Journal of Biomedical Informatics 108 (2020) 103496

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models

Enrico Longato[a], Martina Vettoretti[a], Barbara Di Camillo[a,*]

[a] Department of Information Engineering, University of Padova, Padova, Italy

ARTICLE INFO

Keywords:
Concordance index
Predictive modelling
Survival analysis
Time-to-event models

ABSTRACT

Developing a prognostic model for biomedical applications typically requires mapping an individual's set of covariates to a measure of the risk that he or she may experience the event to be predicted. Many scenarios, however, especially those involving adverse pathological outcomes, are better described by explicitly accounting for the timing of these events, as well as their probability. As a result, in these cases, traditional classification or ranking metrics may be inadequate to inform model evaluation or selection. To address this limitation, it is common practice to reframe the problem in the context of survival analysis, and resort, instead, to the concordance index (C-index), which summarises how well a predicted risk score describes an observed sequence of events. A practically meaningful interpretation of the C-index, however, may present several difficulties and pitfalls. Specifically, we identify two main issues: i) the C-index remains implicitly, and subtly, dependent on time, and ii) its relationship with the number of subjects whose risk was incorrectly predicted is not straightforward. Failure to consider these two aspects may introduce undesirable and unwanted biases in the evaluation process, and even result in the selection of a suboptimal model. Hence, here, we discuss ways to obtain a meaningful interpretation in spite of these difficulties. Aiming to assist experimenters regardless of their familiarity with the C-index, we start from an introductory-level presentation of its most popular estimator, highlighting the latter's temporal dependency, and suggesting how it might be correctly used to inform model selection. We also address the nonlinearity of the C-index with respect to the number of correct risk predictions, elaborating a simplified framework that may enable an easier interpretation and quantification of C-index improvements or deteriorations.

1. Introduction

Predictive models have demonstrated a great potential for shaping a wide variety of processes in the field of biomedicine [1–4], ranging from medical decision-making [5,6] to proactive community-wide intervention [7,8]. Their perceived usefulness, combined with the continuous removal of the technological and methodological barriers that had historically hindered their application, has led to the development of a plethora of new models and prognostic indicators. While, on the one hand, this is a testament to the outstanding steps taken towards integrating analytical models in the management of patients and populations, in and out of the clinics; on the other, it underlines the importance of proper model evaluation. In fact, out of dozens, or even hundreds of proposed models, only a very limited number eventually become part of clinical guidelines or of real-life infrastructure, and those that do have passed several rounds of rigorous testing and validation [9].

One of the main difficulties in evaluating prognostic models is their reliance on time-to-event data, i.e., on data that pertain to both whether an outcome of interested has happened, and to when it occurred [9]. Longitudinal health data are usually collected in this fashion, traditionally during clinical trials, whose participants undergo periodic visits to check for changes in their health states [10]. More recently, owing to the increased focus on the secondary use of data, predictive models have also been based on the content of big data repositories [11]. Nevertheless, despite the different scope and challenges of this emerging approach compared to the clinical trial setting, the most natural prediction framework remains based on time-to-event data, obtained by considering each record in the database together with its timestamp.

In this setting, where timing is crucial, indicators that quantify performance in terms of classification ("event" vs. "no event") or ranking ("high risk" vs. "low risk") may not be sufficient to characterise all notable properties of a prognostic model. Hence, to obtain a comprehensive picture, it is customary to adopt the point of view of survival

* Corresponding author at: Department of Information Engineering, University of Padova, 35131 Padova, Italy.
E-mail addresses: enrico.longato@dei.unipd.it (E. Longato), martina.vettoretti@dei.unipd.it (M. Vettoretti), barbara.dicamillo@unipd.it (B. Di Camillo).

https://doi.org/10.1016/j.jbi.2020.103496
Received 22 October 2019; Received in revised form 12 May 2020; Accepted 24 June 2020
Available online 09 July 2020

analysis [12–14]. Specifically, the most popular alternative to standard metrics, such as accuracy and the area under the ROC curve (AUROC), is the concordance index (C-index) [15,16]. The C-index can quantify the correlation between risk predictions and event times in such a way that maximising this metric means improving the quality of discrimination between early events, associated with higher-risk subjects, and later occurrences. Its ability to summarise three different dimensions of predictions (risk, event occurrence, and time) in a single number enables to distinguish, at a glance, between well-behaved models and quasi-random ones. However, the concise nature of the C-index also renders its practical interpretation more difficult than that of classification and ranking metrics. Furthermore, although this is not immediately apparent from its standard formulation, the C-index maintains an implicit dependency on time, which, albeit sometimes overlooked, is fundamental to extract meaningful information on model performance [17].

In light of these considerations, here, we present an overview of the C-index with particular emphasis on its applications to the evaluation and selection of prognostic models in terms of discrimination ability (calibration, though also crucial, cannot be measured with the C-index and is, thus, out of scope for this work). Our primary objective is enabling an easier and more constructive interpretation of the C-index in practical scenarios. Hence, after briefly recalling the definition of C-index, we will discuss its most popular estimator, and show its properties referencing the standard algorithm used to compute it. Specifically, we will characterise the residual temporal dependency of the C-index, highlighting how experimenters could use it to their advantage. Furthermore, we will show that, unlike accuracy and the AUROC, the numerical value of the C-index is not linearly dependent on the number of subjects whose risk has been correctly predicted. To overcome this limitation, we will derive a direct relationship with the number of errors, which, albeit based on a fairly strict set of hypotheses, we believe may be useful to quantify the entity of improvements or deteriorations of the C-index across different models. Finally, we will apply the concepts we introduced throughout the text to discuss an illustrative and a real-data examples of model selection.

2. The concordance index

2.1. Definition

The C-index [15,16] is the most widely used metric for the global evaluation of prognostic models in survival analysis. In analogy with the AUROC, the C-index is a summary of model performance across a wide array of possible configurations. Its distinguishing feature, however, is that it can characterise a model's prognostic ability by accounting for both outcome occurrence and timing.

Before giving the standard definition of C-index, let us recall the formal meaning of some key terms used in survival analysis. For the purposes of the present work, a "model" is simply a map from the subject space (usually, a set of individual covariates) to a subset of $\mathbb{R}$, which establishes a relationship between the i-th subject and the corresponding score $M_i \in \mathbb{R}$. Specifically, we deal with "risk" scores, i.e., with measures of the probability (or a proxy thereof) that an event might happen to a subject. In particular, an "event" (also called "death" or "failure") is an inevitable occurrence that permanently alters the status of a subject. Conversely, if we stop observing a subject before he or she experiences an event, we perform the act of "censoring," and we end up with a "censored" data point (or "partial observation"). In the dataset, we record all event and censoring times as observation times T_i^{obs}. To distinguish between the two, we introduce the binary variable Δ_i, which is equal to 1 if and only if the i-th subject experienced an event at some point in time T_i, and to 0 if he or she was censored before then. Notably, if $\Delta_i = 1$, then $T_i^{obs} = T_i$; on the contrary, if $\Delta_i = 0$, we can only say that the event happens after the censoring time T_i^{obs}, but the exact time $T_i > T_i^{obs}$ is unknown.

Ultimately, treating the scores and event times of two generic subjects (i and j) as random variables, we can define the C-index as the following probability, conditioned on the relative order of events.

$$C = P(M_j > M_i | T_j < T_i) \tag{1}$$

Hence, the C-index tries to describe how well a model conforms to the following logic: "Greater scores should be attributed to the subjects who are at a higher risk of experiencing an event." More accurately, a "good" model according to the C-index ($C = 1$) is one that always assigns higher scores to the subjects who experience the earlier events. Note that this may not always be the most appropriate definition of goodness of fit (e.g., when the highest risk is, in fact, related with long-term outcomes); nevertheless, it is the most common in survival analysis, where many techniques (e.g., Cox regression [17,18]) assume the existence of a monotonic map between event probabilities and onset times.

