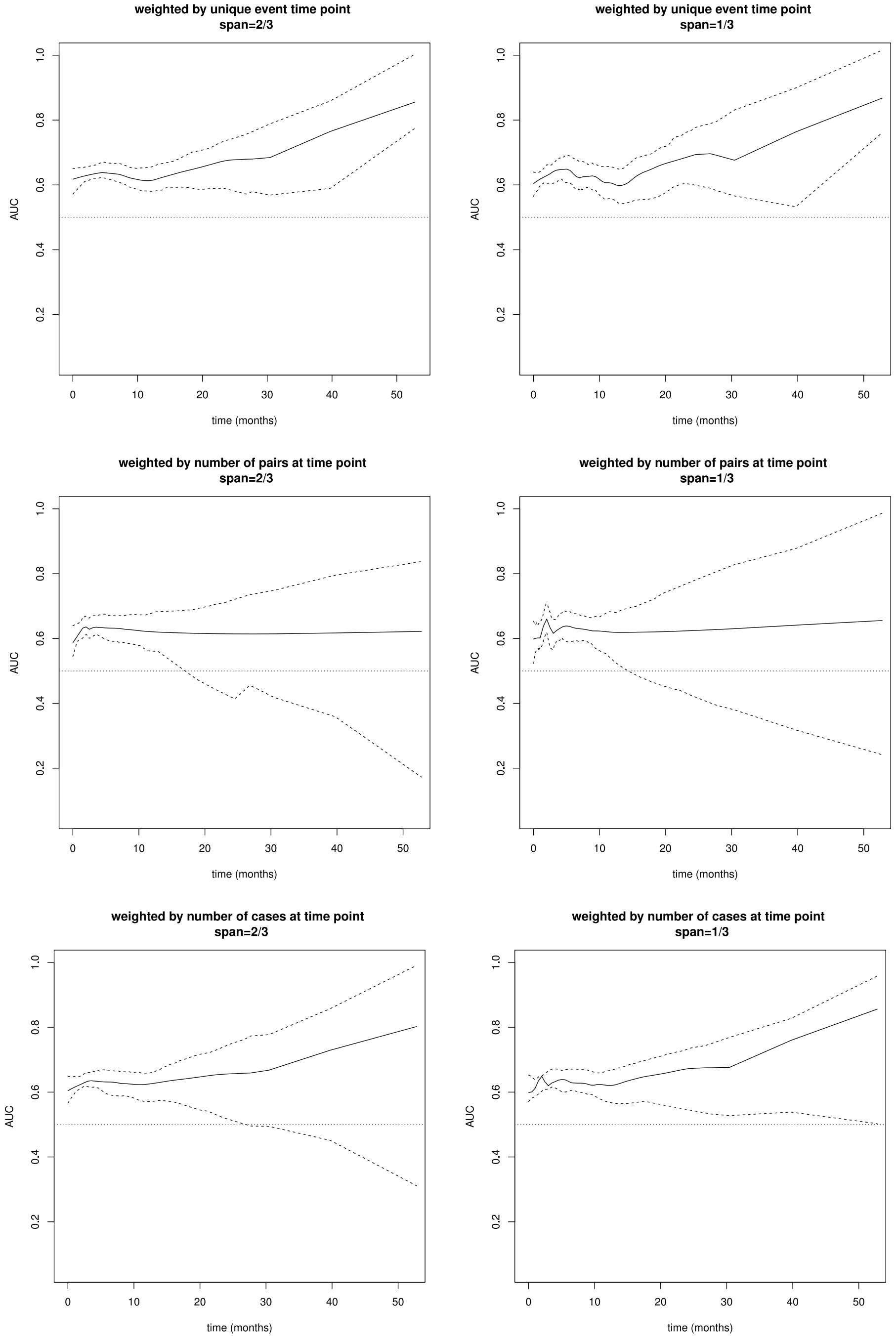


Fig. 2. I/D AUC according to different Van Houwelingen and Putter algorithms.

