Quality of Life Research (2019) 28:1315–1325
https://doi.org/10.1007/s11136-018-02097-2

Modeling strategies to improve parameter estimates in prognostic factors analyses with patient-reported outcomes in oncology

Francesco Cottone[1] · Nina Deliu[1] · Gary S. Collins[2] · Amelie Anota[3,4] · Franck Bonnetain[3,4] · Kristel Van Steen[5,6] · David Cella[7] · Fabio Efficace[1]

Accepted: 29 December 2018 / Published online: 18 January 2019
© Springer Nature Switzerland AG 2019

Abstract

Purpose The inclusion of patient-reported outcome (PRO) questionnaires in prognostic factor analyses in oncology has substantially increased in recent years. We performed a simulation study to compare the performances of four different modeling strategies in estimating the prognostic impact of multiple collinear scales from PRO questionnaires.

Methods We generated multiple scenarios describing survival data with different sample sizes, event rates and degrees of multicollinearity among five PRO scales. We used the Cox proportional hazards (PH) model to estimate the hazard ratios (HR) using automatic selection procedures, which were based on either the likelihood ratio-test (Cox-PV) or the Akaike Information Criterion (Cox-AIC). We also used Cox PH models which included all variables and were either penalized using the Ridge regression (Cox-R) or were estimated as usual (Cox-Full). For each scenario, we simulated 1000 independent datasets and compared the average outcomes of all methods.

Results The Cox-R showed similar or better performances with respect to the other methods, particularly in scenarios with medium–high multicollinearity ($\rho = 0.4$ to $\rho = 0.8$) and small sample sizes ($n = 100$). Overall, the Cox-PV and Cox-AIC performed worse, for example they did not select one or more prognostic collinear PRO scales in some scenarios. Compared with the Cox-Full, the Cox-R provided HR estimates with similar bias patterns but smaller root-mean-squared errors, particularly in higher multicollinearity scenarios.

Conclusions Our findings suggest that the Cox-R is the best approach when performing prognostic factor analyses with multiple and collinear PRO scales, particularly in situations of high multicollinearity, small sample sizes and low event rates.

Keywords Health-related quality of life · Multicollinearity · Patient-reported outcomes · Prognostic factor analysis · Ridge regression

Introduction

Prognostic models in cancer research have traditionally included clinical and laboratory tumor markers. However, over the last two decades, the inclusion of patient-reported

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11136-018-02097-2) contains supplementary material, which is available to authorized users.

✉ Francesco Cottone
f.cottone@gimema.it

[1] Data Center and Health Outcomes Research Unit, Italian Group for Adult Hematologic Diseases (GIMEMA), Rome, Italy

[2] Centre for Statistics in Medicine, NDORMS, University of Oxford, Oxford, UK

[3] Methodology and Quality of Life in Oncology Unit (INSERM UMR 1098), University Hospital of Besançon, Besançon, France

[4] French National Platform Quality of Life and Cancer, Besançon, France

[5] GIGA-R Medical Genomics Unit, University of Liège, Liège, Belgium

[6] Department of Human Genetics – Systems Medicine, University of Leuven, Leuven, Belgium

[7] Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

outcomes (PROs) in prognostic factor analyses has substantially increased [1, 2]. Several studies have found specific PRO domains, for example, global quality of life (QoL), physical functioning, or specific symptoms such as pain or fatigue, that independently predict overall survival (OS) beyond the well-known biomedical prognostic markers. Importantly, such evidence has been replicated across a wide range of cancer populations (including solid tumors and hematologic malignancies) and using a number of different PRO questionnaires [3–17].

However, the identification and evaluation of prognostic PRO domains entail some important methodological issues that must be carefully considered [18, 19]. Indeed, PRO data are often collected from a patient via standardized questionnaires, which typically include multiple scales (domains). For example, the European Organization for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire Core 30 (QLQ-C30) [20] generates fifteen different subscales including five different functional scales, three symptom scales, six single-item scales and an overall global health status/QoL scale. Although each of them measures a specific aspect of a patient's health status, typically they are mutually correlated to different extents and this might negatively influence the stability of the final multivariate prognostic model [19]. In addition, even with a correct model specification, the simultaneous inclusion of multiple collinear scales in a prognostic model might alter the direction/magnitude of their estimated impacts on the response variable [21–23], depending both on their number and the extent of multicollinearity.

Therefore, dealing with multiple collinear scales from a given questionnaire is one of the main challenges when analyzing the potential prognostic value of PROs. When designing studies to identify possible prognostic PRO domains, investigators face challenging decisions about which and how many scales to consider from a given PRO questionnaire to investigate their possible prognostic ability for the outcome of interest. Therefore, the choice is often between either improving effect estimates by excluding potential prognostic PRO domains or retaining all available PROs information although impairing the effect estimate.

A typical approach is to identify, in the design phase, a number of "primary" scales to be investigated, based both on previous clinical evidence (if available) and statistical considerations. However, the evidence and actual extent of multicollinearity among the selected scales would be evident only after data collection, in the analysis phase. Thus, although the inclusion of all selected primary scales could still impair the results, excluding some of them to lessen multicollinearity would not be feasible since all of them had been defined by design as potentially relevant prognostic factors. The problem is even exacerbated when no *a priori* information is available about the potential prognostic importance of PRO scales. In such a case, a very common approach to identify the prognostic PROs is the use of an automatic variable selection procedure, either forward selection, backward elimination or stepwise [18, 24, 25]. However, it is known that such approaches can lead to large variability of the factors included in the final model [26] with upward biased coefficient estimates and corresponding downward biased standard deviations and p values [27]. In addition, automatic variable selection procedures could lead to the exclusion of potentially relevant PROs from the final prognostic model [28].

Alternative approaches for handling multicollinearity aim at retaining all potentially prognostic PROs in the analysis. One of these approaches is respecifying the model by combining all collinear variables into a new summary measure [29]. However, this approach might not be desirable in the case of PRO-based prognostic factor analysis since the scales from PRO questionnaires measure distinct components of the same construct, with each potentially contributing some unique prognostic information. Another approach is to generate and analyze new mutually orthogonal factors from the original collinear variables, such as using principal component analysis (PCA) [29]. Unfortunately, such new outcomes would be difficult to interpret from a clinical perspective.

Therefore, methods including all potentially relevant but collinear variables in a prognostic model, which also lessen the impact of multicollinearity on parameter estimates, might be of value. This goal might be achieved using penalized methods such as the Ridge regression [30]. This approach has been developed in the linear regression framework, providing markedly improved model stability, coefficient estimates and corresponding standard errors [19, 26, 29]. However, to the best of our knowledge, no previous work has investigated the performance of the Ridge regression in the context of PRO-based prognostic factors analyses.

