Contemporary Clinical Trials Communications 15 (2019) 100402

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

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

An alternative trial-level measure for evaluating failure-time surrogate endpoints based on prediction error

Shaima Belhechmi[a,b], Stefan Michiels[a,b,*], Xavier Paoletti[a,b], Federico Rotolo[c]

[a] Université Paris-Saclay, Univ. Paris-Sud, UVSQ, CESP, INSERM, U1018 ONCOSTAT, F-94805, Villejuif, France
[b] Gustave Roussy, Service de biostatistique et d'épidémiologie, F-94805, Villejuif, France
[c] Innate Pharma, Biostatistics and Data Management Unit, F-13009, Marseille, France

ARTICLE INFO

Keywords:
Survival analysis
Surrogate endpoint evaluation
Bivariate models
Copula models
Cox model
Simulation studies. *2010 MSC:* 00–01
99–00

ABSTRACT

To validate a failure-time surrogate for an established failure-time clinical endpoint such as overall survival, the meta-analytic approach is commonly used. The standard correlation approach considers two levels: the individual level, with Kendall's τ measuring the rank correlation between the endpoints, and the trial level, with the coefficient of determination R^2 measuring the correlation between the treatment effects on the surrogate and on the final endpoint. However, the estimation of R^2 is not robust with respect to the estimation error of the trial-specific treatment effects.

The alternative proposed in this article uses a prediction error based on a measure of the weighted difference between the observed treatment effect on the final endpoint and a model-based predicted effect. The measures can be estimated by cross-validation within the meta-analytic setting or external validation on a set of trials. Several distances are presented, with varying weights, based on the standard error of the observed treatment effect and of its predicted value. A simulation study was conducted under different scenarios, varying the number and the size of the trials, Kendall's τ and R^2. These measures have been applied to individual patient data from a meta-analysis of trials in advanced/recurrent gastric cancer (20 randomized trials of chemotherapy, 4069 patients).

The distance-based measures appeared to be robust with respect to different values of simulation parameters in several scenarios (such as Kendall's τ, size and number of clinical trials). The absolute prediction error can be an alternative to the trial-level R^2 for evaluation of candidate time-to-event surrogates.

1. Introduction

The main objective of confirmatory phase-III clinical trials is to determine whether a new treatment is effective. In oncology, the final endpoint of randomized phase-III clinical trials is often the overall survival (OS). However, the use of a final endpoint like OS requires a large number of patients, a long follow-up period, and high research and development costs to achieve the statistical power required in phase-III trials. It is therefore interesting to use intermediate criteria that can be evaluated earlier and used as a surrogate for the final endpoint. Surrogate endpoints may allow shorter trial duration, smaller number of patients, and reduced costs. A surrogate endpoint can also be of interest if it can be measured in a simpler, more reproducible and accurate or less invasive way for patients compared to the reference endpoint.

Before using a surrogate endpoint as a substitute for the reference endpoint, it must be validated. The clinical conclusions must be consistent between the two criteria. In this perspective, several methodological works have been developed in the last twenty years within the meta-analytical approach, which is based on the correlation between the surrogate endpoint and the final endpoint [1–3]. The meta-analytical approach considers two levels of validation for a surrogate endpoint: the individual level and the trial level. Validity at the individual level means that, for each patient, the surrogate endpoint is correlated to the final endpoint. Individual surrogacy is straigthforward to verify and requires only data from a single trial. Validity at the trial level means that, for each trial, the treatment effect on the surrogate endpoint correlates with the treatment effect on the final endpoint. At the individual level, the Kendall's τ can be used as a measure of validation. At the trial level, for time-to-event endpoints, Burzykowski et al. [4] developed a two-stage copula-based approach, with an adjusted trial-level surrogacy measure R^2 calculated in the second stage, which takes

* Corresponding author. Postal address: Service de Biostatistique et d'épidémiologie, Gustave Roussy, B2M, RdC. 114, rue Edouard Vaillant, 94805, Villejuif, France.
E-mail address: stefan.michiels@gustaveroussy.fr (S. Michiels).

https://doi.org/10.1016/j.conctc.2019.100402
Received 18 January 2019; Received in revised form 13 June 2019; Accepted 24 June 2019
Available online 05 July 2019

estimation error of the treatment effects at the first stage into account. This approach can also provide, for a new trial, a prediction of the effect of the treatment on the final endpoint on the basis of the intermediate endpoint. However, the estimation of the coefficient of determination R^2 is not robust with respect to the estimation error of the trial-specific treatment effects, is sensitive to trials with extreme treatments effects, has some convergence issues and often comes with a large standard error [5,6].

Recently, Gabriel et al. [7] have proposed to use an absolute prediction error instead of R^2 to compare trial-level surrogates in a binary setting and formalized the use of cross-validation in this case. In the time-to-event context, Baker and Kramer [8] used the difference in survival at a given time as outcome and proposed a prediction error estimate using leave-one-out cross-validation.

In this paper, we propose the use of an absolute prediction error as an alternative measure to the trial level R^2 for the case of two failure-time endpoints. The new surrogacy measures proposed here are based on the distance between the effect of the treatment observed on the final endpoint in each trial and its predicted value obtained from the effect observed on the surrogate endpoint, based on the prediction model fitted using data from the other trials. We present several weights for distance measurement based on the standard error of the observed treatment effect and the standard error of its predicted value. A leave-one-trial-out cross-validation scheme is used to estimate the prediction error in the context of an internal validation. The potential benefits of this type of measure are twofold: first, it avoids computing the R^2, for which convergence problems are frequently encountered; second, it uses metrics directly measured on the scale of the treatment effect and quantifies the error of its prediction.

