A comparision of R^2-measures for evaluation of the prognostic value in survival models

Christian Hentschke

May 24, 2011

Background

The development of prognostic models is an important issue in different fields of clinical research [1, 2]. Well confirmed prognostic models can be particularly useful for decisions on appropriate treatment strategies, the allocation of limited resources, defining patient populations for entry to clinical trials and risk adjustment within randomized clinical trials as well as for comparisons in observational studies [3, 4, 5]. Methods of survival analysis, especially the Cox Model, are frequently used to discover prognostic models for various diseases, where the endpoint is the time to death or to another event of interest. Examples are the prediction of time until the occurrence of metastases after cancer therapy or the prediction of survival time after a transplantation surgery. Of high medical interest thereby is the explanatory value of single prognostic factors as well as the prognostic model in total. The identification of prognostic classification schemes and the prediction of individual survival time from covariates is notably relevant. From this perspective diagnosis and prognosis constitute a comparable challenge: "the clinician has some information and wants to know how this relates to the true patient state, whether this can be known currently (diagnosis) or only at some point in the future (prognosis)" [5].

Despite the importance, improvement in prognostic assessment from encountered models often is discussed controversial [4, 6, 7]. Particularly the prediction accuracy of individual survival time is critical. P-values and estimated effects of single covariates do not necessarily yield substantial improvements in prediction accuracy [7, 8, 9, 10], mainly because even well performing predictors on the fitted data could fail seriously in external validation with new data [11, 12]. Hence, the relation on these alone could be misleading [13]. This problem exhibits even more prominently in a situation of high dimensional data, where many more variables (p) than samples (n) are available (p>> n problem) [14, 15]. Gene expression microarray data are an example therefore. "As a result of high dimension, it is almost always possible to find a combination of molecular predictors that are associated with the outcome in the considered data set, independently of the true predictive power"[12]. This is of serious practical relevance, when taking the increasing research with high dimensional data into account. For this context Boulesteix and Sauerbrei [12] suggest to consider the predictive power that can be derived from gene expression predictors in additon to clinical predictors rather than the modles total predictive power. Anyway prognostic models have to be both accurate and precise in their predictions in order to derive useful clinical decisions [16]. For that reason an appropriate evaluation of prediction accuracy is foremost important for

prognostic models. However, although different approaches have been proposed yet, the assessment of prediction accuracy in continuous survival models is still not straightforward [17]. This may have several reasons, but is mainly caused by the fact that the application of traditional performance measures like the multiple correlation coefficient R^2 of explained variation from normal multiple linear regression leads to bias in the case of censored data [8]. No equally accepted approach that inherits all statistical properties of traditional R^2 is availible among generalizations or adaptions of R^2-measures to the situation of survival models. In consequence the different approaches have specific drawbacks and advantages, which affects introduced approaches to be still in dispute [17, 18]. Additionally in a high dimensional data situation it is even less obvious how the prognostic value should be assessed, because models have to incorporate methods for variable selection or regularization [15, 16, 18]. Above this, studies that intend to compare the property and performance of methods for variable selection or regularization may be of limited value without reliable knowledge about whether and how "the choice of evaluation criterion may affect the conclusions made [...]" [18].

Up to now proposed measures of model performance for censored survival data can be distinguished in three different approaches [17]: (a) likelihood based approaches, (b) ROC based approaches, (c) distance based approaches. In likelihood-based approaches the log likelihood of a prediction model is compared to the analogous log likelihood of the (nested) "null model" with no covariate information [19, 20, 21, 22, 23, 24, 25]. For Cox proportional hazards models these measures are predominantly calculated on basis of the partial log likelihood rather than on full log likelihood. ROC based approaches utilize the fact that the outcome in survival models is a combination of a binary and a continuous information. They apply concepts of time-dependent missclassification rates (sensitivity and specificity) and time-dependent AUC-criterions of ROC curves to compare levels of interest of the prognostic score (= additive predictor) of the model [26, 27]. Distance-based measures estimate a quantity of prediction error in form of the difference between predicted and observed survival curves [8, 17, ?, 28, 29, 30]. Each of these approaches accentuates another aspect. Performance measures have therefore also been classified in measures of explained variation, measures of discrimination and measures of calibration elsewhere [5]. These classifications of the exising performance measures contribute to distinguish them theoretically and to organize them in a meaningful order. But indeed, an inconsistent and not uniquely used terminology for existing approaches in the according literature complicates the classification of the different types of measures. Therefore some other usage of corresponding terminology or other classification in literature is possible.

However, actually neither one of the proposed performance measures nor one of the general approaches have been universally recognized as a standard for evaluating survival predictions yet [17, 18]. The most commonly applied performance measures for continuous outcomes in practice are likelihood based generalizations of coefficient of explained variation, like Nagelkerke's R^2 [5]. However, the practical suitability of likelihood based measures to assess the prognostic performance of survival models is only insufficiently investigated so far. Some previous comparisons of performance measures, for example from Korn and Simon [31] or Pepe et al. [27], solely focused on distance based measures or ROC based approaches, respectively. Schemper and Stare (1996) [8] examined some distance based and several likelihood based performance measures, but certainly could not include newer proposals of likelihood based mesures. O'Quigley, Xu and Stare

[24] also compaired several likelihood based measures, and some simplified estimators for them, by simulation subsequent to their own proposal of a measure. At least they investigated different strength of regression effects, different censoring percentages and different covariate distributions for two different scenarios of sample size in one simulation run for each combination of these factors. Steyerberg et al. [5] surveyed the properties of different kind of measures to evaluate prediction rules for time-to-event models with a real data set. Although some likelihood based measures were discussed, they layed greater stress on investigating some new measures like reclassification indices and measures to quantify clinical usefulness of a prediction model directly. However, none of these comparisions examined the behavior of likelihood based approaches under deviation from the assumptions of underlying regression model, although this is certainly a frequently occuring case for practical application of R^2-measures. Apart from theoretical considerations particularly their real sensitivity against violation of underlying assumptions is largely unknown. These assumptions, like the assumption that the model is correctly specified with a complete set of covariates, the proportional hazards assumption, or the assumption of not informative random censoring, are inherently incurred in calculation of Cox model's partial likelihood. Because likelihood based R^2-measures are directly worked out on model's partial likelihood, they are supposed to be especially prone against the violation of these assumptions. However, the possible influence on likelihood based R^2-measure that are calculated in presence of violation of these assumptions has not been thoroughly investigated yet. As a consequence, up to now it is not clear which of the proposed likelihood based R^2-measure is most robust against model misspecifications and shold be recommended in situations, where is not known, if assumptions are violated. Incidently, it has already been advised by Schemper and Stare [8] to investigate this. Additionally, it comes along that it is actually not finally clear, which likelihood based measure is most convenient for measuring the performance under a correctly specified model in general or in a high-dimensional data situation. Hence, the examination of the true performance of likelihood based measures regarding these violations and the exploration of additional moderating effects of sample size, presence of non-informative covariates, the amount of censoring and different strength of regression effects gains in interest. Indeed, Hielscher et al. [16] compared the suitability of five likelihood based and two distance based measures to assess the prognostic performance of survival models for a situation of high-dimensional data including misspecified models. However, their work only focuses on missing covariates and degree of censoring as investigated model misspecification. Beyond this, the influence of different other possible situations and misspecifications on the performance of likelihood based R^2-measures remains unclear. To our knowledge, other systematic comparisons of likelihood-based measures concerning their prediction accuracy in different situations of survival analysis have not been conducted yet.

