Stat Methods Appl
https://doi.org/10.1007/s10260-017-0410-2

ORIGINAL PAPER

Prognostic assessment of repeatedly measured time-dependent biomarkers, with application to dilated cardiomyopathy

Giulia Barbati[1] · **Alessio Farcomeni**[2]

Accepted: 11 November 2017

Abstract We propose new time-dependent sensitivity, specificity, ROC curves and net reclassification indices that can take into account biomarkers or scores that are repeatedly measured at different time-points. Inference proceeds through inverse probability weighting and resampling. The newly proposed measures exploit the information contained in biomarkers measured at different visits, rather than using only the measurements at the first visits. The contribution is illustrated via simulations and an original application on patients affected by dilated cardiomiopathy. The aim is to evaluate if repeated binary measurements of right ventricular dysfunction bring additive prognostic information on mortality/urgent heart transplant. It is shown that taking into account the trajectory of the new biomarker improves risk classification, while the first measurement alone might not be sufficiently informative. The methods are implemented in an `R` package (`longROC`), freely available on `CRAN`.

Keywords AUC · NRI · ROC · Prognostic scores

1 Introduction

The evaluation of prognostic and diagnostic biomarkers is a primary issue in medical statistics. Medical diagnoses, indication of treatments, and risk assessment is grow-

✉ Alessio Farcomeni
alessio.farcomeni@uniroma1.it

Giulia Barbati
gbarbati@units.it

[1] Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute, Università di Trieste, Strada di Fiume, 447, 34149 Trieste, Italy

[2] Dipartimento di Sanità Pubblica e Malattie Infettive, Università di Roma "La Sapienza", Piazzale Aldo Moro, 5, 00185 Rome, Italy

Published online: 18 November 2017

ingly based on the evaluation of scores, like CHA_2DS_2-VASc and HAS-BLED in atrial fibrillation, MELD and Child-Pugh in epathology, SOFA in intensive care unit stay, etc. The performance of these scores is continuously evaluated and possibly improved by the addition of new markers and/or modification of the scoring system. Also single biomarkers are continuously evaluated as independent diagnostic or prognostic factors, which is something that can help the doctor in making cost-effective and simple assessments. Notable examples in cardiology are homocysteine, troponin and myoglobin serum levels.

There is a huge literature on the evaluation of prognostic and diagnostic markers. For a general overview we point the reader to the excellent book by Pepe (2003). Our work targets the issue of assessing how a prognostic time-dependent biomarker is related to a time-dependent and possibly censored event. Our purpose is to evaluate the prognostic accuracy of a pre-specified function of a trajectory of the time dependent marker, or score, up to a pre-specified visit time t_s. Time-dependent ROC curves with markers measured at baseline were introduced in Heagerty et al. (2000). In Zheng and Heagerty (2004) the evaluation of a time-dependent marker is considered, see also Zheng and Heagerty (2005), Uno et al. (2007), Gerds et al. (2013) and Li et al. (2017). The last approaches are in the spirit of our work, but are based on evaluating the marker at a single given time point. The entire trajectory of the longitudinal biomarker is considered in other works [e.g., Schoop et al. (2008) and Rizopoulos (2011)], but mostly through the prediction of the biomarker at a given time point. It is therefore difficult to interpret directly the prognostic significance of a given trajectory. We will propose a procedure which directly evaluates how an observed trajectory (i.e., repeated measurements) can predict, in a specific sense, the occurrence (or non-occurrence) of the time-to-event outcome. For example, we will evaluate the probability that a marker is below a threshold at times t_1, t_2 and t_3, if the patient will survive at least $t > t_3$. In our opinion our approach has a direct interpretability and could be useful in using all of the information obtained with repeatedly measured biomarkers.

Our motivation comes from an original study on a cohort of patients affected by Dilated Cardiomyopathy (DCM), a primary myocardial disease characterized by left ventricular systolic dysfunction and dilation, measured mainly by the continuous parameter LVEF (Left Ventricular Ejection Fraction). In the last decades the long-term prognosis of DCM has impressively improved mainly through the effectiveness of pharmacological and non-pharmacological treatments on left ventricular reverse remodeling (basically the improvement of LVEF). See Merlo et al. (2011) for a discussion from a medical perspective. LVEF improvement indeed has emerged as an important prognostic predictor in DCM, highlighting the importance of the systematic re-evaluation of patients during the follow-up. Nevertheless, in clinical practice the prognostic stratification of DCM still remains particularly difficult, and substantially based on basal LVEF. Recent data suggested that right ventricular systolic function (RV-d) is also relevant in the prognostic assessment of DCM patients. RV-d dysfunction, assessed by cardiac magnetic resonance, was found in a sizeable number of DCM patients and showed an incremental prognostic value in addition to LV remodeling evaluation (Gulati et al. 2013). However, these data are limited to few and highly selected populations. To date, there are no data on the prognostic significance of RV-d re-evaluation during follow-up. Therefore, the clinical aim of the

study was to assess the impact on DCM prognosis of the regular RV-d evaluation over time.