2.2. Harrell's estimator of the C-index

The de-facto standard way to compute the C-index is using Harrell's estimator, i.e., the naïve estimator of the probability in Eq. (1) [15]. Harrell's estimator works by computing the percentage of comparable pairs within the dataset whose relative risk was correctly identified by the model. A pair of subjects (i, j) is "comparable" if we can determine which of them (i or j) was the first to experience an event. A comparable pair (i, j) is also "concordant" if the subject who experiences the earlier event is identified as having the greater risk, and "discordant" otherwise. Harrell's well-known algorithm (see Supplementary Material for a full description using our notation) yields the C-index in the form of a ratio between the number of concordant and comparable pairs:

$$\widehat{C} = \frac{\sum_{i=1}^{N} \Delta_i \sum_{j=i+1}^{N} \left[I(T_i^{obs} < T_j^{obs}) + (1 - \Delta_j) I(T_i^{obs} = T_j^{obs}) \right] \left[I(M_i > M_j) + \frac{1}{2} I(M_i = M_j) \right]}{\sum_{i=1}^{N} \Delta_i \sum_{j=i+1}^{N} \left[I(T_i^{obs} < T_j^{obs}) + (1 - \Delta_j) I(T_i^{obs} = T_j^{obs}) \right]} \tag{2}$$

where $I(\cdot)$ is the indicator function that is equal to 1 if its argument is true, and 0 otherwise.

If we further hypothesise tied times ($T_i^{obs} = T_j^{obs}$) and tied scores ($M_i = M_j$) to be impossible, we can obtain a simplified version of Eq. (2). This does not change the emergent properties of the estimator, but makes for a much more approachable formulation, i.e.,

$$\widehat{C} = \frac{\sum_{i=1}^{N} \Delta_i \sum_{j=i+1}^{N} I(T_i^{obs} < T_j^{obs}) I(M_i > M_j)}{\sum_{i=1}^{N} \Delta_i \sum_{j=i+1}^{N} I(T_i^{obs} < T_j^{obs})} \tag{3}$$

2.3. Properties of Harrell's C-index

In this section, referencing Eq. (3) (extension to Eq. (2) is trivial), we will discuss several properties of Harrell's estimator, which are critical for a correct interpretation of the C-index, and, most importantly, for its practical application.

Let us start by showing that $\widehat{C}$ is a reasonable estimator for a probability, i.e., that $0 \leq \widehat{C} \leq 1$, and by specifying when equality holds. We can do that by comparing the numerator and denominator in Eq. (3). The only difference between them is that all terms in the numerator are multiplied by an indicator function in the form $I(M_i > M_j)$, which, by definition, can be equal to either 0 or 1. Hence, if all comparable pairs (i, j) have been assigned scores that reflect the correct order of events, then $I(M_i > M_j) = 1$ for all (i, j), and . On the contrary, if all pairs are discordant ($M_i < M_j$, even though $T_i^{obs} < T_j^{obs}$), the numerator collapses to 0, and so does $\widehat{C}$. In all intermediate situations, the indicator functions will be equal to 1 only for the concordant portion of the comparable pairs, guaranteeing that $\widehat{C}$ is bounded by the extreme cases discussed above (perfect concordance and anti-concordance).

Now that we have established that $\widehat{C}$ is estimating a probability, we should verify whether it matches the definition of C-index reported in Eq. (1). For a rigorous discussion of this problem, we refer the reader to previous works by Pencina and D'Agostino [17], Uno et al. [18], and Heagerty and Zheng [19], where it was demonstrated that $\widehat{C}$, in fact, converges to

$$\widehat{C} \rightarrow \mathrm{P}(\mathrm{M_i} > \mathrm{M_j} | \mathrm{T_i} < \mathrm{T_j}, \Delta_\mathrm{i} = 1, \mathrm{T_i} < \tau) \tag{4}$$

Eq. (4) states that, although Harrell's estimator calculates the probability of a model correctly ordering two subjects' scores (M_i and M_j) according to the corresponding event times (T_i and T_j), the final estimate is also conditioned on the censoring distribution (Δ_i), and on the fact that any realistic dataset is finite (τ; see below). This relationship has crucial repercussions on the interpretation of the C-index. To show why, we need to determine $\widehat{C}$'s update schedule. Preliminarily, recall that we assume the dataset to be sorted according to T_i^{obs} (and, thus, T_j^{obs}), and that only the pairs (i, j) where i experiences an event contribute to the numerator and denominator. This is reflected, in Eq. (3), by the fact that all terms in the outer summations (the ones for $i = 1, \cdots, N$) are pre-multiplied by Δ_i. Consequently, the subjects who experience an event play the role of i in as many comparisons as there are subjects who follow them in chronological order, and of j in as many as there are events that happen before them; whereas, censored observations can only be compared to the events that predate them, and, thus, can only play the role of j. Let us now consider what happens if the last event in the dataset occurs to the K-th subject, with $K < N$ at time $\tau = T_K = T_K^{obs}$, and the remaining $N - K$ observations are censored. In such a situation, we can rewrite Eq. (3) given that $\Delta_i = 0$ for $i > K$.

$$\begin{aligned}\widehat{C_\tau} &= \frac{\sum_{i=1}^{K} \Delta_i \sum_{j=i+1}^{N} I(T_i^{obs} < T_j^{obs}) I(M_i > M_j)}{\sum_{i=1}^{K} \Delta_i \sum_{j=i+1}^{N} I(T_i^{obs} < T_j^{obs})} \\ &= \frac{\sum_i \Delta_i \sum_i I(T_i^{obs} < T_j^{obs}, T_i^{obs} \leq \tau) I(M_i > M_j)}{\sum_i \Delta_i \sum_j I(T_i^{obs} < T_j^{obs}, T_i^{obs} \leq \tau)}\end{aligned} \tag{5}$$

In the second half of Eq. (5), we have made the dependence on the last event's time τ explicit and removed the now-redundant summation limits, which are univocally determined by τ. We can conclude that Harrell's estimator of the C-index receives its final update at the time τ of the last event in the dataset. Furthermore, the relative order and exact timing of the trailing censored observations are immaterial, i.e., it only matters that $T_j^{obs} > T_K^{obs}$ for $j = K + 1, \cdots, N$. This is a consequence of applying the commutative property of addition to all terms with $j > K$, and of the way the logical condition $I(T_i^{obs} < T_j^{obs})$ is formulated. Similar considerations apply with respect to the relative order and exact timing of a group of censored observations that fall between two events: as long as the group's relationships with the immediately preceding and immediately following events are preserved as a whole, $\widehat{C_\tau}$ is unchanged.

To summarise, we have shown that the K-th subject, who experiences the last event in the dataset at time τ, is the latest contributor to $\widehat{C_\tau}$. In other words, consistently with Eq. (5), when the algorithm finishes the K-th iteration, it yields the final value of $\widehat{C_\tau}$, regardless of how many other censored observations may follow. This, in turn, suggests that the C-index is truly indicative of model performance in a practical sense only when it is accompanied by a clear indication of τ. Specifically, while the numerical value of the C-index relates to "how well" a model has been shown to work, τ specifies whether the model has been tested "far enough" into the future. Absent τ, we are unable to determine whether a model performs adequately at the desired prediction horizon. For example, if τ turned out to be much shorter than needed or initially thought, model performance at the desired time range would be effectively unknown. Most importantly, in a typical time-to-event scenario, τ can only be found by directly inspecting the data used to compute the C-index and finding the time of the last event (index K above), which, in principle, can happen any time before the last recorded data point; hence, the burden of declaring τ falls entirely on the experimenters.

A useful generalisation of the concept of τ comes from considering it as a halting point for Harrell's algorithm. Recall that, so far, τ has been the time of the last event in the dataset, which happened to the K-th subject and was followed only by censored observations. However, while this choice of τ is implicit in the way Harrell's C-index is computed, it does not hinge on any particular property of the K-th subject in question, other than the fact he or she experiences an event. It follows that any other subject $K^* < K$ with event time could be used to set a halting point and obtain an intermediate value $\widehat{C_{\tau^*}}$ ($\tau^* < \tau$) of the estimator. In light of these considerations and of the comments we have made regarding Eq. (5), we can confirm that, in practice, Harrell's C-index is only updated at event times, remains constant between events, and reaches its final value at the time τ of the last event in the dataset. Perhaps more importantly, however, explicitly marking τ^* as a halting point (after which all observations are censored) establishes a dichotomy between an "earlier" ($T_i^{obs} \leq \tau^*$) and a "later" portions of the data ($T_i^{obs} > \tau^*$). In the "earlier" portion, identifying an exact correspondence between risks (M_i) and event times (T_i) is paramount; in the "later" portion, the relative order of events is not as important as the correct determination that they happened after a certain cut-off time τ^*.