Cox' model suggested by Heagerty and Zheng showed to be the most accurate method in many cases. Only when there are many observed event time points (in our simulation settings about 800), more flexible methods can be considered. In such cases our extended version of Shen's method using natural cubic splines as base function is a promising method. The non-parametric methods of Van Houwelingen and Putter and Saha-Chaudhuri can be expected to be a bit more volatile, but at the same time more robust than Shen's method in case of sparse data regions. We advise to use these methods with a relatively large bandwidth/span.

All estimation methods showed reduced accuracy at higher time range where there were less events. To warrant valid interpretation, we advise to truncate I/D AUC curves at a point where there are still a sufficient number of events that can be used for estimation of the I/D AUC, in line with the suggestion made by Shen et al. (2015).

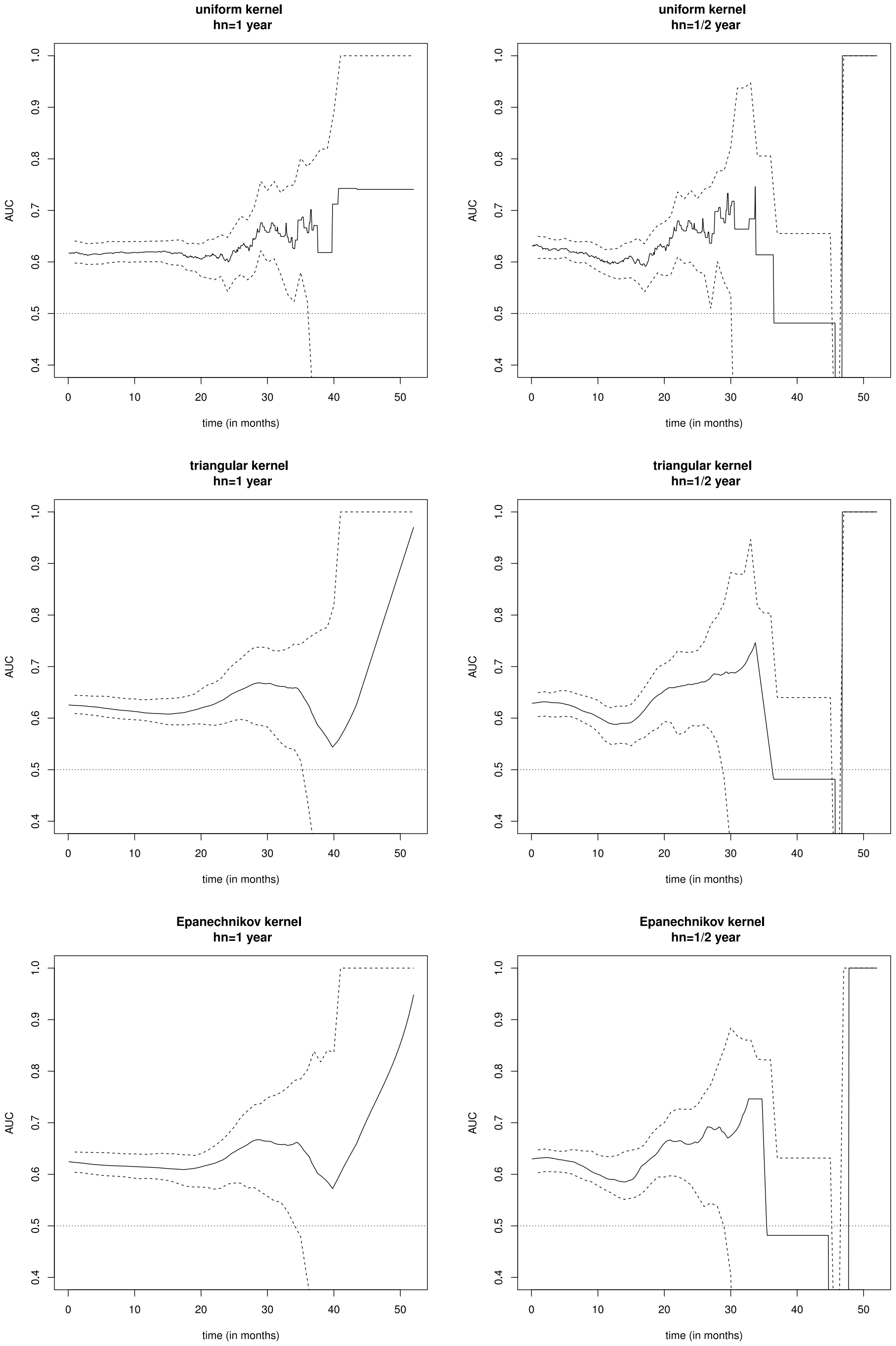


