Received: 17 November 2022 | Revised: 21 April 2023 | Accepted: 19 June 2023

DOI: 10.1002/jrsm.1655

RESEARCH ARTICLE

Research Synthesis Methods WILEY

Accounting for time dependency in meta-analyses of concordance probability estimates

Matthias Schmid[1] | **Tim Friede**[2] | **Nadja Klein**[3] | **Leonie Weinhold**[1]

[1]Department of Medical Biometry, Informatics, and Epidemiology, University Hospital Bonn, Bonn, Germany

[2]Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany

[3]Research Center for Trustworthy Data Science and Security, UA Ruhr/ Department of Statistics, Technische Universität Dortmund, Dortmund, Germany

Correspondence
Matthias Schmid, Department of Medical Biometry, Informatics, and Epidemiology, University Hospital Bonn, VenusbergCampus 1, 53127 Bonn, Germany.
Email: matthias.c.schmid@uni-bonn.de

Funding information
Volkswagen Foundation, Grant/Award Number: 98 948

Abstract

Recent years have seen the development of many novel scoring tools for disease prognosis and prediction. To become accepted for use in clinical applications, these tools have to be validated on external data. In practice, validation is often hampered by logistical issues, resulting in multiple small-sized validation studies. It is therefore necessary to synthesize the results of these studies using techniques for meta-analysis. Here we consider strategies for meta analyzing the concordance probability for time-to-event data ("*C*-index"), which has become a popular tool to evaluate the discriminatory power of prediction models with a right-censored outcome. We show that standard meta-analysis of the *C*-index may lead to biased results, as the magnitude of the concordance probability depends on the length of the time interval used for evaluation (defined e.g., by the follow-up time, which might differ considerably between studies). To address this issue, we propose a set of methods for random-effects meta-regression that incorporate time directly as covariate in the model equation. In addition to analyzing nonlinear time trends via fractional polynomial, spline, and exponential decay models, we provide recommendations on suitable transformations of the *C*-index before meta-regression. Our results suggest that the *C*-index is best meta-analyzed using fractional polynomial meta-regression with logit-transformed *C*-index values. Classical random-effects meta-analysis (not considering time as covariate) is demonstrated to be a suitable alternative when follow-up times are small. Our findings have implications for the reporting of *C*-index values in future studies, which should include information on the length of the time interval underlying the calculations.

KEYWORDS

concordance probability, fractional polynomials, meta-regression, prognostic factor research, restricted cubic splines, time-to-event data

Highlights

What is already known?

Prognostic factor research is a rapidly evolving field with an increased need for meta-analysis. In this field, studies aim at analyzing the associations of risk factors with a time-to-event outcome. The concordance probability for time-to-event data ("*C*-index") has become a popular tool to evaluate the discriminatory power of prognostic models. It is often used in the meta-analysis of validation studies.

What is new?

We show that standard meta-analysis of the *C*-index may lead to biased results, as the magnitude of the *C*-index depends on the length of the time interval used for evaluation. To address this issue, we propose a set of methods for random-effects meta-regression incorporating time directly as covariate in the model equation. Our methods are able to account for nonlinear relationships between the *C*-index and time.

Potential impact for readers outside the authors' field

The proposed methods improve the interpretability of meta-analyzed prognostic models, enabling users of these models to better judge their prognostic power. They also point at a possible overoptimism in the interpretation of prognostic models, which—if evaluated by the *C*-index—could indicate a strong performance simply because the time interval for evaluation has been (too) short. Furthermore, our results have implications for the reporting of future validation studies. These should present the full *C*-index curve, along with the time interval used for evaluation.

1 | INTRODUCTION

During the past decades, the volume of published research has increased dramatically.[1] Even before the COVID-19 pandemic, the number of research articles has been estimated to grow by 8%–9% each year, including more than 1 million papers per year in the biomedical field alone.[2] At the same time, hundreds of newly ranked journals have appeared, with the estimated total amount of active peer-reviewed journals exceeding 30,000.[1,3] In view of this "information overload",[2] there is an obvious need for evidence synthesis to "clarify what is known from research evidence to inform policy, practice and personal decision making and improved methods for meta-analysis".[4]

Prognostic factor research[5] is a rapidly evolving field with an increased need for meta-analysis. In this field, studies aim at analyzing the associations of one or several factors (often termed "risk factors") with a time-to-event outcome $T \in \mathbb{R}^+$. In medicine and epidemiology, for instance, prognostic factors are often given by patient characteristics (e.g. age, sex, smoking behavior, blood pressure) collected at the baseline examination of a longitudinal study. These variables might then be used to predict the occurrence of events such as death, tumor progression, or adverse events. Often, several prognostic factors are summarized by a multivariable risk score (defined, e.g., by a linear combination of the factors). Popular examples of risk scores are the European System for Cardiac Operative Risk Evaluation (EuroSCORE) II to predict mortality after cardiac surgery and the Framingham Risk Score for predicting coronary heart disease.[6] Score development is usually based on a statistical modeling technique applied to a set of training data, yielding a prediction model that is defined by a (univariable or multivariable) prognostic score $\eta \in \mathbb{R}$.

A key issue for the acceptability of a prognostic score is its repeated validation on externally collected test data.[7,8] These validation steps have become a gold standard in prognostic modeling, as they provide a much more realistic assessment of the score's performance than would have been possible using the training data only. Importantly, the results of external validation steps are often found to be heterogeneous, showing a high variability in prognostic performance. Validation studies involving external test data might, for instance, be affected by small sample sizes and differences in the characteristics of the patient population compared to the training data.[6,9] As a consequence, systematic reviews and meta-

analyses are "urgently needed to summarize [the] evidence [of prediction models] and to better understand under what circumstances developed models perform adequately or require further adjustments".[6]

In this paper, we consider strategies for meta analyzing the *concordance probability for time-to-event data* ("*C*-index"), which is a widely used measure to evaluate prediction models with a time-to-event outcome.[10–12] The *C*-index is a discrimination measure that compares the rankings of the individual score values η_i and the event times T_i, $i = 1, ..., n$, in a test sample of size n. It is defined[12] by

$$C(\tau) = \mathrm{P}\left(\eta_i > \eta_j | T_i < T_j, T_i \leq \tau\right), \quad (1)$$

where i,j denote two independent observations in the test data and $\tau > 0$ is a truncation time (e.g., the maximum follow-up time of a clinical study). Setting $\tau = \infty$ yields the *unrestricted concordance probability* $\mathrm{P}\left(\eta_i > \eta_j | T_i < T_j\right)$. Generally, $C(\infty)$ takes the value 1 if the rankings of $-\eta_i$ and T_i agree perfectly. Conversely, $C(\infty) = 0.5$ if η does not predict better than chance alone. In the absence of censoring, the concordance probability can readily be evaluated by comparing all pairs T_i, T_j and by estimating the conditional probability in (1) by its respective relative frequency in the test data. If censoring is present, however, a comparison of all pairs T_i, T_j is no longer possible, and estimation of the concordance probability requires additional assumptions on the data-generating process,[11–14] also see Schmid and Potapov[15] for a comparison of estimators.

For meta-analysis, Debray et al.[6] recently introduced a framework that includes, among other techniques, a method to summarize estimates of the *C*-index obtained from multiple validation studies. Based on earlier work by Snell et al., who meta-analyzed the QRISK2 model to predict 10-year cardiovascular disease risk,[16] the authors proposed to transform estimates to the logit scale before meta-analysis. This strategy has also been adopted in several recent systematic reviews and meta-analyses,[17–19] including evaluations of the CHA2DS2-VASc rule for estimating stroke risk in patients with atrial fibrillation and various scores for the prediction of survival outcomes in colorectal cancer with surgical resection. In other studies, which included meta-analyses of models for survival after resection of intrahepatic cholangiocarcinoma, the *C*-index was meta-analyzed on the original (untransformed) probability scale.[20,21] Meta-analysis of the *C*-index using individual participant data has been studied by Pennells et al.[22] Hattori and Zhou[23] proposed to construct a synthesized *C*-index from an estimate of the summary cumulative ROC curve obtained by analyzing study-specific Kaplan–Meier curves.

Despite numerous methodological advances, which have led to the publication of several guidance papers,[6,9] meta-analysis of prognostic validation studies remains a challenging task. This is, in particular, due to the fact that measures of prediction accuracy in prognostic research are often related to a specific time point or time span.[5] Consequently, meta-analysis of validation studies becomes intrinsically difficult when study-specific performance estimates refer to different time points or spans. As seen from (1), this time dependency also affects the *C*-index studied in this paper: Since the magnitude of C depends on the truncation time τ, *C*-index estimates may not be comparable across studies if they relate to different values of τ. Specifically, since the value of τ is often determined by the duration of the study generating the test data, different study durations may implicitly lead to systematic differences between the resulting *C*-index estimates. Consider, for instance, the simulated meta-analysis in Figure 1: In this example, we considered 30 hypothetical studies, generating event times from a Weibull accelerated failure time model of the form $\log(T) = X - \epsilon$. In each study, X was a normally distributed covariate with zero mean and standard deviation 0.5, and ϵ followed a standard Gumbel distribution. Censoring times were independent of T and followed an exponential distribution with rate 0.5 each. Sample sizes of the studies were generated randomly and ranged between 100 and 1000. After data generation, the observed event times were truncated at study-specific truncation (=maximum follow-up) times τ_k, $k = 1, ..., 30$, which were sampled from a uniform distribution on [0.1,2]. Study-specific *C*-index estimates were calculated using the estimator by Uno et al.[14] As demonstrated in Figure 1, $C(\tau)$ is seen to decrease with τ, and the pooled estimate obtained from standard meta-analysis relates to an implicitly defined truncation time. Thus, if not accounted for, the time dependency of $C(\tau)$ may compromise both the specification of a properly defined estimand *and* the validity of the pooled estimate. While the simulated meta-analysis in Figure 1 serves as an illustrative example to motivate the methods proposed here, very similar effects have also been found in real-world studies (see e.g., our analysis of the German Chronic Kidney Disease Study[24] data in Section 4 of this paper). Another recent example is given by the *C*-index estimates presented in Zacharias et al.,[25] who evaluated two prediction models for the progression of chronic kidney disease to kidney failure. Although the authors considered a slightly different setting than the one introduced above (allowing for the presence of a competing event), the resulting *C*-index estimates (computed in three external validation studies and presented in table 2 of Zacharias et al.) show a clear negative association with the truncation time.