The ability of $\widehat{C_{\tau^*}}$ to capture this kind of distinction and quantify performance at relevant values of τ^* can be extremely valuable for the evaluation and selection of prognostic models. As we have recalled, it is usually crucial that a good model identify a subject who is at a high risk of experiencing a sudden event, and that it be able to predict the timing of such an occurrence. Conversely, depending on the application, it might be misleading to penalise a model just because it failed to establish the exact ordering between low-risk subjects, whose events are predicted to happen far into the future. In this sense, $\widehat{C_{\tau^*}}$ intuitively appears to be a suitable metric to obtain meaningful results, provided that we set an appropriate value of τ^*, e.g., based on domain knowledge or on project requirements. Sections 4 present two examples of how to use $\widehat{C_{\tau^*}}$ for model selection.

3. Interpretation of the numerical value of the C-index

In the literature, introductory presentations of the C-index are usually accompanied by a remark on some notable values of $\widehat{C}$. Specifically, starting from the seminal work by Harrell [15], it is common practice to recall that $\widehat{C} = 1$ implies perfect concordance between risks and event times, $\widehat{C} = 0$ describes the opposite situation of perfect anti-concordance (easily fixable by considering $-M_i$ instead of M_i as the predicted score), and $\widehat{C} = 0.5$ denotes completely random assignments. Conversely, to the best of our knowledge, the interpretation of the intermediate values of $\widehat{C}$ ($0.5 < \widehat{C} < 1$) has not been given the same amount of attention. Hence, there is no clear indication on the relationship between an arbitrary value of $\widehat{C}$ and the practical situation at hand, nor on whether an improvement of 0.05 in terms of C-index would be as substantial when going from $\widehat{C} = 0.70$ to $\widehat{C} = 0.75$ vs. from $\widehat{C} = 0.90$ to $\widehat{C} = 0.95$. Note that, among standard performance metrics, this issue is peculiar to the C-index. For instance, in binary classification, a difference in accuracies of 0.05 is easily interpretable, in absolute or relative terms, because it corresponds to an increase in the number of correctly classified subjects that is equal to 5% of the total sample size. Similarly, an AUROC that improves by 0.05 signals that a greater number of subjects (again, 5% of the total) has been ranked correctly, and further information is discernible from a direct comparison between the two ROC curves. On the contrary, if we applied a similar reasoning to the C-index we could only claim that a greater number of pairs was concordant, relative to the total number of comparable pairs. However, the relationship between $\widehat{C}$ and the number of subjects with correct or incorrect scores would remain opaque.

To address this problem, we propose a simplified view of the C-

index that establishes a direct relationship between $\widehat{C}$ (or, similarly, $\widehat{C_\tau}$ or $\widehat{C_{\tau^*}}$) and the number of subjects whose scores are in the correct order, relative to their event times. Abstracting from the actual composition of the dataset under scrutiny, we consider the simplest way a given value of $\widehat{C}$ could have been generated. We are, then, able to derive a simplified formula to explicitly convert between $\widehat{C}$ and the number of subjects with incorrectly arranged scores. Given the extremely general nature of the C-index, and its ability to capture time-dependent information even in the presence of censored data, we proceed as follows: first, we devise a useful reformulation of the C-index in the absence of censoring (Section 3.1); second, we use this reformulation to derive a direct relationship between $\widehat{C}$ and the number of prediction errors in the simplest, most intuitive scenario (Section 3.2); finally, we discuss the effect of censoring on the reformulation (Section 3.3).

3.1. A reformulation of the C-index in the absence of censoring

In this Section, to highlight the determining factors that cause a deterioration or improvement of the C-index, we present a reformulation of Harrell's C-index, hinging on the following four hypotheses.

H1. All events have been observed ($\Delta_i = 1$ for $i = 1, \cdots, N$) and there are no ties. We will discuss the consequences of relaxing this hypothesis in Section 3.3.
H2. The scores M_i are sorted chronologically and organised in N_R contiguous regions R_h ($h = 1, \cdots, N_R$), comprising a known number N_h of subjects such that $\sum_{h=1}^{N_R} N_h = N$.
H3. The scores in each region R_h exhibit a consistent internal behaviour in the sense measured by the C-index. We quantify the degree of concordance within region R_h with a coefficient α_{hh}. For instance, all scores may perfectly follow the corresponding event times ($\alpha_{hh} = 1$), or their relative position within R_h might be random ($\alpha_{hh} = 0.5$).
H4. The regions R_h are externally concordant, i.e., with a slight abuse of notation, it holds that $\min_{i \in R_h}\{M_i\} > \max_{j \in R_{h+1}}\{M_j\}$. Note that, although this is a strong constraint, its impact on generalisation ability is reduced by two factors: 1) a well-trained model, at a minimum, should be able to identify a risk trend in the data; 2) it is always possible to specify a set of regions that conform to this hypothesis (different sets will have different degrees of internal concordance α_{hh}).

An example of such a dataset is reported in Fig. 1, where the first 30% of the scores is compatible with perfect concordance, the second 40% have been assigned in a random order, and the remaining 30% is, again, perfectly sorted. Using the symbols introduced above, we have three regions $\{R_1, R_2, R_3\}$ with $N_1 = 0.3{\cdot}N$, $N_2 = 0.4{\cdot}N$, $N_3 = 0.3{\cdot}N$, and $\alpha_{11} = 1$, $\alpha_{22} = 0.5$, $\alpha_{33} = 1$. Equivalently, we could have specified only two regions $\{R_1, R_2\}$ with $N_1 = 0.7{\cdot}N$ and $N_2 = 0.3{\cdot}N$, and $\alpha_{11} = 0.837$, $\alpha_{22} = 1$.

Under hypotheses H1-H3, we can reframe the formula for $\widehat{C}$ (Eq. (3)) in terms of internal and external comparisons among the regions R_h ($h = 1, \cdots, N_R$). For the sake of conciseness, let $R_h \times R_l$ denote the number of comparable pairs (i, j) such that $i \in R_h$ and $j \in R_l$ ($h \leq l$). Also let α_{hl} be the proportion of concordant pairs resulting from the comparison between R_h and R_l. Harrell's estimator, then, becomes

$$\widehat{C} = \frac{\sum_{h=1}^{N_R}\sum_{l=h}^{N_R} \alpha_{hl} R_h \times R_l}{\sum_{h=1}^{N_R}\sum_{l=h}^{N_R} R_h \times R_l} \tag{6}$$

We can now split the summations in Eq. (6) to distinguish between internal ($h = l$) and external ($h < l$) comparisons, thus obtaining

$$\widehat{C} = \frac{\sum_{h=1}^{N_R} \alpha_{hh} R_h \times R_h + \sum_{h=1}^{N_R-1}\sum_{l=h+1}^{N_R} \alpha_{hl} R_h \times R_l}{\sum_{h=1}^{N_R} R_h \times R_h + \sum_{h=1}^{N_R-1}\sum_{l=h+1}^{N_R} R_h \times R_l} \tag{7}$$

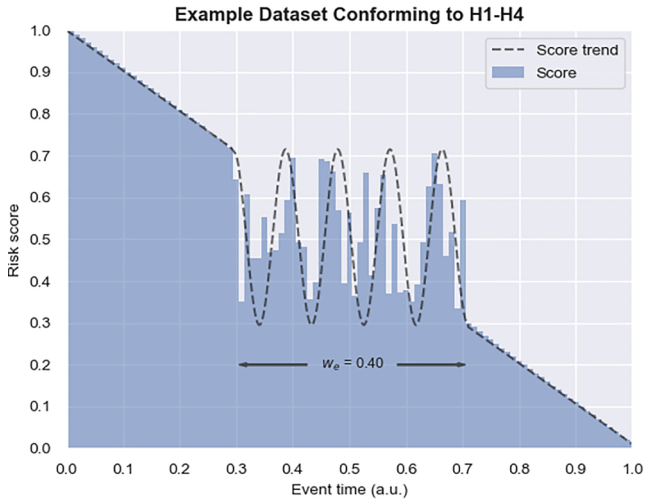

Fig. 1. Example dataset. Schematic representation of a dataset for which hypotheses H1-H4 hold. The height and position of the blue bars denote, respectively, the risk score (M_i) and observation/event time ($T_i^{obs} = T_i$) associated with each subject (i). The dashed outline follows the main trend in risk assignment (straight line for perfect sorting, sinusoidal wave for random relative order).