Because currently only one comprehensive and systematic comparison of likelihood based performance measures that considers model misspecifications is existing in literature, this work is going to compare their ability to measure the prediction accuracy in a survival model situation. Moreover the sensitivity of these measures concerning deviations from underlying model assumptions, like violation of proportional hazards assumption, different degrees of non-informative and informative censoring, missing (unknown) covariates, different sample sizes and different amount of non-informative covariates are examined. Further, an overview of all six studied likelihood-based R^2-measures is given and R-Code for their calculation is provided in the appendix uniformly using event time, event indicator and the vector of linear predictors.

Considered Model and Notation

Througout this paper we refer to the framework of Cox proportional hazards regression model [32] for intended comparisons within a simulatoin study. The Cox proportional hazards regression model is the most often used framework to model censored survival data and the predominant model in survival analysis [33]. Despite the fact that likelihood-based R^2-measures are also defined for parametric survival models, we proceed from the Cox model, because of its high recognition and frequent application. Thereto suppose a survival time $T \in \mathbb{R}^+$and an independent random censoring time $C \in \mathbb{R}^+$. The right censored event time then is described by $Y := \min(T, C)$ with $\mathrm{i} = 1, ..., \mathrm{n}$ realizations $\mathrm{y_i}$ of n independent observations. An according censoring indicator is given by $\Delta := \mathrm{I}(T \leq C)$ with realizations $\delta_\mathrm{i} = 1$ for complete observations and $\delta_\mathrm{i} = 0$ for censored observations. Further assume a vector of time-invariant random covariates $X \epsilon \mathbb{R}^p$ with realizations $\mathbf{X_i} = (\mathrm{x_{i1}}, \ldots, \mathrm{x_{i}}p)$, $\mathrm{i} = 1, ..., \mathrm{n}$. The unknown true conditional survival function is then given by $\mathrm{S(t|\ x)} = \mathrm{P(T > t \mid } X = \mathrm{x})$ with corresponding true conditional hazard function $\mathrm{h(t \mid x)} = -\frac{\partial}{\partial \mathrm{t}} \ln(\mathrm{S(t \mid x)})$ [34]. Under a proportional hazards assumption over all occurring ordered failure times $\mathrm{t_1} < \mathrm{t_2} < ... < \mathrm{t_k}$ the hazard ratio ψ is modeled through

$$\psi(\mathbf{x_i}) = \frac{\mathrm{h_i(t)}}{\mathrm{h_0(t)}} = \exp(\boldsymbol{\beta}^\mathrm{T} \mathbf{X_i}), \tag{1}$$

where $\mathrm{h_0(t)}$ is a time dependent but arbitrary baseline hazard and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_\mathrm{p})$ is the parameter vector to be estimated. This model could also be equivalently specified with the probability density function of t given x:

$$f\left(\mathrm{t} \mid \mathbf{x}; \boldsymbol{\beta}\right) = \mathrm{h_0(t)} \cdot \exp\left(\boldsymbol{\beta}^\mathrm{T} \mathbf{X} - \exp\left(\boldsymbol{\beta}^\mathrm{T} \mathbf{X}\right) \cdot \int_0^\mathrm{t} \mathrm{h_0(u)du}\right). \tag{2}$$

Usually the parameter estimation for parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_\mathrm{p})^\mathrm{T}$ is done from n available data triples $(\mathrm{y_i}, \mathrm{X_i}, \delta_\mathrm{i})$ without specification of $\mathrm{h_0(t)}$ through maximization of the partial likelihood

$$L(\boldsymbol{\beta}) = \prod_{\mathrm{j}=1}^{\mathrm{r}} \frac{\exp(\boldsymbol{\beta}^\mathrm{T} \mathbf{X_j})}{\sum_{\mathrm{l} \epsilon \mathrm{R(t_j)}} \exp(\boldsymbol{\beta}^\mathrm{T} \mathbf{X_l})}, \tag{3}$$

where $\mathrm{x_j}$ is the vector of covariates for the jth individual, who dies at jth-failure time, and $\mathrm{R(t_j)} = \{\mathrm{i} : \mathrm{y_i} \geq \mathrm{t_j}\}$ denotes the set of individuals that are at risk at failure time $\mathrm{t_j}$. Let $\mathrm{Z_i(t_j)} := \mathrm{I(y_i \geq t_j)}$ be the indicator function of $\mathrm{i} \epsilon \mathrm{R(t_j)}$, indicating, whether the individual i is at risk at time $\mathrm{t_j}$. Then the partial likelihood $L(\boldsymbol{\beta})$ could also be expressed as the product over the conditional probabilities $\pi_\mathrm{j}(\mathrm{t_j}; \beta)$ of choosing individual j that fails at $\mathrm{t_j}$, given all individuals at risk at time $\mathrm{t_j}$ [23, 24], with

$$\pi_\mathrm{j}(\mathrm{t_j}; \beta) = \frac{\mathrm{Z_j(t_j)} \cdot \exp(\boldsymbol{\beta}^\mathrm{T} \mathbf{X_j})}{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{Z_i(t_j)} \cdot \exp(\boldsymbol{\beta}^\mathrm{T} \mathbf{X_i})}. \tag{4}$$

The partial log likelihood is often preferred for computation yet due to better numerical performance:

$$l(\boldsymbol{\beta}) = \ln\left(L(\boldsymbol{\beta})\right) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \delta_\mathrm{i} \left[\boldsymbol{\beta}^\mathrm{T} \mathbf{X_i} - \sum_{\mathrm{l} \epsilon \mathrm{R(t_i)}} \exp(\boldsymbol{\beta}^\mathrm{T} \mathbf{X_l})\right]. \tag{5}$$

However, parameter estimation based on these traditional partial likelihood methods in the majority of cases does not supply satisfying results regarding interpretability and prediction accuracy in high dimensional data situations with a vast number of features [10]. Higher precision in prediction could be obtained through utilization of an appropiate regularization method [15, 18]. To account for over-fitting and the n >> p problem, frequently cross-validated LASSO (Least Absolute Shrinkage and Selection Operator) method is used for parameter estimation and variable selection [35, 36, 37], because of its thoroughly examined and convenient statistical properties [15, 38]. If applied to the Cox model, the penalized partial log likelihood [10, 35, 36, 39]

$$l(\boldsymbol{\beta}) - \lambda \sum_{\mathrm{j}=1}^{p_g} \mid \boldsymbol{\beta}_{\mathrm{j}}^{\mathrm{g}} \mid \tag{6}$$

is maximized instead of equation 5. In the considered L_1-penalization λ denotes a tuning parameter that is chosen on the data and $\boldsymbol{\beta}_{\mathrm{j}}^{\mathrm{g}}$ represents the parameter vector of high-dimensional features, which is treated separately from additional clinical covariates. High dimensional features are penalized by their L_1-norm, while clinical covariates are included unpenalized. An advatage of the LASSO-estimation is that it combines parameter estimation and parameter selection. Because a penalization term is introduced, estimated parameters are shrunken towards zero. In consequence they also yield smaller standard errors, which leads to more precise predictions [10, 36, 37]. The LASSO estimation includes variable selection, because it is likely that several coefficients are shrunken to zero [37]. A computationally efficient algorithm for maximizing the LASSO-penalized log likelihood (equation 6) has been proposed by Goeman [40].