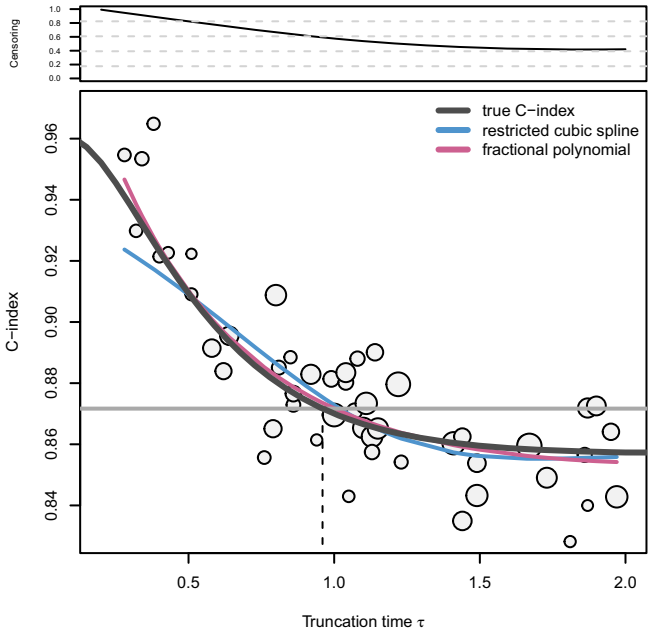

FIGURE 1 Exemplary meta-analysis with simulated test data. The upper panel shows the expected censoring rate at each value of the truncation time. The lower panel shows the study-specific *C*-index estimates. The sizes of the bubbles are proportional to the inverse variances of the *C*-index estimates. The solid black line refers to the true *C*-index according to the data-generating process whereas the horizontal gray line refers to the pooled *C*-index estimate that would have been obtained from a standard random effects meta-analysis ignoring time dependency. The vertical dashed line shows the "implicit" truncation time corresponding to the pooled estimate. Obviously, this model lacks a well-defined estimand, and it is unclear how the pooled estimate should be interpreted. The blue and red lines refer to the meta-regression curves obtained from fitting a restricted cubic spline and a fractional polynomial model to the logit-transformed *C*-index estimates. For details on model specification, see Section 3. [Colour figure can be viewed at wileyonlinelibrary.com]

To address these issues and to improve the interpretability of pooled *C*-index estimates, we consider a set of statistical techniques that incorporate the time dependency of $C(\tau)$ directly in a suitably specified meta-regression model. Our proposed model is based on the frequentist modeling approach with restricted maximum likelihood (REML) estimation, as recommended in the recent guidance paper by Debray et al.[6] Specifically, due to the above-mentioned heterogeneity of external validation results, we will focus throughout on random-effects models. We propose to model the time dependency of $C(\tau)$ by either a restricted cubic spline (RCS) or a 2nd degree fractional polynomial (FP2), thereby accounting for nonlinearities in the regression curve (cf. Figure 1). Using simulation studies, we will compare the RCS and FP2 models to standard random-effects meta-analysis not including τ as covariate, and also to linear meta-regression. Furthermore, we will investigate whether meta-regression can be improved by transforming the *C*-index estimates before model fitting (for instance, using a logit transformation).

The rest of the paper is organized as follows: After starting with the definition of relevant quantities (Section 2.1), we provide a brief overview of existing techniques to estimate the *C*-index (Section 2.2). The proposed methodology is described in Sections 2.3 and 2.4. Section 3 contains a comprehensive simulation study on the properties of the proposed approach, including a comparison to existing methods. A real-world illustration on data collected for the German Chronic Kidney Disease Study is presented in Section 4. The final section summarizes the main findings of the article.

2 | METHODS

2.1 | Derivation and properties of the *C*-index

Consider a validation study with n observations and a time-to-event outcome that might be subject to right censoring. The observations are assumed to be independent and identically distributed. The score values and observed event times are denoted by η_i and $\widetilde{T}_i = \min(T_i, Z_i)$, $i = 1,...,n$, respectively, where $(Z_1,...,Z_n)^T$ is a vector of continuous censoring times. The binary variables $\Delta_i = I(T_i \le Z_i)$, $i = 1,...,n$, indicate whether observations are censored ($\Delta_i = 0$) or not ($\Delta_i = 1$). Assumptions on the censoring process are given below. We further assume that there are no tied observations, i.e. all sample values of T_i and Z_i are assumed to be unique. As shown by Heagerty and Zheng,[11] the concordance probability in (1) can be derived from a set of time-dependent sensitivities and specificities, which, at each time point t, relate the current survival status to the event that η exceeds a given threshold $c \in \mathbb{R}$. More specifically, following the *incident/dynamic* approach,[11] one defines incident cases by observations experiencing an event at t (i.e., $T_i = t$) and dynamic controls by observations having the event after t (i.e., $T_i > t$). With these definitions, time-dependent sensitivities and specificities are given by

$$\text{sens}_t^I(c) = \text{P}(\eta_i > c \,|\, T_i = t) \quad \text{and}$$

$$\text{spec}_t^D(c) = \text{P}(\eta_i \le c \,|\, T_i > t),$$

where the superscripts I and D refer to the terms "incident" and "dynamic", respectively. At each time point,

$\text{sens}_t^I(c)$ and $\text{spec}_t^D(c)$ can be summarized by an incident/dynamic receiver operating characteristic (ROC) curve, which is defined as

$$\text{ROC}_t^{I/D}(p) = \text{sens}_t^I\left[\left(1-\text{spec}_t^D\right)^{-1}(p)\right], \quad p \in [0,1]. \qquad (2)$$

Incident/dynamic ROC curves can further be summarized by the incident/dynamic AUC curve

$$\text{AUC}_t^{I/D} = \int_0^1 \text{ROC}_t^{I/D}(p)\,dp,$$

which equals the probability $\text{P}\left(\eta_i > \eta_j \,|\, T_i = t, T_j > t\right)$ for independent observations i and j. Finally, denoting the probability density function of T by $f(t)$, the concordance probability $C(\tau)$ is derived as the area under a weighted version of the incident/dynamic AUC curve. More specifically, it can be shown that

$$C(\tau) = \text{P}\left(\eta_i > \eta_j \,|\, T_i < T_j, T_i \le \tau\right) = \int_0^\tau w_t^\tau \cdot \text{AUC}_t^{I/D}\,dt \qquad (3)$$

with weights $w_t^\tau = f(t)\cdot \text{P}(T>t)/\int_0^\tau f(u)\cdot P(T>u)\,du$ (see Heagerty and Zheng[11] for a formal proof).

A related quantity is the *cumulative/dynamic* ROC curve, which is defined in the same way as (2) but with $\text{sens}_t^I(c)$ replaced by time-dependent sensitivities of the form $\text{sens}_t^C(c) = \text{P}(\eta_i > c \,|\, T_i \le t)$, where the superscript C refers to the term "cumulative". With this approach, *cumulative cases* are defined by observations experiencing an event at or before t (i.e., $T_i \le t$). Correspondingly, the cumulative/dynamic AUC curve is given by the areas under the cumulative/dynamic ROC curves, i.e. by $\text{AUC}_t^{C/D} = \text{P}\left(\eta_i > \eta_j \,|\, T_i \le t, T_j > t\right)$. Defining a generalized version of $\text{AUC}_t^{C/D}$ by $\text{AUC}_{s,t}^{C/D} = \text{P}\left(\eta_i > \eta_j \,|\, T_i \le s, T_j > t\right)$, it can further be shown[23] that

$$\text{AUC}_t^{I/D} = \frac{\partial}{\partial s}\text{AUC}_{s,t}^{C/D}\Big|_{s=t} \cdot \frac{\text{P}(T \le t)}{f(t)} + \text{AUC}_t^{C/D}. \qquad (4)$$

Thus, combining Equations (3) and (4), the C-index can be derived using either incident or cumulative case definitions.

In practice, $C(\tau)$ is often observed to decrease monotonically with τ (e.g., Figure 1). This behavior could, for example, be caused by a monotonically decreasing AUC curve, which tends to take smaller values as t increases.[26] Note, however, that the monotonicity of $C(\tau)$ does not hold in general and that it is possible to construct scenarios where $C(\tau)$ shows a distinctly non-monotonic behavior (see Figure 2).

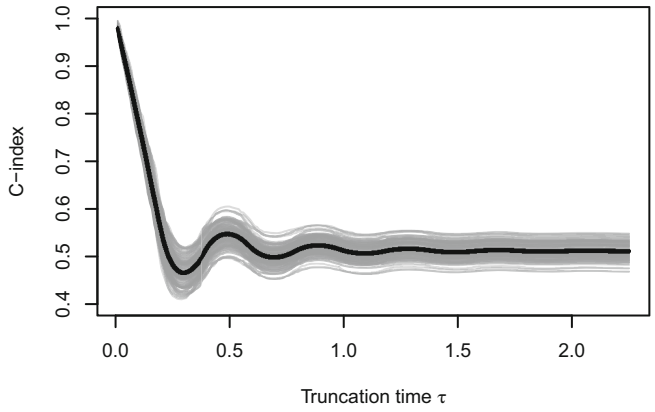

FIGURE 2 Example of a concordance probability with non-monotonic behavior. The black line (depicting the C-index as a function τ) was derived by averaging 100 estimates of $C(\tau)$ using the method of Uno et al.[14] Estimates were obtained from 100 independent samples with exponentially distributed event times ($n = 1000$, rate $= 1$, no censoring). The true underlying model was given by $\eta = -\sin(8\cdot T)^2$. The gray lines refer to the 100 sample-specific curves. Although this example has been designed for illustrative purposes only and would be rarely encountered in practice, it shows that monotonicity of $C(\tau)$ cannot be assumed in general.

2.2 | Estimation of the *C*-index

In the absence of censoring, $C(\tau)$ is naturally estimated by the relative frequency

$$\widehat{C}_{\text{RF}}(\tau) = \frac{\sum_{i\neq j} \text{I}\left(\eta_i > \eta_j\right)\cdot \text{I}\left(\widetilde{T}_i < \widetilde{T}_j\right)\cdot\left(\widetilde{T}_i \le \tau\right)}{\sum_{i\neq j} \text{I}\left(\widetilde{T}_i < \widetilde{T}_j\right)\cdot \text{I}\left(\widetilde{T}_i \le \tau\right)},$$

which compares the orderings of $\widetilde{T}_i$ and η_i in an observation-wise manner. When applied to right-censored data, this approach is no longer appropriate, as pairs of observations where the shorter observed event time is censored ($\widetilde{T}_i < \widetilde{T}_j$ and $\Delta_i = 0$) cannot be compared in a meaningful way. An obvious way to incorporate censoring is to discard all pairs of non-comparable observations, yielding the estimator

$$\widehat{C}_{\text{Harrell}}(\tau) = \frac{\sum_{i\neq j} \text{I}\left(\eta_i > \eta_j\right)\cdot \text{I}\left(\widetilde{T}_i < \widetilde{T}_j\right)\cdot \text{I}\left(\widetilde{T}_i \le \tau\right)\cdot \Delta_i}{\sum_{i\neq j} \text{I}\left(\widetilde{T}_i < \widetilde{T}_j\right)\cdot \text{I}\left(\widetilde{T}_i \le \tau\right)\cdot \Delta_i}.$$

During the past decades, this estimator (also termed "Harrell's C") has become the most popular way to evaluate $C(\tau)$. However, it shows a notable upward bias if censoring rates are high.[12,15] To address this issue, Uno et al. proposed an inverse-probability-of-censoring-weighted

version of Harrell's C (termed "Uno's C") that is defined by

$$\widehat{C}_{\text{Uno}}(\tau) = \frac{\sum_{i \neq j} \mathrm{I}\left(\eta_i > \eta_j\right) \cdot \mathrm{I}\left(\widetilde{T}_i < \widetilde{T}_j\right) \cdot \mathrm{I}\left(\widetilde{T}_i \leq \tau\right) \cdot \Delta_i / \widehat{G}\left(\widetilde{T}_i\right)^2}{\sum_{i \neq j} \mathrm{I}\left(\widetilde{T}_i < \widetilde{T}_j\right) \cdot \mathrm{I}\left(\widetilde{T}_i \leq \tau\right) \cdot \Delta_i / \widehat{G}\left(\widetilde{T}_i\right)^2}, \tag{5}$$

where $\widehat{G}(\cdot)$ is a consistent estimator of the censoring survival function $G(t) = \mathrm{P}(Z_i > t)$ obtained from the validation data.[14] Usually, $G(\cdot)$ is estimated by the Kaplan–Meier method, although more complex models (e.g. depending on a set of covariates) might be considered. Assuming conditionally independent censoring (i.e. independence of T_i and Z_i $\forall i$ given the covariates) and a correctly specified censoring model with $G(t) > \delta > 0$ $\forall t$, Uno et al.[14] showed that $\widehat{C}_{\text{Uno}}(\tau)$ is weakly consistent for $C(\tau)$ as $n \to \infty$.