Fig. 3. I/D AUC according to the Saha-Chaudhuri and Heagerty algorithm using three different kernels and two different neighborhood spans h_n: neighborhood span one year back and forward, neighborhood span half a year back and forward.

7. Discussion

In this paper, we compared the behavior of four recently proposed estimators of I/D AUC examining the discriminative ability of a dynamic prediction model over time. We conclude that one should be very cautious when interpreting such a curve based on a single method, especially in case of sparse data. The modeling choices necessary to make these curves allow many ‘researcher degrees of freedom’. The recommendations given in our study can aid in making these choices in a less data-driven manner.

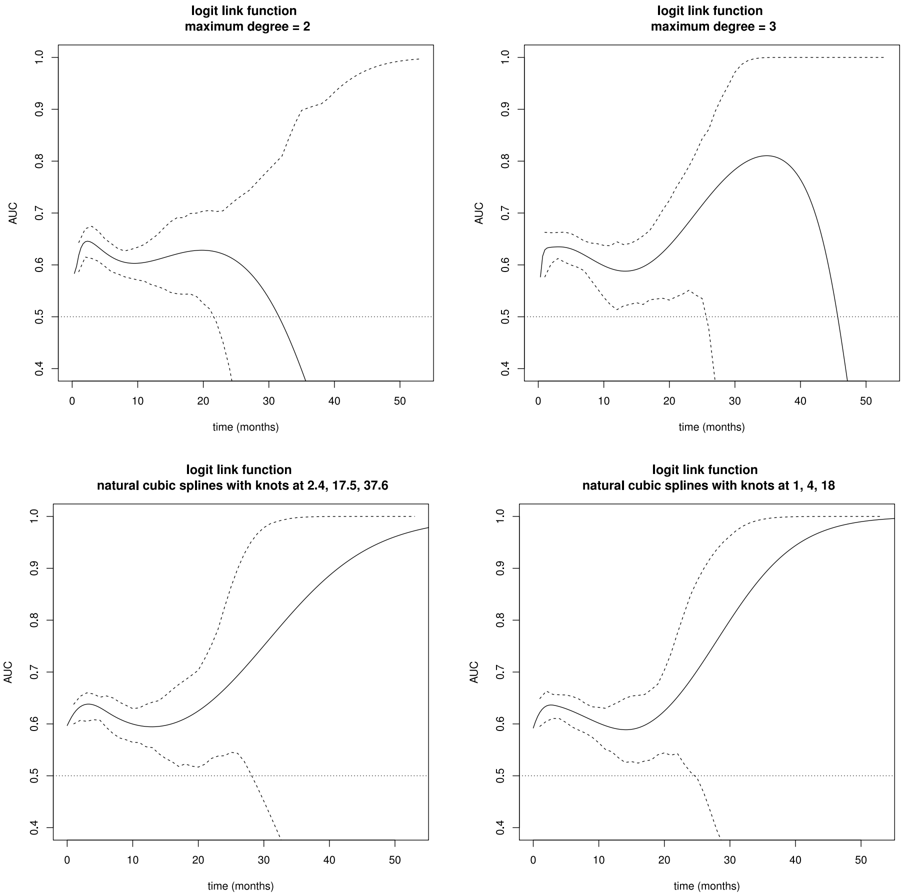


Fig. 4. I/D AUC according to the Shen et al. algorithm using fractional polynomials as base functions with different maximum degree of the set of polynomials (top panels) and using natural cubic splines as base functions with knots at different locations (bottom panels).