In order to evaluate the prediction accuracy of a prognostic model for future observations R^2-measures consequently should be obtained from new data, which have not been used to derive the model [41].

Overview of Likelihood-based R^2-Measures

The predictive value of a model could be described in terms of amount of variation that is explained by a set of covariates included in a model compared to a (nested) reference model without these variables or a null model[1], respectively. In normal linear regression with unsecored observations this is expressed by the (unadjusted) coefficient of determination R^2_{OLS}, which is defined as one minus the amount of not explained variation (SSE = residual sum of squares) divided by the amount of total variation (SST = total sum of squares):

$$R^2_{\mathrm{OLS}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} = \frac{\sum_{\mathrm{i}}(\mathrm{y_i} - \bar{\mathrm{y}})^2 - \sum_{\mathrm{i}}(\mathrm{y_i} - \hat{\mathrm{y}})^2}{\sum_{\mathrm{i}}(\mathrm{y_i} - \bar{\mathrm{y}})^2}. \tag{7}$$

As is well-known, in survival analysis the application of a corresponding coefficient is not equally genuine [25, 31, 42]. A direct utilization of the defined R^2_{OLS} - measure in survival analysis is not appropriate due to dependency on observation time and bias in the presence of censoring [16]. Generalizations of R^2-measures for survival analysis have to take this into account pertinently. The same is true for the adjusted R^2_{adj}-measure of linear regression, which is adjusted by the

[1] Throughout this paper it is referred to a "null model", when the vector of parameter coefficients is set to zero ($\boldsymbol{\beta} = 0$). This corresponds to a Cox model without covariates.

according degrees of freedom (df) and therefore calculated on basis of variances instead of sum of squares. With intending a similar but more general R^2-coefficient, different likelihood based R^2-type measures for use with survival data and generalized linear models were suggested by several authors [19, 20, 21, 22, 23, 24, 25]. In the context of generalized R^2-measures, however, the term "explained randomness" has been adopted for likelihood based approaches, following a suggestion of Kent and O'Quigley [20]. In contrast, some authors distinguish the term "explained variation", which has been maintained for distance based generalized R^2- measures. For subsequent description of likelihood based R^2-measures the uniform denotation introduced by Hielscher et al. [16] is used throughout this paper rather than their original denotation.

R^2_L-Measure of Cox, Snell and Magee

Likelihood based concepts of R^2-measures take advantage of the equivalence of the likelihood ratio statistic and the corresponding F statistic in normally distributed linear regression models, when testing the same hypothesis [21]. Commonly the F statistic with $(p-1)$ and $(n-p)$ degrees of freedom can then be defined directly or also as a monotonic increasing function of R^2:

$$F = \frac{(SST - SSE)\,/\,(p-1)}{SSE/\,(n-p)} = \frac{R^2/\,(p-1)}{(1-R^2)\,/\,(n-p)}. \tag{8}$$

With growing F the explanatory value of the model would increase. The same hypothesis could be tested by a likelihood ratio test with a log likelihood ratio of

$$LR = 2 \cdot \ln\left(\frac{L(\hat{\boldsymbol{\beta}})}{L(0)}\right) = n \cdot \ln\left(\frac{SST}{SSE}\right) = -n \cdot \ln\left(1 - \left(1 - \frac{SSE}{SST}\right)\right) = -n \cdot \ln\left(1 - R^2\right), \tag{9}$$

where $L(\hat{\beta})$ and $L(0)$ denote the likelihoods of the unrestricted and the restricted model. An early measure of model performance for generalized linear models that directly applies this analogy between F statistic and likelihood ratio statistic in normal linear regression has been presented by Maddala [43] and proposed for use with survival analysis by Cox and Snell [19] and independently by Magee [21]:

$$R^2_L = 1 - \left(\frac{L(\hat{\boldsymbol{\beta}})}{L(0)}\right)^{-\frac{2}{n}} = 1 - \exp\left(-\frac{2}{n} \cdot \left(l(\hat{\boldsymbol{\beta}}) - l(0)\right)\right). \tag{10}$$

For use in Cox regression the log likelihoods of unrestricted ($l(\hat{\beta})$) and restricted ($l(0)$) model are replaced by their log partial likelihoods (see equation 5).

R^2_N-Measure of Nagelkerke

An apparent drawback of R^2_L-measure is, however, that it cannot reach a maximum value of 1 for models, where the likelihood is calculated from a product of discrete probabilities instead from densities. Nagelkerke [22] argued that this concerns logistic models as well as Cox's model, to a lesser degree. The reachable maximum results in

$$\max(R^2_L) = 1 - \exp\left(\frac{2}{n} \cdot l(0)\right) = 1 - L(0)^{\frac{2}{n}}. \tag{11}$$

Thus a clear interpretation of R^2_L as a proportion of explained variation or explained randomness, respectively, will be less obvious. In order to account for this disadvantage Nagelkerke [22] suggested a modification, which results in Nagelkerke's R^2_N:

$$R^2_N = \frac{R^2_L}{\max(R^2_L)}. \tag{12}$$

R^2_{OXS}-Measure of O'Quigley, Xu and Stare

Though their negative correlation with the amount of censoring has been regarded as a serious weakness of both of these approaches within the considered framework of Cox proportional hazards regression model [8]. To correct therefore O'Quigley, Xu and Stare [24] recommended to divide by the number of events k rather than by number of observations n, yielding the R^2-measure by O'Quigley, Xu and Stare:

$$R^2_{OXS} = 1 - \left(\frac{L(\hat{\boldsymbol{\beta}})}{L(0)}\right)^{-\frac{2}{k}} = 1 - \exp\left(-\frac{2}{k} \cdot \Big(l(\hat{\boldsymbol{\beta}}) - l(0)\Big)\right). \tag{13}$$

As an additional positive effect this correction exploits a larger interval of values than R^2_L, although it still does not reach a possible maximum of 1.