In this setting, this method could reduce the impact of harmful multicollinearity on the parameter estimates of a given set of potentially prognostic (collinear) PROs. However, we note that the Ridge regression cannot help in selecting which and/or how many scales should be investigated in a given research setting. Indeed, the selection of which PRO domains to investigate should be based on *a priori* clinical knowledge in the specific research context. The Ridge regression might help in achieving more reliable estimates than those obtained by simply including all available PRO scales into the model. Therefore, the Ridge regression could be particularly helpful when no *a priori* knowledge is available for estimating the parameters from a model with a large number of collinear PRO scales. Although other shrinkage methods such as the lasso regression or elastic-net would also allow for the selection among collinear variables, we will not consider such approaches in this work since we focused on the reliable estimation of parameters in prognostic models including a given set of collinear PRO scales.

In this analysis, we investigate and compare the performances of different modeling strategies (methods) based on the Cox PH regression for estimating the parameters of interest in prognostic factors analysis involving collinear PRO scales. Such methods consist of Cox PH models including all variables, which are either penalized using the Ridge regression or are estimated as usual, and the Cox PH models stemming from automatic stepwise selection procedures. When comparing such methods, we also assess the bias/variance trade-off. This is an inherent limitation of any model. i.e., a model can be refined to lessen the bias of estimates only at the cost of increasing the corresponding variance and vice versa. We base our analyses on simulated data reflecting realistic settings in PRO-based prognostic factor analysis and covering a wide range of possible scenarios according to the different degrees of multicollinearity, event rates and sample sizes.

Theoretical framework and statistical methods

All analyses in this work are based on the Cox proportional hazards model (Cox PH) with the hazard function of the i th individual defined as

$$h\left(t|\boldsymbol{x}_i\right) = h_0(t)\exp(\beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}) = h_0(t)\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)$$

where t is the time, $h_0(t)$ is the baseline hazard function, $\boldsymbol{x}_i = [x_{i,1}, \ldots, x_{i,p}]'$is the i th specific set of p observed covariates and $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]'$is the set of corresponding coefficients. Unless differently specified, the maximum partial likelihood estimator for $\boldsymbol{\beta}$ is defined as

$$\hat{\boldsymbol{\beta}} = argmax\, l(\boldsymbol{\beta}) \tag{1}$$

where $l(\boldsymbol{\beta})$ is the usual Cox log-partial likelihood function.

We considered the following methods.

Stepwise Cox PH based on the likelihood ratio-test (Cox-PV)

The automatic stepwise selection procedure is based on the likelihood ratio test, including all observed covariates as candidate variables with each entering and exiting the model at each iteration according to $\alpha = 0.05$. The final selected model included all variables such that no other additional candidate was significant or improved the model fit at the $\alpha = 0.05$ level.

Stepwise Cox PH based on the Akaike information criterion (Cox-AIC)

This employed the same model selection procedure as described above and is based on the Akaike Information Criterion (AIC) [31]. The final selected model was that with the lowest AIC that balanced both model fit and size [31].

Full Cox PH model (Cox-full)

This was a Cox PH model including all variables and collinear PRO scales.

Penalized Cox-PH Ridge (Cox-R)

A penalized Cox PH model estimates $\boldsymbol{\beta}$ using the Ridge regression or L^2-norm estimator defined as

$$\widehat{\boldsymbol{\beta}_{\boldsymbol{R}}} = argmax\left(l(\boldsymbol{\beta}) - \frac{1}{2}\lambda\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{\beta}\right),$$

where $\lambda > 0$ is the shrinkage parameter. The penalized regression is an alternative to traditional regression modeling where a constant is added to (1) [32]. The Ridge regression [30] uses an L^2-norm penalty (the sum of squares of regression coefficients multiplied by a penalty factor λ) [32].

Simulation settings

We used Monte Carlo (MC) simulations to generate data reflecting the characteristics of those real world studies investigating the prognostic significance of PROs in cancer clinical trials [1]. We defined four independent baseline demographic and clinical variables representing age (range of 18–90 years), sex, current comorbidity (yes vs. no), and a generic variable X_4 (continuous, range of 0–1). In addition, we considered five typically collinear self-reported scales (continuous, range of 0–100) from the EORTC QLQ-C30 questionnaire, i.e., Global health status/QoL (QL), physical functioning (PF), fatigue (FA), pain (PA) and appetite loss (AP). The EORTC QLQ-C30 is one of the most widely used PRO questionnaires in prognostic factor analyses in oncology [1]. Each patient's self-reported scale is scored according to the standard EORTC procedures [33] using ordered numerical responses to one or more items (Likert-type scale). For each scale, the actual distribution of the scores strongly depends on the corresponding number of items and responses. Thus, we chose to simulate the five scales described above since they are representative of all possible items-per-scale quantities in the QLQ-C30 questionnaire.

Higher scores indicate better outcomes for QL and PF scales and a higher symptom severity for FA, PA and AP scales.

Data-generating process

Age and X_4 were drawn from Beta distributions ($\alpha = 3, \beta = 2$), while binomial distributions were used for sex ($p = 0.4$) and current comorbidity ($p = 0.5$). We set the true hazard ratios (HR) as 1.03, 1.07, 1.16 and 1.00 (no effect), respectively for age, sex, current comorbidity and X_4 (see Table 1).

To resemble realistic patterns QoL, we generated individual EORTC QLQ-C30 scales using the partial credit model (PCM) [34] (see Appendix for details). For each subject and scale, we first simulated the individual latent trait from a normal distribution ($\mu = 0, \sigma = 0.25$). Then, we computed the corresponding multinomial probabilities of each item per scale according to prespecified category difficulty parameters, i.e., $\delta_7 = (-1, 0.6, -0.2, 0.2, 0.6, 1)$ for the QL items and $\delta_4 = (-0.7, 0, 0.7)$ for the remaining scales. These probabilities were used to generate the item responses from a multinomial distribution. Finally, we computed the score of each scale following the EORTC QLQ-C30 scoring manual [33]. We ensured, on average, low (0.2), medium (0.4) and high (0.8) degrees of multicollinearity between all QLQ-C30 scales.