It shall be underlined that even though we are evaluating a longitudinal biomarker possibly censored by an associated process, our work has very limited relationship with e.g. the literature on joint models for longitudinal and survival data. We recommend researchers to use the most appropriate model to compute a longitudinal score based on longitudinal biomarkers, matching the model with the research aims. In this regards, a thorough and detailed discussion can be found in Kurland et al. (2009). In this paper we discuss how to evaluate scores once these have been computed.

The rest of the paper is as follows: in the next section we give a more detailed description of our data set. In Sect. 3 we give a very brief overview of methods for evaluation of prognostic biomarkers, and provide our innovative definitions of sensitivity, specificity, and reclassification. In Sect. 4 we show how to estimate our proposed measures. A synthetic evaluation based on simulated data is provided in Sect. 5, and we apply our approach on the motivating example in Sect. 6. Concluding remarks are given in Sect. 7.

The methods are implemented in an `R` package (`longROC`), freely available on `CRAN` for download.

2 Data

We retrospectively analyzed a group of patients enrolled in the Trieste Heart Muscle Disease Registry from 1993 to 2008 (Merlo et al. 2014). Patients who underwent extensive clinical and laboratory evaluation at baseline and with at least one available short-time evaluation (i.e., at 6 months) were considered for the present analysis. Other visits were scheduled at 24, 48, and 72 months. Follow-up ended on 31 December 2014 or at the time of death/urgent (status I) heart transplantation (D/HT), thus each patient had a potential minimal follow-up of at least 72 months (last patient was included at the end of 2008). Data come from a referral center for cardiomyopathies where enrolled patients are regularly followed-up; the censoring mechanism could therefore be defined as 'administrative' censoring.

The institutional ethical board approved the study and the informed consent was obtained under the institutional review board policies of hospital administration. The data we observe consist of multiple parameters measured at the different follow up times, plus time-to-event data. For the present analysis, we selected the most relevant clinical and instrumental measures:

- Two parameters fixed at baseline: Age at enrollment and Heart Failure (HF) duration;
- One continuous longitudinal marker: LVEF;
- One discrete longitudinal marker: RV-d as a binary parameter indicating presence or absence of the dysfunction.

We did not include gender as it is not significant at univariate or multivariate Cox modeling for the data at hand, and it would bring no improvement to the prognostic indices considered below. Note that the role of gender in this disease needs clarifica-

tion, and is often not predictive of time-to-events after adjustment for other relevant predictors [e.g., Merlo et al. (2016)].

3 Prognostic accuracy of functions of trajectories of longitudinal biomarkers

Given a marker X and time-to-event data (T, Δ), where Δ is an indicator of event, time-dependent sensitivity and specificity are commonly defined as

$$Sp(t, c) = \Pr(X \leq c | T > t)$$

and

$$Se(t, c) = \Pr(X > c | T \leq t).$$

Here t is a time-horizon of interest, which is specified by the user, and so is threshold c. Examples in different areas of clinical research are reported in Pignatelli et al. (2015), Iacovelli et al. (2015), Riggio et al. (2015), Basili et al. (2017) and Cardellini et al. (2017).

Suppose now that we are actually measuring a time-dependent marker $X(t)$, which has been repeatedly assessed for n subjects at pre-specified visit times $t_1 = 0, t_2, \ldots, t_{s_i}$. Let $S = \max_i s_i$. This framework encompasses several scenarios, including our motivating example. In prospective studies (but also in certain retrospective ones, as in our case) follow-up occasions are pre-specified and, hence, patients will be visited at approximately the same time intervals between one visit and another.

We now define time-dependent sensitivity and specificity for time-dependent biomarkers or scores as

$$Sp(t, c, s, u) = \Pr(X(t_1) \leq c, X(t_2) \leq c, \ldots, X(t_{s_i}) \leq c | T > t) \quad (1)$$

$$Se(t, c, s, u) = \Pr\left(\cup_{j=1}^{s_i} X(t_j) > c | u \leq T \leq t\right), \quad (2)$$

where t, c, s and u are parameters that are pre-specified by the user. The parameter t, much like the usual definition of time-dependent measures, is a time-horizon for events (i.e., no events are of interest beyond t). The parameter $\boldsymbol{s} = (s_1, \ldots, s_n)$ is a vector of number of visits to be evaluated for each patient. An implicit requirement is that $t_{s_i} \leq u < t$. While in general one might want to set $s = s_1 = s_2 = \cdots = s_n$, and discard patients who have not reached s visits for evaluation of the quantities above, in more general scenarios one could consider a different number of visits for each patient. An underlying assumption when s_i is not constant is that sensitivity and specificity with s_i number of visits are equal to sensitivity and specificity with s_j number of visits for patients i and j, with $s_i \neq s_j$. For this reason we could define sensitivity and specificity above as a function of only one patient, even if we will use all data to estimate them.