R^2_{KO}-Measure of Kent and O'Quigley

A very general likelihood based approach to measure the dependence between Y and X has been proposed by Kent and O'Quigley [20]. Instead of relying on likelihoods, they focus on the *information* gained through a specified model of interest in contrast to a reference or null model. This information gain is quantified by the Kullback-Leibler distance Γ [44, 45] between the considered nested hypothesis that are compared in a scaled form, for instance "$H_0 : \boldsymbol{\beta} = 0_p$" and "$H_1$: no restrictions on $\boldsymbol{\beta}$". For different parametrization of a model the Kullback-Leibler information gain is given for a general vector of model parameters θ by twice the difference in the expectations I of the log likelihood function

$$\Gamma(\boldsymbol{\theta}) = 2[I(\boldsymbol{\theta}_1;\boldsymbol{\theta}_1) - I(\boldsymbol{\theta}_0;\boldsymbol{\theta}_1)], \tag{14}$$

where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_0$ define the values that maximize the expectations of log likelihood function, satisfying H_0 or H_1, respectively, over the parameter space of $\boldsymbol{\theta}$. The Kullback-Leibler information gain is here evaluated at θ_1, which is considered as the true value of $\boldsymbol{\theta}$. The expectations of the log likelihood function for θ under θ_1 generally take the form

$$I(\boldsymbol{\theta};\boldsymbol{\theta}_1) = E\left[\ln f(\mathrm{Y} \mid \mathrm{X};\theta)\right] = \int_{\mathbb{R}^p} \int_0^{t_k} \ln\left[f(\mathrm{y} \mid \mathbf{X};\boldsymbol{\theta})\right] f(\mathrm{y} \mid \mathbf{X};\boldsymbol{\theta}_1) dt dG(\mathbf{X}), \tag{15}$$

where $f(\mathrm{y} \mid \mathbf{X};\boldsymbol{\theta})$ denote the conditional distribution functions of y given x, and $G(\boldsymbol{X})$ is the unconditional distribution of $\mathbf{X}$. Finally, Kent and O'Quigley define their R^2- coefficient as

$$R^2_{KO} = 1 - \exp(-\Gamma) \tag{16}$$

to keep it ranging between 0 (for no information gained) and 1 (for perfect explanation of response variable).

Applying this concept to a model with normally distributed probability density function of the residual term ε would result in the ordinary multiple correlation coefficient R^2 [20, 45]. However, to be able to work out the resulting integrals in 15 explicitly a specified probability density function of the error term is a necessary condition within this approach [46, 47]. Due to its semi-parametric nature with an unspecified baseline hazard function $h_0(t)$ the error distribution of Cox proportional hazards model is unidentified. To overcome this problem for implementation of R^2_{KO} Kent and O'Quigley [20] used the circumstance that the conditional probability distribution function of T given X in Cox models is specified only up to a monotone transformation of T. So for any strictly monotone increasing function $\phi(.)$, $T^* = \phi(T)$ leads to the same vector of estimated regression coefficients $\hat{\boldsymbol{\beta}}$ in a Cox regression model. "Each Cox model result can be seen as a member of a class of equivalent results" [46]. To facilitate the calculation of R^2_{KO}-measure, Kent and O'Quigley [20] assume a Weibull distribution for the conditional distribution of T^* given X as a natural choice of a general parametric distribution within the same distribution family. More precise the Weibull model is a special case of the Cox model. Heinzl [46] and Heinzl, Stare and Mittlböck [47] formulate a modified version of the underlying considered linear regression model for $Y^* = \log(T^*)$ that simplifies computation:

$$Y^* = -\frac{\mu}{\alpha} - \frac{\boldsymbol{\beta}^T \mathbf{X}}{\alpha} + \frac{\varepsilon^*}{\alpha}, \tag{17}$$

where μ denotes the regression parameter for the constant term and α denotes a scaling parameter with $\alpha > 0$. The baseline hazard function is then proportional to a power of t with $h_0^*(t) = \alpha \cdot \exp(\mu) \cdot t^{\alpha-1}$ for any choice of μ and $\alpha > 0$. The error term of the model ε^* follows a standard extreme value distribution with density $f(u) = \exp(u - \exp(u))$, say. The relevant model parameters for estimation of R^2_{KO}-measure under the considered model are $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^T, \mu_0, \alpha_0)^T$, with $\boldsymbol{\beta}_0 = 0_p$, and $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1^T, \mu_1, \alpha_1)^T$, generally with $\boldsymbol{\beta}_1 \neq 0$. Because the conditional probability density function $f(y \mid \mathbf{X}; \boldsymbol{\theta})$ is related to the error density by $f(y \mid \mathbf{X}; \boldsymbol{\theta}) = \alpha f(\alpha y + \mu + \boldsymbol{\beta}^T \mathbf{X})$, the expected log likelihood under the assumption of the Weibull model takes the form

$$I(\boldsymbol{\theta}; \boldsymbol{\theta}_1, \mathbf{x}) = \int_{-\infty}^{+\infty} \ln\left[\alpha f(\alpha y + \mu + \boldsymbol{\beta}^T \mathbf{X})\right] \alpha_1 f(\alpha_1 y + \mu_1 + \boldsymbol{\beta}_1^T \mathbf{X}) dy =$$

$$= \ln(\alpha) + \frac{\alpha}{\alpha_1}\gamma'(1) + b - \exp(b)\gamma\left(\frac{\alpha}{\alpha_1} + 1\right), \tag{18}$$

where $b = \mu + \boldsymbol{\beta}^T \mathbf{X} - (\alpha/\alpha_1)(\mu_1 + \boldsymbol{\beta}_1^T \mathbf{X})$ and $\gamma(.)$ denotes the gamma function and $\gamma'(.)$ its first derivate. To find an estimate for $\Gamma(\boldsymbol{\theta})$ one has to insert appropriate estimates for the considered parameters $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^T, \mu_0, \alpha_0)^T$ and $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1^T, \mu_1, \alpha_1)^T$. Because Γ does not depend on the choice of μ_1 and $\alpha_1 > 0$ [20, 47], they can be given arbitrary but convenient values, for instance $\mu_1 = 0$ and $\alpha_1 = 1$. An appropriate estimate for $\boldsymbol{\theta}_1$ can then be obtained by inserting the estimated vector of coefficients $\hat{\boldsymbol{\beta}}_{Cox}$ from the fitted Cox model under H_1 for $\boldsymbol{\beta}_1$, so that $\boldsymbol{\theta}_1 = (\hat{\boldsymbol{\beta}}_{Cox}^T, 0, 1)^T$. According to Heinzl [46] and Heinzl, Stare and Mittlböck [47] an estimator for $\boldsymbol{\theta}_0$ can be found by numerically maximizing the empirical expected log likelihood function $I(\boldsymbol{\theta}; \boldsymbol{\theta}_1) = \frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}_1, \mathbf{X}_i)$ over all $\boldsymbol{\theta} = (0_p, \mu, \alpha)^T$ satisfying H_0. An explicit solution for $\hat{\mu}_0$ with

$$\hat{\mu}_0 = -\ln\left(\gamma(\hat{\alpha}_0 + 1)\right) - \ln\left(\frac{1}{n}\sum_{i=1}^{n} \exp\left(-\hat{\alpha}_0 \hat{\boldsymbol{\beta}}_1^T \mathbf{X}_i\right)\right) \tag{19}$$

and an implicit solution for $\hat{\alpha}_0$ defined as

$$\xi(\alpha) := \psi(1) - \psi(\alpha) + \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{\exp(-\alpha \mathrm{z_i})}{\sum_{\mathrm{j}=1}^{\mathrm{n}} \exp(-\alpha \mathrm{z_j})} \mathrm{z}_i = 0 \tag{20}$$

is offered by Heinzl [46], where $\mathrm{z_i} = \hat{\boldsymbol{\beta}}_1^{\mathrm{T}} \mathbf{X}_\mathrm{i} - \hat{\boldsymbol{\beta}}_1^{\mathrm{T}} \bar{\mathbf{X}}$ and $\psi(.)$ denotes the digamma function.