We generated individual follow-up times from baseline to either the event of interest (death) or the exit from the study with a maximum follow-up period of 104 weeks (2 years). Individual death times t_d were independently generated by the inverse transform method from a Cox PH model with a Weibull-distributed baseline hazard [35, 36] as a function of all covariates but X_4. The HR of each scale reflected a ten-point increase on a 0–100 range. We set QL as having the largest impact on survival (HR = 0.85), followed by PA (HR = 1.13), FA (HR = 1.11), PF (HR = 0.89) and AP (HR = 1.07). We drew individual noninformative censoring times t_c from a Uniform distribution. The distribution parameters of death and censoring times were chosen through iterations to reach predefined event rates (30%, 50% and 70%). For each subject, we defined a status indicator d based on which event occurred first ($d = 1$ for death and 0 otherwise). All individuals with either $d = 0$ or $t_c > 104$ were right-censored.

Table 1 Simulated covariates with distribution, range and hazard ratio

Variable	Distribution	Range	HR
Age	Beta ($\alpha = 3, \beta = 2$)	[18, 90]	1.03
Sex	Binom ($p = 0.4$)	{0, 1}	1.07
Comorbidity	Binom ($p = 0.5$)	{0, 1}	1.16
X4	Beta ($\alpha = 3, \beta = 2$)	[0, 1]	1.00
Global health status[a]	Partial credit model	[0, 100]	0.85[c]
Physical functioning[a]	Partial credit model	[0, 100]	0.89[c]
Fatigue[b]	Partial credit model	[0, 100]	1.11[c]
Pain[b]	Partial credit model	[0, 100]	1.13[c]
Appetite loss[b]	Partial credit model	[0, 100]	1.07[c]

HR hazard ratio

[a] A higher score represents a better health status

[b] A higher score represents a higher symptom burden

[c] HRs coefficients of QLQ-C30 scales reflect a 10-point increase

Scenarios and performance evaluation criteria

We generated twenty-seven different scenarios according to the combinations of three degrees of multicollinearity between five PRO scales (low, 0.2; medium, 0.4; and high, 0.8), three event rates (30%, 50% and 70%) and three sample sizes (100, 300 and 500 patients). For each scenario, we simulated 1000 independent datasets and compared the average outcomes of all methods. When considering stepwise-based methods (I–II, Theoretical framework and statistical methods), we investigated the average outcomes of the most selected (out of 1000) multivariable models [37, 38]. When using the Ridge regression penalty, we performed a grid search to choose the optimal regularization parameter λ in the range of (0, 50] over 1000 independent replicates for each scenario. Then, we applied the penalty factor λ that minimized the average root-mean-squared error (RMSE) over the 1000 samples.

We applied the Cox PH-based methods I–IV as described above to each of the 1000 generated samples and compared their performances in terms of the estimated HRs and corresponding 95% confidence intervals (CIs), standardized biases (B_s, bias/standard error of estimates) and root-mean-squared errors (RMSEs). We used the Efron approximation to handle ties [39]. All analyses were performed using the R Statistical Software [40], v. 3.3.1.

Results

When sample size was $n \leq 300$ and multicollinearity was $\rho > 0.4$, the most selected models by both the Cox-PV and Cox-AIC did not include one or more collinear PROs (see Table 2). The number of excluded PROs increased as multicollinearity increased from 0.2 to 0.8 with some scenarios where the Cox-PV and Cox-AIC were not able to select any PROs (sample size $n = 100$, multicollinearity $\rho = 0.8$ and 50% or lower event rate).

When selected, the estimates of PRO parameters provided by either methods were more biased with respect to those from the Cox-Full and Cox-R, as shown in Fig. 1, although they did not have different variability with respect to the Cox-Full. In addition, the Cox-PV and

Table 2 Top selected models out of 1000 generated data sets for all scenarios

	Top selected model[a]											
	True HR	1.03	1.07	1.16	1.00	0.85	0.89	1.11	1.13	1.07		
	Event rate	Age	Sex	Com	X4	QL	PF	FA	PA	AP	Cox-PV (%)[b]	Cox-AIC (%)[c]
$N = 500$												
Low MCL ($\rho = 0.2$)	30%	√✗				√✗	√✗	√✗	√✗	√✗	44	37
	50%	√✗				√✗	√✗	√✗	√✗	√✗	58	37
	70%	√✗				√✗	√✗	√✗	√✗	√✗	60	33
Medium MCL ($\rho = 0.4$)	30%	√✗				√✗	√✗	√✗	√✗	√✗	30	31
	50%	√✗				√✗	√✗	√✗	√✗	√✗	53	36
	70%	√✗				√✗	√✗	√✗	√✗	√✗	58	33
High MCL ($\rho = 0.8$)	30%	√✗				√✗	√	√	√✗		9	8
	50%	√✗				√✗	√✗	√✗	√✗	√	12	14
	70%	√✗		√		√✗	√✗	√✗	√✗	√✗	15	18
$N = 300$												
Low MCL ($\rho = 0.2$)	30%	√✗				√✗	√✗	√✗	√✗	√✗	21	26
	50%	√✗				√✗	√✗	√✗	√✗	√✗	45	37
	70%	√✗				√✗	√✗	√✗	√✗	√✗	59	38
Medium MCL ($\rho = 0.4$)	30%	√✗				√✗	√✗	√✗	√✗	√	16	17
	50%	√✗				√✗	√✗	√✗	√✗	√✗	32	31
	70%	√✗				√✗	√✗	√✗	√✗	√✗	45	38
High MCL ($\rho = 0.8$)	30%	√✗				√✗	√	√	√✗		10	5
	50%	√✗				√✗	√	√	√✗		9	7
	70%	√✗				√✗	√✗	√✗	√✗	√	9	10
$N = 100$												
Low MCL ($\rho = 0.2$)	30%	√✗				√	√	√	√		14	5
	50%	√✗				√✗	√✗	√✗	√✗		6	8
	70%	√✗				√✗	√✗	√✗	√✗	√	13	14
Medium MCL ($\rho = 0.4$)	30%	√✗				√			√		19	4
	50%	√✗				√	√	√	√		8	7
	70%	√✗				√✗	√✗	√✗	√✗		9	11
High MCL ($\rho = 0.8$)	30%	√✗									37	9
	50%	√✗									24	4
	70%	√✗				√			√		18	5

MCL multicollinearity, *Com* comorbidity, *QL* Global Health Status/QoL, *PF* physical functioning, *FA* fatigue, *PA* pain, *AP* appetite loss, *Cox-PV* stepwise Cox PH with *p* value of 0.05 as entry/stay criterion, *Cox-AIC* stepwise Cox PH with AIC as entry/stay criterion

[a] Top selected model refers to the most selected model in the scenario

[b,c] This percentage refers to the number of times a given model was selected out of 1000 independent generated datasets

√The variable was selected in the model by the Cox-AIC stepwise procedure

✗The variable was selected in the model by the Cox-PV stepwise procedure

Cox-AIC were more sensitive to the event rate than other methods, in terms of bias. For the sample size of $n = 100$, such results were consistent across all scenarios and PROs and the Cox-R provided the best performance. Indeed, although the biases were similar between the Cox-Full and Cox-R, the latter produced less variable estimates than the Cox-Full with markedly better performances for higher levels of multicollinearity (Fig. 1). For example, with $\rho = 0.8$ and $n = 100$, the RMSE ranges were, respectively, [0.07, 0.14] for the Cox-AIC, [0.10, 0.25] for the Cox-Full and [0.07, 0.12] for the Cox-R. In this setting, no RMSE results were available for the Cox-PV since it did not select any PRO scales in the final model.