2. Methods

The meta-analytic approach allows estimating the relation between the effect of the treatment on the surrogate endpoint and on the final endpoint. A valid surrogate endpoint must predict the final effect precisely, and the difference between the predicted and the observed final effects quantifies the predictive value of the surrogate.

2.1. Two-step surrogacy model

The proportional hazard model [9] is often used in the analysis of survival data to evaluate the effect of various covariates on the event of interest. Let S_{ij} and T_{ij} be the surrogate and final endpoint time variables, respectively, for patient $j = 1, ...,n_i$ in trial $i = 1, ...,N$. Burzykowski et al. [4] proposed the following two step model to validate S as surrogate of T:

$$\begin{cases} h_{Sij}(s; Z_{ij}) = h_{Si}(s)\exp\{\alpha_i Z_{ij}\}, \\ h_{Tij}(t; Z_{ij}) = h_{Ti}(t)\exp\{\beta_i Z_{ij}\}, \\ S_{ij}(s, t; Z_{ij}) = C_\delta(S_{Sij}(s; Z_{ij}), S_{Tij}(t; Z_{ij})), \end{cases} \quad (1)$$

with trial-specific baseline hazards $h_{Si}(s)$ and $h_{Ti}(s)$, treatment effects α_i and β_i and the treatment indicator Z (0 for the control arm, 1 for the experimental arm). The copula function $C_\delta(S_{Sij}(s), S_{Tij}(t))$ [10] is used to model the individual dependence, with $S_{Sij}(s)$ and $S_{Tij}(t)$ the survival functions of S_{ij} and T_{ij}. In our work, we choose to use the Clayton Copula [11].

$$C_\delta(u, v) = (u^{-\delta} + v^{-\delta} - 1)^{-1/\delta}, \quad (2)$$

with $\delta > 0$ and Kendall's $\tau = \delta/(\delta + 2)$ [12]. The marginal survival functions are based on the proportional hazards assumption for each of the two endpoints, with a Weibull parametric baseline hazard function.

The estimation of this model is carried out in two steps and consists in: estimating the treatment effects on both endpoints for each trial using the copula model (1) and thus calculating the Kendall's τ; then, performing a linear regression that possibly takes into account the estimation errors at the first step, to estimate the relationship between the treatment effect on the surrogate endpoint and the treatment effect on the final endpoint. Indeed, the estimates of the treatment effects obtained in the first stage are assumed to follow the mixed model

$$\begin{pmatrix} \hat{\alpha}_i \\ \hat{\beta}_i \end{pmatrix} = \begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} + \begin{pmatrix} \varepsilon_{\alpha_i} \\ \varepsilon_{\beta_i} \end{pmatrix},$$

with true treatment effects

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \begin{pmatrix} d_\alpha^2 & d_\alpha d_\beta \rho_{\text{trial}} \\ d_\alpha d_\beta \rho_{\text{trial}} & d_\beta^2 \end{pmatrix} \right),$$

and estimation errors

$$\begin{pmatrix} \varepsilon_{ai} \\ \varepsilon_{bi} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{\Omega}_i = \begin{pmatrix} \omega_{\alpha_i}^2 & \omega_{\alpha_i}\omega_{\beta_i}\rho_{\varepsilon_i} \\ \omega_{\alpha_i}\omega_{\beta_i}\rho_{\varepsilon_i} & \omega_{\beta_i}^2 \end{pmatrix} \right).$$

The measure of the surrogacy at the trial level is then the determination coefficient $R^2_{trial} = \rho^2_{trial}$, obtained by fixing the estimation errors Ω_i at their values estimated in the first step [13]. In real-life examples often the algorithm used to estimate the coefficients of the model does not converge and an "**unadjusted**" model is used in which Ω_i is fixed to zero. We denote the model which fully accounts for measurement error as "**adjusted**".

2.2. Cross-validation

In the meta-analytic context, we have previously proposed to compare the observed treatment effect on the final endpoint with an independent prediction through leave-one-(trial)-out cross-validation (LOOCV) [14]. LOOCV consists in excluding one trial at a time, estimating the two step model (1) from the remaining trials, and then applying this prediction model to the excluded trial. This last step consists in plugging the observed treatment effect on the surrogate endpoint in the left-out trial into the prediction model in order to obtain the predicted treatment effect on the final endpoint for the left-out trial.

2.3. Motivating example

Our motivating example was the advanced GASTRIC meta-analysis [15], which included individual data of 4069 patients with advanced/recurrent gastric cancer from 20 randomized trials of chemotherapy. The main endpoint was OS and the candidate surrogate was progression-free survival (PFS). In previous works [6,16], the individual-level association between PFS and OS, as measured by the rank correlation coefficient, was given by a value of Kendall's τ of 0.85 (95% confidence interval [CI]: 0.85–0.85). The association at the trial level between the treatment effects on OS $\log(HR_{OS})$ and on PFS $\log(HR_{PFS})$ was moderate, with a coefficient of determination, R^2, adjusted for the estimation errors, of 0.61 ([CI]: 0.04–1.00). As in many applications, the width of the confidence interval suggests that there was a large uncertainty in the prediction model and that conclusions should be drawn with great caution. Fig. 1 shows, for this example, the cross-validated prediction intervals for the treatment effect of each trial in the advanced GASTRIC meta-analysis, together with the observed effects.