Because it depends on methods of numerically optimization, the computation of this measure has been regarded difficult [25, 46].

$\mathrm{R}^2_{\mathrm{XO}}$-Measure of Xu and O'Quigley

A further approach, motivated by the intention to simplify the computation of $\mathrm{R}^2_{\mathrm{KO}}$ while keeping its presumed good statistical properties, has been introduced by Xu and O'Quigley [23]. To reduce the computational effort this R^2-measure aims at estimating gained information only from routinely calculated quantities in proportional hazards analysis, but which on the other hand restricts this measure to use with Cox proportional hazards regression. It is founded on the basic idea that the dependency between survival time T and covariates X is reflected equally well by the conditional distribution of X given T compared to the other way round. Therefore an alternative definition of the expected log likelihood is considered in this approach [23, 24]

$$I(\boldsymbol{\theta}, \boldsymbol{\theta}_1) = \int_0^{\mathrm{t_k}} \int_{\mathbb{R}^\mathrm{p}} \ln\left[g(\mathbf{X} \mid \mathrm{t}; \boldsymbol{\theta})\right] g(\mathbf{X} \mid \mathrm{t}; \boldsymbol{\theta}_1) d\mathrm{X} dF(\mathrm{t}) \tag{21}$$

where $g(\mathbf{X} \mid \mathrm{t}; \cdot)$ denotes the conditional probability density function of X given T and $F(\mathrm{t})$ is the marginal distribution function of T. A resulting definition of information gained could then be expressed through

$$\Gamma(\boldsymbol{\beta}) = 2\left[I(\boldsymbol{\beta}_1, \boldsymbol{\beta}_1) - I(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)\right] =$$

$$= 2 \int_0^{t_k} \int_{\mathbb{R}^\mathrm{p}} \ln\left(\frac{g(\mathbf{X} \mid \mathrm{t}; \boldsymbol{\beta})}{g(\mathbf{X} \mid \mathrm{t}; 0)}\right) g(\mathbf{X} \mid \mathrm{t}; \boldsymbol{\beta}) d\mathbf{X} dF(\mathrm{t}) \tag{22}$$

Xu and O'Quigley [23] suggest a semi-parametric estimator of gained information that does not need the assumption of an underlying Weibull model from routinely calculated quantities by

$$\hat{\Gamma}(\hat{\boldsymbol{\beta}}) = 2 \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{W}(\mathrm{y_i}) \cdot \sum_{\mathrm{j}=1}^{\mathrm{n}} \pi_\mathrm{j}(\mathrm{y_i}; \hat{\boldsymbol{\beta}}) \ln\left(\frac{\pi_\mathrm{j}(\mathrm{y_i}; \hat{\boldsymbol{\beta}})}{\pi_\mathrm{j}(\mathrm{y_i}; 0)}\right) \tag{23}$$

This involves the conditional probabilities $\pi_\mathrm{j}(\mathrm{y_i}; \beta)$ of choosing the individual that fails over all observed times $\mathrm{y_i}$ given the risk sets at a time as an estimator for conditional distribution of X given T (see equation 4). The marginal distribution of T is estimated by the Kaplan-Meier estimate, so that $\mathrm{W}(\mathrm{y_i}) = \hat{F}(\mathrm{Y_i}+) - \hat{F}(\mathrm{Y_i})$ indicates the jump of the Kaplan-Meier curve at observed time $\mathrm{y_i}$. The $\mathrm{R}^2_{\mathrm{XO}}$- measure is then also defined as

$$\mathrm{R}^2_{\mathrm{XO}} = 1 - \exp(-\Gamma) \tag{24}$$

Hielscher et al. [16] point out the relation of this concept to Schoenfeld residuals[48].

Because of the inverse interprtation of this measure and the fact that the computation still

seemed quite difficult, O'Quigley, Xu and Stare [24] proposed another simplyfication that should approximate R^2_{XO} as well as R^2_{KO}. This simplification resulted in R^2_{OXS}, which is already described above.

R^2_R-Measure of Roysten

Despite of a close relationship in underlying ideas and a asymptotically numerical equivalence for normal linear regression models, measures of explained randomness and measures of explained variation can sizable differ for proportional hazards models [16, 25]. While measures of explained randomness can be approximated by

$$R^2 \approx \frac{\text{var}(\mathbf{X}\hat{\boldsymbol{\beta}})}{1 + \text{var}(\mathbf{X}\hat{\boldsymbol{\beta}})} \tag{25}$$

measures of explained variation are rather expressed by

$$R^2 = \frac{\text{var}(\mathbf{X}\hat{\boldsymbol{\beta}})}{\sigma^2_{\text{res}} + \text{var}(\mathbf{X}\hat{\boldsymbol{\beta}})} \tag{26}$$

To introduce a likelihood based measure more close to explained variation, Roysten [25] modified the R^2_{OXS}-measure to resemble equation 26 more readily:

$$R^2_R = \frac{R^2_{OXS}}{R^2_{OXS} + (\pi^2/6) \cdot (1 - R^2_{OXS})} \tag{27}$$

He argued that measures of explained randomness always exceed measures of explained variation. The suggested correction assumes an underlying Weibull model that is equivalent to linear regression model with a log-link and a Gumbel (standard extreme) distributed error term, for which the residual variance is estimated by $\sigma^2_{\text{res}} = \pi^2/6$.

Interconnection of R^2-Measures

A summary of the interconnection of considered R^2-measures outlined so far in the description above could also be allegorised graphically as suggested by Hielscher et al. [16]. Because Hielscher et al. [16] consider additional measures, we present a modified version of their diagram in figure 1.

Design of Simulation Study

General Approach

The conducted comparison of described likelihood based performance measures was implemented on basis of computer simulations by using the R software environment for statistical computing [49]. In order to systematically compare the particular effects of examined misspecifications on the calculated values of considered R^2-measures we assessed their behavior under a certain misspecification in contrast to a correctly specified model with otherwise same conditions. In the first instance this assessment was done graphically by compering the distributional properties of the R^2-measures within two series simulation runs under both conditions: correctly specified vs. misspecified. Namely we