When sample size was $n = 300$ or $n = 500$, the stepwise methods selected all PROs in almost all scenarios. However, both failed when a higher event rate (ER) and /or larger sample size were not sufficient to compensate for the extent of multicollinearity. For example, 70% ER was necessary for

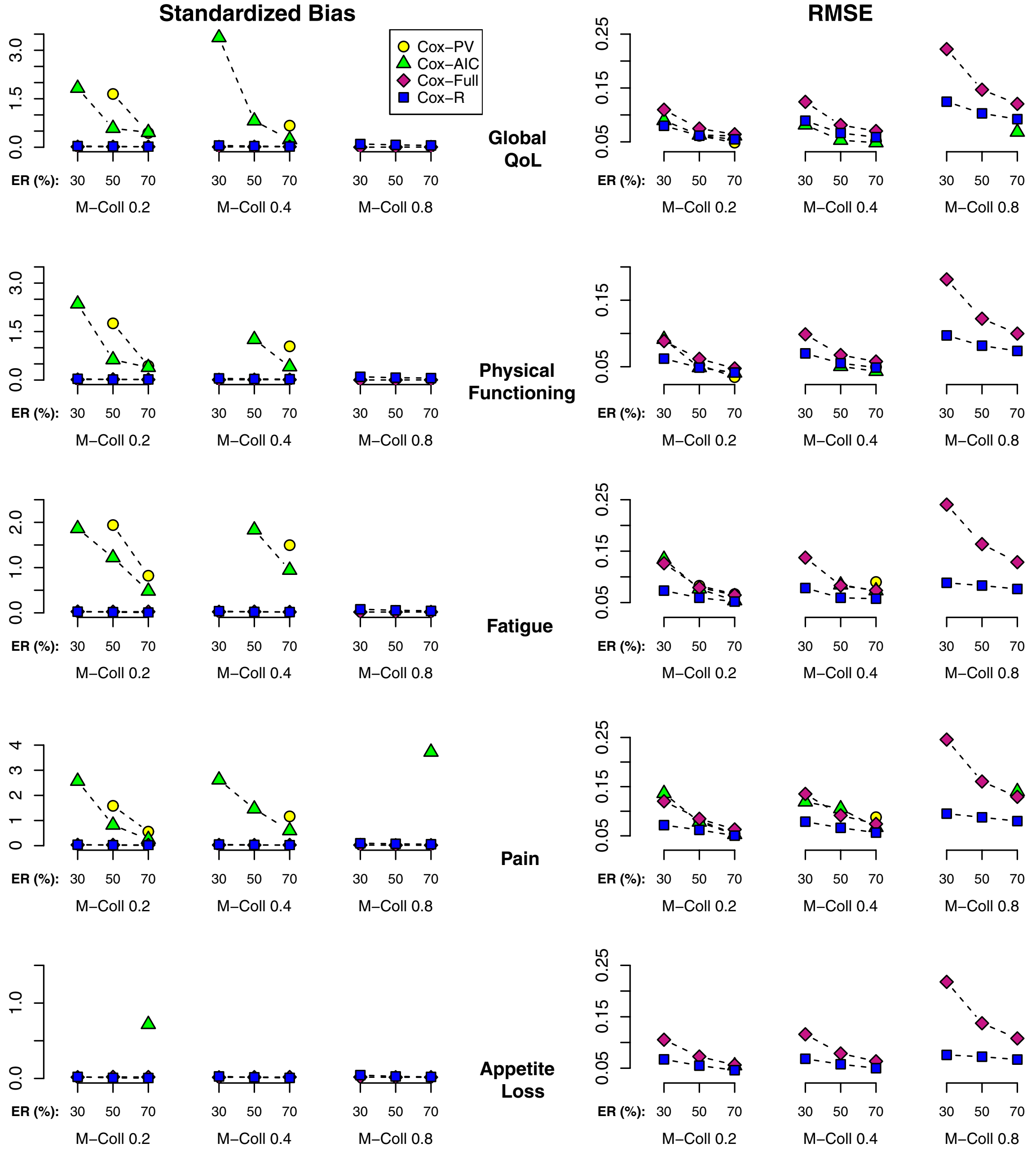

Fig. 1 Standardized bias and root-mean-squared error for all PROs and scenarios ($n = 100$). *RMSE* root-mean-squared error, *M-Coll* multicollinearity degree, *ER* event rate, *Cox-PV* stepwise Cox PH with 0.05 *p* value as entry/stay threshold, *Cox-AIC* stepwise Cox PH with AIC as entry/stay criterion, *Cox-Full* Cox PH with all variables, *Cox-R* Cox PH with Ridge penalty and all variables

the Cox-AIC to include all PROs with $\rho = 0.8$ and $n = 300$, whereas 50% ER was sufficient with $n = 500$ (Table 2). When $n = 300$, the biases of the estimates were roughly similar among all methods only when the multicollinearity was $\rho = 0.2$ (see Fig. 2, first left column), except for appetite loss (AP). For this scale, the Cox-PV and the Cox-AIC showed greater sensitivity to ER than the other methods in terms of bias. However, as the multicollinearity increased, such sensitivity also came up for other scales. In addition, for $n = 300$, the biases were similar across all scenarios for the Cox-R and Cox-Full, with both performing the same or better than the Cox-PV and the Cox-AIC. However, the differences in the RMSEs of the Cox-R and the Cox-Full were overall lower with respect to $n = 100$, as they remained markedly different only for a high degree of multicollinearity, i.e., $\rho = 0.8$ (Fig. 2).