Remark: The estimator considered by Gerds et al. is slightly different from (5) in that is $\widehat{G}\left(\widetilde{T}_i\right)^2$ replaced by $\widehat{G}\left(\widetilde{T}_i\right) \cdot \widehat{G}\left(\widetilde{T}_i-\right)$ in both the numerator and the denominator, where $\widetilde{T}_i-$ refers to a time point that is infinitesimally smaller than $\widetilde{T}_i$.[12] Clearly, this difference is only relevant when $\widehat{G}(\cdot)$ is not continuous in t (for instance when the Kaplan–Meier method is used to estimate $\widehat{G}(\cdot)$). In our analysis we will use the R add-on package *pec*[27] that implements the method by Gerds et al.[12] but refer to this estimator as "Uno's C".

A major advantage of Harrell's C and Uno's C is that both estimators are non-parametric in the sense that they do not make any assumptions on the distribution of T. An alternative way to deal with non-comparable pairs of observations is to specify a parametric or semi-parametric working model for T (e.g., a Cox regression model) and to derive estimators of $C(\tau)$ based on the characteristics of this model.[11,13,28] It is also possible to apply a model-free estimator of the incident/dynamic AUC curve[29] and to estimate the C-index via numerical integration of the AUC estimate. In this paper we will consider Harrell's C and Uno's C throughout.

2.3 | On the role of the truncation time τ

As stated in Section 1, the unrestricted C-index $P(\eta_i > \eta_j | T_i < T_j)$ comes with an intuitive probabilistic interpretation, comparing the rankings of the values η_i and T_i, $i = 1,...,n$. This interpretation is considerably less intuitive if an additional truncation time $\tau < \infty$ is included in the definition of the C-index. Nonetheless, there exist both conceptual *and* technical reasons to prefer a restricted version of the concordance probability over the unrestricted one: First, the sample values $\left(\widetilde{T}_i, \Delta_i, \eta_i\right)$, $i = 1,...,n$, are often obtained from a validation study with a limited follow-up time. In this case, the maximum possible time horizon τ is naturally given by the length of the follow-up time, implying that *any* estimate of the concordance probability derived from the validation data is a restricted one.[30] Second, the censoring model used in the definition of Uno's C usually assumes $G(t) > \delta > 0$ $\forall t$, posing a problem if non- or semi-parametric methods are applied to estimate $G(\cdot)$ beyond $\tau := \max_i\left(\widetilde{T}_i\right)$. In particular, the Kaplan–Meier estimator (being the predominant estimator of $G(\cdot)$ in practice) is zero beyond $\max_i\left(\widetilde{T}_i\right)$ if the longest observed event time corresponds to a censored observation (and does not even exist beyond $\max_i\left(\widetilde{T}_i\right)$ if this observation has $\Delta_i = 1$). These problems can be avoided if a restricted version of the C-index (with a suitably defined value of $\tau < \max_i\left(\widetilde{T}_i\right)$) is considered for analysis.

2.4 | Meta-regression of C-index estimates

In this section we describe a set of models to account for the time dependency of the restricted C-index in meta-regression. We start with the classical model for time-independent meta-analysis, also discussing possible transformations of C-index estimates before model fitting.

Random-effects meta-analysis. Consider a set of K independent validation studies with reported study-specific estimates $\widehat{C}_1,...,\widehat{C}_K$ and reported variance estimates $\widehat{\sigma}_1^2,...,\widehat{\sigma}_K^2$. As argued above, each of these estimates relates to a study-specific truncation time τ_k, $k = 1,...,K$. Classical parametric meta-analysis ignores this time dependency, assuming that $\widehat{C}_1,...,\widehat{C}_K$ are estimates of some study-specific unrestricted concordance probabilities $C_1,...,C_K$. We further assume (here and in all other models, following standard procedures) that each $\widehat{\sigma}_k^2$ corresponds to the true variance σ_k^2 of the respective residual term $\epsilon_k = \widehat{C}_k - C_k$. The corresponding model is given by

$$\widehat{C}_k = C_k + \epsilon_k,\ \epsilon_k \sim N\left(0, \sigma_k^2\right),$$

$$C_k = C_{\text{pop}} + a_k, a_k \sim N\left(0, \sigma_a^2\right),\ k = 1,...,K, \tag{6}$$

where the aim is to obtain a "pooled" estimate of the population value C_{pop}. The study-specific deviations a_k are

assumed to be independent of ϵ_k and to follow a normal distribution with between-study variance σ_a^2.

If homogeneity of studies is assumed, i.e. $\sigma_a^2 = 0$ and $C_1 = \ldots = C_k = C_{\text{pop}}$, then this is referred to as common-effect meta-analysis. In contrast, random-effects meta-analysis assumes $\sigma_a^2 \neq 0$, accounting for study-specific heterogeneity. Results of validation studies are usually expected to vary between studies, as these may differ in sample selection and many other design aspects. Therefore, and in line with the recommendation of Debray et al.,[6] we will restrict our analysis to random-effects models for the purpose of our study. In the literature, numerous methods to estimate σ_a^2 have been proposed.[31] Here we follow the recommendation by Debray et al.[6] and consider methods based on restricted maximum likelihood (REML) estimation. With this approach, estimation of σ_a^2 and C_{pop} is performed jointly using a model with inverse variance weights $1/\widehat{\sigma}_k^2$.

Transformations of C-index estimates. The classical approach to meta-analyze *C*-index values is based on the untransformed estimates $\widehat{C}_1, \ldots, \widehat{C}_K$. This approach, which relies on the asymptotic normality of estimators like Uno's *C*, has been followed e.g. by Büttner et al.[20] and Waldron et al.[32] Other authors have argued that the concordance probability is bounded between 0 and 1, so that the normality and homoscedasticity assumptions in (6) are unlikely to hold. To address these issues, they transformed *C*-index estimates before meta-analysis, using e.g. the logistic transformation $g\left(\widehat{C}_k\right) = \log\left(\widehat{C}_k / \left(1 - \widehat{C}_k\right)\right)$[33] or the arcsine square root transformation $g\left(\widehat{C}_k\right) = \sin^{-1}\left(\widehat{C}_k^{1/2}\right)$.[34] After model fitting, the estimate of C_{pop} is usually back-transformed to the original probability scale.

Linear meta-regression. As argued above, classical meta-analysis does not account for the implicit time dependency of the estimates $\widehat{C}_1, \ldots, \widehat{C}_K$. As a consequence, it is unclear how to interpret the population value C_{pop} in Equation (6). In particular, C_{pop} will not be a meaningful approximation of the unrestricted *C*-index if $C(\tau)$ decreases with τ (see Figure 1).

A more appropriate approach to account for the time dependency of $C(\tau)$ is to consider a meta-regression model of the form

$$g\left(\widehat{C}_k\right) = f(\tau_k;\gamma) + a_k + \epsilon_k, \quad a_k \sim N\left(0, \sigma_a^2\right), \quad \epsilon_k \sim N\left(0, \sigma_k^2\right), \tag{7}$$

$k = 1, \ldots, K$, where $g(\cdot)$ is a pre-specified transformation (for instance, the logistic transformation) and τ_k is included as a covariate. The relationship between $\widehat{C}_k$ and τ_k is modeled by the (possibly nonlinear) function $f(\cdot)$ depending on a coefficient vector $\gamma \in \mathbb{R}^p$. Instead of calculating a one-dimensional pooled estimate of C_{pop}, the idea is to first estimate the coefficient vector γ and to subsequently approximate the full curve $C(\tau)$ by the estimated function $g^{-1}(f(\tau;\widehat{\gamma}))$.

The simplest way of specifying a model of the form (7) is to consider the linear function $f(\tau_k;\gamma) = \gamma_0 + \tau_k \cdot \gamma_1$, yielding the *linear meta-regression model* with $\gamma = (\gamma_0, \gamma_1)^T \in \mathbb{R}^2$. Estimation of γ is performed in the same way as above, i.e. using REML with inverse variance weights $1/\widehat{\sigma}_k^2$.

Spline meta-regression. Although the linear meta-regression model accounts for the time dependency of $C(\tau)$, it does not capture nonlinear functional relationships as the ones presented in Figures 1 and 2. This might be a problem even when the values of $\widehat{C}_k$ are transformed before model fitting. A convenient approach to address nonlinearity is to represent $f(\tau_k;\gamma)$ by a restricted cubic spline, as implemented in the R packages *metafor* and *rms*.[35,36] With this approach, $f(\tau_k;\gamma)$ is specified as a weighted sum of truncated power basis functions (defined using a pre-specified set of *interior knots*), and γ is set equal to the vector of weights. Regarding the number and placement of the knots, we follow the recommendations in section 2.4.6 of Harrell Jr.,[37] using four knots (= basis functions) if $K \geq 30$ and three knots if $K < 30$. Obviously, the spline meta-regression model depends on a larger number of coefficients than the linear meta-regression model, increasing its flexibility but also being more prone to overfitting (especially when the number of studies is small).