The parameter $u < t$ is a minimal time horizon for defining events. Often one might want to set $u = \max_i t_{s_i}$, that is, record events immediately after the last active visit.

On the other hand in some cases it might be interesting to fix $u > \max_i t_{s_i}$, for instance when visits occur in the hospital and u is set as the time of discharge (as in-hospital events might occur to fragile or terminal patients). Note that u does not explicitely appear in the definition for Sp, we clarify in the next section why it does implicitely.

The definition of sensitivity and specificity above involve a definition of positive diagnosis as soon as the biomarker exceeds the threshold at least once during the follow-up. A more general definition can be given by

$$Se(t, c, \boldsymbol{s}, u) = \Pr(f_c(X(t_1), X(t_2), \ldots, X(t_{s_i}))|u \leq T \leq t),$$
$$Sp(t, c, \boldsymbol{s}, u) = 1 - \Pr(f_c(X(t_1), X(t_2), \ldots, X(t_{s_i}))|T > t)$$

for some fixed f_c, where c is a threshold. Here f_c is a general function for defining positive diagnoses. For instance, f_c can be defined as the event that the mean or median of the biomarker exceeds the threshold (e.g., $f_c(X(t_1), X(t_2), \ldots, X(t_{s_i})) = \{\sum_{j=1}^{s_i} X(t_i)/s_i > c\}$), as the event that the threshold is exceeded at least twice, as the event that the biomarker has increased at least c units (e.g., $f_c(X(t_1), X(t_2), \ldots, X(t_{s_i})) = \{X(t_{s_i}) - X(t_1) > c\}$) or that it has done so at least once between visits (e.g., $f_c(X(t_1), X(t_2), \ldots, X(t_{s_i})) = \{X(t_2) - X(t_1) > c \cup X(t_3) - X(t_2) > c \cup \cdots \cup X_{t_{s_i}} - X_{t_{s_i-1}} > c\}$). Several definitions of f_c are possible, and in general we suggest to choose the appropriate f_c according to the underlying clinical mechanism under study. Of course, several possibilities can be compared in terms of prognostic performance.

The corresponding ROC curves are given by a plot of $\{Se(t, c, s, u), 1 - Sp(t, c, s, u)\}$ for all possible values of c, and there will be one for each admissible value of (u, t, s). The area-under-the-curve (AUC) is defined as usual as the area under the ROC curve as a function of c. We note here that if $s_1 = s_2 = \cdots = s_n = 1$ and $u = 0$, our definitions reduce to the usual definitions for time-dependent sensitivity, specificity and ROC curves as introduced by Heagerty et al. (2000), hence our proposal can be seen as a direct generalization of commonly used time-dependent ROC curves with baseline markers. Note furthermore that summaries, depending on the data configuration and f_c definition might be insensitive to changes in u, s_i and/or t.

We are also interested in the added value of an additional marker to an already available score. Many methods were recently proposed, see Pencina et al. (2008, 2011) and Uno et al. (2013). Here we focus on the Net Reclassification Index (NRI), where two events U_s and D_s denote the fact that use of the additional marker lead to an increase (U_s) or decrease (D_s) of the predicted risk when considering the trajectory up to time s. Also call $E_{ut} = \{u \leq T \leq t\}$. Using formula (4) in Pencina et al. (2011) in our context, we obtain

$$NRI(t, s) = \frac{(\Pr(E_{ut}|U_s) - \Pr(E_{ut}))\Pr(U_s) + (\Pr(E_{ut}) - \Pr(E_{ut}|D_s))\Pr(D_s)}{\Pr(u \leq T \leq t)\Pr(T > t)}.$$

This formula does not explicitely show what the NRI is, as it is a reformulation apt at expressing the NRI in a form which can be evaluated for time-to-event data. The basic idea behind the NRI is that when two markers are compared, their difference can be summarized by considering subjects that are reclassified (e.g., risk increased

or risk decreased). When comparing the new marker with respect to the baseline one, we would like the risk of events (here, patients experiencing the event between times u and t) to be increased and the risk of non-events (here, patients experiencing the event after time t) to be decreased. Hence, in general NRI can be thought as the sum of two components, first one being the relative increase in the risk for subjects who experience the event, and second one being the relative decrease for subjects who do not. The two addends are often also evaluated separately and referred to as "NRI for events" and "NRI for non-events". As it is expressed, $-2 \leq NRI \leq 2$, where $NRI < 0$ indicates that the new biomarker or score is worse than its competitor. An interpretation of 1/2 NRI has been outlined in Pencina et al. (2012), where thresholds of 10, 20 and 30% have been indicated as weak, moderate and strong evidence of improvement.

4 Inference

Evaluation of sensitivity, specificity, and NRI as defined in the previous section is complicated by the presence of censored subjects, that is, subjects lost at follow-up (possibly because of administrative censoring) before time t, and hence for which we do not know the true status between time u and t. We note that we always condition (explicitely or implicitely) on having survived up to u. This has some relations with the literature on residual life [e.g., Jeong et al. (2008) and Jung et al. (2009)]. Subjects with an event before time t_{s_i} do not contribute to the estimates.

