Modeling strategies to improve estimates of prognostic factor analyses with patient reported outcomes in oncology

Francesco Cottone ¹, Nina Deliu ¹, Gary S. Collins ², Amelie Anota ³, Franck Bonnetain ³, Kristel Van Steen ^{4,5}, David Cella ⁶, Fabio Efficace ¹

Affiliations:

1. Italian Group for Adult Hematologic Diseases (GIMEMA), Data Center and Health Outcomes Research Unit. Rome, Italy.
2. Centre for Statistics in Medicine, NDORMS, University of Oxford, UK
3. Methodology and Quality of Life in Oncology Unit (INSERM UMR 1098), University Hospital of Besançon, Besançon, France; French National Platform Quality of Life and Cancer, France.
4. GIGA-R Medial Genomics Unit, University of Liège, Liège, Belgium
5. Department of Human Genetics – Systems Medicine, University of Leuven, Leuven, Belgium
6. Northwestern University, Feinberg School of Medicine, Department of Medical Social Sciences, Chicago, USA.

Corresponding author:

Francesco Cottone, PhD
Statistician, Health Outcomes Research Unit
Italian Group for Adult Hematologic Diseases (GIMEMA)
GIMEMA Data Center
Phone: +39 06 441 639830; Fax: +39 06 4402516; E-mail: f.cottone@gimema.it;
ORCID: 0000-0001-6240-8317

Abstract

Purpose

Inclusion of patient-reported outcomes (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 according to the combination of different sample sizes, event rates and degrees of multicollinearity between five PRO scales. For each scenario, we used the Cox proportional hazards (PH) model to estimate hazard ratios (HR), by four approaches. Two of these were automatic stepwise selection procedures, based on either likelihood ratio-test (Cox-PV) or the Akaike Information Criterion (Cox-AIC). Other methods consisted in Cox PH models including all variables, either penalized by Ridge-regression (Cox-R) or estimated as usual (Cox-Full). For each scenario, we simulated 1000 independent datasets and compared all methods on averaged outcomes.

Results

Cox-R showed similar or better performances with respect to other methods, in particular with medium-high multicollinearity ($\rho=0.4$ to $\rho=0.8$) and small sample size ($n=100$). Overall, Cox-PV and Cox-AIC performed worse, e.g. not selecting one or more prognostic collinear PRO scales in some scenarios. With respect to Cox-Full, Cox-R provided HR estimates with similar bias patterns but smaller root mean squared errors, particularly in higher multicollinearity scenarios.

Conclusions

Our findings suggest Cox-R as 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 outcomes (PRO) 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, to independently predict overall survival (OS), beyond 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 also by using a number of different PRO questionnaires. [3-17].

However, the identification and evaluation of prognostic PRO domains entails 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 impact 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 PRO. 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, as all 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 prognostic PROs is the use of an automatic variable selection procedure, either forward selection, backward elimination or stepwise [24,25,18]. However, it is known that such approaches can lead to large variability of 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 re-specifying the model by combining all collinear variables in a new summary measure [29]. However, this approach might not be desirable in case of PRO-based prognostic factors analysis, as scales from PRO questionnaires measure distinct components of the same construct, each potentially contributing unique prognostic information. Another approach is to generate and analyze new mutually orthogonal factors from the original collinear variables, for example by Principal Component Analysis (PCA) [29]. Unfortunately, such new outcomes would be difficult to interpret from a clinical perspective.

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

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

In this analysis, we investigate and compare the performances of different modeling strategies (methods) based on Cox PH regression, in estimating parameters of interest in prognostic factors analysis involving collinear PROs scales. Such methods consist in Cox PH models including all variables, either penalized by Ridge-regression or estimated as usual, and 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 PROs-based prognostic factor analysis, covering a wide range of possible scenarios according to 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 hazard function of the i -th individual defined as

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

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

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

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

We considered the following methods:

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

An automatic stepwise selection procedure based on the likelihood ratio test, including all observed covariates as candidate variables, 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 model fit at $\alpha = 0.05$ level.

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

The same model selection procedure as described above, based on the Akaike's Information Criterion (AIC) [31]. The final selected model was that with the lowest AIC, balancing both model fit and size [31].