When discussing the AUC curves with our clinical collaborators, focus quickly turned towards the interpretation of any increasing or decreasing trend that was visualized in the curves. The underlying question seems very relevant: does the dynamic prediction model under study gain (or lose) accuracy over time? The danger of such a focus is that spurious trends can be expected if there are few observed event time points. Especially at higher time range where typically there is much censoring, the curves will be unstable and a spurious increase or decrease can easily occur.

The simulation results suggested that using larger bandwidth leads to improved accuracy. When looking at the curves from the application, the bumps generated by the smaller bandwidth methods indeed look unrealistic. It is not expected that the true discriminative ability of the evaluated prediction model is so volatile. However, despite being less accurate, a bumpy picture might also have an advantage: it protects against over-interpretation of trends while still capable of capturing true trends if they exist. For presentation purposes, one might be willing to accept a bit lower pointwise accuracy and choose a method that results in a more bumpy curve as trends from bumpy curves are less prone to over-interpretation. For example, the curves resulting from using a uniform kernel function in the method of Saha-Chaudhuri (upper panels Fig. 3) seem less vulnerable to misinterpretation of time trends than the more smoothly varying curves produced by the method of Shen et al. in Fig. 4. There seems to be a trade-off between more accurate smoothly varying curves and protecting against over-interpretation of trends. We argue that future work should focus on objective measures to determine whether the I/D AUC increases or decreases over time.

Saha-Chaudhuri and Heagerty suggested that alternative types of neighborhoods could be employed, using a fixed number of neighboring event times instead of a fixed time span (Saha-Chaudhuri and Heagerty, 2013). This could be effective in balancing the sensitivity of the method to picking up time trends at early time points where there is sufficient data and the stability at later time points where there is less data. We have not yet explored this suggestion further.

Due to the smoothing intrinsic to all the algorithms, noise trends in the sparse data regions can also influence the estimated trend at earlier time points where there may still be a reasonable amount of data. Presenting confidence intervals around the curves should offer some protection against over-interpretation of trends. However, weighing these objectively in a visual manner can be challenging.

Some authors have argued that ROC or AUC measures that use dynamic controls (compared to static controls Heagerty and Zheng, 2005) are not well suited to study trends over time since such trends may be due to a combination of changing detection properties of the test or marker under evaluation and changing control groups (Pepe et al., 2008; Blanche et al., 2013b). While this may be true for situations where the main interest is in the diagnostic value of a certain test or marker, in the here studied setting of dynamic survival models we believe using dynamic controls is the most natural way of evaluation. Changing risk sets are intrinsic to dynamic prediction and we aim to evaluate the accuracy of such a model based on patients under study and information available at that time point. In particular, in the motivating fertility prediction model, no time varying covariates were used in the development of the dynamic prediction model. The updated predictions are due to the changing patient selection and we wished to evaluate how well the model performs with this population that changes over time (Van Eekelen et al., 2017). The methods compared in our study are equally applicable to settings where true longitudinal markers exist, i.e. where new measurements are taken over time that lead to updated predictions.

The difficulty in estimating the incident/dynamic AUC lies in the choice of defining the accuracy based on event and control status at one single time point. As pointed out in the introduction, for the cases this leads to low numbers per time point requiring some degree of smoothing that is implemented in different ways in each of the methods discussed. One could argue that due to the smoothing, the estimates of all methods are not strictly based on data at a single time point anymore. In particular, for all of the presented methods the I/D AUC estimate at a certain time point t is also influenced by data at earlier time points. This may be an undesirable property when the focus is on the evaluation of models that predict the future and not the past. In the method of Saha-Chaudhuri this could be overcome by choosing a ‘forward’ neighborhood, i.e. covering only the right neighborhood of the time point of interest. Also for the lowess and spline based methods choosing a span only covering future time points could be considered further.