In order to make inference on the quantities of interest we use Bayes theorem to write:

$$Se(t,c,s,u) = \frac{\Pr(u \leq T \leq t | T \geq u, f_c(X(t_1),\ldots,X(t_{s_i})) > 0) \Pr(f_c(X(t_1),\ldots,X(t_{s_i})) > 0)}{\Pr(u \leq T \leq t | T \geq u)},$$

and

$$Sp(t,c,s,u) = \frac{\Pr(T > t | T \geq u, f_c(X(t_1),\ldots,X(t_{s_i})) \leq 0) \Pr(f_c(X(t_1),\ldots,X(t_{s_i})) \leq 0)}{\Pr(T > t | T \geq u)}.$$

Once we express sensitivity and specificity as above, estimates are readily available. In order to estimate $\Pr(f_c(X(t_1),\ldots,X(t_{s_i})) \leq 0)$ we might use the empirical proportion among subjects being observed at least up to time t_{s_i}. The probabilities of events linked with T are readily available via the Kaplan-Meier (KM) product limit estimator. An underlying assumption is that of independent censoring. In order to estimate

$$\Pr(T > t | f_c(X(t_1),\ldots,X(t_{s_i})) \leq 0, T \geq u)$$

and

$$\begin{aligned}&\Pr(T \leq t | f_c(X(t_1),\ldots,X(t_{s_i})) > 0, T \geq u)\\ &\quad = 1 - \Pr(T > t | f_c(X(t_1),\ldots,X(t_{s_i})) > 0, T \geq u)\end{aligned}$$

we can use KM estimates stratified by the conditioning events. Note that $NRI(t, s)$ can be estimated with a similar strategy.

A ROC curve, as outlined above, is simply the curve plotting one minus sensitivities and specificities. Obviously, it must be non-decreasing. On the other hand, given that sensitivities and specificities are separately estimated for each threshold, this is not guaranteed when drawing a ROC curve based on raw estimates. In order to guarantee monotonicity of the estimated ROC curve we perform isotonic regression, which is a non-parametric method to estimate a monotone function in order to describe the non-linear relationship between two variables. We regress sensitivity as a function of one minus specificity. The isotonic regression algorithm involves (i) computing cumulative sums of sensitivities along the order given by one minus specificities, (ii) determining the greatest convex minorant (gcm) of the cumulative sum, that is, a *convex* function which is at every point at most equal to the cumulative sum, but larger than any other convex function with this property, (iii) taking the first differences of the gcm found at step (ii). By definition of convex functions, the first differences of the gcm are monotonically non-decreasing. Since we have used the *greatest* convext minorant, the estimated non-decreasing function will be as close as possible to the estimated (non necessarily monotone) ROC curve. Furthermore, in case the raw estimated ROC curve is monotone, isotonic regression estimates will coincide with the raw estimates. For more details on non-parametric isotonic regression see Robertson et al. (1988) and references therein.

Finally, to obtain standard errors and (parametric or non-parametric) confidence intervals we rely on resampling. For standard errors and confidence intervals we perform the bootstrap, that is, we repeatedly sample the data with replacement and compute the statistic of interest (e.g., the AUC) on the resampled data. The standard error can be estimated as the standard deviation of the resampled AUCs. A 95% non-parametric confidence interval corresponds to the 2.5 and 97.5% quantiles of the resampled statistics, while parametric confidence intervals (which are more appropriate when the number of replicates is small) assume that the resampled statistics are approximately Gaussian distributed. In order to preserve the dependency structure we resample units, rather than single measurements. Hypothesis testing on the AUC and NRI are based on Wald statistics after a Gaussian approximation of the bootstrap resamples. For comparison of two AUCs we use instead permutation testing as in Venkatraman (2000). To this end, we relabel at random each biomarker value as marker 1 or marker 2, independently of its true label. It is straightforward to check that this corresponds to a random permutation of the two markers. We then compute the two AUC values. A non-parametric p value for the test that two AUC values are the same is obtained as the proportion of resamples (obtained with random relabeling) with a difference in absolute value that is larger than the observed one.

5 Simulations

In order to illustrate the newly proposed indeces, and assess the performance of the proposed inferential procedure, we generate three scores, S_1, S_2, S_3. The first score is generated independently of survival times, therefore being completely irrelevant. The

second one is a time-fixed score repeatedly measured at each visit occasion. The third is a time-varying version of the previous one. The idea is that monitoring time-varying biomarkers provides additional prognostic information. For simplicity, in this and the next section we let $s_1 = \cdots = s_n = s$ and $u = t_s$.