3. Proposed distance measures

Fig. 1, which was obtained by LOOCV, prompted us to propose new measures based on the distance between the observed value $\hat{\beta} = \log(HR_{observed})$ and the predicted value $\tilde{\beta} = \log(HR_{predicted})$ to validate a surrogate endpoint, in the context of internal or external validation. In particular, we assume that the prediction error in terms of the distance between the predicted and observed values is the most important measure. At the same time, trials with narrower prediction

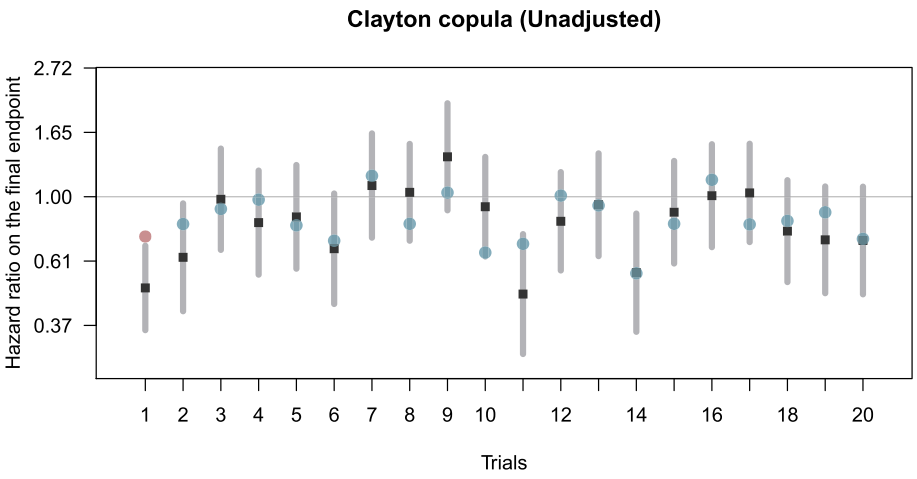

Fig. 1. Leave-one-(trial)-out cross-validation results for the advanced GASTRIC meta-analysis. The black squares and the vertical gray lines are the predicted values of the treatment effect on overall survival (OS), with the 95% prediction intervals (PI). Dots are the observed treatment effects on OS (green = within the PI, magenta = out of the PI). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

intervals and more precise $\hat{\beta}$ should have a greater weight than those with less precise estimates and predictions. In the next section, we propose novel measures based on different weighted prediction errors in which the weights are inspired from this figure.

3.1. Weighted prediction errors

Let $\hat{\alpha}_{(i)}$ and $\hat{\beta}_{(i)}$ be the treatment effects on the surrogate and final endpoints estimated separately in the two marginal proportional hazard models for trial *i*. Let $\tilde{\beta}_{(i)}$ be the prediction of $\hat{\beta}_{(i)}$ obtained by injecting the observed effect on the surrogate endpoint $\hat{\alpha}_{(i)}$ in the prediction model estimated on the $N-1$ remaining trials. The prediction of the effect on the final endpoint can be obtained as

$$\tilde{\beta}_{(i)} = \tilde{\gamma}_0 + \tilde{\gamma}_1 \hat{\alpha}_{(i)}, \tag{3}$$

with

$$\tilde{\gamma}_1 = \frac{(N-1)S_{ab} - \sum_{i' \neq i} \hat{\omega}_{\alpha i'} \hat{\omega}_{\beta i'} \hat{\rho}_{\text{trial}}}{(N-1)S_{aa} - \sum_{i' \neq i} \hat{\omega}^2_{\alpha i'}}, \tag{4}$$

and

$$\tilde{\gamma}_0 = \frac{\sum_{i' \neq i} (\hat{\beta}_{i'} - \tilde{\gamma}_1 \hat{\alpha}_{i'})}{(N-1)}, \tag{5}$$

where S_{aa} and S_{ab} are the sample variance and covariance of the estimates $\hat{\alpha}_{(i)}$ and $\hat{\beta}_{(i)}$ in the $N-1$ trials used for estimating the prediction model [17].

Let $diff_i$ denote the absolute difference between the observed effect on the final endpoint $\hat{\beta}_{(i)}$ in the trial i and its prediction $\tilde{\beta}_{(i)}$:

$$diff_i = |\hat{\beta}_{(i)} - \tilde{\beta}_{(i)}|. \tag{6}$$

We defined different weighted averages of the (cross-validated) absolute distances to account for the accuracy of the predicted and of the observed treatment effects, which can be different across trials:

$$\overline{diff}(w_i) = \frac{\sum_{i=1}^{N} diff_i \times w_i}{\sum_{i=1}^{N} w_i}, \tag{7}$$

with w_i the trial-specific weight. We considered the following possible weights:

- 1, to calculate the mean absolute distance;
- n_i, to give more importance to bigger trials;
- $1/SE^2(\hat{\beta}_{(i)})$, the inverse of the squared standard error of $\hat{\beta}_{(i)}$, to give more weight to the trials with more precise estimation of the observed treatment effect;
- $1/SE^2(\tilde{\beta}_{(i)})$, the inverse of the squared standard error of $\tilde{\beta}_{(i)}$, to give more weight to the trials with more precise predictions of the treatment effect;
- $1/(SE^2(\hat{\beta}_{(i)}) + SE^2(\tilde{\beta}_{(i)}))$, the inverse of the sum of the squared standard errors of the distance $diff_i$, to take into account the precision of both the prediction and the observed value;
- $1/(SE(\tilde{\beta}_{(i)})/SE(\hat{\beta}_{(i)}))^2$, this weighting represents the precision of the prediction with respect to the estimation precision [8].