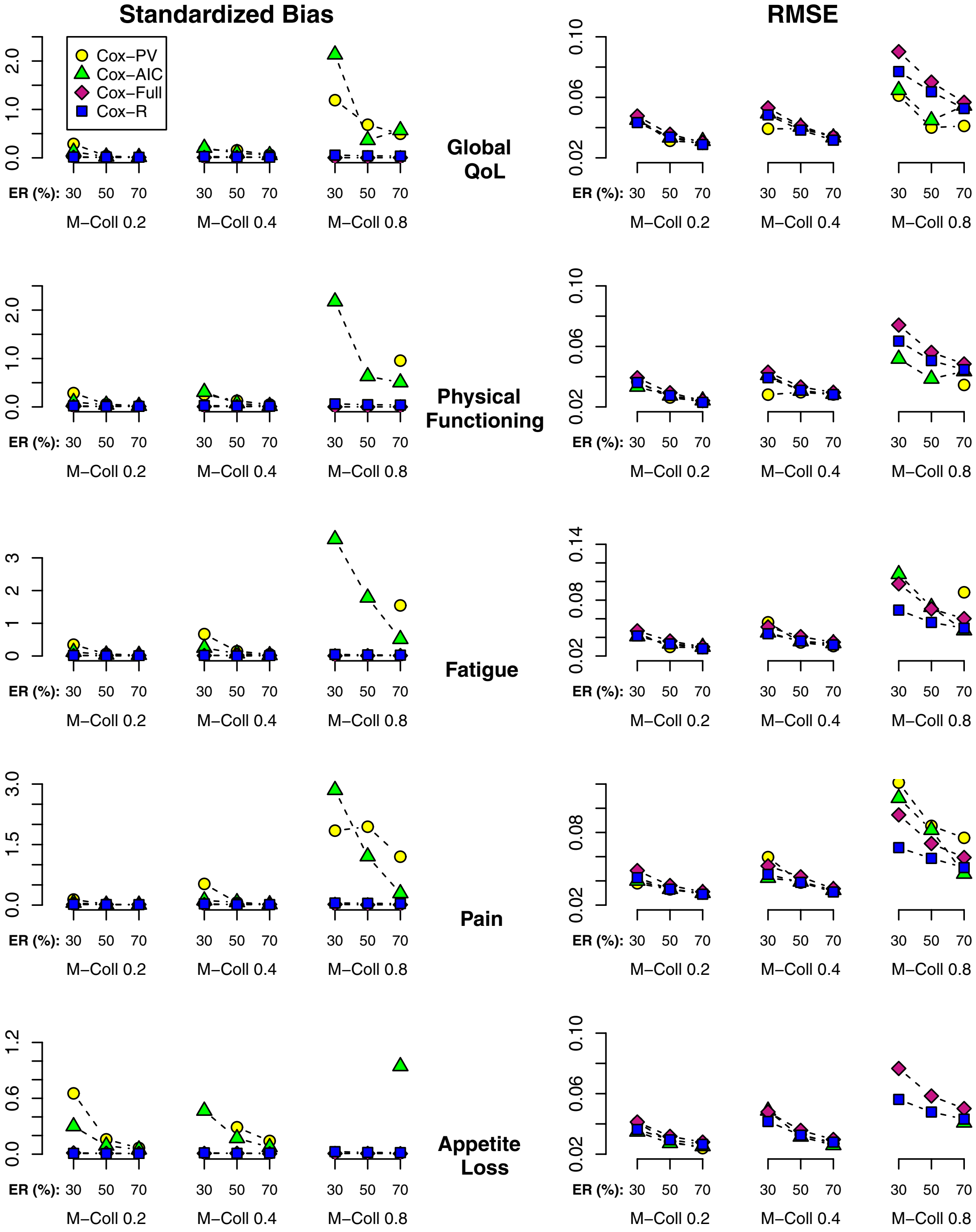

Fig. 2 Standardized bias and root-mean-squared error for all PROs and scenarios ($n=300$). *RMSE* root-mean-squared error, *M-Coll* multicollinearity degree, *ER* event rate, *Cox-PV* stepwise Cox PH with 0.05 *p* value as entry/stay threshold, *Cox-AIC* stepwise Cox PH with AIC as entry/stay criterion, *Cox-Full* Cox PH with all variables, *Cox-R* Cox PH with Ridge penalty and all variables

The stepwise methods also had RMSEs similar to those of the Cox-R and the Cox-Full except when $\rho=0.8$. For $n=500$, analogous results can be observed in terms of the bias among all methods (Fig. 3) for lower degrees of multicollinearity (except for AP), while the Cox-PV and the Cox-AIC provided larger biases when $\rho=0.8$. Furthermore, with respect to the RMSE, the same considerations can be drawn as for $n=300$. To illustrate, for high multicollinearity ($\rho=0.8$) and $n=500$, the RMSE values of the PROs ranges were, respectively, [0.03, 0.09] for the Cox-PV, [0.03, 0.08] for the Cox-AIC, [0.03, 0.08] for the