We generate survival through an accelerated failure time model with time-dependent covariates, where

$$\Pr(T_i > t) = \exp\left[-\int_0^t h(x)\, dx\right],$$

and

$$h(x) = \exp\left\{\beta_1 X_{1x} + \beta_2 X_{2x}\right\},$$

where $\beta_1 = 0$, $\beta_2 = 1$. We fix four visit times $(0, 1, 2, 5)$. The markers are generated as follows: X_{1x} is sampled from a standard white noise Gaussian process. The second marker X_{2x} is a step function with initial value uniformly sampled from the set $\{0, 1, 2\}$, a change point uniformly sampled in the time-interval $(0, 1.5)$, and second value corresponding to $X_{2t_1} + 2(-1)^U$, where $U = \{0, 1\}$ is a binary random variable with $\Pr(U = 1) = 0.5$. In this way some subjects have a constant value for the first two visits. This happens if the uniformly sampled change point from the interval $(0, 1.5)$ yields a value in $(1, 1.5)$, hence for approximately 33% of the subjects. It shall be noted that these subjects represent a worst-case scenario for our purposes when $s = 2$ is evaluated.

We assume that these processes are measured without error at visit times $t_1, \ldots, t_s$, but of course all of their trajectory influences the survival time as outlined above. Our scores are, for $j > 0$, $S_{1j} = X_{1t_j}$, $S_{2j} = X_{2t_1}$, $S_{3j} = X_{2t_j}$. Censoring is generated independently as a uniform random variable, so to obtain a proportion of censoring of about 80% to match the low event rate of the motivating application.

We generate data for $n = \{200, 800\}$, and evaluate AUC and NRI for $s = 2, 3$ and $t = 3, 8$. We repeat the operation $B = 1000$ times and report the average AUC and NRI in Table 1. Boxplots (for the 1000 replicates) of selected scenarios can be found in Fig. 1.

It can be seen that regardless of s, n and t, S_1 has an AUC of approximately 50%. The time-constant S_2 has an AUC of approximately 70% when $t = 3$ and slightly lower when $t = 8$. This reflects the fact that baseline values might be less informative as the time horizon is increased. The improvement of S_3 over S_2 and S_1 is also testified by the 1/2 NRI, which is positive and above the threshold of 20% in most cases. When assessing S_3 we see that AUC is on average the largest in all cases, indicating that we can in general expect a slight improvement in terms of AUC when taking into account time-dependent markers. It can also be noted that the AUC for S_3 increases slightly when considering $s = 3$ over $s = 2$, because then the time-dependent marker has definitely changed and is providing more (and more recent) information towards the chance of observing the event.

Table 1 Simulation results for different sample size n, number of visits s and time-horizon t for an irrelevant (S_1), a time-constant (S_2) and a time-dependent (S_3) marker

			AUC			1/2 NRI		
n	s	t	S_3	S_2	S_1	S_3 versus S_1	S_3 versus S_2	S_2 versus S_1
200	2	3.0	0.791	0.699	0.503	0.274	0.029	0.224
800	2	3.0	0.787	0.697	0.499	0.274	0.027	0.222
200	3	3.0	0.860	0.700	0.501	0.277	0.080	0.225
800	3	3.0	0.858	0.699	0.499	0.277	0.076	0.224
200	2	8.0	0.791	0.659	0.499	0.283	0.127	0.179
800	2	8.0	0.789	0.655	0.500	0.280	0.126	0.176
200	3	8.0	0.906	0.658	0.500	0.300	0.220	0.177
800	3	8.0	0.907	0.657	0.501	0.297	0.221	0.176