III. Full Cox PH Model (Cox-full)

A Cox PH model including all variables including collinear PROs scales.

IV. Penalized Cox-PH Ridge (Cox-R)

A penalized Cox PH model estimating β by the Ridge or L^2 -norm estimator

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

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

Simulations setting

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 18-90 years), sex, current comorbidity (yes vs. no), and a generic variable X_4 (continuous, range 0-1). In addition, we considered five typically collinear self-reported scales (continuous, range 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 PROs questionnaires in prognostic factor analyses in oncology [1]. Each patient's self-reported scale is scored according to the standard EORTC procedures[33], from ordered numerical responses to one or more items (Likert-type scale). For each scale, the actual distribution of scores strongly depends on the own number of items and corresponding responses. Thus, we chose to simulate the five scales described above, as they are representative of all possible items-per-scale quantities in the QLQ-C30 questionnaire. Higher scores indicate a better HRQoL for QL and PF, a higher symptom burden for FA, PA and AP.

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 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 HRQoL patterns, 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 computed the corresponding multinomial probabilities of each item per scale according to pre-specified category difficulty parameters, i.e. $\delta_7 = (-1, 0.6, -0.2, 0.2, 0.6, 1)$ for QL items and $\delta_4 = (-0.7, 0, 0.7)$ for the remaining scales. Such probabilities were used to simulate 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 event (death) or exit from study, with a maximum follow-up period of 104 weeks (two 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 HRQoL 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 non-informative censoring times t_c from a Uniform distribution. The parameters of distributions of death and censoring times were chosen by 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, 0 otherwise). All individuals with either $d=0$ or $t_c > 104$ were right-censored.

[Insert Table 1 here]

Scenarios and performance evaluation criteria

We generated twenty-seven different scenarios, according the combination of three degrees of multicollinearity between five PROs scales (low, 0.2, medium, 0.4, 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 all methods on averaged outcomes. When considering stepwise-based methods (I-II, section 2), we investigated the average outcomes of the most selected (out of 1000) multivariable models [37,38]. When using Ridge penalty, we performed a grid search to choose the optimal regularization parameter λ_2 in the range (0, 50] over 1000 independent replicates, for each scenario. Then, we applied the penalty factor λ_2 that minimized the average root mean squared error (RMSE) over 1000 samples.

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

Results

When sample size was $n \leq 300$ and multicollinearity was $\rho > 0.4$, all most selected models by both Cox-PV and Cox-AIC did not include one or more collinear PROs (see table 2). The number of excluded PROs increased as multicollinearity raised from 0.2 to 0.8, with some scenarios where 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).

[insert Table 2 here]

When selected, the estimates of PROs provided by either methods were more biased with respect to those from Cox-Full and Cox-R, as shown in figure 1, although not different in variability with respect to Cox-Full. In addition, Cox-PV and Cox-AIC were more sensitive to event rate than other methods, in the provided amount of bias. For sample size $n=100$, such results were consistent across all scenarios and PROs, where Cox-R provided the best performance. Indeed, although the bias was similar between Cox-Full and Cox-R, the latter produced less variable estimates with respect Cox-Full, with markedly better performances for higher levels of multicollinearity (figure 1). For example, with $\rho=0.8$ and $n=100$, RMSE ranges were, respectively, [0.07, 0.14] for Cox-AIC, [0.10, 0.25] for Cox-Full and [0.07, 0.12] for Cox-R. In this setting, no results about RMSE were available for Cox-PV, as it did not select any PROs scales in the final model.

[insert Figure 1 here]

When sample size was $n=300$ or $n=500$, the stepwise methods selected all PROs in almost all scenarios. However, both failed when 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 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 bias of estimates was roughly similar among all methods only when multicollinearity was $\rho=0.2$ (see Figure 2, first left column), except for Appetite Loss (AP). For this scale, Cox-PV and Cox-AIC showed a greater sensitivity to ER than other methods, with respect to provided bias. However, as multicollinearity level raised, such sensitivity came up also for other scales. In addition, for $n=300$, biases were similar across all scenarios as provided by Cox-R and Cox-Full, both performing the same or better than Cox-PV and Cox-AIC. However, differences in RMSE between Cox-R and Cox-Full were overall lower with respect to $n=100$, as they remained markedly different only for a high degree of multicollinearity, i.e. $\rho=0.8$ (Figure 2).