Since we have not applied H4 yet, Eqs. (6) and (7) are strictly equivalent to Eq. (3) in the absence of censoring, provided that the regions and the corresponding coefficients are correctly specified.

Leveraging hypothesis H4, we observe that all terms α_{hl} with $h < l$ are, in fact, equal to 1. In other words, if H4 holds, all pairs (i, j) such that $i \in R_h$ and $j \in R_l$ ($h < l$) must be concordant. Incorporating this result into Eq. (7), after slightly rearranging it to highlight the dependency on the degree of internal concordance ($h = 1, \cdots, N_R$), yields

$$\widehat{C} = 1 - \frac{\sum_{h=1}^{N_R} (1 - \alpha_{hh}) R_h \times R_h}{\sum_{h=1}^{N_R} R_h \times R_h + \sum_{h=1}^{N_R-1}\sum_{l=h+1}^{N_R} R_h \times R_l} \tag{8}$$

Note that a correct specification of regions and coefficients continues to be possible even after applying H4: although more interesting representations are usually obtainable, we can always define a single region R_1 with $\alpha_{11} = \widehat{C}$. Most importantly, however, Eq. (8) highlights that the deterioration of $\widehat{C}$ is determined only by the proportion of discordant pairs $(1 - \alpha_{hh})$ within each region R_h.

This relationship can also be expressed in terms of the sample size N. Specifically, the denominator of Eqs. (7) and (8) is the total number of comparable pairs, which, according to hypothesis H1, is

$$Denominator\,[\widehat{C}] = \frac{N(N+1)}{2} \simeq \frac{N^2}{2} \tag{9}$$

Similarly, the number of comparable pairs obtained by comparing with each other the N_h subjects who belong to a generic region R_h can be approximated as

$$R_h \times R_h = \frac{N_h(N_h+1)}{2} \simeq \frac{N_h^2}{2} \tag{10}$$

Plugging Eqs. (9) and (10) into Eq. (8) yields

$$\hat{C} \simeq 1 - \frac{\sum_{h=1}^{N_R} (1 - \alpha_{hh}) \frac{N_h^2}{2}}{\frac{N^2}{2}} = 1 - \sum_{h=1}^{N_R} (1 - \alpha_{hh}) w_h^2 \tag{11}$$

where we have defined $w_h = \frac{N_h}{N}$, which we can interpret as the relative width of a generic region R_h expressed as a fraction of the sample size N.

3.2. Establishing a direct relationship between the C-index and the number of errors

In Section 3.1, we have obtained a fairly general formulation of $\widehat{C}$, valid under hypotheses H1-H4. The main advantage of Eq. (11) is that it explicitly links the value of the C-index to the number of samples within each region of the dataset through w_h ($h = 1, \cdots, N_R$). However, the formula also retains a direct dependency on the proportion of concordant pairs α_{hh}, which, as we have discussed, is an obstacle to $\widehat{C}$'s interpretation because it is not directly related to the number of subjects with correct or incorrect risk scores. To solve this problem, we need to further narrow the possible range of situations we intend to describe with Eq. (11). On the one hand, this will inevitably result in a loss of generality and flexibility; on the other, it will enable us to intuitively grasp the relationship between an improvement or deterioration of the C-index and the corresponding increase or decrease in correctly attributed risk scores. The process hinges on the idea that any value of $\widehat{C}$ such that $0.5 < \widehat{C} < 1$ can be obtained either from the evaluation of the actual risk-event relationship in the dataset, or by creating an equivalent scenario that conforms to hypotheses H1-H4. The simplest way to construct such a scenario is postulating the presence of a single region R_e of width w_e where the relative order between the scores M_i such that $i \in R_e$ is random, while all pairs involving the remaining subjects ($i \notin R_e$) are perfectly concordant. Using the symbols introduced in Eq. (6), this means that $\alpha_{ee} = 0.5$, while $\alpha_{hh} = 1$ for $h \neq e$ (recall that H4 implies that all the other α coefficients are equal to 1). Note that the position of R_e does not matter, i.e., $\widehat{C}$ does not change between the three scenarios $\{R_1, R_e\}$, $\{R_e, R_2\}$, and $\{R_1, R_e, R_3\}$, as long as the width w_e is consistent. We can compute the resulting value of $\widehat{C}$ as a special case of Eq. (11), thus obtaining

$$\widehat{C} \simeq 1 - 0.5 w_e^2 \tag{12}$$

This formula establishes a direct relationship between $\widehat{C}$ and the number (technically, the percentage) of subjects who, in our ideal scenario, have been assigned incorrect scores. Furthermore, since we are interested in $0.5 < \widehat{C} < 1$, the relationship is also invertible, i.e., we can express the width of the unsorted region R_e as a function of the C-index.

$$w_e \simeq \sqrt{2(1 - \widehat{C})} \tag{13}$$

Eqs. (12) and (13) are especially useful to quantify, although in an idealised scenario, the nonlinear relationship between $\widehat{C}$ and the number of errors, represented by w_e. The implications of such nonlinearity are, perhaps, more easily apparent from Table 1 and Fig. 2, where we have listed some values of $\widehat{C}$ and the corresponding w_e's. Specifically, we observe the following.

1. If $\widehat{C} = 1$, the region R_e does not exist (technically, $w_e = 0$), which is consistent with how we have constructed the scenario used to derive Eqs. (12) and (13).
2. $\widehat{C} = 0.5$ corresponds to R_e being the only region with $w_e = 1$, i.e., to a dataset where the order of the scores M_i is completely random with respect to the event times T_i.
3. Seemingly minor improvements of excellent values of $\widehat{C}$ are, in fact, extremely substantial when viewed in terms of number of unsorted samples. Case in point, great effort is required to retune a model so that $\widehat{C}$ increases from 0.95 to 0.99 (region R_e needs to be much narrower, from $w_e = 0.32$ to $w_e = 0.14$); whereas going from $\widehat{C} = 0.75$ to $\widehat{C} = 0.80$ does not involve as drastic a change in risk-assignment capabilities (from $w_e = 0.71$ to $w_e = 0.63$).
4. The number of unsorted subjects might seem alarmingly high even for reasonably satisfactory values of $\widehat{C}$ (e.g., if $\widehat{C} = 0.80$, $w_e = 0.63$). However, this should not be a cause for concern, because it is but a by-product of the simplifications leading to Eqs. (12) and (13). In fact, hypothesis H4 implies that the model we are ideally considering is able to capture the general trend of the relationship between risk and event times. Hence, it is unsurprising that such a model would exhibit good performance, even for moderately large values of w_e.

Table 1
Relationship between the C-index $\widehat{C}$ and the proportion of subjects with incorrectly assigned scores w_e. The values of w_e (second column of each set) have been computed by plugging the corresponding values of $\widehat{C}$ (first column of each set) into Eq. (13).

$\widehat{C}$	w_e	$\widehat{C}$	w_e
1.00	0	0.75	0.71
0.99	0.14	0.74	0.72
0.98	0.20	0.73	0.73
0.97	0.24	0.72	0.75
0.96	0.28	0.71	0.76
0.95	0.32	0.70	0.77
0.94	0.35	0.69	0.79
0.93	0.37	0.68	0.80
0.92	0.40	0.67	0.81
0.91	0.42	0.66	0.82
0.90	0.45	0.65	0.84
0.89	0.47	0.64	0.85
0.88	0.49	0.63	0.86
0.87	0.51	0.62	0.87
0.86	0.53	0.61	0.88
0.85	0.55	0.60	0.89
0.84	0.57	0.59	0.91
0.83	0.58	0.58	0.92
0.82	0.60	0.57	0.93
0.81	0.62	0.56	0.94
0.80	0.63	0.55	0.95
0.79	0.65	0.54	0.96
0.78	0.66	0.53	0.97
0.77	0.68	0.52	0.98
0.76	0.69	0.51	0.99

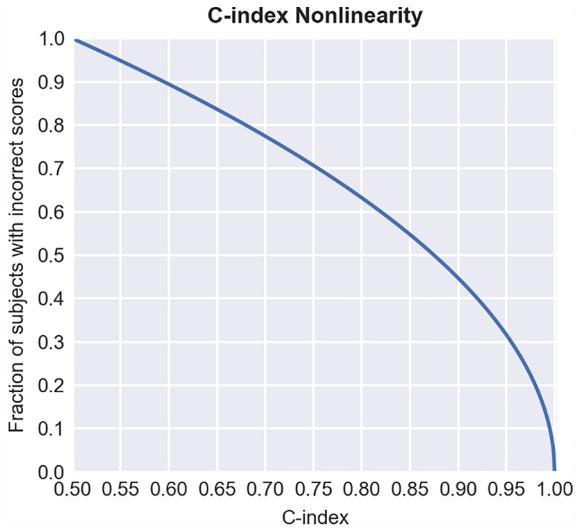

Fig. 2. Graphical representation of the nonlinear relationship between the C-index $\widehat{C}$ and the proportion of subjects with incorrectly assigned scores w_e. The values of w_e (y axis) have been computed by plugging the corresponding values of $\widehat{C}$ (x axis) into Eq. (13).