Results are overaged over $B = 1000$ replicates

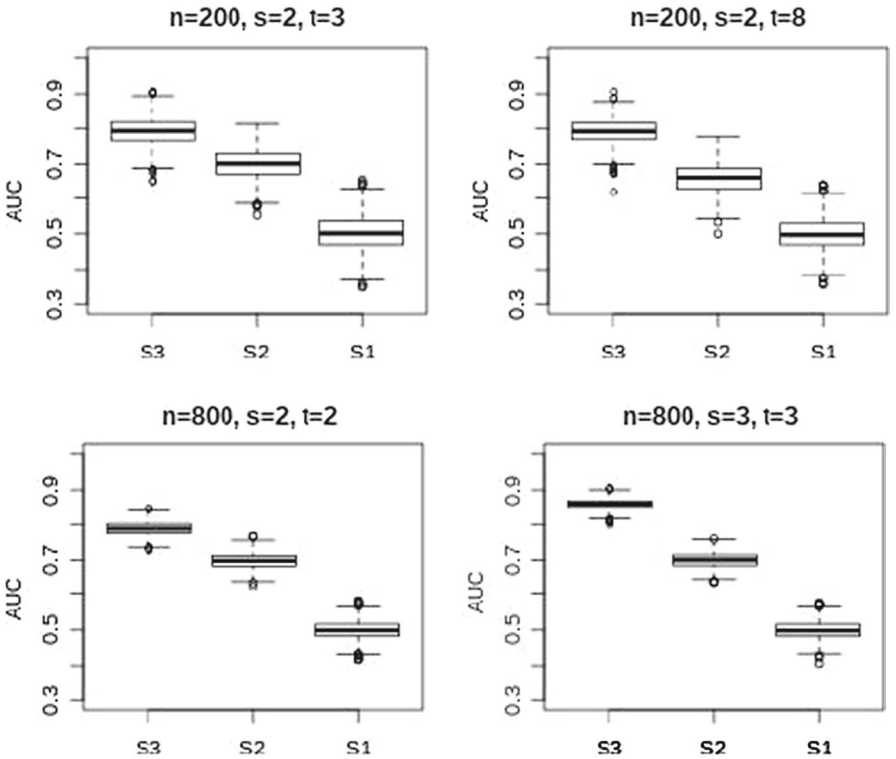

Fig. 1 Simulation results for different sample size n, number of visits s and time-horizon t for an irrelevant (S_1), a time-constant (S_2) and a time-dependent (S_3) marker. We report boxplots over $B = 1000$ replicates

6 Data analysis

The study population in our motivating data example included 452 patients, with at least one short-time evaluation during the follow-up period (on average at 6 months) and with a median follow-up of 147 months (IQR 91–217). The median age at enrollment was 46 years (IQR 36–55) and 68% were males. The median baseline LVEF was 31% (IQR 25–39%). RV-d dysfunction (*RV*) was observed in 96 patients at enrollment.

In our context, the clinical focus is on the RV-d parameter and its prognostic significance at a relatively long-term. We also focus on the information gained by monitoring this parameter over repeated visits, hence evaluating the effect of this variation. If a change in RV dysfunction is observed between two visits, a significant effect is expected on the survival status even taking into account other covariates.

The analysis of pre-specified follow-up evaluations were performed in 452, 327, 304, and 243 patients, respectively, on average at 6, 24, 48, and 72 months. Globally we observed 78 events of D/HT during the follow up, and nearly 50% of them within 72 months. RV-d promptly improved after the initiation of optimal treatment. While about 20% of patients have RV dysfunction at baseline, only 10% have the dysfunction at the first follow-up visit, and about 5% in the subsequent visits. Transitions between states, nevertheless, are frequent and patients generally tend to improve (the transition probability from RV-d to absence of RV-d is 72%). Noteworthy, RV normalization occurred earlier compared to LVEF improvement, as outlined in Merlo et al. (2016).

To evaluate the new marker *RV* we compute the following scores: *RV* alone ($S_1(t)$), $S_2(t) = 0.1 * age + 0.02 * HFd - 0.06 * LVEF(t)$, where age denotes age at enrollment and HFd the time between diagnosis of HF and baseline time (HF duration). Finally, we evaluate the complete score $S_3(t) = RV(t) - 0.06 * LVEF(t) + 0.1 * age + 0.02 * HFd$. All scores are time-varying, and time-dependent markers $RV(t)$ and $LVEF(t)$ have been evaluated at each of a maximum of four visits. We fix a time-horizon of $T = 72$ months for evaluation of prognostic accuracy, and compare the use of a single visit (usual time-dependent ROC) and multiple (up to four) visits. Weights in the scores above are far from being optimal and simply based on the Cox regression coefficients for the covariates fixed at their baseline measurements (i.e., as if everything was computed only once).

AUCs and 95% CIs are reported in Table 2, where bold indicates significance at the 5% level for the hypothesis $H_0 : AUC = 50\%$.

Table 2 AUC and 95% CIs in parentheses for three scores predicting occurrence of D/HT at $t = 72$ months for the DCM data

	S_1	S_2	S_3
$s = 1$	0.55 (0.48–0.63)	**0.73** (0.64–0.82)	**0.75** (0.66–0.83)
$s = 2$	**0.61** (0.52–0.70)	**0.75** (0.65–0.83)	**0.77** (0.69–0.86)
$s = 3$	0.64 (0.49–0.79)	**0.71** (0.51–0.90)	**0.76** (0.56–0.95)
$s = 4$	0.71 (0.44–0.97)	0.72 (0.40–1.00)	**0.81** (0.54–1.00)

AUCs in bold are significant at the 5% level

Table 3 1/2 NRI, NRI for events and for non-events when comparing $S_3(t)$ with $S_2(t)$ using a different number of visits; with 95% confidence intervals in parentheses and p values

	1/2 NRI	NRI events	NRI non-events
$s = 1$	0.126 (−0.047; 0.284) $p = 0.153$	−0.336 (−0.632; 0.442) $p = 0.186$	0.590 (0.446; 0.733) $p <$ 0.001
$s = 2$	0.200 (0.014; 0.385) $p = 0.0348$	−0.161 (−0.562; 0.239) $p = 0.429$	0.561 (0.469; 0.653) $p <$ 0.001
$s = 3$	0.200 (−0.108; 0.509) $p = 0.203$	−0.198 (−1.141; 0.746) $p = 0.681$	0.598 (0.375; 0.821) $p <$ 0.001
$s = 4$	0.529 (0.164; 0.893) $p = 0.004$	0.316 (−0.402; 1.035) $p = 0.388$	0.742 (0.561; 0.922) $p <$ 0.001

It could be noted that all AUCs are globally increasing with the number of visits, and that there is also at each visit an improvement in accuracy by adding RV-d to the LVEF score adjusted by age and HF duration. It seems like the best performing score is S_3, at each visit, and interestingly RV-d-alone AUC is significant at the second visit, which as noted above corresponds to the time when most RV improvements happened.