For each estimator, the standard error was defined as

$$SE(\overline{diff}(w_i)) = \frac{\sum_{i=1}^{N} (diff_i - \overline{diff}(w_i))^2 \times w_i}{\sum_{i=1}^{N} w_i} \tag{8}$$

4. Simulation study

We conducted a simulation study under different scenarios to evaluate the operational characteristics of the distance measures presented in Section 3. All the methods for model fitting and data generation are implemented in the R package surrosurv, publicly available from the CRAN [18].

4.1. Simulation of data

The individual data of the surrogate and final endpoints were generated from exponential distributions with baseline rates fixed to 4/log(2) and 8/log(2), in order to obtain median survival times of 4 years and 8 years respectively. We generated treatment effects with mean $\alpha = \beta = \log(0.75)$, variance equal to $d_\alpha^2 = d_\beta^2 = 0.1$ and various correlations ρ_{trial} depending on the scenario (see below). Independent administrative censoring at 15 years was added and data was generated from a mixed proportional hazard model and from a Clayton copula model as in [6].

4.2. Simulation scenarios

In our simulations, we varied the following parameters: R^2, the trial-level coefficient of determination between the effects of the treatment; N, the number of trials; n_i, the number of patients per trial; and Kendall's τ, the dependence at the individual level between the two criteria.

We generated databases, each containing a meta-analysis of N= 10, 20, or 40 randomized trials of average size n= 400, 200, or 100. The actual size of each trial n_i was randomly generated from a uniform distribution between $0.5n$ and . These various choices allowed us to study the possible effect on the distance measures of the number of trials included in a meta-analysis and their size.

Four scenarios were considered in terms of dependence at the

Table 1
Simulation scenarios. N and n are the number and average size of the trials, respectively.

Scenario	1	2	3	4
R^2	0.2	0.2	0.8	0.8
Kendall's τ	0.4	0.6	0.4	0.6
N (n)	10 (400)	10 (400)	10 (400)	10 (400)
	20 (200)	20 (200)	20 (200)	20 (200)
	40 (100)	40 (100)	40 (100)	40 (100)

individual level and at the trial level (see Table 1): a moderate ($\tau = 0.4$) and a high value ($\tau = 0.6$) of individual dependence and a low ($R^2 = 0.2$) and a high value ($R^2 = 0.8$) at the trial level. Thus we took two extreme settings with a poor trial-level surrogate and a very good trial level-surrogate.

We generated 500 databases per scenario and we computed the proposed mean distances using the two-step model with and without adjustment for estimation error in the second step. To summarize the results, we reported the mean distances, the empirical standard error (ESE), and the average standard error (ASE) across the 500 distances obtained for each scenario.

4.2.1. Results

Table 2 summarizes the results for the adjusted Clayton copula models in all the trials, where the term adjusted is intended for accounting for estimation error in the first-step copula model (see end of Section 2.1). Fig. 2 illustrates the empirical distribution of the mean distances according to the correlation at the trial level (R^2) and the correlation at the individual level (Kendall's τ).

Comparing scenarios 1 vs. 3 or scenarios 2 vs. 4 ($R^2 = 0.2$ vs. $R^2 = 0.8$) shows that the prediction was more accurate (i.e. the mean distances were smaller) and precise (i.e. the standard errors were lower) with a high correlation at the trial level ($R^2 = 0.8$). Furthermore, the comparison of scenarios 1 vs. 2 or scenarios 3 vs. 4 ($\tau = 0.4$ vs. $\tau = 0.6$) indicates that the distances were largely independent of the correlation at the individual level. In addition, the comparison of the three lines of plots does not suggest any impact of the number and size of trials on the distance measures.

For a high correlation at the trial level $R^2 = 0.8$ (scenarios 3 and 4) the mean distances were around 0.2 and for a low correlation at trial level $R^2 = 0.2$ (scenarios 1 and 3) the mean distances were around 0.3. These values mean that if, for instance, the estimated $\hat{HR}$ is 0.75 and the prediction error is $\pm$ 0. 2, the average prediction is $0.75\exp(-0.2) = 0.61$ (equivalent to $\tilde{\beta} = \log(0.75) - 0.2 = \log(0.57)$) or $0.75\exp(0.2) = 0.92$ (equivalent to $\tilde{\beta} = \log(0.75) + 0.2 = \log(0.92)$); on the other hand, if the estimated $\hat{HR}$ is 0.75 and the prediction error is $\pm$ 0. 3, the average prediction is 0.56 or 1.01.

The ASE was often higher (in general two-fold higher) than the ESE, suggesting that the standard error of these distances was overestimated, which means that the estimated confidence intervals are expected to be too wide.

The mean distance weighted by the sum of the standard deviations $\overline{diff}(1/(SE^2(\hat{\beta}_{(i)}) + SE^2(\tilde{\beta}_{(i)}))$ showed in general the highest difference of all distances between $R^2 = 0.2$ and $R^2 = 0.8$, especially for $N = 10$ and $n = 400$ (equal to 0.137 for $\tau = 0.6$). Nevertheless, the results are not very different across the weighting schemes.

The results obtained using unadjusted copula models remained barely unchanged (Supplementary Table A1 and Figure A1). The mean distances and the standard errors were generally smaller than for the adjusted copula models.