In summary, we have shown that a direct interpretation of the C-index in terms of the number of subjects with correct or incorrect risk scores presents a number of difficulties. This is at variance with other traditional metrics such as accuracy and the area under the ROC curve, which can easily characterise performance differences. To overcome this issue, we analysed the C-index in a simplified scenario, establishing a relationship between $\widehat{C}$ and the percentage of subjects w_e with

improperly assigned scores. In this way, we quantified the nonlinear behaviour of $\widehat{C}$ at different points of the performance scale. Specifically, we highlighted that improving a $\widehat{C}$ that is close to 1 is difficult not only because of the obviously great effort needed to outperform an extremely satisfactory model; but also due to the inherent need to correctly assign risks to more subjects to obtain the same numerical improvement at the top end of the C-index scale.

3.3. Extension to censored data

In Sections 3.1 and 3.2 we operated under the assumption of absence of censoring (hypothesis H1). While this turned out to be advantageous to characterise the main drivers of C-index deteriorations, it could be argued that real-life datasets are seldom devoid of censoring. Hence, in this Section, we will demonstrate the repercussions of relaxing hypothesis H1 to allow censored data within each region R_h, thus deriving analogues to Eqs. (7), (8), and (11). A complete discussion of the relationship between the C-index and censoring is outside of the scope of the present, introductory-level work. Nevertheless, the content of this Section might serve as a useful starting point, e.g., to develop other simplified models of $\widehat{C}$ to quantify its nonlinearity, as we have done in Section 3.2.

Let β_h be the fraction of censored data points within region R_h (defined according to hypotheses H2 and H3), and γ_h the effect of censoring on the degree of internal concordance within region R_h. While β_h and γ_h are obviously related, their relationship is not deterministic, but it depends on the position of the censored observations within R_h. Particularly, γ_h's lower and upper bounds are reached when all the censored observations are, respectively, at the beginning and end of region R_h. These bounds can be approximated via Eq. (10) as

$$(1-\beta_h)^2 \lesssim \gamma_h \lesssim 1-\beta_h^2 \tag{14}$$

Recall that, as per Eq. (3), censored data points are never compared to following observations. Hence, we can rewrite Eq. (7) as

$$\widehat{C} = \frac{\sum_{h=1}^{N_R}\gamma_h\alpha_{hh}R_h \times R_h + \sum_{h=1}^{N_R-1}\sum_{l=h+1}^{N_R}(1-\beta_h)\alpha_{hl}R_h \times R_l}{\sum_{h=1}^{N_R}\gamma_h R_h \times R_h + \sum_{h=1}^{N_R-1}\sum_{l=h+1}^{N_R}(1-\beta_h)R_h \times R_l} \tag{15}$$

where the number of comparable (and, consequently, concordant) pairs are modulated by the effect of censoring. In turn, this means that the α coefficients ought to be interpreted as the residual degree of (internal or external) concordance that remains after censoring.

Similarly to what we have done in Section 3.1, we can leverage hypothesis H4 (technically, a looser version of it whereby the constraint only applies to the events in R_h, i.e., $\min_{i\in R_h,\Delta_i=1}\{M_i\} > \max_{j\in R_{h+1}}\{M_j\}$) to impose $\alpha_{hl}=1$ for all $h \neq l$.

$$\widehat{C} = 1 - \frac{\sum_{h=1}^{N_R}\gamma_h(1-\alpha_{hh})R_h \times R_h}{\sum_{h=1}^{N_R}\gamma_h R_h \times R_h + \sum_{h=1}^{N_R-1}\sum_{l=h+1}^{N_R}(1-\beta_h)R_h \times R_l} \tag{16}$$

Finally, following the same process we used to obtain Eq. (11) from Eq. (8), we can rewrite Eq. (16) in terms of region widths:

$$\widehat{C} \simeq 1 - \frac{\sum_{h=1}^{N_R}\gamma_h(1-\alpha_{hh})w_h^2}{\sum_{h=1}^{N_R}\gamma_h w_h^2 + 2\sum_{h=1}^{N_R-1}\sum_{l=h+1}^{N_R}(1-\beta_h)w_h w_l} \tag{17}$$

The tight relationship between Eqs. (8) and (16) (and, by extension, between Eqs. (11) and (17)) is immediately discernible: we can obtain the former from the latter by removing the influence of censoring, i.e., by setting $\gamma_h=1$ and $\beta_h=0$ for $h=1,\cdots,N_R$ (and adjusting α_{hh}, if needed). Interestingly, looking at Eq. (17), censoring does not allow, in general, for as neat a simplification as the one presented in Eq. (11). This is mainly due to the former's effect on both the numerator and denominator of $\widehat{C}$, i.e., ultimately, on the number of comparable pairs, whose number depends not only on the proportion of censored subjects in each region (β_h), but also on their position within the region (γ_h). Consequently, the denominator in Eq. (17) cannot be approximated as per Eq. (9). Nevertheless, assuming sufficient information (e.g., on expected follow-up), one could construct a simplification of $\widehat{C}$ that is more complex, but more relevant to the available data than Eq. (11). An interesting, albeit unrealistic (usually, censoring is more likely at the tail end of the observation window), simplification happens in case of a constant censoring prevalence within regions ($\beta_h=\beta^*$ for $h=1,\cdots,N_R$) and an average (i.e., middle point of the bounds presented in Eq. (14)) modulation effect of censoring on concordance ($\gamma_h = 1-\beta^*$ for $h=1,\cdots,N_R$). In this situation, Eqs. (16) and (8) (and Eqs. (11) and (17)) share the same form.

4. Examples of model selection using the C-index

4.1. Illustrative example on synthetic data

To illustrate the practical effectiveness of the concepts we have discussed in Sections 2 and 3, we will now present an example of a simple model selection task. Specifically, we will consider three hypothetical models, whose individual properties would be impossible to discern through a naïve analysis, and we will show that the ambiguity between them is easily resolved by focusing on the time-dependent properties of the C-index (see Eq. (5), Section 2). The simulated dataset comprises $N=60$ subjects, who have all experienced an adverse pathological event within 180 days of the index date. All event times T_i ($i=1,\cdots,N$) are distinct and have been generated as $T_i = 0.0343\cdot(i^2+25i)$. This means that one third of them have been observed to happen within the first 34 days, one third between 34 and 93 days, and the remaining one third after 93 days. Three different models have been proposed to predict the incidence of these events, and we are tasked with determining the best performing one. We are also aware of an additional constraint: due to the time-sensitive nature of the intervention required to manage the pathological events, it is critical that the most urgent cases be accurately identified so that appropriate resources may be allocated. Fig. 3 shows how each of the three models would assign the risk scores M_i to the N subjects. Note that, as the C-index only requires information on the order of risk scores vs. the order of events, and not on the generating model (e.g., on whether it is a Cox model or whether it accounts for time-dependent covariates), in this illustrative example, we simply specified a score pattern for each model (Section 4.2 also includes model training). Particularly, we divided the data in three regions such that, in the concordant ones, $M_i = i + 1.08G$ with G sampled from a standard normal distribution; whereas, in the random ones, M_i was sampled from a uniform distribution that preserved hypothesis H4. All the scores were, then, rescaled to lie between 0 and 1. The three models almost perfectly rank two thirds of the data according to their event times, but they are incapable of determining the correct relative order of the remaining subjects. Specifically, model 1 seems to fail towards the end of the observation interval, model 2 around the middle, and model 3 right at the start.