As often happens, there is no significance when comparing AUC values. As noted by Pencina et al. (2011), "*area under the curve (AUC) or C statistic hardly moves after a few good risk factors are already included in the model*". For this reason, in order assess the importance of $RV(t)$ we additionally evaluate 1/2 NRI to measure the reclassification improvement. We do so by evaluating how much information is added by $RV(t)$ when included in the best score without it, that is, $S_2(t)$. Results are reported in Table 3.

Here we could note that globally the reclassification improvement is significant at the second visit (i.e. when the maximal variation in the RV function is observed with respect to the baseline), and at the fourth visit, when more events are cumulatively observed. The improvement in reclassification seems to be driven by the non-events group, i.e., RV-d seems to allow to better classify patients who will not experience the event before $t = 72$. This is explained as at baseline RV dysfunction prevalence is high, and while some patients improve others do not. Here RV evolution is considered conditionally on the well-known LVEF marker path. These findings in our opinion are of remarkable importance in the long-term risk stratification and management of patients with DCM, especially in light that specific subgroups of patients often switch to more aggressive therapies, during follow up, when they are classified as high risk.

7 Conclusions

We have proposed a time-dependent ROC curve and time-dependent NRI that can take into account time-dependent biomarkers (or scores) repeatedly measured up to a certain time, $t_s \leq t$, where t is the time-horizon for observing the occurrence of events. Inference proceeds rather naturally through inverse probability weighting by the Kaplan–Meier estimator, and standard errors are obtained via resampling.

Our proposed measures require the specification of a desired time-horizon t and lower bound u, a number of visits to use s_i, $i = 1, \ldots, n$, and in general of a function $f_c(\cdot)$ of the trajectory (based on a threshold c). In our examples we always have used the union function, that is, we had a positive diagnosis if the marker was observed to be above a threshold in one or more of the first s_i visits. Other functions are possible and might be based for instance on persistent crossing (that is, a positive diagnosis if and only if two or more consecutive visits give a value above the threshold c) or simple increment (that is, a positive diagnosis if and only if in the second or later visits there is an increment above the threshold), etc.

It is important to underline that interpretation of $Se(t, c, s, u)$ and $Sp(t, c, s, u)$ is strongly dependent on s_i, as all survival probabilities up to time t are conditional on having survived up to time t_{s_i}. This makes it difficult to compare performance measures for different number of visits, as the additional marker measurements are not the only information used. Overcoming this issue is not at all straightforward, though.

Our approach can be used to assess the prognostic significance of scores and/or of new biomarkers. In our real data example we have obtained simple scores based on the Cox regression coefficients for the baseline covariates. We have done so for simplicity, overlooking that optimal weights might be better approximated when considering time-dependent covariates in Cox modeling. We additionally note that optimal weights can be directly obtained by maximization of AUC, rather than of Cox partial likelihood. A function to do so is available in the accompanying `R` package `longROC`. Our purpose was to show that RV dysfunction might bring about independent information for risk prediction even after considering LVEF and possible confounders, rather than to obtain an optimal score for risk prediction. In further work our time-dependent measures will be made resistant to the presence of outliers [see Farcomeni and Ventura (2012) for a detailed discussion on robust ROC curves]. To do so, we will exploit robust survival analysis methodologies [e.g., Farcomeni and Viviani (2011) and references therein].

Acknowledgements The authors are grateful to an AE and two referees for kind comments that helped improve the presentation.

References

Basili S, Loffredo L, Pastori D, Proietti M, Farcomeni A, Vestri AR, Pignatelli P, Davi G, Hiatt WR, Lip GY, Corazza GR, Perticone F, Violi F (2017) Carotid plaque detection improves the predictive value of CHA2DS2-VASc score in patients with non-valvular atrial fibrillation: the ARAPACIS study. Int J Cardiol 231:143–149

Cardellini M, Farcomeni A, Ballanti M, Morelli M, Davato F, Cardolini I, Grappasonni G, Rizza S, Gugliemi V, Porzio O, Pecchioli C, Menghini R, Ippoliti A, Federici M (2017) C-peptide: a predictor of cardiovascular mortality in subjects with established atherosclerotic disease. Diabetes Vasc Dis Res 4:395–399

Farcomeni A, Viviani S (2011) Robust estimation for the Cox regression model based on trimming. Biomet J 53:956–973

Farcomeni A, Ventura L (2012) An overview of robust methods in medical research. Stat Methods Med Res 21:111–133