Fractional polynomial meta-regression. An alternative to spline regression is fractional polynomial (FP) modeling, which is based on transformations of τ_k by a weighted sum of power functions. Following Royston and Sauerbrei,[38] we consider fractional polynomials of degree 2 ("FP2"), which are defined by $f(\tau_k;\gamma) = \gamma_0 + \gamma_1 \cdot \tau_k^{p_1} + \gamma_2 \cdot \tau_k^{p_2}$, where p_1 and p_2 are chosen from the predefined set of powers $S = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ with $\tau_k^0 := \log(\tau_k)$. In case $p_1 = p_2 =: p^*$, the function $f(\cdot)$ is defined by $f(\tau_k;\gamma) = \gamma_0 + \gamma_1 \cdot \tau_k^{p^*} + \gamma_2 \cdot \tau_k^{p^*} \cdot \log(\tau_k)$. As demonstrated by Royston and Sauerbrei[38] and Royston and Altman,[39] FP2 models are able to capture a wide variety of nonlinear trends, providing "enough flexibility for modelling many of the types of continuous function we encounter in the health sciences and elsewhere" (Royston & Sauerbrei,[38] p. 73). For meta-regression of the *C*-index we propose to use a function of the form $f(\tau_k;\gamma) = \gamma_0 + \gamma_1 \cdot \tau_k^{-0.5} + \gamma_2 \cdot \tau_k^{0.5}$, i.e. p_1 and p_2 are set to -0.5 and 0.5, respectively. The latter values are inspired by the "typical" shape of $C(\tau)$ in Figure 1 and by the fact that this shape closely resembles the respective FP2 plot in Figure 1 of Royston.[40] Section 4

presents a detailed empirical analysis of the choice of the power values.

Exponential decay meta-regression. In addition to the aforementioned meta-regression models, we consider the *exponential decay meta-regression model*, which employs an alternative regression function that requires the time-restricted concordance index to be monotone decreasing with τ. This approach might be suitable when there is strong evidence of a monotonic trend in $C(\tau)$ (as the one shown in Figure 1). The exponential decay meta-regression model is specified as

$$g\left(\widehat{C}_k\right) = \theta + a_k + (R_0 - (\theta + a_k)) \cdot \exp(-\exp(\beta) \cdot \tau_k) + \epsilon_k,\ a_k \sim N\left(0, \sigma_a^2\right),\ \ \epsilon_k \sim N\left(0, \sigma_k^2\right), \tag{8}$$

with parameter vector $\gamma = (\theta, \beta, R_0)^T$. By definition, $g\left(\widehat{C}_k\right)$ converges to $\theta + a_k + \epsilon_k$ as $\tau_k \to \infty$, implying that the unrestricted *C*-index might be estimated by the fitted value of θ. (Note that this is not possible with the linear, spline, and fractional polynomial regression approaches described above.) The value of R_0 corresponds to an approximation of $C(\tau_k)$ at $\tau_k = 0$, and β determines the rate of decay. Further note that the random-effects structure of the exponential decay model is slightly different from the respective structure in (7), as the random effect a_k enters (8) in a nonlinear way.

3 | SIMULATION STUDY

3.1 | Experimental setup

Here we present the results of a simulation study that we conducted to analyze the properties of the models discussed in Section 2. The aims of our study were (i) to investigate the benefit of incorporating the truncation times τ_k in meta-regression models for the concordance probability, (ii) to compare the performance of the meta-regression approaches discussed in Section 2 with regard to estimation accuracy and numerical stability, and (iii) to investigate the use of variable transformations before model fitting.

Our simulation study was based on a Weibull model of the form

$$\log(T_i) = \eta_i - \sigma W_i,\ \ \eta_i \sim N\left(0, 0.5^2\right),\ \ i = 1, \ldots, n, \tag{9}$$

with normally distributed score values η_i and noise variables W_i that followed a standard Gumbel distribution (independent of η_i). The parameter σ was set to 0.5, yielding the *C*-index curve presented in Figure S1. For example, we obtained $C(\tau_k) = 0.79$, 0.77 and 0.74 for $\tau_k = 0.2$, 0.7 and 1.5, respectively.

Based on Model (9), we considered four scenarios for meta-regression, setting the number of validation studies to $K = 10$, 15, 30, and 50. For each K we simulated study data sets with n_k observations ($k = 1, \ldots, K$), generating the sample sizes n_k randomly from the grid $\{100, 110, 120, \ldots, 990, 1000\}$. The censoring times Z_i were sampled from an exponential distribution with rate parameter 0.5. The truncation times τ_k of the studies were generated as follows: First, we defined a joint maximum follow-up time (denoted by $\tau_{\max}$) for all studies. Afterwards we sampled the values τ_k from a truncated gamma distribution on the interval $[0.1; \tau_{\max}]$. The shape and rate parameters of this distribution were set to 1.5 and 1, respectively. Subsequently, event times with $\widetilde{T}_i > \tau_k$ were censored at τ_k (study-wise). We considered three values of $\tau_{\max}$, namely $\tau_{\max} = 0.7$ ("short follow-up"), $\tau_{\max} = 0.9$ ("medium follow-up"), and $\tau_{\max} = 2$ ("long follow-up"), yielding average censoring rates of 0.92, 0.86 and 0.64, respectively. Estimates of the *C*-index were obtained using Uno's *C*, as implemented in the R package **pec**. To introduce study-specific heterogeneity, we added normally distributed random numbers a_k to the *C*-index estimates. These numbers were drawn from a normal distribution with zero mean and variance σ_a^2. Again we considered three values, setting $\sigma_a^2 = 0$ ("no heterogeneity"), $\sigma_a^2 = 0.01^2$ ("moderate heterogeneity"), and $\sigma_a^2 = 0.03^2$ ("large heterogeneity"). The choice of these numbers was inspired by our work in Zacharias et al.,[25] where differences in *C*-index values varied between 0.002 and 0.06 across validation studies (the latter number corresponding to two standard deviations of our "large heterogeneity" setting).

In each of the $4 \times 3 \times 3 = 36$ scenarios (defined by the values of K, $\tau_{\max}$ and σ_a^2) we set the number of Monte Carlo replications to 1000 and fitted the following models to the simulated study data: (i) meta-analysis, (ii) linear meta-regression, (iii) spline meta-regression, (iv) fractional polynomial meta-regression, and (v) exponential decay meta-regression, as described in Section 2. Model fitting was carried out using the metamean, metareg and rma functions of the R packages *meta* and *metafor*,[35,41] except for the exponential decay model for which the function nlme of the R package *nlme*[42] was used. For sensitivity analysis, we additionally carried out random-effects meta-analyses using the 30% and 50% of studies with largest values of τ_k only. This approach was inspired by the shape of $C(\tau)$ in Figure 1, assuming that studies with a long follow-up time would be less affected by time dependency due to the convergence behavior of $C(\tau)$. Standard errors of the *C*-index estimates (needed to calculate the weights $1/\widehat{\sigma}_k^2$)

were computed using 1000 bootstrap samples with replacement.

Transformation functions included the identity transformation (id), the logistic transformation (logit), and the arcsine square root transformation (asin). Another candidate transformation would have been the double arcsine transformation; however, we did not consider this transformation because it has recently been found unsuitable for meta-analysis purposes.[43] Comparisons of the respective C-index estimates were carried out at the population level (setting $\sigma_a = 0$) after back transforming the fitted values to the original scale.

The following criteria were used to evaluate the results of the simulation study:

1. To investigate the numerical stability of the methods, we calculated the proportion of simulation runs in which the respective R fitting functions issued errors and/or warnings indicating convergence issues. These assessments were necessary since each of the studies entered the models with a separate random effect a_k and a separate variance term σ_k^2, potentially leading to some instabilities in the REML procedure.
2. To investigate the estimation accuracy of the meta-regression models at a fixed time point, we evaluated the pooled estimates of the restricted concordance probability at $t = 0.8 \cdot \tau_{\max}$ and compared these estimates (including their 95% confidence intervals) to the respective true values of $C(0.8 \cdot \tau_{\max})$.
3. For all methods we computed the areas enclosed by the true and the estimated C-index curves, using $\min_k(\tau_k)$ and $\max_k(\tau_k)$ as interval limits. All areas were divided by the interval length $(\max_k(\tau_k) - \min_k(\tau_k))$, see Figure S2 for an illustration.

3.2 | Results

Tables S1–S4 contain a summary of the failure rates that is, the percentages of the simulation runs in which the R fitting functions issued either an error or a warning. It is seen that fitting the exponential decay model resulted in a large number of convergence issues. For example, in the scenario with $K = 30$ studies (Table S3), failure rates were as high as 74.3% of the simulation runs. Generally, failure rates tended to decrease with the length of follow-up, which can be explained by the more pronounced curvature of the C-index curve in these scenarios (showing stronger support for the shape of the exponential decay function). Still, failure rates were high even in the most favorable settings. We conclude that the exponential decay method may not be recommended for meta-regression of the concordance index, and we therefore did not consider this model further. The failure rates of the other methods were throughout close to zero.

In the remainder of this section, we present the results obtained from the scenario with $K = 30$ studies. The results of the other three scenarios ($K = 10$, $K = 15$, $K = 50$) are presented in the Supporting Information.

Figure 3 presents the pooled concordance probability estimates at the fixed truncation time $0.8 \cdot \tau_{\max}$ (logistic transformation, $K = 30$). It is seen that ignoring the time-dependency of $C(\tau)$ resulted in a bias of classical random-effects meta-analysis. In line with Figure 1, this bias was positive in most of the scenarios and was most pronounced when the follow-up time was long. It was close to zero on average when the follow-up time was short. As expected, the estimates obtained from the sensitivity analyses (corresponding to random-effects analyses of the 30% and 50% of studies with largest values of τ_k) were almost unbiased in the scenarios with long follow-up. The meta-regression methods performed well in all settings, with spline meta-regression showing a higher variability than linear and fractional polynomial meta-regression. As expected, the variance of the estimates increased as the heterogeneity between studies became larger.

Similar results were obtained in the scenarios with $K = 10$, $K = 15$ and $K = 50$ (Figures S3–S5, respectively). The results obtained from the untransformed and arcsine-square-root-transformed estimates ($K = 30$) are presented in Figures S6 and S7, respectively. Compared to the logit-transformed estimates, these estimates showed a slightly increased bias, especially in the scenarios with short follow-up. Again, spline meta-regression had a higher variability than fractional polynomial meta-regression.

Figure 4 presents the estimated coverage probabilities of the 95% Hartung-Knapp confidence intervals at the fixed truncation time $0.8 \cdot \tau_{\max}$ (logistic transformation, $K = 30$). It is seen that the confidence intervals obtained from the meta-analysis model (ignoring follow-up time) did not reach the desired coverage probability in the scenario with long follow-up. All other coverage probability estimates were close to the 95% level. The results obtained from the models with untransformed and arcsine-square-root-transformed C-index estimates ($K = 30$) showed similar patterns (Figures S8 and S9, respectively), except that the meta-analysis model performed generally worse than the other models when fitted to the untransformed estimates (regardless of the length of follow-up). The estimated coverage probabilities obtained from the scenarios with $K = 10$, $K = 15$ and $K = 50$ showed similar patterns as well, again suggesting that the meta-analysis model is inferior to the meta-regression models in the scenarios with long follow-up (data not shown).