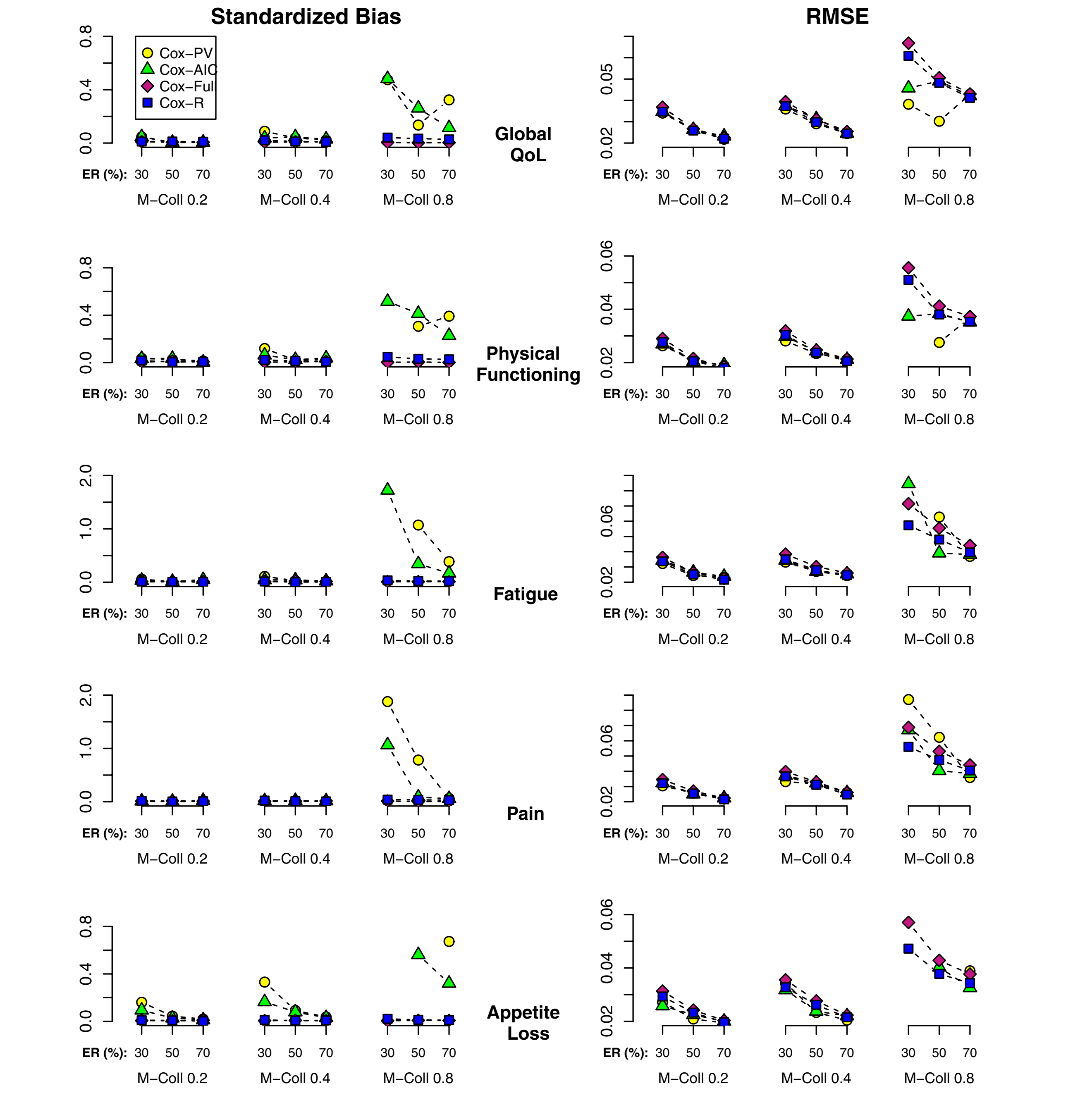

Fig. 3 Standardized bias and root-mean-squared error for all PROs and scenarios ($n=500$). *RMSE* root-mean-squared error, *M-Coll* multicollinearity degree, *ER* event rate, *Cox-PV* stepwise Cox PH with 0.05 *p*-value as entry/stay threshold, *Cox-AIC* stepwise Cox PH with AIC as entry/stay criterion, *Cox-Full* Cox PH with all variables, *Cox-R* Cox PH with Ridge penalty and all variables

Cox-Full and [0.02, 0.06] for the Cox-R. For each method, the patterns of the parameter estimates of noncollinear covariates did not show substantial improvements (results not shown). Overall, the Cox-R provided estimates similar to those of the Cox-Full, which were only slightly less biased, particularly in the presence of higher multicollinearity; in addition, the Cox-R produced narrower confidence intervals for the estimates than the Cox-Full (see Figs. 4–8 in the Appendix).

Discussion

When conducting PRO-based prognostic factor analyses, the researcher often has to make challenging decisions on the scales entering the analysis while limiting the risks associated with “harmful” multicollinearity as much as possible. Indeed, the simple inclusion of all collinear scales in a prognostic model would be unsatisfactory and provide overly biased estimates. However, excluding one or more collinear PRO scales from the prognostic model would waste potentially relevant prognostic information since each scale measures distinct facets of a subject's quality of life or symptomatic burden. Most of the strategies that were previously proposed to lessen the impact of harmful multicollinearity [29] are unsuitable for prognostic factor analysis, including PRO data. For example, a recombination of scales in mutually orthogonal factors using PCA would produce outcomes that would be difficult to interpret.

Indeed, given an initial set of collinear multiple PRO scales, the Ridge regression would be a potentially useful approach for balancing the completeness of PROs information and the accuracy of estimates. Previous studies dealing with multicollinearity in prognostic factor analysis showed better performances of the penalized regression estimators (shrinkage methods) compared to the maximum likelihood estimators [29, 41–44]. However, most of these studies focused on linear or logistic regression models and the generalization of their findings to time-to-event analysis is not straightforward. To the best of our knowledge, the only work evaluating shrinkage methods in the presence of multicollinearity in this framework [45] only considered the sample size of $n=50$, 4 to 6 covariates, and 3° of multicollinearity (0.2, 0.4 and 0.8) with no censored data, and reported results for just one covariate. In addition, there is no previous research on the application of the Ridge regression to prognostic factor analysis with collinear PROs.

In this work, we investigated and compared the performance of different modeling strategies based on the Cox PH regression in terms of the completeness of the retained PROs information in the prognostic model and the accuracy of parameter estimation. The first consisted of a Cox PH model including all variables and was penalized using the Ridge regression (Cox-R). The other (Cox-Full) was also a Cox PH model including all variables, but it was estimated as usual. The other two modeling strategies were automatic stepwise selection procedures, which were based on either the likelihood ratio test (Cox-PV) or the Akaike Information Criterion (Cox-AIC). We considered twenty-seven different scenarios that were designed according to different sample sizes, degrees of multicollinearity and event rates, and there were five PRO collinear variables and a right censoring scheme in the follow-up data. In general, the performances of all methods worsened with the increase of the multicollinearity and improved with higher event rates and larger sample sizes; however, this occurred to different extents depending on the combination of the multicollinearity, sample size and event rate.

For all scenarios, we found that the Cox-PV performed worse than the Cox-AIC in the model selection, which also provided less-biased coefficient estimates with overall lower RMSEs. However, both the Cox-PV and the Cox-AIC failed in selecting all significant collinear PRO scales for a small sample size ($n=100$) and a high degree of multicollinearity ($\rho=0.8$). In addition, we note that both methods were not able to select the full true model since they excluded some noncollinear variables across all scenarios.

The procedures including all variables outperformed those based on automatic selection. The Cox-R provided less variable HR estimates than the Cox-full, although at the cost of (slightly) higher biases. Such a bias/variance trade-off favored the Cox-R, particularly in those scenarios with higher degrees of multicollinearity ($\rho=0.4$ and $\rho=0.8$), smaller sample sizes ($n=100$ and $n=300$) and lower event rates (30% and 50%), where the Cox-R had lower RMSEs than the Cox-Full but similar bias patterns.