Gerds TA, Kattan M, Schumacher M, Yu C (2013) Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. Stat Med 32:2173–2184

Gulati A, Ismail T, Jabbour A, Alpendurada F, Guha K, Ismail N, Raza S, Khwaja J, Brown T, Morarji K, Liodakis E, Roughton M, Wage R, Pakrashi T, Sharma R, Carpenter J, Cook S, Cowie M, Assomull

R, Pennell D, Prasad S (2013) The prevalence and prognostic significance of right ventricular systolic dysfunction in nonischemic dilated cardiomyopathy. Circulation 128:1623–1633
Heagerty P, Lumley T, Pepe M (2000) Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 56:337–344
Iacovelli R, Farcomeni A, Sternberg CN, Carteni G, Milella M, Santoni M, Cerbone L, Di Lorenzo G, Verzoni E, Ortega C, Sabbatini R, Ricotta R, Procopio G (2015) Prognostic factors in patients receiving third-line targeted therapy for metastatic renal cell carcinoma. J Urol 193:1905–1910
Jeong J-H, Jung S-H, Costantino JP (2008) Nonparametric inference on median residual life function. Biometrics 64:157–163
Jung S-H, Jeong J-H, Bandos H (2009) Regression on quantile residual life. Biometrics 65:1203–1212
Kurland BF, Johnson LL, Egleston BL, Diehr PH (2009) Longitudinal data with follow-up truncated by death: match the analysis method to research aims. Stat Sci 24:211–222
Li L, Greene T, Hu B (2017) A simple method to estimate the time-dependent receiver operating characteristic curve and the area under the curve with right censored data. Stat Methods Med Res. http://doi.org/10.1177/0962280216680239
Merlo M, Pyxaras S, Pinamonti B, Barbati G, Di Lenarda A, Sinagra G (2011) Prevalence and prognostic significance of left ventricular reverse remodeling in dilated cardiomyopathy receiving tailored medical treatment. J Am Coll Cardiol 57:1468–1476
Merlo M, Pivetta A, Pinamonti B, Stolfo D, Zecchin M, Barbati G, Di Lenarda A, Sinagra G (2014) Long-term prognostic impact of therapeutic strategies in patients with idiopathic dilated cardiomyopathy: changing mortality over the last 30 years. Eur J Heart Fail 16:317–324
Merlo M, Gobbo M, Stolfo D, Losurdo P, Ramani F, Barbati G, Pivetta A, Di Lenarda A, Anzini M, Gigli M, Pinamonti B, Sinagra G (2016) The prognostic impact of the evolution of right ventricular function in idiopathic dilated cardiomyopathy. J Am Coll Cardiol Cardiovasc Imaging 9:1034–1042
Pencina MJ, D'Agostino RB, Vasan RS (2008) Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 27:157–172
Pencina M, D'Agostino R, Steyerberg E (2011) Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med 30:11–21
Pencina MJ, D'Agostino RB, Pencina KM, Janssens CJW, Greenland P (2012) Interpreting incremental value of markes added to risk prediction models. Am J Epidemiol 176:473–481
Pepe MS (2003) The statistical evaluation of medical tests for classification and prediction. Oxford University Press, Oxford
Pignatelli P, Pastori D, Carnevale R, Farcomeni A, Cangemi R, Nocella C, Bartimoccia S, Vicario T, Saliola M, Lip GYH, Violi F (2015) Serum NOX2 and urinary isoprostanes predict vascular events in patients with atrial fibrillation. Thromb Haemost 113:617–624
Riggio O, Amodio P, Farcomeni A, Merli M, Pasquale C, Nardelli S, Pentassuglio I, Gioia S, Onori E, Piazza N, Montagnese S (2015) A model for the prediction of overt hepatic encephalopathy in patients with cirrhosis. Clin Gastroenterol Hepatol 13:1346–1352
Rizopoulos D (2011) Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. Biometrics 67:819–829
Robertson T, Wright FT, Dykstra RL (1988) Order restricted statistical inference. Wiley, New York
Schoop R, Graf E, Schumacher M (2008) Quantifying the predictive performance of prognostic models for censored survival data with time-dependent covariates. Biometrics 64:603–610
Uno H, Cai T, Tian L, Wei L (2007) Evaluating prediction rules for t-year survivors with censored regression models. J Am Stat Assoc 102:527–537
Uno H, Tian L, Cai T, Kohane I, Wei LJ (2013) A unified inference procedure for a class of measures to assess improvement in risk prediction systems with survival data. Stat Med 32:2430–2442
Venkatraman ES (2000) A permutation test to compare receiver operating characteristic curves. Biometrics 56:1134–1138
Zheng Y, Heagerty P (2004) Semiparametric estimation of time-dependent ROC curves for longitudinal marker data. Biostatistics 5:615–632
Zheng Y, Heagerty P (2005) Partly conditional survival models for longitudinal data. Biometrics 61:379–391