5. Application

Table 3 reports the distance measures estimated on individual data from the advanced GASTRIC meta-analysis presented in Section 2.3. These distances were all around 0.2. Although these distances are not intended to be a means to recalculate R^2, comparing these results with a similar simulated scenario ($\tau = 0.6$, $N = 20$ and $n = 200$) shows that the mean distances fall between the two scenarios $R^2 = 0.2$ and $R^2 = 0.8$. This suggests a moderate correlation at the trial level in the advanced GASTRIC meta-analysis data, which is consistent with the previously published estimate of $R^2 = 0.6$ [6] but which came with a confidence interval covering almost the entire range from 0 to 1. In order illustrate the magnitude of a prediction error of 0.2 on the treatment effect scale, for an estimated $\hat{HR} = 0.70$ and a mean absolute distance $\overline{diff}(w_i) = 0.2$, the predicted $\widetilde{HR}$ will be on average about $0.57 = 0.7\exp(-0.2)$ (equivalent to a) in case of overprediction, or $0.85 = 0.7\exp(0.2)$ (equivalent to a $\tilde{\beta} = \log(0.85) = \log(0.7) + 0.2$) for underprediction of the treatment effect.

6. Discussion

The meta-analytic approach to surrogate endpoint validation is intended for estimating a determination coefficient at the trial level. Nevertheless, it also provides the researcher with a prediction model for the treatment effect on the final endpoint of new trials for which the effect on the surrogate endpoint has been estimated. In the context of leave-one-(trial)-out cross-validation, the estimated and the (independently) predicted effects can be compared for each trial.

In previous works [14,19], we reported the rate of trials with the observed treatment effect falling within its prediction interval as a measure of goodness-of-prediction. It can be noted that the simple fact that the observed value falls in the prediction interval or not is important, but it does not summarize all the information concerning the coherence between the prediction and the final estimate. As mentioned in the introduction, the standard adjusted copula approach does not converge in many applications and, when it does converge, confidence intervals of the estimated R^2 are often extremely wide, which prompted us to propose alternative measures based on an absolute prediction error.

The distance measures presented in this article allow measuring the correlation strength between failure-time endpoints in terms of the prediction error, i.e. the mean difference between the observed effect and the predicted effect of treatment on the final endpoint.

The results of the simulations show that the proposed distances are rather independent of the correlation at the individual level (Kendall's τ) across scenarios (number of trials $N = 10, 20, 40$ and size of trials $n = 400, 200, 100$), which is a desirable property. They are lower and more accurate for a high correlation at the trial level ($R^2 = 0.8$) than with a low correlation ($R^2 = 0.2$). Overall, there is not much difference between the results of the adjusted and unadjusted copula models.

The weighting of distances with $1/SE(\hat{\beta}_{(i)})$ gives more importance to big trials with more precise estimations of the treatement effect. Weighting with $1/SE(\tilde{\beta}_{(i)})$ gives more weight to trials with more precise prediction, which can mean those trials with an effect on surrogate nearer to the mean across trials but can also mean more weight on small trials for which the prediction model is estimated by leaving out fewer patients. Based on the results of the simulation study, there is no obvious difference between the weighting strategies, but the inverse of the sum of the standard squared errors $1/(SE^2(\tilde{\beta}_{(i)}) + SE^2(\hat{\beta}_{(i)}))$ seems the most discriminating between high and low trial-level correlation.

Our simulation study has some limitation. We studied two main scenarios in the simulation study: a small number of large trials and a large number of small trials. We did not study a scenario with small number (≤ 10) of small trials (≤ 100 patients) because it would be very uninformative and could be scarcely useful for the evaluation of surrogacy, we also did not consider meta-analyses with a large number (≥ 40) of big trials (≥ 400 patients) because simulations would be too computationally intensive. Also, we did not vary the variance of treatments effects as we did not expect a major impact of this parameter on the findings.

Table 2
Results with data simulated with 500 repetitions for the adjusted copula models. N and n are the number and average size of the trials, respectively. $\overline{diff}\,(w_i)$ is the weighted mean absolute distance; ESE: empirical standard error and ASE: average standard error.