As we are interested in selecting a model that correctly prioritises subjects at a high risk of experiencing an unfavourable event, evaluating performance with the C-index is a natural first step. Thus, we start by computing $\widehat{C}$ for each model. However, according to Eq. (5), we should first specify the value of τ that is implicit in the C-index's formulation. Here, $\tau = 171$ days, because, even though we have observed all the events, including the one at time $T_N = 180$ days with $\Delta_N = 1$, the last subject that contributes to $\widehat{C_\tau}$ is, in fact, the $(N-1)$-th. Having established that, we can proceed in the evaluation, thus obtaining the following values for $\widehat{C} = \widehat{C_{171}}$: 0.947 (model 1), 0.939 (model 2), and 0.942 (model 3). Unsurprisingly, these are practically the same. In fact, given the composition of the dataset, we could have used Eq. (12) to predict that

$$\widehat{C_{171}} \simeq 1 - 0.5w_e^2 \simeq 1 - 0.5(0.33)^2 \simeq 0.94 \tag{18}$$

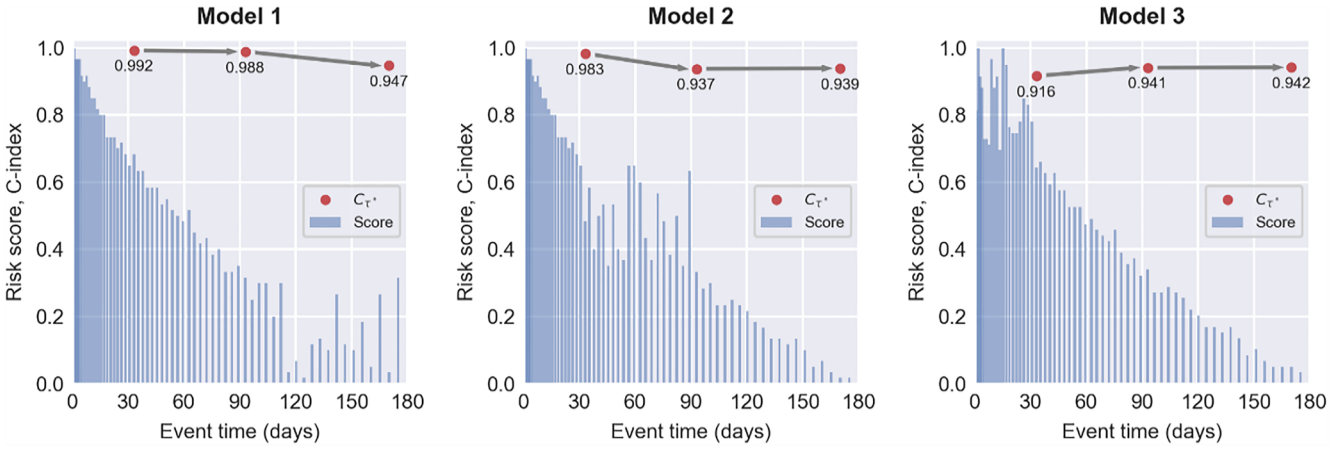

Fig. 3. Illustrative example of model selection. The three panels show the relationships between event time and risk captured by three hypothetical models (1, 2, and 3, from left to right). The height and position of each of the blue bars denote, respectively, the risk score (M_i) and observation/event time ($T_i^{obs} = T_i$) associated with each subject (i). The red dots are placed at coordinates (τ^*, $\widehat{C_{\tau^*}}$) for the same set of three values of τ^*. The grey arrows highlight the relationships between pairs of subsequent values of $\widehat{C_{\tau^*}}$.

in all three cases.

Hence, it is apparent that a naïve analysis of the C-index, as if it were completely time-independent, is insufficient to select one of the three models. However, in Section 2.3, we have introduced the concept of $\widehat{C_{\tau^*}}$, i.e., of the intermediate values of $\widehat{C_\tau}$, obtained by considering all subjects with $T_i^{obs} > \tau^*$ as if they had been censored. We have also remarked that τ^* represents an ideal cut-off between a portion of the data where the relative order between subjects is extremely important, and another where it is enough to have correctly characterised the subjects' risks as low. This property of $\widehat{C_{\tau^*}}$ can help us gain insight into the individual identities of the three models, by quantifying their performances at different cut-offs. Here, the natural choice is to select 34 and 93 days as potentially interesting values for τ^*, and compute $\widehat{C_{34}}$ and $\widehat{C_{93}}$ accordingly. As it is apparent from Table 2, where we have also expressed $\widehat{C_{34}}$ and $\widehat{C_{93}}$ in terms of w_e to ease their comparison, this approach allows us to distinguish between the three models. Particularly, it is immediately clear that the short-term performance of model 3 is incompatible with the stated need to resolve the relative order of the most urgent cases ($\widehat{C_{34}} = 0.916$, i.e., $w_e = 0.41$). On the contrary, both model 1 and model 2 are particularly adept at this task ($\widehat{C_{34}} = 0.992$ and 0.983, respectively, i.e., $w_e = 0.13$ and 0.18). However, given that the specifications do not precisely quantify the concept of "urgency," we should err on the side of caution and identify model 1 as the most appropriate one ($\widehat{C_{93}} = 0.988$ vs. 0.937 for model 2, with an absolute improvement of 0.20 in terms of w_e). Thus, we have succeeded in selecting a suitable model despite the initial premise that $\widehat{C}$ was approximately the same for models 1, 2, and 3.

4.2. Real-data example: The Primary Biliary Cirrhosis dataset

In this Section, we will show that the same general approach used in the synthetic case of Section 4.1 can be applied to real datasets, resulting in very similar considerations. Specifically, we will study the problem of comparing prediction models of adverse outcomes (death or liver transplant) in the well-known Primary Biliary Cirrhosis (PBC) dataset [20]. To contextualise this example, we imagine that the model would be used as a risk-prediction tool within an internal medicine ward. Our hypothetical ward's objective is two-fold, as is often the case in real life: on the one hand, PBC's life expectancy is measured in the order of several years [21]; on the other, resource allocation in the short term (2 years as per hospital policy) is also a pressing concern.

We follow a basic model development pipeline, starting from the pre-processing of the dataset described in [22] (312 subjects; 5 selected covariates: presence of edema, age in years, logarithm of prothrombin time, logarithm of serum bilirubin, and logarithm of serum albumin). We randomly split the dataset between a training and a test sets comprising 67% and 33% of the data. We use the training set data to estimate the parameters of a Cox proportional hazard model and of a logistic regression model, tuning the latter on a binarised version of the outcome variable that is equal to one if and only if an outcome is recorded within 2 years of the baseline. Note that both models output scalar scores that are compatible with the C-index.

The results of the evaluation of the two models on the test set ($N = 103$, 50 events, of which 15 within 2 years) are reported in Table 3, and the pattern of assigned scores is shown in Fig. 4. The table also reports 95% confidence intervals, computed as per [17]. In the following, we will state the p values associated with differences in C-index [23], assuming a significance level set to 0.05.

Table 2
Comparison between the models (illustrative example). Each cell of the second, third, and fourth columns, respectively corresponding to models 1, 2, and 3, contains the values of $\widehat{C_{\tau^*}}$ (with the corresponding w_e between brackets) computed at the τ^*s reported in the first column. The values in the last row are also the global estimates ($\widehat{C_\tau} = \widehat{C}$) for the dataset.

τ^*	$\widehat{C_{\tau^*}}(w_e)$		
	Model 1	**Model 2**	**Model 3**
34 days	0.992 (0.13)	0.983 (0.18)	0.916 (0.41)
93 days	0.988 (0.15)	0.937 (0.35)	0.941 (0.34)
171 days	0.947 (0.33)	0.939 (0.35)	0.942 (0.34)

Table 3
Comparison between the models (real-data example). Each cell of the second, and third columns, contains the values of $\widehat{C_{\tau^*}}$ with 95% confidence intervals, and the corresponding w_e computed at the τ^*s reported in the first column. The values in the last row are also the global estimates ($\widehat{C_\tau} = \widehat{C}$) for the dataset.