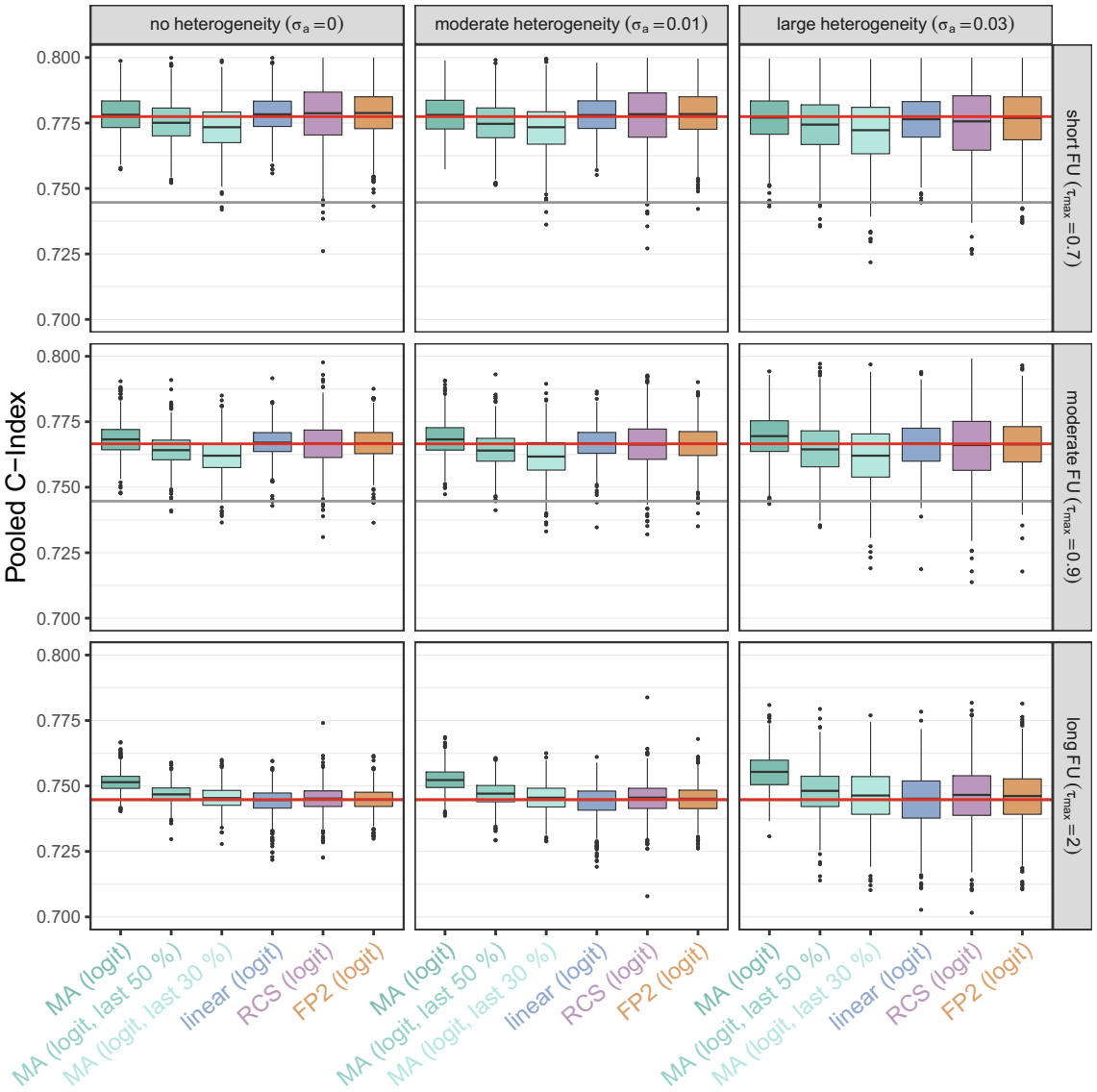

FIGURE 3 Results of the simulation study ($K = 30$). The boxplots summarize the pooled estimates of the restricted concordance index at $0.8 \cdot \tau_{\max}$. All *C*-index estimates were transformed using a logistic transformation before model fitting. The red and the gray lines refer to the true values of $C(0.8 \cdot \tau_{\max})$ and the unrestricted values of the concordance index, respectively. Note that the gray lines coincide with the red lines in the lower three panels. [Colour figure can be viewed at wileyonlinelibrary.com]

Table 1 presents the areas enclosed by the true and the estimated *C*-index curves. It is seen that the areas obtained from time-independent random-effects meta-analysis tended to increase with increasing follow-up time, whereas the respective areas obtained from the meta-regression models tended to *de*crease with increasing follow-up time. Random-effects meta-analysis was the overall best method in settings with short follow-up. By contrast, linear meta-regression of logit-transformed *C*-index estimates and fractional polynomial meta-regression of logit-transformed *C*-index estimates tended to perform best in the scenarios with moderate and long follow-up times, respectively. These results clearly suggest that the time-constant functions obtained from random-effects meta-analysis are reasonable approximations to $C(\tau)$ in settings with a short follow-up. By contrast, the benefits of modeling *C*-index values by a regression function become apparent when follow-up times are "long enough" to demonstrate possible time dependencies and nonlinear shapes of $C(\tau)$ (for a discussion on how to assess the relative length of the follow-up time, see Section 4). In most cases, the areas between the true and the estimated curves were smallest when *C*-index estimates were transformed by the logistic transformation before model fitting. Similar results were obtained in the scenarios with $K = 10$, $K = 15$ and $K = 50$ (Tables S5–S7, respectively), except that the areas obtained from time-independent meta-analysis were considerably less dependent on follow-up time when $K = 10$.

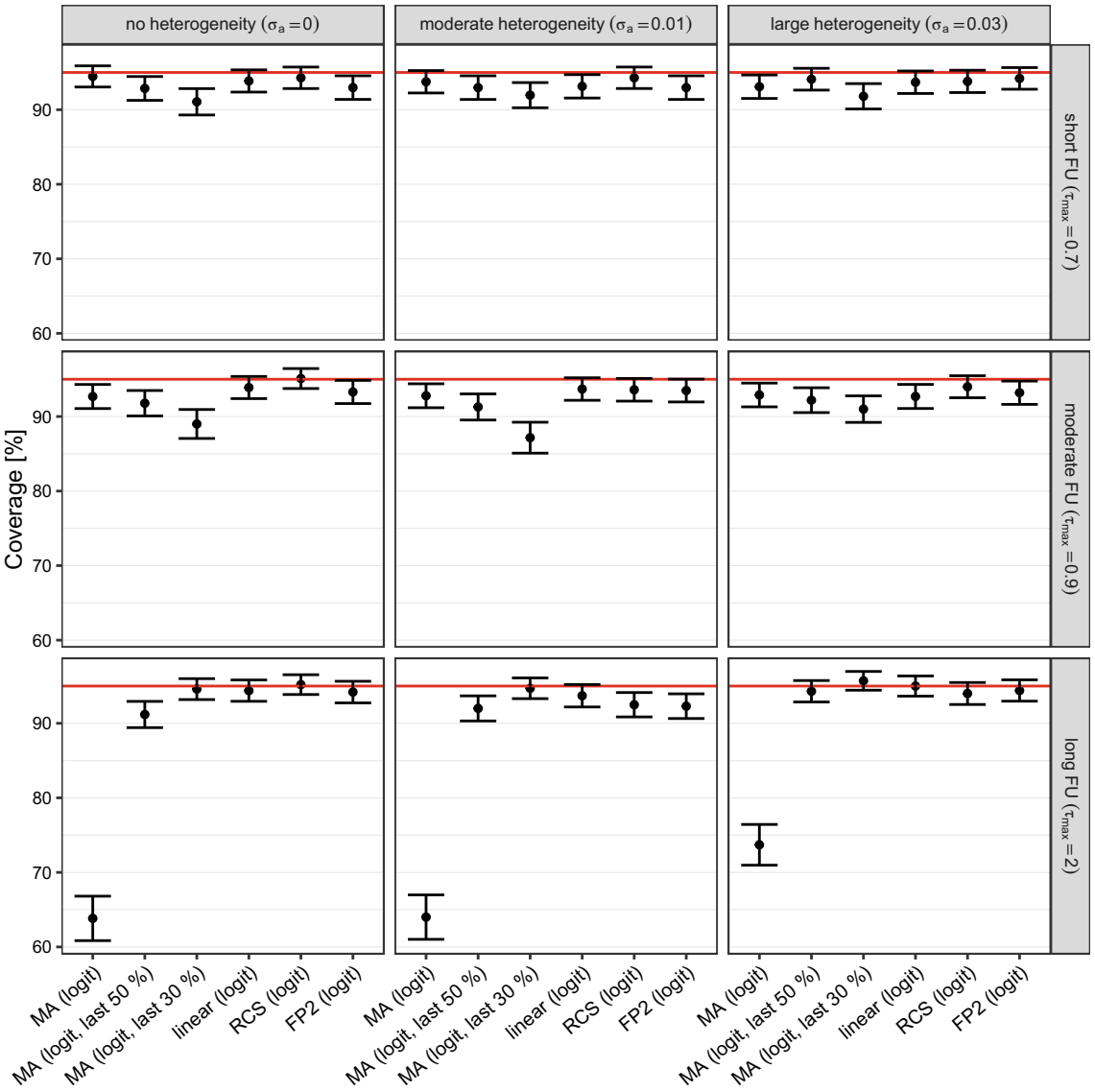

FIGURE 4 Results of the simulation study ($K = 30$). The plot shows the estimated coverage probabilities (%) that is, the proportion of simulation runs in which the 95% Hartung-Knapp confidence intervals contained the true value of $C(0.8 \cdot \tau_{\max})$. Confidence limits (represented by the black lines) were computed as $\left[\widehat{p} \pm 1.96 \cdot \sqrt{\widehat{p} \cdot (1-\widehat{p})/1000}\right]$, where $\widehat{p}$ denotes the point estimate of the coverage probability. The red lines refer to the 95% confidence level. All *C*-index estimates were transformed by a logistic transformation before model fitting. [Colour figure can be viewed at wileyonlinelibrary.com]

In addition, linear meta-regression of logit-transformed *C*-index estimates tended to perform best for long follow-up lengths in this scenario (Table S5).

In summary, our simulation study suggests that (i) spline and fractional polynomial meta-regression should be preferred over the exponential decay approach to model nonlinearities in $C(\tau)$, (ii) pooled estimates obtained from time-independent meta-analysis are reasonable approximations of $C(\tau)$ as long as follow-up times short, whereas it is beneficial to consider increasingly complex meta-regression models with increasing values of τ and K, and (iii) models with logit-transformed *C*-index estimates showed the overall best performance (compared to untransformed and arcsine-square-root-transformed estimates).

We further note that, compared to spline meta-regression, fractional polynomial meta-regression showed a slightly better overall performance in terms of the areas enclosed between the true and the fitted *C*-index curves. Importantly, the performance of fractional polynomial meta-regression could be improved further by optimizing the power values p_1 and p_2 (instead of considering the fixed values -0.5 and 0.5, as done in this section). We will investigate this issue further in Section 4. For details on variable and power selection in fractional polynomial regression, see Royston and Sauerbrei.[38]

TABLE 1 Results of the simulation study ($K = 30$).