		scenario 1 ($\tau = 0.4$; $R^2 = 0.2$)			**scenario 2** ($\tau = 0.6$; $R^2 = 0.2$)			**scenario 3** ($\tau = 0.4$; $R^2 = 0.8$)			**scenario 4** ($\tau = 0.6$; $R^2 = 0.8$)		
	w_i	$\overline{diff}\,(w_i)$	ESE	(ASE)	$\overline{diff}\,(w_i)$	ESE	(ASE)	$\overline{diff}\,(w_i)$	ESE	(ASE)	$\overline{diff}\,(w_i)$	ESE	(ASE)
$N = 10$;	1	0.321	0.373	(0.302)	0.285	0.102	(0.220)	0.191	0.125	(0.163)	0.163	0.054	(0.125)
$n = 400$	n_i	0.326	0.481	(0.291)	0.283	0.101	(0.206)	0.187	0.123	(0.151)	0.161	0.054	(0.118)
	$1/SE^2(\hat{\beta}_{(i)})$	0.327	0.535	(0.290)	0.279	0.104	(0.204)	0.181	0.093	(0.145)	0.158	0.056	(0.115)
	$1/SE^2(\tilde{\beta}_{(i)})$	0.347	0.093	(0.218)	0.326	0.095	(0.206)	0.222	0.077	(0.132)	0.198	0.069	(0.118)
	$1/(SE^2(\hat{\beta}_{(i)}) + SE^2(\tilde{\beta}_{(i)}))$	0.324	0.087	(0.214)	0.306	0.089	(0.202)	0.191	0.061	(0.129)	0.169	0.051	(0.115)
	$1/(SE(\tilde{\beta}_{(i)})/SE(\hat{\beta}_{(i)}))^2$	0.351	0.107	(0.217)	0.330	0.103	(0.204)	0.230	0.085	(0.134)	0.202	0.074	(0.118)
$N = 20$;	1	0.290	0.052	(0.216)	0.282	0.056	(0.215)	0.202	0.038	(0.157)	0.175	0.034	(0.134)
$n = 200$	n_i	0.287	0.053	(0.208)	0.280	0.061	(0.208)	0.198	0.038	(0.149)	0.172	0.034	(0.128)
	$1/SE^2(\hat{\beta}_{(i)})$	0.280	0.052	(0.203)	0.273	0.065	(0.203)	0.191	0.037	(0.144)	0.166	0.032	(0.123)
	$1/SE^2(\tilde{\beta}_{(i)})$	0.314	0.051	(0.223)	0.302	0.053	(0.217)	0.233	0.049	(0.163)	0.195	0.038	(0.138)
	$1/(SE^2(\hat{\beta}_{(i)}) + SE^2(\tilde{\beta}_{(i)}))$	0.298	0.052	(0.214)	0.289	0.053	(0.209)	0.202	0.039	(0.149)	0.174	0.034	(0.128)
	$1/(SE(\tilde{\beta}_{(i)})/SE(\hat{\beta}_{(i)}))^2$	0.324	0.067	(0.226)	0.314	0.069	(0.223)	0.244	0.060	(0.170)	0.203	0.047	(0.143)
$N = 40$;	1	0.327	0.119	(0.268)	0.306	0.039	(0.238)	0.249	0.033	(0.199)	0.211	0.027	(0.168)
$n = 100$	n_i	0.320	0.117	(0.259)	0.300	0.039	(0.229)	0.240	0.033	(0.189)	0.205	0.027	(0.160)
	$1/SE^2(\hat{\beta}_{(i)})$	0.307	0.113	(0.247)	0.288	0.037	(0.218)	0.229	0.031	(0.178)	0.194	0.025	(0.149)
	$1/SE^2(\tilde{\beta}_{(i)})$	0.335	0.040	(0.254)	0.316	0.039	(0.240)	0.267	0.041	(0.202)	0.223	0.029	(0.170)
	$1/(SE^2(\hat{\beta}_{(i)}) + SE^2(\tilde{\beta}_{(i)}))$	0.316	0.040	(0.240)	0.300	0.039	(0.227)	0.235	0.033	(0.182)	0.198	0.026	(0.152)
	$1/(SE(\tilde{\beta}_{(i)})/SE(\hat{\beta}_{(i)}))^2$	0.365	0.066	(0.276)	0.344	0.065	(0.261)	0.295	0.059	(0.223)	0.247	0.048	(0.190)