[insert Figure 2 here]

The stepwise methods also showed RMSE levels similar to those of Cox-R and Cox-Full, except when $\rho=0.8$. For $n=500$, analogous results can be observed in terms of bias among all methods (Figure 3) for lower degrees of multicollinearity (except for AP), while Cox-PV and Cox-AIC provided larger bias for $\rho=0.8$. Also with respect to RMSE, the same considerations can be drawn as for $n=300$. To illustrate, for high multicollinearity ($\rho=0.8$) and $n=500$, RMSE values of PROs ranges were, respectively [0.03, 0.09] for Cox-PV, [0.03, 0.08] for Cox-AIC, [0.03, 0.08] for Cox-Full and [0.02, 0.06] for Cox-R. For each method, patterns of parameter estimates of non-collinear covariates did not show substantial improvements (results not shown). Overall, Cox-R provided estimates similar to those of Cox-Full, which were only slightly less biased, particularly in presence

of higher multicollinearity; in addition, Cox-R produced narrower confidence intervals for estimates than Cox-Full (see Figures 4-8 in Appendix).

[insert Figure 3 here]

Discussion

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

Indeed, given an initial set of collinear multiple PROs scales, Ridge regression would be a potentially useful approach, in balancing completeness of PROs information and accuracy of estimates. Previous studies dealing with multicollinearity in prognostic factor analysis showed better performances of penalized regression estimators (shrinkage methods) compared to maximum likelihood estimators [41,29,42-44]. 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 our knowledge, the only work evaluating shrinkage methods in presence of multicollinearity in this framework [45] only considered sample size $n=50$, 4 to 6 covariates, three degrees of multicollinearity (0.2, 0.4 and 0.8) with no censored data, reporting results for just one covariate. In addition, there is no previous research on the application of Ridge regression to prognostic factors analysis with collinear PROs.

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

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

The procedures including all variables outperformed those based on automatic selection. Cox-R provided less variable HR estimates than Cox-full, although at the cost of (slightly) higher bias.

Such bias/variance trade-off favored Cox-R particularly in those scenarios with higher degree 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 Cox-R showed lower RMSEs than Cox-Full but similar bias patterns.

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

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

clinical relevance. In addition, the application of all methods described in this work is feasible the most common statistical programs, such as R, SAS and SPSS.

This study first highlights the impact of harmful multicollinearity in PROs-based prognostic factor analysis, encouraging researchers to carefully check for its presence before performing any modeling procedure. In addition, it underlines the drawbacks of model identification by automatic stepwise selection procedures. Overall, our work shows the added value of using the Cox-R approach in PROs-based prognostic factor analysis. Overall, Cox-R allows achieving more accurate estimates than other methods, of the prognostic importance of multiple collinear PROs, while retaining all of these in the multivariable model. This result was consistent throughout all scenarios we explored, according to different issues such as sample size, event rate and degree of multicollinearity, suggesting penalized Ridge regression approach as the most appropriate, particularly when covariates are affected by at least $\rho=0.4$ multicollinearity and 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.

Contributions

Conception and design: FC, FE

Statistical analyses: FC, ND, FE

Interpretation of results: all authors

Manuscript writing: all authors

Disclosure statement

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

Supplementary data

Supplementary material includes the formal presentation of the Partial Credit Model and additional figures showing current findings.

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 : official journal of the American Society 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. *Eur J 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. *Eur J 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 : official journal of the American Society 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. *Int J Radiat Oncol Biol Phys*, 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 : official journal of the American Society 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. *Eur J 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. *Eur J 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 : official journal of the American Society 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 : official journal of the American Society 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. *Ann Oncol*, 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 Oncol*, 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 Rev Pharmacoecon Outcomes Res*, 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. *Stat Med*, 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 *Encyclopedia of Statistical Sciences* (Vol. 2, pp. 639-643): S. Kotz, N.L. Johnson, and C.B. Read (eds). Wiley: New York.
 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), R1-12.
 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* (Second edition. ed., Springer series in statistics,).
 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, N.J. ; London: Erlbaum.
 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. a. C. Petrov, F (Ed.), *Second International Symposium on Information Theory* (pp. 267-281): Akademiai Kiado, Budapest.
 32. Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The elements of statistical learning : data mining, inference, and prediction* (Springer series in statistics). New York: Springer.
 33. Fayers, P., Aaronson, NK, 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 Edition). *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., Peter He, 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.