	$\sigma_a = 0$			$\sigma_a = 0.01$			$\sigma_a = 0.03$		
	Short	**Moderate**	**Long**	**Short**	**Moderate**	**Long**	**Short**	**Moderate**	**Long**
MA (id)	8.9 (4.9)	10.0 (2.0)	12.0 (1.5)	9.1 (5.1)	10.1 (2.1)	13.4 (2.5)	10.2 (6.1)	10.9 (3.3)	13.4 (2.5)
MA (id, last 50%)	8.4 (4.0)	11.7 (3.2)	11.7 (1.6)	8.6 (4.4)	11.9 (3.6)	14.1 (4.5)	10.7 (6.4)	13.4 (5.7)	14.1 (4.5)
MA (id, last 30%)	9.4 (5.0)	13.5 (4.4)	12.5 (2.5)	9.8 (5.6)	13.9 (5.1)	15.6 (5.8)	12.6 (8.3)	15.8 (8.0)	15.6 (5.8)
MA (logit)	7.9 (3.6)	10.8 (2.5)	11.4 (1.1)	8.2 (3.9)	10.8 (2.7)	12.9 (2.2)	9.7 (5.5)	11.5 (3.9)	12.9 (2.2)
MA (logit, last 50%)	9.2 (4.6)	12.7 (3.6)	11.9 (1.8)	9.5 (4.9)	12.9 (4.0)	14.1 (4.5)	11.4 (6.9)	14.1 (6.2)	14.1 (4.5)
MA (logit, last 30%)	10.4 (5.6)	14.4 (4.8)	12.7 (2.6)	10.8 (6.1)	14.7 (5.5)	15.5 (5.8)	13.5 (8.7)	16.4 (8.3)	15.5 (5.8)
MA (asin)	8.1 (4.2)	10.2 (2.1)	11.6 (1.3)	8.3 (4.4)	10.3 (2.2)	13.1 (2.3)	9.6 (5.6)	10.9 (3.3)	13.1 (2.3)
MA (asin, last 50%)	8.6 (4.2)	12.2 (3.4)	11.8 (1.7)	8.9 (4.5)	12.4 (3.8)	14.1 (4.5)	10.9 (6.5)	13.6 (5.9)	14.1 (4.5)
MA (asin, last 30%)	9.8 (5.3)	13.9 (4.6)	12.6 (2.5)	10.2 (5.8)	14.3 (5.3)	15.5 (5.8)	12.9 (8.4)	16.0 (8.1)	15.5 (5.8)
linear (id)	13.4 (7.6)	10.0 (6.2)	7.0 (2.2)	13.3 (7.6)	10.2 (6.2)	9.7 (3.8)	13.5 (7.9)	11.5 (6.4)	9.7 (3.8)
linear (logit)	9.5 (5.7)	7.4 (4.3)	6.9 (1.6)	9.7 (5.9)	7.8 (4.4)	9.5 (3.6)	11.4 (6.8)	10.2 (5.4)	9.5 (3.6)
linear (asin)	11.2 (6.7)	8.4 (5.2)	6.8 (1.7)	11.3 (6.8)	8.7 (5.3)	9.6 (3.7)	12.1 (7.2)	10.6 (5.8)	9.6 (3.7)
RCS (id)	16.0 (6.9)	12.6 (5.7)	6.9 (3.0)	16.2 (6.8)	12.9 (5.7)	11.5 (4.4)	17.6 (7.2)	15.1 (6.0)	11.5 (4.4)
RCS (logit)	13.4 (5.8)	10.8 (4.6)	6.2 (2.5)	13.7 (5.9)	11.2 (4.6)	11.2 (4.3)	16.0 (6.6)	14.1 (5.5)	11.2 (4.3)
RCS (asin)	14.5 (6.4)	11.5 (5.1)	6.4 (2.7)	14.8 (6.4)	11.9 (5.1)	11.3 (4.3)	16.6 (6.8)	14.5 (5.7)	11.3 (4.3)
FP2 (id)	14.4 (6.9)	11.2 (5.6)	6.3 (2.7)	14.6 (6.9)	11.5 (5.6)	9.8 (4.1)	15.5 (7.3)	13.2 (6.0)	9.8 (4.1)
FP2 (logit)	11.4 (5.7)	9.3 (4.4)	5.7 (2.4)	11.7 (5.9)	9.7 (4.5)	9.7 (4.1)	13.6 (6.7)	12.2 (5.3)	9.7 (4.1)
FP2 (asin)	12.7 (6.3)	10.0 (5.0)	5.9 (2.5)	12.9 (6.4)	10.4 (5.0)	9.7 (4.1)	14.3 (7.0)	12.6 (5.6)	9.7 (4.1)

Note: The table summarizes the areas enclosed by the true and the estimated C-index curves (mean [sd]), as obtained from the meta-regression models described in Section 2. All areas were divided by the interval lengths $(\max_k(\tau_k) - \min_k(\tau_k))$ and multiplied by 1000.

4 | ILLUSTRATION

To illustrate the proposed methods, we analyzed data from the German Chronic Kidney Disease (GCKD) Study, which is an ongoing multi-center cohort study that enrolled 5217 patients with chronic kidney disease (CKD). The aim of the study is to identify risk factors associated with CKD progression, cardiovascular events and death. For details on the inclusion/exclusion criteria and the design of the study, see Eckardt et al.[24] Baseline data collection took place between March 2010 and March 2012; it comprised measurements on clinical and lifestyle variables (e.g. coronary heart disease, smoking) and biomarker measurements obtained from blood and urine samples. Follow-up data are collected annually. The laboratory measurements collected for the GCKD Study have been used previously for predictive modeling and score development.[25]

An important characteristic of the GCKD Study is its wide geographical coverage. Altogether, there are nine study centers, each representing a specific German region with a distinct patient population. During the past years, it has become increasingly popular to account for such heterogeneity by synthesizing center-specific estimates via meta-analysis techniques.[44–47] Here we followed this approach and used the GCKD data to evaluate a prognostic model in each of the nine centers, illustrating our proposed methodology by meta analyzing the respective center-specific C-index estimates ($K = 9$). Note that the availability of individual patient data allowed us to estimate $C(\tau)$ in each center at arbitrary time horizons.

For model building and evaluation we considered the endpoint "time to cardiovascular death". Data were exported from the GCKD database after the 8th follow-up examination (maximum follow-up time 2933 days, median = 2554 days, first quartile = 2060 days, third quartile = 2591 days, cardiovascular death rate = 200/4455 = 4.5% after listwise deletion of patients with a missing value in at least one of the covariates). In the first step, we split the data randomly into three equally sized parts: The first part was used as *training data* for model building, the second part was used as *analysis data* for prediction and meta-regression, and the third part was used as *test data* for evaluating the performance of the meta-regression models. In the second step, we derived a prediction model for cardiovascular death by fitting a Cox regression model to the training data. The following (pre-selected) baseline covariates were included in the model: C-reactive protein

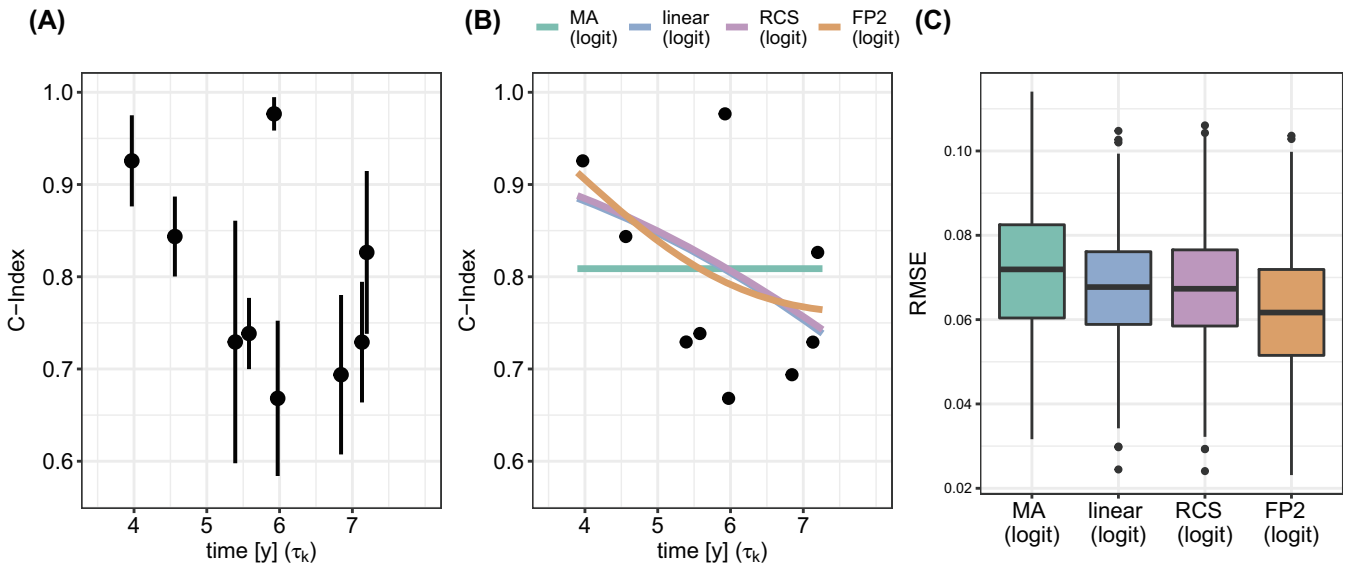

FIGURE 5 Analysis of the German Chronic Kidney Disease (GCKD) Study data. (a) Nine center-specific *C*-index estimates at randomly generated truncation times τ_k, $k = 1,...,9$. Estimates and 95% confidence intervals (represented by bars) were obtained by application of Uno's *C* to the GCKD analysis data. (b) The colored lines refer to the back-transformed meta-analysis and -regression curves obtained by fitting the models of Section 2.4 to the logit-transformed *C*-index estimates (FP2, fractional polynomial meta-regression; linear, linear meta-regression; MA, standard random effects meta-analysis ignoring time dependency; RCS, restricted cubic spline meta-regression). (c) Root mean squared error values obtained from the bootstrapped test data (1000 replications). [Colour figure can be viewed at wileyonlinelibrary.com]

(mg/L), cholesterol (mg/dL), calcium (mmol/L), phosphate (mmol/L), albumin (g/L), cystatin C (mg/L), age (years), sex (male/female), urine albumin-to-creatinine ratio (mg/g), hypertension (yes/no), previous coronary heart disease (yes/no), smoking (non-smoker, former smoker, current smoker), and estimated glomerular filtration rate (mL/min/1.73 m^2). Furthermore, we generated a random truncation time τ_k, $k = 1,...,9$, for each of the study centers, restricting τ_k to be larger than 3.5 years in order to obtain sufficiently large event counts. In the third step, we used the coefficients of the Cox model to predict the values of η in the analysis data. These values were subsequently used to estimate the restricted concordance probability in each study center at the center-specific truncation times τ_k. The resulting estimates $\widehat{C}_k$ (obtained by application of Uno's *C*) are visualized in Figure 5a. In the fourth step, we meta-analyzed the center-specific *C*-index estimates by applying the methods presented in Section 2.4 to the pairs of values $\left(\tau_1, \widehat{C}_1\right),...,\left(\tau_9, \widehat{C}_9\right)$. Based on the results of our simulation study, *C*-index estimates were logit-transformed before model fitting. In the fifth step, we generated 1000 bootstrap samples from the test data and re-estimated the concordance probabilities $C(\tau_k)$, $k = 1,...,9$, as well as the fitted values $g^{-1}(f(\tau_k;\widehat{\gamma}))$ (obtained from meta-regression) in each of the samples. Furthermore, we computed the weighted root mean squared error (RMSE, defined by $\left[\sum_{k=1}^{K} n_k/n\left(\widehat{C}_k - g^{-1}(f(\tau_k;\widehat{\gamma}))\right)^2\right]^{1/2}$), which was used to evaluate and compare the performance of the meta-regression models.