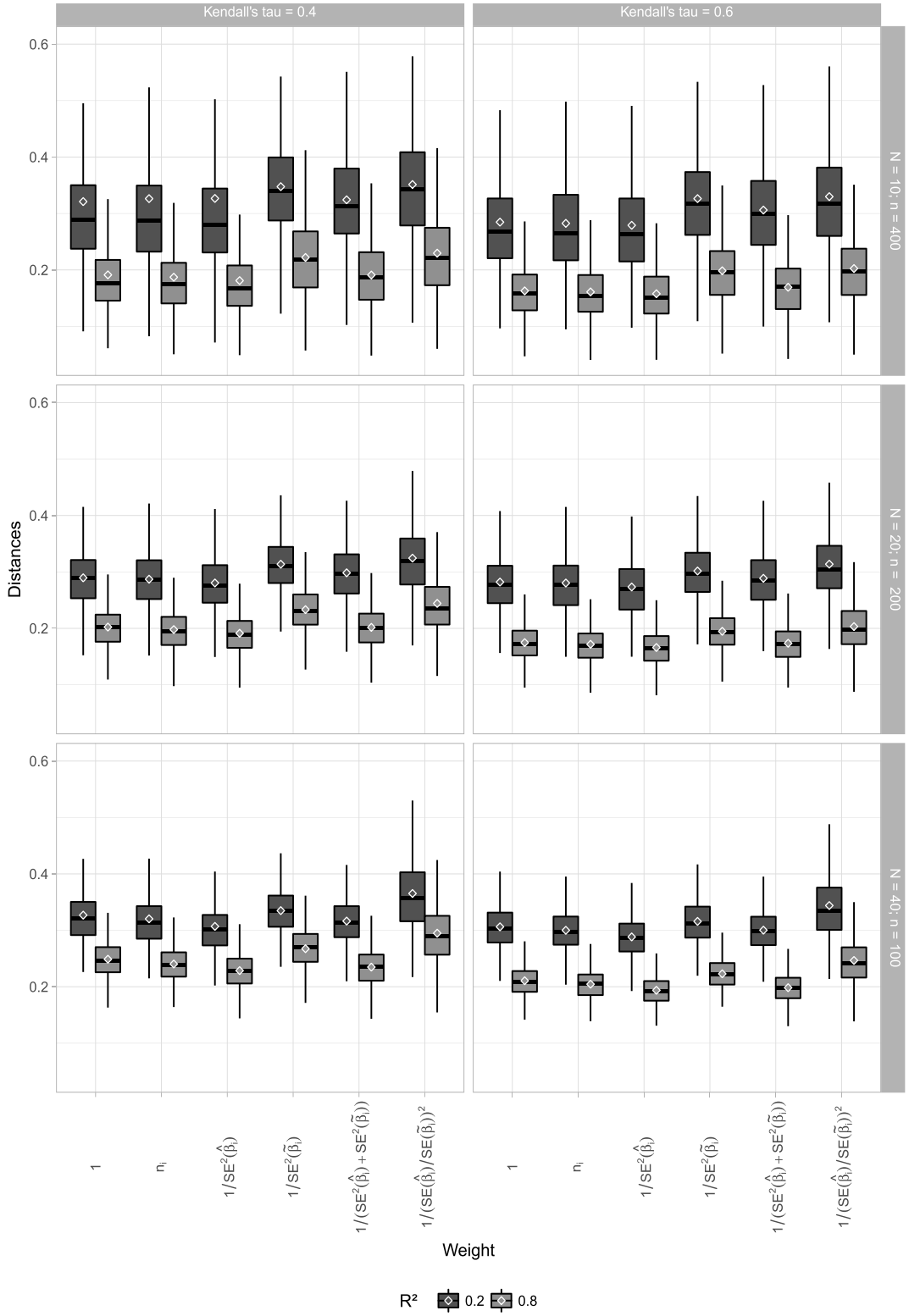

Figure 2. Boxplots of distances according to the different scenarios with the data generated with the adjusted Clayton copula. *N*: number of trials; *n*: average size of trials; The box bounds the interquartile range (IQR) and contains a horizonatal line corresponding to the median; outside of the box, the Tukey-style whiskers extend to a maximum of 1.5*IQR beyond the box. The white diamonds are the mean distances.

Note that, the precision of both the R^2 and the proposed prediction-error metrics is often low. This issue largely reflects the inherent scarcity of information due to the limited number of trials that are usually available in practice. We have illustrated the correlation between the estimation uncertainty (Empirical Standard Error (ESE)) of the R^2 and the proposed distances for the adjusted copula models across all scenarios (see figures in the supplementary appendix). We did not observe a clear correlation between the two uncertainty measures.

The sample size of a new trial (much larger or smaller than historical trials) may have an impact on the SE of the estimated α and thus

Table 3
Mean distances between observed and predicted treatment effect on overall survival for advanced GASTRIC meta-analysis. $\overline{diff}(w_i)$ is the weighted mean absolute distance and SE is the standard error.

w_i	$\overline{diff}(w_i)$	$SE(\overline{diff}(w_i))$
1	0.234	0.178
n_i	0.202	0.154
$1/SE^2(\hat{\beta}_{(i)})$	0.201	0.154
$1/SE^2(\tilde{\beta}_{(i)})$	0.213	0.159
$1/(SE^2(\hat{\beta}_{(i)}) + SE^2(\tilde{\beta}_{(i)}))$	0.201	0.155
$1/(SE(\tilde{\beta}_{(i)})/SE(\hat{\beta}_{(i)}))^2$	0.251	0.173

on the prediction precision. This potential issue is common to surrogacy measures such as the surrogate threshold effect [20] which is usually provided for a new trial of infinite size (e.g. as in Mauguen et al. [21]).

Despite consistent results across scenarios, recommending thresholds for final decision is as difficult as for the R^2. Nevertheless, the obtained mean distance can be interpreted more easily and mean predicted *HR*'s can be computed and compared to the mean estimated ones. In addition, our proposed alternative measures are more stable and more easily interpretable for a clinician than the coefficient of determination R^2. In our experience, clinicians find it difficult to interpret the amount of explained variation to validate a surrogate endpoint. Our proposed distance metrics are measured on the scale of the treatment effect and we propose to illustrate to clinicians the prediction error for an estimated hazard ratio in terms of average values for over- or underprediction of treatment effects.

The purpose of this article is to propose an alternative method to evaluate failure-time surrogate endpoints based on prediction error for time-to-event outcomes. Although this framework is the most common in oncology, these metrics can be applied more generally, whatever the type of the marginal distributions.

Declarations of interest

The authors report no conflicts of interest.

Acknowledgements

The authors gratefully acknowledge financial support from the Ligue Nationale Contre le Cancer. The authors are also grateful to the GASTRIC (Global Advanced/Adjuvant Stomach Tumor Research International Collaboration) Group for permission to use their data and the Institut National du Cancer (INCa).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.conctc.2019.100402.