The only way to avoid the smoothing challenge intrinsic to the incident/dynamic AUC is to use a different case definition. The prospective cumulative case definition, where events over some predefined future time period are considered cases (Cai et al., 2006; Zheng and Heagerty, 2007), seems an obvious candidate. Drawbacks of using cumulative cases are that there is less distinction between earlier and later events and the AUCs of adjacent time points are based on partly redundant information over time (Blanche et al., 2013b). This redundancy might make it less suitable for studying time trends in accuracy of dynamic prediction models. However, if it is possible to define a relevant short future time period which still captures a sufficient number of events for numerically stable results, it might be a valid alternative.

Acknowledgments

We are grateful to the CECERM study group (Collaborative Effort for Clinical Evaluation in Reproductive Medicine) who collected the motivating data presented here with grant support from ZonMw, the Netherlands Organization for Health Research and Development, The Hague, the Netherlands, grant 945/12/002 (Van der Steeg et al., 2007). We also thank Ilaria Prosepe for proofreading this manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2020.107095.

References

Blanche, P., Dartigues, J.F., Jacqmin-Gadda, H., 2013a. Review and comparison of ROC estimators for a time-dependent outcome with marker-dependent censoring. Biometric J. 55 (5), 687–704.

Blanche, P., Latouche, A., Viallon, V., 2013b. Time-dependent AUC with right-censored data: a survey. In: Lee, M.-L.T., Gail, M., Pfeiffer, R., et al. (Eds.), Risk Assessment and Evaluation of Predictions. In: Lecture Notes in Statistics, Springer, New York, pp. 239–251.

Cai, T., Pepe, M.S., Zheng, Y., et al., 2006. The sensitivity and specificity of markers for event times. Biostatistics 7 (2), 182–197.

Harrell, F.E., Califf, R.M., Pryor, D.B., et al., 1982. Evaluating the yield of medical tests. JAMA 247, 2543–2546.

Heagerty, P.J., Zheng, Y., 2005. Survival model predictive accuracy and ROC curves. Biometrics 61 (1), 92–105.

Pepe, M.S., Zheng, Y., Jin, Y., et al., 2008. Evaluating the ROC performance of markers for future events. Lifetime Data Anal. 14, 86–113.

Rizopoulos, D., 2011. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. Biometrics 67 (3), 819–829.

Rizopoulos, D., 2012. Joint Models for Longitudinal and Time-to-Event Data: With Applications in R. Chapman and Hall/CRC.

Royston, P., Altman, D.G., 1994. Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. J. R. Stat. Soc. B 43, 429–467.

Saha-Chaudhuri, P., Heagerty, P.J., 2013. Non-parametric estimation of a time-dependent predictive accuracy curve. Biostatistics 14 (1), 42–59.

Schmid, M., Potapov, S., 2012. A comparison of estimators to evaluate the discriminatory power of time-to-event models. Stat. Med. 31 (23), 2588–2609.

Shen, W., Ning, J., Yuan, Y., 2015. A direct method to evaluate the time-dependent predictive accuracy for biomarkers. Biometrics 71 (2), 439–449.

Song, X., Zhou, X.H., Ma, S., 2012. Nonparametric receiver operating characteristic-based evaluation for survival outcomes. Stat. Med. 31, 2660–2675.

Van der Steeg, J.W., Steures, P., Eijkemans, M.J., et al., 2007. Pregnancy is predictable: a large-scale prospective external validation of the prediction of spontaneous pregnancy in subfertile couples. CECERM study group (Collaborative Effort for Clinical Evaluation in Reproductive Medicine). Hum. Rep. 22 (2), 536–542.

Van Eekelen, R., Scholten, I., Tjon-Kon-Fat, R.I., et al., 2017. Natural conception: repeated predictions over time. Hum. Rep. 32 (2), 346–353.

Van Houwelingen, H.C., Putter, H., 2011. Dynamic Prediction in Clinical Survival Analysis. In: Monographs on Statistics and Applied Probability, Chapman and Hall/CRC.

Xu, R., O'Quigley, J., 2000. Proportional hazards estimate of the conditional survival function. J. R. Stat. Soc. B 62, 667–680.

Zheng, Y., Heagerty, P.J., 2007. Prospective accuracy for longitudinal markers. Biometrics 63, 332–341.