The first limitation of this study is that we did not investigate a real case study in which more than 10–20 variables might be highly correlated and time-varying variables might affect the outcome. In addition, we note that the Cox-R does not replace the careful selection of potential clinically relevant factors, including PROs. Such selection process should be performed previously to the Cox-R modeling, ideally including all available *a priori* knowledge about the topic of interest and/or based on statistical considerations. Indeed, the Cox-R can provide more reliable estimates than other methods based on a given set of variables. In this work, we did not consider other popular shrinkage methods, i.e., the lasso regression or elastic-net. However, we did not consider these approaches since we focused on a statistical approach allowing us to simultaneously retain all collinear PRO scales while reducing the effects of multicollinearity on the parameter estimates.

Our paper also has key strengths. The simulation-based approach allows for knowing the true model underlying the data, which can be used as a benchmark to compare the performances of the different modeling approaches. Previous studies investigating the effects of multicollinearity on model stability and estimation variability typically used real datasets and performance measures were obtained using resampling techniques such as bootstrapping [37, 46]. However, these procedures cannot rely on the knowledge of the real data-generating process, and so they can evaluate only how often a certain model is selected without providing any insight into its correctness. In addition, we investigated the performances of methods in several different scenarios, including those with less than ten events available per variable [47, 48]. Our findings might be useful in those PRO studies with insufficient sample sizes compared to the number of potentially meaningful collinear PRO scales that can be included in the model, particularly when no *a priori* knowledge is available about their clinical relevance. In addition, the application of all the methods described in this work is feasible using the most common statistical programs, such as R, SAS and SPSS.

This study further emphasizes the impact of harmful multicollinearity in PRO-based prognostic factor analysis, thereby encouraging researchers to carefully check for its presence before performing any modeling procedure. In addition, this study underlines the drawbacks of the model identification using automatic stepwise selection procedures and shows the benefits of using the Cox-R approach in PRO-based prognostic factor analysis. Overall, the Cox-R achieves more accurate estimates of the prognostic importance of multiple collinear PROs than other methods while retaining all of these in the multivariable model. This result was consistent throughout all the scenarios we explored, according to different issues such as the sample size, event rate and degree of multicollinearity. This suggested that the penalized Ridge regression approach was the most appropriate, particularly when covariates are affected by multicollinearity of at least $\rho = 0.4$ and the sample size is less than $n = 300$. Although this work was focused on overall survival, we are confident that our findings might be generalized to prognostic modeling for any outcome of interest in time to event analysis.

Author contributions FC, FE: Conception and design, FC, ND, FE: Statistical analyses, all authors: Interpretation of results, all authors: Manuscript writing.

Compliance with ethical standards

Conflict of interest No potential conflict of interest for this paper was reported by the authors.

References

1. Gotay, C. C., Kawamoto, C. T., Bottomley, A., & Efficace, F. (2008). The prognostic significance of patient-reported outcomes in cancer clinical trials. *Journal of Clinical Oncology, 26*(8), 1355–1363.
2. Secord, A. A., Coleman, R. L., Havrilesky, L. J., Abernethy, A. P., Samsa, G. P., & CELLA, D. (2015). Patient-reported outcomes as end points and outcome indicators in solid tumours. *Nature Reviews Clinical oncology, 12*(6), 358–370.
3. Efficace, F., Gaidano, G., Breccia, M., Voso, M. T., Cottone, F., Angelucci, E., et al. (2015). Prognostic value of self-reported fatigue on overall survival in patients with myelodysplastic syndromes: A multicentre, prospective, observational, cohort study. *The Lancet Oncology, 16*(15), 1506–1514.
4. Efficace, F., Bottomley, A., Coens, C., Van Steen, K., Conroy, T., Schoffski, P., et al. (2006). Does a patient's self-reported health-related quality of life predict survival beyond key biomedical data in advanced colorectal cancer? *European Journal of Cancer, 42*(1), 42–49.
5. Quinten, C., Martinelli, F., Coens, C., Sprangers, M. A., Ringash, J., Gotay, C., et al. (2014). A global analysis of multitrial data investigating quality of life and symptoms as prognostic factors for survival in different tumor sites. *Cancer, 120*(2), 302–311.
6. Efficace, F., Biganzoli, L., Piccart, M., Coens, C., Van Steen, K., Cufer, T., et al. (2004). Baseline health-related quality-of-life data as prognostic factors in a phase III multicentre study of women with metastatic breast cancer. *European Journal of Cancer, 40*(7), 1021–1030.
7. Maisey, N. R., Norman, A., Watson, M., Allen, M. J., Hill, M. E., & Cunningham, D. (2002). Baseline quality of life predicts survival in patients with advanced colorectal cancer. *European Journal of Cancer, 38*(10), 1351–1357.
8. Efficace, F., Innominato, P. F., Bjarnason, G., Coens, C., Humblet, Y., Tumolo, S., et al. (2008). Validation of patient's self-reported social functioning as an independent prognostic factor for survival in metastatic colorectal cancer patients: results of an international study by the Chronotherapy Group of the European Organisation for Research and Treatment of Cancer. *Journal of Clinical Oncology, 26*(12), 2020–2026.
9. Fang, F. M., Tsai, W. L., Chiu, H. C., Kuo, W. R., & Hsiung, C. Y. (2004). Quality of life as a survival predictor for esophageal squamous cell carcinoma treated with radiotherapy. *International Journal of Radiation Oncology, Biology, Physics, 58*(5), 1394–1404.
10. Chau, I., Norman, A. R., Cunningham, D., Waters, J. S., Oates, J., & Ross, P. J. (2004). Multivariate prognostic factor analysis in locally advanced and metastatic esophago-gastric cancer–pooled analysis from three multicenter, randomized, controlled trials using individual patient data. *Journal of Clinical Oncology, 22*(12), 2395–2403.
11. de Graeff, A., de Leeuw, J. R., Ros, W. J., Hordijk, G. J., Blijham, G. H., & Winnubst, J. A. (2001). Sociodemographic factors and quality of life as prognostic indicators in head and neck cancer. *European Journal of Cancer, 37*(3), 332–339.
12. Chiarion-Sileni, V., Del Bianco, P., De Salvo, G. L., Lo Re, G., Romanini, A., Labianca, R., et al. (2003). Quality of life evaluation in a randomised trial of chemotherapy versus bio-chemotherapy in advanced melanoma patients. *European Journal of Cancer, 39*(11), 1577–1585.
13. Dubois, D., Dhawan, R., van de Velde, H., Esseltine, D., Gupta, S., Viala, M., et al. (2006). Descriptive and prognostic value of patient-reported outcomes: the bortezomib experience in relapsed and refractory multiple myeloma. *Journal of Clinical Oncology, 24*(6), 976–982.
14. Eton, D. T., Fairclough, D. L., Cella, D., Yount, S. E., Bonomi, P., & Johnson, D. H. (2003). Early change in patient-reported health during lung cancer chemotherapy predicts clinical outcomes beyond those predicted by baseline report: Results from Eastern Cooperative Oncology Group Study 5592. *Journal of Clinical Oncology, 21*(8), 1536–1543.
15. Bottomley, A., Coens, C., Efficace, F., Gaafar, R., Manegold, C., Burgers, S., et al. (2007). Symptoms and patient-reported well-being: Do they predict survival in malignant pleural mesothelioma? A prognostic factor analysis of EORTC-NCIC 08983: Randomized phase III study of cisplatin with or without raltitrexed in patients with malignant pleural mesothelioma. *Journal of Clinical Oncology, 25*(36), 5770–5776.
16. Cella, D., Traina, S., Li, T., Johnson, K., Ho, K. F., Molina, A., et al. (2018). Relationship between patient-reported outcomes and clinical outcomes in metastatic castration-resistant prostate cancer: post hoc analysis of COU-AA-301 and COU-AA-302. *Annals of Oncology, 29*(2), 392–397.
17. Movsas, B., Hu, C., Sloan, J., Bradley, J., Komaki, R., Masters, G., et al. (2016). Quality of life analysis of a radiation dose-escalation study of patients with non-small-cell lung cancer: A secondary analysis of the radiation therapy oncology group 0617 randomized clinical trial. *JAMA Oncology, 2*(3), 359–367.