References

[1] M. Buyse, G. Molenberghs, Criteria for the validation of surrogate endpoints in randomized experiments, Biometrics 54 (3) (1998) 1014–1029, https://doi.org/10.2307/2533853.
[2] M. Buyse, G. Molenberghs, T. Burzykowski, D. Renard, H. Geys, The validation of surrogate endpoints in meta-analyses of randomized experiments, Biostatistics 1 (1) (2000) 49–67.
[3] G.M. Blumenthal, S.W. Karuri, H. Zhang, L. Zhang, S. Khozin, D. Kazandjian, S. Tang, R. Sridhara, P. Keegan, R. Pazdur, Overall response rate, progression-free survival, and overall survival with targeted and standard therapies in advanced non–small-cell lung cancer: us food and drug administration trial-level and patient-level analyses, J. Clin. Oncol. 33 (9) (2015) 1008–1014.
[4] T. Burzykowski, G. Molenberghs, M. Buyse, H. Geys, D. Renard, Validation of surrogate end points in multiple randomized clinical trials with failure time end points, J. R. Stat. Soc.: Series C (Applied Statistics) 50 (4) (2001) 405–422, https://doi.org/10.1111/1467-9876.00244.
[5] L.A. Renfro, Q. Shi, D.J. Sargent, B.P. Carlin, Bayesian adjusted r2 for the meta-analytic evaluation of surrogate time-to-event endpoints in clinical trials, Stat. Med. 31 (8) (2012) 743–761.
[6] F. Rotolo, X. Paoletti, T. Burzykowski, M. Buyse, S. Michiels, A poisson approach to the validation of failure time surrogate endpoints in individual patient data meta-analyses, Stat. Methods Med. Res. 28 (2019) 170–183.
[7] E.E. Gabriel, M.J. Daniels, M.E. Halloran, Comparing biomarkers as trial level general surrogates, Biometrics 72 (4) (2016) 1046–1054.
[8] S.G. Baker, B.S. Kramer, Evaluating Surrogate Endpoints, Prognostic Markers, and Predictive Markers: Some Simple Themes, Clinical trials, London, England), 2014, https://doi.org/10.1177/1740774514557725.
[9] D.R. Cox, Regression models and life-tables, J. R. Stat. Soc. Ser. B 34 (1972) 187–220 URL http://www.jstor.org/stable/2985181.
[10] R.B. Nelsen, An Introduction to Copulas, Springer Science & Business Media, 2007.
[11] D. Clayton, A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, Biometrika 65 (1) (1978) 141–151, https://doi.org/10.1093/biomet/65.1.141.
[12] M.G. Kendall, A new measure of rank correlation, Biometrika 30 (1/2) (1938) 81–93 URL http://www.jstor.org/stable/2332226.
[13] H.C. van Houwelingen, L. Arends, T. Stijnen, Advanced methods in meta-analysis: multivariate approach and meta-regression, Stat. Med. 21 (2002) 589–624, https://doi.org/10.1002/sim.1040.
[14] S. Michiels, A. Le Maître, M. Buyse, T. Burzykowski, E. Maillard, J. Bogaerts, J.B. Vermorken, W. Budach, T.F. Pajak, K.K. Ang, J. Bourhis, J.-P. Pignon, Surrogate endpoints for overall survival in locally advanced head and neck cancer: meta-analyses of individual patient data, Lancet Oncol. 10 (4) (2009) 341–350, https://doi.org/10.1016/S1470-2045(09)70023-3.
[15] GASTRIC group, Role of chemotherapy for advanced/recurrent gastric cancer: an individual-patient-data meta-analysis, Eur. J. Cancer 49 (7) (2013) 1565–1577, https://doi.org/10.1016/j.ejca.2012.12.016.
[16] X. Paoletti, K. Oba, Y.-J. Bang, H. Bleiberg, N. Boku, O. Bouché, P. Catalano, N. Fuse, S. Michiels, M. Moehler, et al., Progression-free survival as a surrogate for overall survival in advanced/recurrent gastric cancer trials: a meta-analysis, J. Natl. Cancer Inst. 105 (21) (2013) 1667–1670.
[17] T. Burzykowski, J. Cortiñas Abrahantes, Validation in the case of two failure-time endpoints, in: T. Burzykowski, G. Molenberghs, M. Buyse (Eds.), The Evaluation of Surrogate Endpoints, Springer, New York, NY, 2005, pp. 163–194.
[18] F. Rotolo, X. Paoletti, S. Michiels, surrosurv: An R package for the evaluation of failure time surrogate endpoints in individual patient data meta-analyses of randomized clinical trials, Methods and Programs in Biomedicine 155 (2018) 189–198, https://doi.org/10.1016/j.cmpb.2017.12.005.
[19] F. Rotolo, J.-P. Pignon, J. Bourhis, S. Marguet, J. Leclercq, W. Tong Ng, J. Ma, A.T.C. Chan, P.-Y. Huang, G. Zhu, D.T.T. Chua, Y. Chen, H.-Q. Mai, D.L.W. Kwong, Y.L. Soong, J. Moon, Y. Tung, K.-H. Chi, G. Fountzilas, L. Zhang, E.P. Hui, A.W.M. Lee, P. Blanchard, S. Michiels, Surrogate end points for overall survival in loco-regionally advanced nasopharyngeal carcinoma: an individual patient data meta-analysis, J. Natl. Cancer Inst. 109 (4) (2017) djw239, https://doi.org/10.1093/jnci/djw239.
[20] T. Burzykowski, M. Buyse, Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation, Pharmaceut. Stat. 5 (3) (2006) 173–186, https://doi.org/10.1002/pst.207.
[21] A. Mauguen, J.-P. Pignon, S. Burdett, C. Domerg, D. Fisher, R. Paulus, S.J. Mandrekar, C.P. Belani, F.A. Shepherd, T. Eisen, H. Pang, L. Collette, W.T. Sause, S.E. Dahlberg, J. Crawford, M. O'Brien, S.E. Schild, M. Parmar, J.F. Tierney, C.L. Pechoux, S. Michiels, Surrogate endpoints for overall survival in chemotherapy and radiotherapy trials in operable and locally advanced lung cancer: a re-analysis of meta-analyses of individual patients' data, Lancet Oncol. 14 (7) (2013) 619–626, https://doi.org/10.1016/S1470-2045(13)70158-X.