τ^*	$\widehat{C_{\tau^*}}(w_e)$	
	Cox Model	**Logistic Regression**
2 years	0.902 (0.44)	0.942 (0.34)
	95% CI: 0.841 – 0.963	*95% CI: 0.912 – 0.973*
10.55 years	0.859 (0.53)	0.813 (0.61)
	95% CI: 0.818 – 0.899	*95% CI: 0.713 – 0.872*

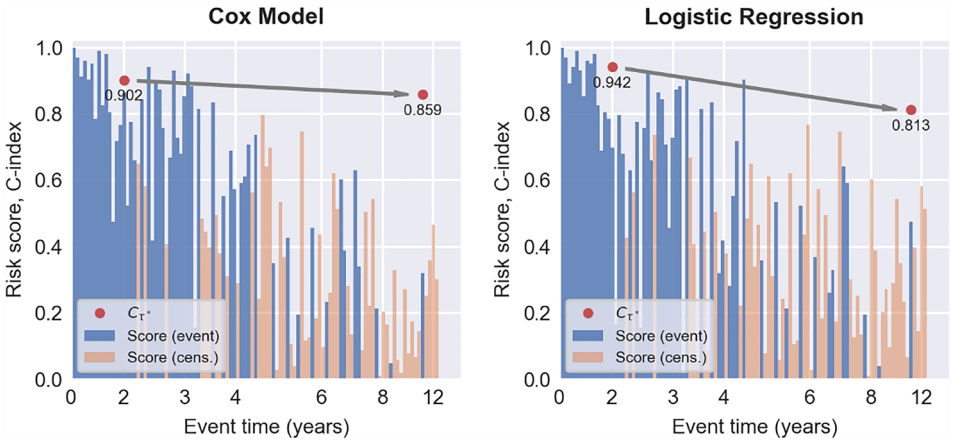

Fig. 4. Real-data example of model selection (PBC dataset). The two panels show the relationships between event time and risk captured by a Cox proportional hazard model (left) and a logistic regression model (right). The colour, height and position of each of the bars denote, respectively, the event/censored status (blue if $\Delta_i = 1$, orange otherwise), the risk score (M_i), and observation time (T_i^{obs}) associated with each subject (i). The red dots are placed at coordinates (τ^*, $\widehat{C_{\tau^*}}$) for the same set of two values of τ^*. The bars are juxtaposed to highlight the lesser importance of exact timing vs. event order (note the uneven time scale).

Before proceeding with the analysis, we must explicitly state that $\tau = 10.55$ years, meaning that the standard computation of the C-index, will yield $\widehat{C_\tau} = \widehat{C_{10.55}}$. According to this metric, the Cox model outperforms logistic regression by a wide margin, i.e., 0.859 vs. 0.813 (statistically significant with p = 0.036). Recall that this 0.046 difference is, in fact, a difference in the fraction of concordant pairs, thus inherently difficult to put in perspective. To partially alleviate this problem and get a better sense of scale, it is useful to refer to Eq. (15) and compute the corresponding values of w_e, i.e., 0.53 and 0.61. Their difference (0.08) can be interpreted as the fraction of additional errors that would have been committed by logistic regression if the obtained $\widehat{C_{10.55}}$s had been the results of model evaluation in a scenario where all the hypotheses presented in Section 3.2 held. Obviously, unlike in Section 4.2, these hypotheses do not hold, and this is not a direct quantification of model behaviour. Nevertheless, this change in perspective remains useful, as it suggests that the Cox model does indeed outperform logistic regression by a wide margin. Thus, we can conclude that, for longer prediction horizons, compatible with PBC's life expectancy, the Cox model should be preferred.

Let us now consider performance in the short term, focusing on the 2-year timeframe indicated by our hypothetical hospital's administration. Given our project's requirements, a natural choice for τ^* is 2 years: only during that period it is critical that we predict the imminence of adverse events, which may require intensive care and higher resource investment. Thus, we evaluate $\widehat{C_2}$ for both models and find that, at variance with the previous comparison, logistic regression exhibits the best performance (0.942 vs. 0.902, statistically significant with p = 0.04) by a margin of 0.040 in terms of C-index, which we can translate (Eq. (15)) to a 0.10 difference in w_e (0.34 vs. 0.44). Interestingly, the latter is only slightly greater than the difference we had found for $\widehat{C_{10.55}}$, suggesting that logistic regression dominates over the Cox model in the short term approximately as much as the opposite is true in the long term. Hence, it would appear that neither model can satisfy both project objectives, and that either a new approach or a redefinition of priorities might be in order. The crucial point, however, is that a naïve, time-agnostic view of the C-index would not have enabled us to reach this conclusion.

4.3. A note on computing $\widehat{C_{\tau^*}}$ with existing software

As we have alluded to in Section 2, computing $\widehat{C_{\tau^*}}$ for any τ^* with any piece of software that implements Harrell's C-index requires minimal effort: we simply run the same algorithm after setting $\Delta_i = 0$ for all subjects with $T_i^{obs} > \tau^*$. In general, C-index functions (e.g., in R: estC from compareC [23], concordance from survival [24], concordance.index from survcomp [25], cindex.CV from compound.cox [26]; in python: concordance_index from lifelines [27], concordance_index_censored from scikit-survival [28]) take three main arguments, i.e., three N-dimensional vectors comprising, respectively, observation times $\{T_i^{obs}\}$, event indicators $\{\Delta_i\}$, and risk scores $\{M_i\}$, aligned according to the subject identifier $i = 1, \cdots, N$. Internally, these arguments are processed to output an estimate of the C-index that we have shown to be $\widehat{C_\tau}$ (see Section 2). To use the same function to compute $\widehat{C_{\tau^*}}$ instead of the default $\widehat{C_\tau}$, we proceed as follows: first, we define a new vector of event indicators $\{\Delta_i'\}$ such that $\Delta_i' = \Delta_i$ if $T_i^{obs} \leq \tau^*$, and $\Delta_i' = 0$ otherwise; then, we simply call the C-index function with arguments $\{T_i^{obs}\}$, $\{\Delta_i'\}$, $\{M_i\}$ instead of $\{T_i^{obs}\}$, $\{\Delta_i\}$, $\{M_i\}$. This invocation of the function will output $\widehat{C_{\tau^*}}$.

5. Discussion and conclusion

In the present work, we have studied the properties and behaviour of the C-index referencing its most popular estimator, developed by Harrell. Our focus has been primarily practice-oriented, i.e., we have chosen to forego a rigorous statistical analysis, in favour of highlighting the concepts that we feel would be most useful in practical model evaluation or selection scenarios. Hence, in Section 2.3, we commented on the implicit residual dependency of $\widehat{C}$ on the time τ of the final event in the dataset, extended this reasoning to arbitrarily-defined cut-off times τ^*, and, in Section 4, applied it to two illustrative examples. In the same spirit, after showing the intrinsic difficulty of interpreting improvements or deteriorations of the C-index, we devised a useful approximation to directly relate its numerical value to the number of risk scores that we would expect to have been correctly or incorrectly assigned in a simplified scenario characterised by the same value of $\widehat{C}$ (Section 3.2). The main limitation of this approach is its reliance on a rather strict set of hypotheses, and the consequent divergence between the idealised scenario used to compute w_e and the reality of the experiment. Nevertheless, we deemed this conceptual shift to be useful to quantify the effect of the nonlinearity of the C-index on performance differences in more relatable terms (general formulations are available in Sections 3.1 and 3.3).

An issue we have not fully discussed, here, is the determination of the best value or set of values for τ^* for any given application. While in the illustrative example of Section 4.1, the simple structure of the dataset lent itself to the immediate identification of interesting cut-off values, in real life, these distinctions may not be as clear-cut. Nevertheless, it should often be possible to reference project requirements or

domain knowledge to identify notable τ^*s (e.g., as we did in Section 4.2). Failing that, a fully time-dependent analysis of $\widehat{C_{\tau^*}}$ could highlight discontinuities in predictive behaviour that might inform the model selection process. In this sense, an approach similar to the one detailed in [29], involving two time-dependent definitions of the AUROC, appears extremely promising. Therein, the authors show how to extend the AUROC to the time-dependent case, and distinguish between an incident/dynamic approach, which is preferable for descriptive continuous-time analytics, and a cumulative/dynamic one that is useful when specific time thresholds are the main focus (i.e., when the equivalent of τ^* is of interest). Notably, they characterise the C-index as merely a "global summary" of the former approach, whereas, here, we have shown that it can be used for time-dependent performance evaluation. In this respect, there is a subtle difference between the two strategies, i.e., while those AUROC variants only capture the distinction between events that happen before and after a varying threshold τ^*, $\widehat{C_{\tau^*}}$ also takes into account the relative order of the earlier events ($T_i^{obs} \leq \tau^*$). Thus, depending on the application at hand, either or, most likely, both of those solutions could be appropriate.