18. Mauer, M., Bottomley, A., Coens, C., & Gotay, C. (2008). Prognostic factor analysis of health-related quality of life data in cancer: A statistical methodological evaluation. *Expert Review of Pharmacoeconomics & Outcomes Research, 8*(2), 179–196.
19. Van Steen, K., Curran, D., Kramer, J., Molenberghs, G., Van Vreckem, A., Bottomley, A., et al. (2002). Multicollinearity in prognostic factor analyses using the EORTC QLQ-C30: identification and impact on model selection. *Statistics in Medicine, 21*(24), 3865–3884.
20. Aaronson, N. K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N. J., et al. (1993). The european organization for research and treatment of cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute, 85*(5), 365–376.
21. Cramer, E. M. (1985). Multicollinearity. In S. Kotz, N. L. Johnson & C. B. Read (Eds.), *Encyclopedia of statistical sciences*. (Vol. 2, pp. 639–643). New York, Wiley.
22. Slinker, B. K., & Glantz, S. A. (1985). Multiple regression for physiological data analysis: The problem of multicollinearity. *The American Journal of Physiology, 249*(1 Pt 2), R1–R12.
23. Sithisarankul, P., Weaver, V. M., Diener-West, M., & Strickland, P. T. (1997). Multicollinearity may lead to artificial interaction: An example from a cross sectional study of biomarkers. *The Southeast Asian Journal of Tropical Medicine and Public Health, 28*(2), 404–409.
24. Ediebah, D. E., Coens, C., Zikos, E., Quinten, C., Ringash, J., King, M. T., et al. (2014). Does change in health-related quality of life score predict survival? Analysis of EORTC 08975 lung cancer trial. *British Journal of Cancer, 110*(10), 2427–2433.
25. Staren, E. D., Gupta, D., & Braun, D. P. (2011). The prognostic role of quality of life assessment in breast cancer. *The Breast Journal, 17*(6), 571–578.
26. Harrell, f. e. jr., Lee, K. L., Matchar, D. B., & Reichert, T. A. (1985). Regression models for prognostic prediction: Advantages, problems, and suggested solutions. *Cancer Treatment Reports, 69*(10), 1071–1077.
27. Harrell, F. E. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis*. Cham: Springer.
28. Simon, R., & Altman, D. G. (1994). Statistical aspects of prognostic factor studies in oncology. *British journal of cancer, 69*(6), 979–985.
29. Cohen, J. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah: Lawrence Erlbaum Associates Publishers.
30. Hoerl, A. E., & Kennard, R. W. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 42*(1), 80–86.
31. Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov, F. Csaki (Ed.), *Second international symposium on information theory* (pp. 267–281): Budapest: Akademai Kiado.
32. Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer.
33. Fayers, P., Aaronson, N. K., Bjordal, K., Groenvold, M., Curran, D., & Bottomley, A. on behalf of the EORTC Quality of Life Group. (2001). The EORTC QLQ-C30 Scoring Manual (3rd Edn). *European Organisation for Research and Treatment of Cancer, Brussels*.
34. Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.
35. Lee, E. T., & Go, O. T. (1997). Survival analysis in public health research. *Annual Review of Public Health, 18*, 105–134.
36. Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine, 24*(11), 1713–1723.
37. Altman, D. G., & Andersen, P. K. (1989). Bootstrap investigation of the stability of a Cox regression model. *Statistics in Medicine, 8*(7), 771–783.
38. Sauerbrei, W., Boulesteix, A. L., & Binder, H. (2011). Stability investigations of multivariable regression models derived from low- and high-dimensional data. *Journal of Biopharmaceutical Statistics, 21*(6), 1206–1231.
39. Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association, 72*, 557–565.
40. Team, R. C. (2016). R: A language and environment for statistical computing. https://www.R-project.org/.
41. Morozova, O., Levina, O., Uuskula, A., & Heimer, R. (2015). Comparison of subset selection methods in linear regression in the context of health-related quality of life and substance abuse in Russia. *BMC Medical Research Methodology, 15*, 71.
42. Steyerberg, E. W., Eijkemans, M. J., Harrell, F. E. Jr., & Habbema, J. D. (2000). Prognostic modelling with logistic regression analysis: A comparison of selection and estimation methods in small data sets. *Statistics in Medicine, 19*(8), 1059–1079.
43. Yoo, W., Mayberry, R., Bae, S., Singh, K., He, P., Q., & Lillard, J. W. Jr. (2014). A study of effects of multicollinearity in the multivariable analysis. *International Journal of Applied Science and Technology, 4*(5), 9–19.
44. Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., et al. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography, 36*(1), 27–46.
45. Xue, X., Kim, M. Y., & Shore, R. E. (2007). Cox regression analysis in presence of collinearity: An application to assessment of health risks associated with occupational radiation exposure. *Lifetime Data Analysis, 13*(3), 333–350.
46. Sauerbrei, W., & Schumacher, M. (1992). A bootstrap resampling procedure for model building: Application to the Cox regression model. *Statistics in Medicine, 11*(16), 2093–2109.
47. Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology, 49*(12), 1373–1379.
48. Harrell, F. E. Jr., Lee, K. L., Califf, R. M., Pryor, D. B., & Rosati, R. A. (1984). Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine, 3*(2), 143–152.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.