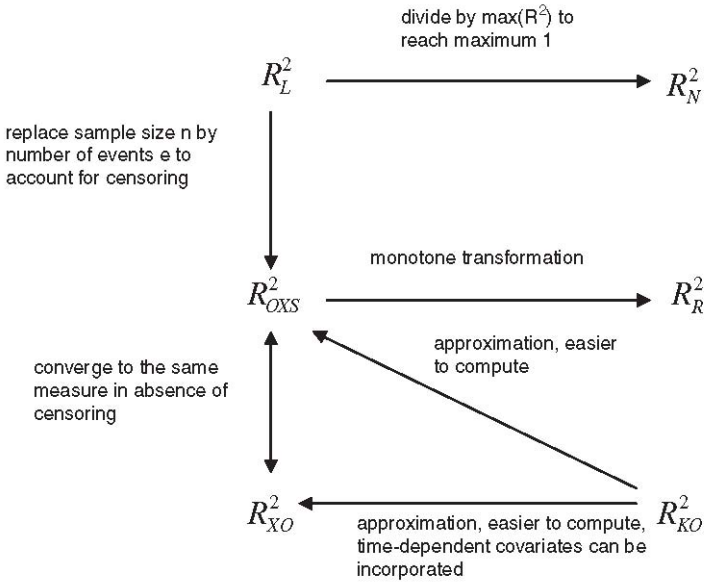

Figure 1: Graphical representation of some inter-relations between R^2-measures (modified from Hielscher et al. [16])

thereby considered missing response-related covariates, informative censoring and violation of the proportional hazards assumption as model misspecifications. These were investigated in the light of three different values of the additional moderating effects: amount of censoring, sample size and presence of non-informative covariates. For each of the considered conditions 100 data sets (= simulation runs) with different numbers of observations, as one of the moderating variables, were simulated. Additionally to graphical comparison the numerical agreement of the measures and the empirical confirmation of the above described relations between measures was rated. To evaluate the effect of considered misspecifications on treated R^2-measures, they had either to be included in model specification or already in data generation for the misspecified simulation runs. Applied data generation is described in detail below. However, the misspecificastion of a missing predictor covariate was the only one to be carried out in model specification. For the according simulation run simply one covariate was left out in model fitting due to be unknown. Subsequently considered R^2-measures were calculated as described above from accordingly estimated penalized partial log likelihood, which is given in equation 6. Appropriate values for the penalty parameter λ in parameter estimation was found by five-fold cross-validation. In order to prevent overly optimistic estimates due to possible over-fitting, all R^2-measures were computed on new data separate from the data that has been used for model fitting [41]. Every generated data set has therefore been randomly splitted 2:1 into training (n_{tr}) and test data set (n_{te}). Model fitting was done on the training data set and R^2-measures were computed from test data set. In the following we refer to the total number of observations in training plus test data set by $N = n_{tr} + n_{te}$.

Data Generation

Due to its definition through the hazard function, the simulation of appropriate survival times for the Cox proportional hazards model is not as straightforward as in linear regression models

[50]. A frequently used distribution for survival times in studies regarding the Cox model is the Weibull distribution [50], which is considered a special case of the Cox model. Random survival times $T \epsilon \mathbb{R}^+$, with $\mathrm{i} = 1, ..., \mathrm{N}$ realizations $\mathrm{t_i}$, were therefore generated according to a Weibull model conditional on a vector of d= 3 time-invariant continuous random predictor variables $X \epsilon \mathbb{R}^d$, using its log-linear representation form, which refers to the ATF (accelerated failure time) property of the Weibull distribution [34]:

$$T = \exp\left(\mu + \tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X} + \sigma\boldsymbol{\varepsilon}\right). \tag{28}$$

Here μ denotes the regression parameter for the constant term (intercept), which was set to zero for data generation. Vector of predictor covariates were previously drawn from a multivariate standard normal distribution $\mathbf{X} \sim N_d(\mu, \boldsymbol{\Sigma})$, with $\mu = 0^d$ and $\boldsymbol{\Sigma} = \mathbf{I}_d$, where $\mathbf{I}_d$ is the d-dimensional identity matrix. The true regression parameters $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\beta}_3)^{\mathrm{T}}$ were set to $\tilde{\beta}_1 = -0.25$, $\tilde{\beta}_2 = 0.25$, $\tilde{\beta}_3 = 0.5$ for all correctly specified and misspecified simulation runs. By choosing a standard extreme value distribution for the error term $\boldsymbol{\varepsilon}$ the resulting survival times follow a Weibull distribution $T \sim W(\lambda_{\mathrm{w}} \mathrm{e}^{-\gamma\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}}, \gamma)$ directly depending on $\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}$, in which λ_{w} and γ are the scale and shape parameter of the Weibull baseline hazard function [34]. Following Bender et al. [50] the needed standard extreme value distributed error term was obtained beforehand by generating a uniformly distributed random variable u, say, and inserting it to the density function $f(\mathrm{u}) = \exp(\mathrm{u} - \exp(\mathrm{u}))$. The vector of true response-related regression parameters $\boldsymbol{\beta}$ in a proportional hazards model directly corresponds to the vector of true regression parameters $\tilde{\boldsymbol{\beta}}$ in the log-linear Weibull model by $\boldsymbol{\beta} = -\tilde{\boldsymbol{\beta}}/\sigma$ [34], where σ denotes a scale parameter of the model in equation 28. For the scale parameter σ we separately implemented the values 0.3, 0.5 and 0.8, but we will only consider its influence on results up to the point that we choose the most convenient value to explore the effects of the other described misspecifications. All results are therefore presented at $\sigma = 0.5$, if not otherwise specified.

Three different rates of censoring (30%, 50% and 70%) were introduced in several simulation runs by generating independently exponential distributed random censoring times $C \sim Exp(\lambda_{exp})$, with $\mathrm{i} = 1, ..., \mathrm{N}$ realizations $\mathrm{c_i}$, and setting the event time to $Y := \min(T, C)$. The censoring rate was approximately controlled by the inserted value for the scale parameter λ_{exp} for each value of σ. Used values for λ_{exp} to produce the according censoring rates are shown in table 1. To assess

	noninformative censoring	informative censoring
$\sigma = 0.3$		
30% censoring	$\lambda_{exp} \approx 0.365$	$\lambda_{exp} \approx 0.33$
50% censoring	$\lambda_{exp} \approx 0.770$	$\lambda_{exp} \approx 0.64$
70% censoring	$\lambda_{exp} \approx 1.525$	$\lambda_{exp} \approx 1.115$
$\sigma = 0.5$		
30% censoring	$\lambda_{exp} \approx 0.385$	$\lambda_{exp} \approx 0.332$
50% censoring	$\lambda_{exp} \approx 0.845$	$\lambda_{exp} \approx 0.65$
70% censoring	$\lambda_{exp} \approx 1.74$	$\lambda_{exp} \approx 1.126$
$\sigma = 0.8$		
30% censoring	$\lambda_{exp} \approx 0.4$	$\lambda_{exp} \approx 0.317$
50% censoring	$\lambda_{exp} \approx 0.935$	$\lambda_{exp} \approx 0.620$
70% censoring	$\lambda_{exp} \approx 2.15$	$\lambda_{exp} \approx 1.075$

Table 1: used λ_{exp}-values for different rates of censoring

the influence of informative censoring, in the according data set a dependency on the survival times was introduced. The exponential distributed informative censoring times C_{inform} were generated depending on the ratio of the mean of previously generated survival times $\bar{t}$ and survival times t_i themselves, so that $C_{\text{infrom}} \sim Exp(\lambda_{exp} \cdot \bar{t}/t_i)$. Adjusted parameter values for λ_{exp} that are required to maintain the intended censoring rates of 30%, 50% and 70% could also be obtained from table 1.

To create an appropriate violation of proportional hazards assumption for according analysis, the generation of survival times directly was altered. Therefore on the one hand a dependency of the error term $\boldsymbol{\varepsilon}$ on the first predictor variable $\mathbf{x}_1$ was introduced in the alternatively used generation mechanism. On the other hand a randomly generated scale parameter σ was used, so that each generated survival time was scaled differently. Both modifications together led to a satisfying violation of proportional hazards assumption.