The fitted curves obtained from the analysis data are presented in Figure 5b. It is seen that the meta-regression models (accounting for the length of follow-up) resulted in very similar fits. Fractional polynomial meta-regression seemed to perform best by visual inspection. The model summaries (given in Table 2) confirm these results, with the values of the estimated between-study standard deviation $\widehat{\sigma}_a$ ranging between 0.708 and 0.821 on the logit scale. Boxplots of the RMSE values (obtained from the bootstrapped test data) are shown in Figure 5c. Again, it can be seen that the models performed similarly, with the highest RMSE value observed for the meta-analysis model and the lowest RMSE value for the fractional polynomial meta-regression model.

In the final step, we investigated whether we could improve the performance of fractional polynomial meta-regression by optimizing the power values of the FP2 model. To this purpose, we repeated the bootstrap analysis of the GCKD test data, this time computing the RMSE

values obtained from all possible combinations of the powers $p_1, p_2 \in \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$. The results of our analysis suggest that the RMSE values were not very sensitive to the choice of powers (Figure 6). In particular, the performance of our initial model from Section 2.4 ($p_1 = -0.5$, $p_2 = 0.5$, Figure 5c) was close to the performance of the optimal model with $p_1 = p_2 = -2$. In simulations (based on the same design as in Section 3), we additionally investigated the use of 3rd degree fractional polynomials for meta-regression of the *C*-index. Our results (not shown) suggest that increasing the degree of the fractional polynomial had very little effect on the median *C*-index estimates at $0.8 \cdot \tau_{\max}$. Instead, this approach tended to increase both the variance of the estimates and the areas enclosed by the true and estimated *C*-index curves.

TABLE 2 Analysis of the GCKD Study data.

	$\widehat{\sigma}_a$	*Q*	df	*p* value
MA (logit)	0.708	22.3	8	0.0043
Linear (logit)	0.711	19.9	7	0.0059
RCS (logit)	0.821	18.9	6	0.0043
FP2 (logit)	0.793	18.1	6	0.0060

Note: The table summarizes the fits of the meta-analysis and -regression models, as obtained by applying the methods of Section 2.4 to the center-specific *C*-index values (estimated from the GCKD analysis data). The logistic transformation was applied to the estimated *C*-index values before model fitting. The table presents the values of the estimated between-study standard deviation $\widehat{\sigma}_a$ (on the logit scale), the test statistic for residual heterogeneity *Q* (following a Chi-squared distribution under the null hypothesis of homogeneous residuals), its degrees of freedom (df), and the corresponding *p* value.

5 | DISCUSSION

The development of prognostic models has become a predominant task in medical and epidemiological research. As noted by Riley et al.,[5] "prognostic factors have many potential uses, including aiding treatment and lifestyle decisions, improving individual risk prediction, providing novel targets for new treatment, and enhancing the design and analysis of randomized trials identify patients for trials". To this end, a large number of prognostic scores has been developed, requiring proper validation to become accepted for eventual use in clinical practice. The gold standard for validation is to analyze the performance of novel scores using large external cohorts; however, for a variety of reasons (including confidentiality and logistical issues), this is often not possible. It is therefore important to combine the results of smaller validation studies using meta-analysis techniques.

In this paper we proposed and evaluated a framework for meta-analyzing the concordance index for time-to-event data, which has become an established measure for the discriminative ability of prognostic scores (see Steyerberg[8] for a comprehensive introduction to predictive modeling, also including other aspects of validation like calibration and clinical usefulness). We analyzed the

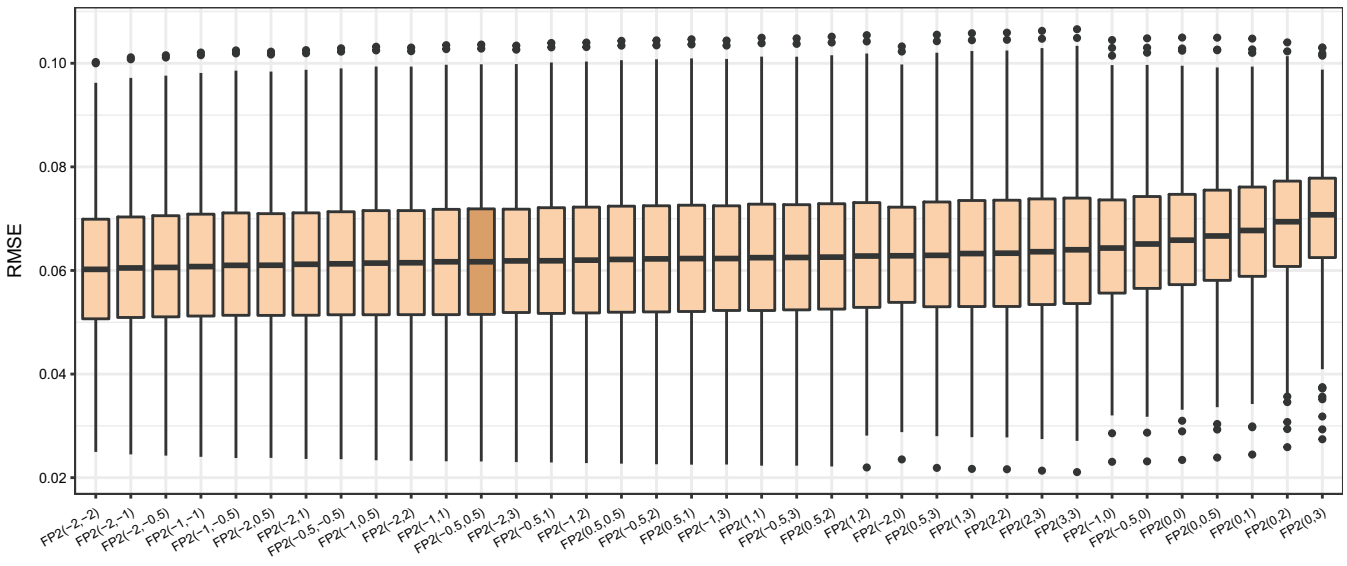

FIGURE 6 Analysis of the German Chronic Kidney Disease (GCKD) data. The boxplots show the root mean squared error (RMSE) values obtained from the bootstrapped test data (1000 replications) when evaluating all possible combinations of the power values $p_1, p_2 \in \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ of the FP2 models. The boxplots are ordered by median RMSE value. The dark orange boxplot corresponds to the powers of the FP2 model from Section 2.4 ($p_1 = -0.5$, $p_2 = 0.5$). [Colour figure can be viewed at wileyonlinelibrary.com]

inherent time-dependency of *C*-index estimates (noted previously by Longato et al.[30]) and proposed methods to account for this time-dependency using meta-regression models. In this respect, our paper connects to Debray et al.[9] who noted that "[...] researchers often refrain from undertaking a quantitative synthesis or meta-analysis of the predictive performance of a specific model. Potential reasons for this pitfall are [...] or simply a lack of methodological guidance."

A key result of this work is that meta-regression models including the study-specific truncation time as covariate perform systematically better than classical random-effects meta-analysis when follow-up times of validation studies are long. Conversely, pooled estimates obtained from classical meta-analysis are reasonable approximations of the restricted concordance probability as long as follow-up times short. We acknowledge that, in practice, it might be challenging to determine whether a time horizon should be considered "short" or "long", in particular when observation times are affected by high dropout rates and/or the presence of competing events. Still, we recommend to carefully investigate this issue, especially since *C*-index values often tend to decrease with τ, implying that studies with a short follow-up time might suggest an overly optimistic discrimination accuracy. Generally, the rate of administrative censoring might be an indicator of whether follow-up times might be considered "long" or "short". Furthermore, we recommend visual inspection of *C*-index estimates in order to examine their dependency on τ.

Based on our numerical experiments, we recommend to transform *C*-index values using a logistic transformation and to employ either restricted cubic splines or fractional polynomials to model the functional relationship between the truncation time and the concordance index. We further recommend to prefer fractional polynomials over splines in settings where the number of studies is "small" ($5 \leq K \leq 10$), as they typically involve fewer degrees of freedom than restricted cubic splines. In case of convergence problems (which might become an issue when the number of studies is smaller than five), our framework readily allows for switching to a simpler model (e.g. a linear meta-regression model). We also note that our proposed models could be extended by additional covariates reflecting different inclusion criteria in the analyzed studies. Along the same lines, our framework could be adapted to models with competing events.[48]

Generally, by building on the framework of Debray et al.,[6,9] our methodology is designed to synthesize *C*-index estimates referring to the *same* score (e.g., the Framingham Risk Score). Technically, one could also imagine using our methods for a synthesis of *C*-index estimates obtained from *different* scores. However, such analyses are rarely done in practice, which is likely due to the typically high heterogeneity between the included models (see e.g. Büttner et al.,[20] who considered a total of 18 prognostic models, but meta-analyzed only those *C*-index estimates referring to the same model). In view of these aspects, our perspective on the analysis of *different* scores is that performance estimates referring to different scores should better not be synthesized directly; instead, comparisons of different scores should be done by computing performance estimates from the same validation data (using the same truncation time for each score). As an alternative, one could borrow information from the other scores when analyzing one of them. This could be achieved by shrinkage estimation, e.g. in a Bayesian framework.[49] A key barrier to meta-analyzing *C*-index values is the huge variety of estimators that have been proposed during the past decades (such as Harrell's *C* and Uno's *C*).[15] Since each of these estimators comes with a different set of assumptions and/or properties, it is challenging to synthesize validation studies relying on different kinds of estimators. Importantly, some of the estimators are known to be systematically biased, e.g. when they rely on a Cox model (whose assumptions might be violated) or when they show a censoring bias (such as Harrell's *C*). We argue that these systematic deviations should not be represented in a meta-regression model by zero-mean random effects. Instead, we suggest to develop methodological guidance on the definition and use of appropriate estimators for the evaluation of discriminatory power, aiming at a unified methodology that would become a standard in future validation studies. Work on such guidance is currently undertaken by the STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative.[50]

Meta-regression of *C*-index estimates is also compromised by the lack of proper reporting. In fact, when searching for a real-world application to be presented in Section 4, we found that most published studies reporting *C*-index estimates did *not* include any information on the respective time horizon. In some cases, we were able to approximate this time horizon by the length of the respective follow-up time; however, in many cases the time horizon was not mentioned at all. Based on the findings presented in Section 3 of this paper, we suggest to always report the time horizons together with *C*-index estimates in future validation studies. We further suggest to report and visualize the whole estimated *C*-index curve whenever a meta-regression has been performed. Ideally, reporting would also include time-dependent sensitivities and specificities in addition to *C*-index estimates, allowing for the application of bivariate meta-analysis techniques.[51]

We finally note that the concordance index is (by far) not the only prognostic measure to be affected by an

inherent time dependency. Another important example are incidence rates, which by definition depend on the time frames under consideration. Clearly, the lengths of these time frames have to be considered when meta-analyzing incidence rates (see Olaciregui-Dague et al.[52] for a recent example). Further research is needed to evaluate possible adaptions of our methodology to these measures.