As a final remark, we should also note that Harrell's estimator is not the only way to compute the C-index, or necessarily the best. In fact, several extensions to this estimator exist, as well as alternative metrics, such as time-dependent estimators for the AUROC, sensitivity, and specificity, or global indicators such as the integrated Brier score [30–33]. Nevertheless, on the one hand these are still affected, at least in some measure, by the issues we have described; and, on the other, Harrell's C-index is the most consistently implemented variant across popular programming libraries, where it is often the default measure of goodness of fit for survival models [24,26–28]. Hence, although experimenters may elect not to insert it into their own model development pipelines, they are still likely to encounter it, e.g., in the analysis of literature benchmarks.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2020.103496.

References

[1] G.S. Collins, K.G.M. Moons, Reporting of artificial intelligence prediction models, The Lancet 393 (2019) 1577–1579, https://doi.org/10.1016/S0140-6736(19)30037-6.

[2] C.V. Cosgriff, L.A. Celi, D.J. Stone, Critical Care, Critical Data, Biomed. Eng. Comput. Biol. 10 (2019), https://doi.org/10.1177/1179597219856564.

[3] A. Harris, Path From Predictive Analytics to Improved Patient Outcomes: A Framework to Guide Use, Implementation, and Evaluation of Accurate Surgical Predictive Models, Ann. Surg. 265 (2017) 461–463, https://doi.org/10.1097/SLA.0000000000002023.

[4] R.B. Parikh, M. Kakad, D.W. Bates, Integrating Predictive Analytics Into High-Value Care: The Dawn of Precision Delivery, JAMA 315 (2016) 651–652, https://doi.org/10.1001/jama.2015.19417.

[5] C.V. Cosgriff, L.A. Celi, C.M. Sauer, Boosting Clinical Decision-making: Machine Learning for Intensive Care Unit Discharge, Ann. ATS 15 (2018) 804–805, https://doi.org/10.1513/AnnalsATS.201803-205ED.

[6] D. Alemayehu, M.L. Berger, Big Data: transforming drug development and health policy decision making, Health Serv. Outcomes Res. Method 16 (2016) 92–102, https://doi.org/10.1007/s10742-016-0144-x.

[7] D.W. Bates, A. Heitmueller, M. Kakad, S. Saria, Why policymakers should care about "big data" in healthcare, Health Policy Technol. 7 (2018) 211–216, https://doi.org/10.1016/j.hlpt.2018.04.006.

[8] A. Giga, How health leaders can benefit from predictive analytics, Healthc. Manage. Forum. 30 (2017) 274–277, https://doi.org/10.1177/0840470417716470.

[9] K.G.M. Moons, D.G. Altman, J.B. Reitsma, J.P.A. Ioannidis, P. Macaskill, E.W. Steyerberg, et al., Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration, Ann. Intern. Med. 162 (2015) W1, https://doi.org/10.7326/M14-0698.

[10] M.W. Kattan, A.J. Vickers, Incorporating predictions of individual patient risk in clinical trials, Urologic Oncol.: Semin. Origin. Invest. 22 (2004) 348–352, https://doi.org/10.1016/j.urolonc.2004.04.012.

[11] J.S. Rumsfeld, K.E. Joynt, T.M. Maddox, Big data analytics to improve cardiovascular care: promise and challenges, Nat. Rev. Cardiol. 13 (2016) 350–359, https://doi.org/10.1038/nrcardio.2016.42.

[12] R. Singh, K. Mukhopadhyay, Survival analysis in clinical trials: Basics and must know areas, Perspect. Clin. Res. 2 (2011) 145–148, https://doi.org/10.4103/2229-3485.86872.

[13] S. Mittal, D. Madigan, R.S. Burd, M.A. Suchard, High-dimensional, massive sample-size Cox proportional hazards regression for survival analysis, Biostatistics 15 (2014) 207–221, https://doi.org/10.1093/biostatistics/kxt043.

[14] P. Wang, V. Tech, Y. Li, Reddy CK. Machine Learning for Survival Analysis: A Survey, ACM Comput. Surv. (1) (n.d.) 43.

[15] F.E. Harrell Jr, R.M. Califf, D.B. Pryor, K.L. Lee, R.A. Rosati, Evaluating the yield of medical tests, Jama 247 (1982) 2543–2546.

[16] M. Schmid, M.N. Wright, A. Ziegler, On the use of Harrell's C for clinical risk prediction via random survival forests, Expert Syst. Appl. 63 (2016) 450–459, https://doi.org/10.1016/j.eswa.2016.07.018.

[17] M.J. Pencina, R.B. D'Agostino, Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation, Stat. Med. 23 (2004) 2109–2123, https://doi.org/10.1002/sim.1802.

[18] H. Uno, T. Cai, M.J. Pencina, R.B. D'Agostino, L.J. Wei, On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data, Stat. Med. 30 (2011) 1105–1117, https://doi.org/10.1002/sim.4154.

[19] P.J. Heagerty, Y. Zheng, Survival Model Predictive Accuracy and ROC Curves, Biometrics 61 (2005) 92–105, https://doi.org/10.1111/j.0006-341X.2005.030814.x.

[20] T.M. Therneau, P.M. Grambsch, Modeling Survival Data: Extending the Cox Model, Springer New York, New York, NY, 2000 http://doi.org/10.1007/978-1-4757-3294-8.

[21] EASL Clinical Practice Guidelines: The diagnosis and management of patients with primary biliary cholangitis, J. Hepatol. 67 (2017) 145–172. http://doi.org/10.1016/j.jhep.2017.03.022.

[22] P. Blanche, M.W. Kattan, T.A. Gerds, The c-index is not proper for the evaluation of t-year predicted risks, Biostatistics 20 (2019) 347–357, https://doi.org/10.1093/biostatistics/kxy006.

[23] L. Kang, W. Chen, N.A. Petrick, B.D. Gallas, Comparing two correlated C indices with right-censored survival outcome: a one-shot nonparametric approach, Stat. Med. 34 (2015) 685–703, https://doi.org/10.1002/sim.6370.

[24] T.M. Therneau, T. Lumley, survival: Survival Analysis 2018. https://CRAN.R-project.org/package=survival (accessed January 28, 2019).

[25] M.S. Schröder, A.C. Culhane, J. Quackenbush, B. Haibe-Kains, survcomp: an R/Bioconductor package for performance assessment and comparison of survival models, Bioinformatics 27 (2011) 3206–3208, https://doi.org/10.1093/bioinformatics/btr511.

[26] T. Emura, H.-Y. Chen, S. Matsui,Y.-H. Chen, compound.Cox: Univariate Feature Selection and Compound Covariate for Predicting Survival, (2018).

[27] Cameron Davidson-Pilon, Jonas Kalderstam, Ben Kuhn, Paul Zivich, Andrew Fiore-Gartland, Luis Moneda, et al. CamDavidsonPilon/lifelines: v0.17.5. Zenodo, (2019). http://doi.org/10.5281/zenodo.2549675.

[28] S. Pölsterl, scikit-survival (2019). https://github.com/sebp/scikit-survival (accessed January 28, 2019).

[29] A. Bansal, P.J. Heagerty, A Tutorial on Evaluating the Time-Varying Discrimination Accuracy of Survival Models Used in Dynamic Decision Making, Med. Decis. Making. 38 (2018) 904–916, https://doi.org/10.1177/0272989X18801312.

[30] A.R. Brentnall, J. Cuzick, Use of the concordance index for predictors of censored survival data, Statist. Methods Med. Res. 27 (2018) 2359–2373, https://doi.org/10.1177/0962280216680245.

[31] M. Schmid, S. Potapov, A comparison of estimators to evaluate the discriminatory power of time-to-event models, Stat. Med. 31 (n.d.) 2588–2609. http://doi.org/10.1002/sim.5464.

[32] L.-P. Kronek, A. Reddy, Logical analysis of survival data: prognostic survival models by detecting high-degree interactions in right-censored data, Bioinformatics 24 (2008) i248–i253, https://doi.org/10.1093/bioinformatics/btn265.

[33] E. Graf, C. Schmoor, W. Sauerbrei, M. Schumacher, Assessment and comparison of prognostic classification schemes for survival data, Statist. Med. 18 (1999) 2529–2545, https://doi.org/10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5.