Because a situation of high-dimensional data should be explored, different data sets with p= 100, p= 1000 and p= 5000 noise-variables or features were generated. To implement a blockwise correlation structure of block size $m = 10$, they were drawn blockwise from a multivariate normal distribution with an accordingly specified covariance matrix Σ. Different data sets were also generated to apply $N = 600$, $N = 450$ and $N = 300$ observations as different values of for each misspecification and the correctly specified model.

Results

General Considerations

Figure 2 shows the distributional properties of all six considered R^2-measures under a correctly specified model in form of grouped boxplots. They are plotted for 30%, 50% and 70% of censored observations with p= 100 noninformative covariates and a sample size of $N = 600$. The plot in the middle of the figure represents the according boxplots for a value $\sigma = 0.5$. Subsequently following graphical comparisons of the distributional properties of R^2-measures for examined misspecifications were done proceeding from this conditions as a reference. We therefore briefly examine this condition first. Graphical assessment of the measures started from the same refernce conditions for the purpose of analysing type and degree of distributional similarities and differences of the measures more systematically. Although we mainly consider a value of $\sigma = 0.5$ for investigation of regarded misspecificastions, at this point we shall also take on a brief examination of the influence of σ on distributional properties of R^2-measures at least for correctly specified model preconditions. Figure 2 therefore also includes one plot for $\sigma = 0.3$ and $\sigma = 0.8$. Scaling parameter σ mainly alters generated strength of regression effects in two ways, as can be seen in equation 28. On the one hand smaller values of σ increase the generated values of the true regression parameters $\boldsymbol{\beta}$ by dividing the fixed values of $\tilde{\boldsymbol{\beta}}$ by a smaller value than one: $\boldsymbol{\beta} = -\tilde{\boldsymbol{\beta}}/\sigma$. On the other hand smaller values for σ result in a smaller random error term $\sigma\boldsymbol{\varepsilon}$ (equation 28). Both leads to a stronger regression effect for smaller values of σ and also results in higher association between response and covariates. All R^2-measures reflected this in subsequent higher values for smaller σ in good average agreement. Regarding the influence of σ should be noted that we cannot clearly confirm the trend pointed out by Hielscher et al. [16] that higher values of σ or lower model fit, respectively, leads to increased variability of R^2-measures from figure 2, because the highest variability seem to appear

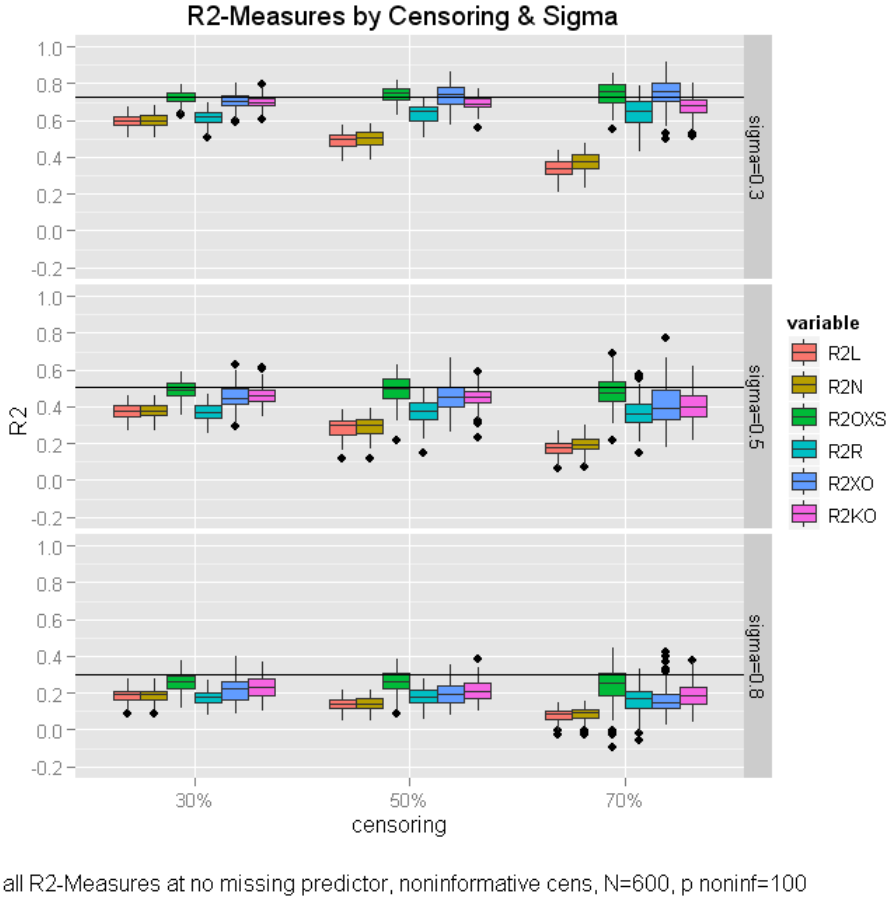

Figure 2: R^2-measures by Censoring an different values of σ

with $\sigma = 0.5$. The continuously drawn line in each plot in figure 2 indicates the corresponding convergence value of R^2-measures. Convergence values were determined empirically by the mean of 500 simulation runs with each $N = 7500$ observations for no censoring and no noise features for each implemented value of σ in a correctly specified model. Strength of implemented regression effects, resulting hazard ratios and convergence value of R^2-measures can be obtained from table 2, where convergence values are denoted by $E(R^2)$. However, while the measures R^2_L, R^2_N, R^2_{OXS} and R^2_{XO} agreed well up to the fifth decimal place, R^2_{KO} as well as R^2_R did not converge to the same value. Therefeore they are represented separately in table 2. While the convergence value of R^2_{KO} was slightly higher, the corresponding value of R^2_R was sizable lower for all three implemented values of σ. Certainly the convergence value of R^2_R could be expected to be lower, because this measure was intended to reproduce a measure of prediction error by including a correction term as described in section "Overview of R^2-Measures". This must be taken into account when evaluating R^2_R under several misspecificatons. In contrast, why the convergence value of R^2_{KO} is slightly but constantly higher is not comparable clear. We suppose a relation to the fact that the true model that was used for data generation is assumed in the calculation of R^2_{KO}.