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($p=.4$)	{0, 1}	1.07
Comorbidity	Binom($p=.5$)	{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

Legend:

^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

Abbreviations: HR = Hazard Ratio

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	1.10	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%	✓ X				✓ X	✓ X	✓ X	✓ X	✓ X	44	37
	50%	✓ X				✓ X	✓ X	✓ X	✓ X	✓ X	58	37
	70%	✓ X				✓ X	✓ X	✓ X	✓ X	✓ X	60	33
Medium MCL ($\rho=0.4$)	30%	✓ X				✓ X	✓ X	✓ X	✓ X	✓ X	30	31
	50%	✓ X				✓ X	✓ X	✓ X	✓ X	✓ X	53	36
	70%	✓ X				✓ X	✓ X	✓ X	✓ X	✓ X	58	33
High MCL ($\rho=0.8$)	30%	✓ X				✓ X	✓	✓	✓ X		9	8
	50%	✓ X				✓ X	✓ X	✓ X	✓ X	✓	12	14
	70%	✓ X		✓		✓ X	✓ X	✓ X	✓ X	✓ X	15	18
N = 300												
Low MCL ($\rho=0.2$)	30%	✓ X				✓ X	✓ X	✓ X	✓ X	✓ X	21	26
	50%	✓ X				✓ X	✓ X	✓ X	✓ X	✓ X	45	37
	70%	✓ X				✓ X	✓ X	✓ X	✓ X	✓ X	59	38
Medium MCL ($\rho=0.4$)	30%	✓ X				✓ X	✓ X	✓ X	✓ X	✓	16	17
	50%	✓ X				✓ X	✓ X	✓ X	✓ X	✓ X	32	31
	70%	✓ X				✓ X	✓ X	✓ X	✓ X	✓ X	45	38
High MCL ($\rho=0.8$)	30%	✓ X				✓ X	✓	✓	✓ X		10	5
	50%	✓ X				✓ X	✓	✓	✓ X		9	7
	70%	✓ X				✓ X	✓ X	✓ X	✓ X	✓	9	10
N = 100												
Low MCL ($\rho=0.2$)	30%	✓ X				✓	✓	✓	✓		14	5
	50%	✓ X				✓ X	✓ X	✓ X	✓ X		6	8
	70%	✓ X				✓ X	✓ X	✓ X	✓ X	✓	13	14
Medium MCL ($\rho=0.4$)	30%	✓ X				✓			✓		19	4
	50%	✓ X				✓	✓	✓	✓		8	7
	70%	✓ X				✓ X	✓ X	✓ X	✓ X		9	11
High MCL ($\rho=0.8$)	30%	✓ X									37	9
	50%	✓ X									24	4
	70%	✓ X				✓			✓		18	5

Legend:^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.

X = the variable was selected in the model by the Cox-PV stepwise procedure.

Abbreviations: 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 0.5 as entry/stay criteria; Cox-AIC = Stepwise Cox PH with AIC as entry/stay criteria.

Fig. 1 Standardized Bias and Root Mean Squared Error for all PROs and scenarios (n=100)

Abbreviations: RMSE, Root Mean Squared Error; M-Coll = multicollinearity degree; ER = Event rate; Cox-PV = Stepwise Cox PH with 0.5 p-value as entry/stay thresholds; 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.

Fig. 2 Standardized Bias and Root Mean Squared Error (RMSE) for all PROs and scenarios (n=300)

Abbreviations: RMSE, Root Mean Squared Error; M-Coll = multicollinearity degree; ER = Event rate; Cox-PV = Stepwise Cox PH with 0.5 p-value as entry/stay thresholds; 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.

Fig 3 Standardized Bias and Root Mean Squared Error for all PROs and scenarios (n=500)

Abbreviations: RMSE, Root Mean Squared Error; M-Coll = multicollinearity degree; ER = Event rate; Cox-PV = Stepwise Cox PH with 0.5 p-value as entry/stay thresholds; 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.