AUTHOR CONTRIBUTIONS

Matthias Schmid: Conceptualization; formal analysis; methodology; writing – original draft; writing – review and editing. **Tim Friede:** Methodology; writing – review and editing. **Nadja Klein:** Writing – review and editing. **Leonie Weinhold:** Conceptualization; formal analysis; methodology; software; visualization; writing – review and editing.

ACKNOWLEDGMENTS

We thank the GCKD Study Investigators for providing data of the GCKD Study for illustrative purposes. The GCKD study was supported by a grant from the KfH Foundation for Preventive Medicine (https://www.kfh-stiftung-praeventivmedizin.de). Tim Friede is grateful for support by the Volkswagen Foundation (Az.: 98 948; "Bayesian and Nonparametric Statistics-Teaming up two opposing theories for the benefit of prognostic studies in Covid-19"). Open Access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

The R scripts for the simulation study are available on GitHub at https://github.com/weinholdl/metac. The GCKD Study data contain information that could compromise the privacy of study participants. Data sharing restrictions imposed by national and transnational data sharing laws prohibit general sharing of these data. Upon submission of a proposal to the principal investigator of the GCKD Study and approval of this proposal by the Steering Committee of the GCKD Study, data collected for the study can be made available to other researchers.

ORCID

Matthias Schmid https://orcid.org/0000-0002-0788-0317
Tim Friede https://orcid.org/0000-0001-5347-7441
Nadja Klein https://orcid.org/0000-0002-5072-5347

REFERENCES

1. Fire M, Guestrin C. Over-optimization of academic publishing metrics: observing Goodhart's law in action. *GigaScience*. 2019;8(6):giz053.
2. Landhuis E. Scientific literature: information overload. *Nature*. 2016;535:457-458.
3. Altbach PG, de Wit H. Too much academic research is being published. *Int Higher Educ*. 2018;96:2-3.
4. Gough D, Davies P, Jamtvedt G, et al. Evidence synthesis international (ESI): position statement. *Syst Rev*. 2020;9:155.
5. Riley RD, Moons KGM, Snell KIE, et al. A guide to systematic review and meta-analysis of prognostic factor studies. *BMJ*. 2019;364:k4597285.
6. Debray TPA, Damen JAAG, Riley RD, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. *Stat Methods Med Res*. 2019;28:2768-2786.
7. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC Med*. 2015;13(1):1.
8. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. 2nd ed. Springer; 2019.
9. Debray TPA, Damen JAAG, Snell KIE, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ*. 2017;356:i6460.
10. Harrell FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modeling strategies for improved prognostic prediction. *Stat Med*. 1984;3:143-152.
11. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics*. 2005;61:92-105.
12. Gerds TA, Kattan MW, Schumacher M, Yu C. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Stat Med*. 2013;32:2173-2184.
13. Gönen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*. 2005;92:965-970.
14. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med*. 2011;30:1105-1117.
15. Schmid M, Potapov S. A comparison of estimators to evaluate the discriminatory power of time-to-event models. *Stat Med*. 2012;31:2588-2609.
16. Snell KIE, Ensor J, Debray TPA, Moons KGM, Riley RD. Meta-analysis of prediction model performance across multiple studies: which scale helps ensure between-study normality for the C-statistic and calibration measures? *Stat Methods Med Res*. 2018;27:3505-3522.
17. van Doorn S, Debray TPA, Kaasenbrood F, et al. Predictive performance of the CHA2DS2-VASc rule in atrial fibrillation: a systematic review and meta-analysis. *J Thromb Haemost*. 2017;15:1065-1077.
18. van den Boorn HG, Engelhardt EG, van Kleef J, et al. Prediction models for patients with esophageal or gastric cancer: a systematic review and meta-analysis. *PLoS One*. 2018;13(2):e0192310.
19. He Y, Ong Y, Li X, et al. Performance of prediction models on survival outcomes of colorectal cancer with surgical resection: a systematic review and meta-analysis. *Surg Oncol*. 2019;29:196-202.
20. Büttner S, Galjart B, Beumer BR, et al. Quality and performance of validated prognostic models for survival after resection of intrahepatic cholangiocarcinoma: a systematic review and meta-analysis. *HPB*. 2021;23:25-36.

21. Kothari G, Korte J, Lehrer EJ, et al. A systematic review and meta-analysis of the prognostic value of radiomics based models in non-small cell lung cancer treated with curative radiotherapy. *Radiother Oncol*. 2021;155:188-203.
22. Pennells L, Kaptoge S, White IR, Thompson SG, Wood AM. Emerging risk factors collaboration. Assessing risk prediction models using individual participant data from multiple studies. *Am J Epidemiol*. 2014;179:621-632.
23. Hattori S, Zhou XH. Summary concordance index for meta-analysis of prognosis studies with a survival outcome. *Stat Med*. 2021;40:5218-5236.
24. Eckardt KU, Bärthlein B, Baid-Agrawal S, et al. The German chronic kidney disease (GCKD) study: design and methods. *Nephrol Dial Transplant*. 2012;27:1454-1460.
25. Zacharias HU, Altenbuchinger M, Schultheiss UT, et al. A predictive model for progression of CKD to kidney failure based on routine laboratory tests. *Am J Kidney Dis*. 2022;79:217-230.
26. Brentnall AR, Cuzick J. Use of the concordance index for predictors of censored survival data. *Stat Methods Med Res*. 2018;27:2359-2373.
27. Gerds TA. pec: Prediction error curves for risk prediction models in survival analysis. R package version 2022.05.04. 2022. Accessed July 8, 2023. https://CRAN.R-project.org/package=pec
28. Song X, Zhou XH. A semiparametric approach for the covariate specific ROC curve with survival outcome. *Stat Sin*. 2008;18:947-965.
29. van Geloven N, He Y, Zwinderman AH, Putter H. Estimation of incident dynamic AUC in practice. *Comput Stat Data Anal*. 2021;154:107095.
30. Longato E, Vettoretti M, Camillo BD. A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models. *J Biomed Inform*. 2020;108:103496.
31. Sinha BK, Hartung J, Knapp G. *Statistical Meta-Analysis with Applications*. Wiley; 2011.
32. Waldron L, Haibe-Kains B, Culhane AC, et al. Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. *J Natl Cancer Inst*. 2014;106:dju049.
33. van Klaveren D, Steyerberg EW, Perel P, Vergouwe Y. Assessing discriminative ability of risk models in clustered data. *BMC Med Res Methodol*. 2014;14:5.
34. Schwarzer G, Chemaitelly H, Abu-Raddad LJ, Rücker G. Seriously misleading results using inverse of freeman-Tukey double arcsine transformation in meta-analysis of single proportions. *Res Synth Methods*. 2019;10:476-483.
35. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw*. 2010;36(3):1-48.
36. Harrell FE. rms: Regression modeling strategies. R package version 6.3-0. 2022. Accessed July 8, 2023. https://CRAN.R-project.org/package=rms
37. Harrell FE. *Regression Modeling Strategies: with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. 2nd ed. Springer; 2015.
38. Royston P, Sauerbrei W. *Multivariable Model-Building: A Pragmatic Approach to Regression Anaylsis Based on Fractional Polynomials for Modelling Continuous Variables*. Wiley; 2008.
39. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *J R Stat Soc Ser C*. 1994;43:429-453.
40. Royston P. Model selection for univariable fractional polynomials. *Stata J*. 2017;17:619-629.
41. Balduzzi S, Rücker G, Schwarzer G. How to perform a meta-analysis with R: a practical tutorial. *Evid Based Ment Health*. 2019;22:153-160.
42. Pinheiro J, Bates D, R Core Team. nlme: linear and nonlinear mixed effects models. R package version 3.1-160. 2022. Accessed July 8, 2023. https://CRAN.R-project.org/package=nlme
43. Röver C, Friede T. Double arcsine transform not appropriate for meta-analysis. *Res Synth Methods*. 2022;13:645-648.
44. Franceschini N, Shara N, Wang H, et al. The association of genetic variants of type 2 diabetes with kidney function. *Kidney Int*. 2012;82:220-225.
45. Beekman M, Blanche H, Perola M, et al. Genome-wide linkage analysis for human longevity: genetics of healthy ageing study. *Aging Cell*. 2013;12:184-193.
46. Jögi NO, Kitaba N, Storaas T, et al. Ascaris exposure and its association with lung function, asthma, and DNA methylation in northern Europe. *J Allergy Clin Immunol*. 2022;149:1960-1969.
47. Collatuzzo G, Visci G, Violante FS, et al. Determinants of anti-S immune response at 6 months after COVID-19 vaccination in a multicentric European cohort of healthcareworkers—ORCHESTRA project. *Front Immunol*. 2022;13:986085.
48. van Geloven N, Giardiello D, Bonneville EF, et al. Validation of prediction models in the presence of competing risks: a guide through modern methods. *BMJ*. 2022;377:e069249.
49. Röver C, Friede T. Dynamically borrowing strength from another study through shrinkage estimation. *Stat Methods Med Res*. 2020;29:293-308.
50. Sauerbrei W, Abrahamowicz M, Altman DG, le Cessie S, Carpenter J, STRATOS Initiative. Strengthening analytical thinking for observational studies: the STRATOS initiative. *Stat Med*. 2014;33:5413-5432.
51. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol*. 2005;58:982-990.
52. Olaciregui-Dague K, Weinhold L, Hoppe C, Schmid M, Surges R. Anti-seizure efficacy and retention rate of carbamazepine is highly variable in randomized controlled trials: a meta-analysis. *Epilepsia Open*. 2022;7:556-569.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Schmid M, Friede T, Klein N, Weinhold L. Accounting for time dependency in meta-analyses of concordance probability estimates. *Res Syn Meth*. 2023;14(6):807-823. doi:10.1002/jrsm.1655