Apart from the systematic shift in absolute R^2-values, which is related to strength of regression effects and results in different convergence values, some formerly recognized effects of the amount

	$\tilde{\boldsymbol{\beta}}$	$\boldsymbol{\beta} = -\tilde{\boldsymbol{\beta}}/\sigma$	$\psi = \exp(\beta)$	$E(\mathrm{R}^2_{\mathrm{L,N,OXS,XO}})$	$E(\mathrm{R}^2_{\mathrm{KO}})$	$E(\mathrm{R}^2_{\mathrm{R}})$
$\sigma = 0.3$						
β_1	-0.25	0.833	2.301			
β_2	0.25	-0.833	0.435			
β_3	0.5	-1.667	0.187			
$\beta_1 + \beta_2 + \beta_3$			0.187	0.61211	0.63676	0.48966
$\sigma = 0.5$						
β_1	-0.25	0.5	1.649			
β_2	0.25	-0.5	0.607			
β_3	0.5	-1.0	0.368			
$\beta_1 + \beta_2 + \beta_3$			0.368	0.43076	0.45447	0.31512
$\sigma = 0.8$						
β_1	-0.25	0.313	1.367			
β_2	0.25	-0.313	0.732			
β_3	0.5	-0.625	0.535			
$\beta_1 + \beta_2 + \beta_3$			0.535	0.25859	0.27240	0.17497

Table 2: relation between generated and estimated β-coefficients, hazard ratios and convergence values

of censoring can be confirmed from figure 2. $\mathrm{R}^2_{\mathrm{L}}$ as well as $\mathrm{R}^2_{\mathrm{N}}$ depend on censoring [8, 16, 24, 25], because their values systematically decrease with higher amount of censoring independently of strength of regression effects. As described above, it was explained by O'Quigley, Xu and Stare [24] to divide by the number of events k rather than by number of observations n in the corresponding statistic. More difficult to confirm in figure 2 is the predicated positively correlation of $\mathrm{R}^2_{\mathrm{OXS}}$ with the quantity of censored observations [16, 24], but a slight correlation seems to be within the range of possibility. From the plot for $\sigma = 0.3$ also a positive correlation of $\mathrm{R}^2_{\mathrm{XO}}$ with the amount of censoring could be suspected but is not confirmed for other values of σ. Also the stated independence from the amount of censoring of $\mathrm{R}^2_{\mathrm{KO}}$ [8, 16, 24] can be approved from figure 2. Additionally, a general trend over all R^2-measures with increasing amount of consoring becomes visible. A higher amount of censoring leads to higher standard deviation of the measures. Some measures are stronger affected than others. E.g. $\mathrm{R}^2_{\mathrm{XO}}$ and $\mathrm{R}^2_{\mathrm{OXS}}$ and in consequence $\mathrm{R}^2_{\mathrm{R}}$ incline to have higher dispersion. The property of $\mathrm{R}^2_{\mathrm{OXS}}$ to have a especially small standard error, as pointed out by O'Quigley, Xu and Stare [24] cannot be confirmed here. Instead $\mathrm{R}^2_{\mathrm{L}}$, $\mathrm{R}^2_{\mathrm{N}}$ and $\mathrm{R}^2_{\mathrm{KO}}$seems to have relatively small standard errors, although differences are not enormous.

Beyond the single influence of the considered conditions "amount of censoring" and "different values of σ" another smaller effect becomes apparent from figure 2. With increasing vlaues for σ R^2-measures diverge downwards from corresponding convergence value. However, all systematic deviations of the measures that are soleley caused by different σ would have been covered by the convergence value. Additionally the downwards divergence gets stronger with higher degree of censoring. An interaction effect of censoring and σ can therefore be ascertained. Hence, likelihood-based R^2-measures tend to get more conservative in their expected value with increasing amount of censoring and decreasing regression effects. But because imprecision also increases, single R^2-values can also be to optimistic. An additional effect of noise features could also not be excluded, because convergence values were determined without noninformative covariates and R^2-measures in figure 2 at $p = 100$ noninformative features.

	R^2_L	R^2_N	R^2_{OXS}	R^2_R	R^2_{XO}	R^2_{KO}
30% censoring						
R^2_L	1.000					
R^2_N	1.000	1.000				
R^2_{OXS}	0.268	0.272	1.000			
R^2_R	0.952	0.950	0.278	1.000		
R^2_{XO}	0.083	0.085	0.154	0.095	1.000	
R^2_{KO}	0.084	0.085	0.212	0.091	0.869	1.000
50% censoring						
R^2_L	1.000					
R^2_N	0.995	1.000				
R^2_{OXS}	0.150	0.159	1.000			
R^2_R	0.458	0.492	0.432	1.000		
R^2_{XO}	0.005	0.005	−0.017	−0.001	1.000	
R^2_{KO}	0.020	0.021	0.052	0.047	0.873	1.000
70% censoring						
R^2_L	1.000					
R^2_N	0.901	1.000				
R^2_{OXS}	0.065	0.083	1.000			
R^2_R	0.140	0.188	0.518	1.000		
R^2_{XO}	0.009	0.011	0.018	0.037	1.000	
R^2_{KO}	0.010	0.014	0.065	0.098	0.830	1.000

Table 3: numerical agreement of the measures by concordance correlation coefficient (ccc) for $\sigma = 0.5$

Numerical agreement of considered R^2-measures under a correctly specified model was assessed by the concordance correlation coefficient (ccc) [51, 52, 53]. According values for $\sigma = 0.5$ and different qunatities of censoring can be obtained from table 3. High numerical agreement was detected between R^2_L and R^2_N, as well as between R^2_{XO} and R^2_{KO}. This confirms in both cases their strong theoretical link, described above. While the perfect agreement between R^2_L and R^2_N diminished slightly with increasing censoring, the association between R^2_{XO} and R^2_{KO} seems to be largely independent of censoring, which could have been led back to the fact that the letter measures are independent from censoring themselves. Possibly, a good agreement between R^2_L and R^2_N, measured by the concordance correlation coefficient, could have been expected, because they are interconnected only by rescaling (see equation 12). The good association between R^2_{XO} and R^2_{KO} on the contrary underpins the theoretical considerations of Xu and O'Quigley [23] that led to their proposition of R^2_{XO} briefly summerized above. Apart from these exceptions numerical agreement among considered R^2-measures generally seems to be low. Initially this could appear quite surprising, because it partly contradicts existing literature. Schmeper and Stare [8] briefly state that the R^2-measure proposed by Maddala [43], which is essentially the same as R^2_L, is generally close to R^2_{KO}. O'Quigley, Xu and Stare [24] outline a good agreement of R^2_{OXS}, R^2_{XO} and R^2_{KO} from a former simulation study, even in the prensence of heavy censoring. This can only be confirmed for R^2_{XO} and R^2_{KO} as described above, when compared with table 3. Admittedly both of these studies made their conclusion from a simulation of large sample sizes with $N = 5000$ and only one predictor covariate. Note, for reasons of comparison, that in our simulation runs that were used to obtain convergence values with $N = 7500$ all concordance correlation coefficients took a value

of one for the measures that converged to the same R^2-value. However, we also did not introduce censoring in these data sets. Summarizing table 3 we agree with the statement of Hielscher et al. [16] to this issue: "The different underlying mechanisms of all the measures result in different values of R^2 coefficients for identical scenarios and prognostic models. Thus one cannot interpret these measures with respect to some absolute scale".

Missing Covariates

Figure 3 depicts the behavior of the distributional attributes of the considered R^2-measures by amount of censoring and sample size for a scenario of a correctly specified model (left side) as well as for a model with a missing predictor covariate (right side). Therefore this figure shows on the upper left side the same grouped boxplot diagramm for 30%, 50% and 70% of censored observations, with p= 100 noninformative covariates, a sample size of N = 600 and a value of $\sigma = 0.5$ as the middle plot in figure 2.

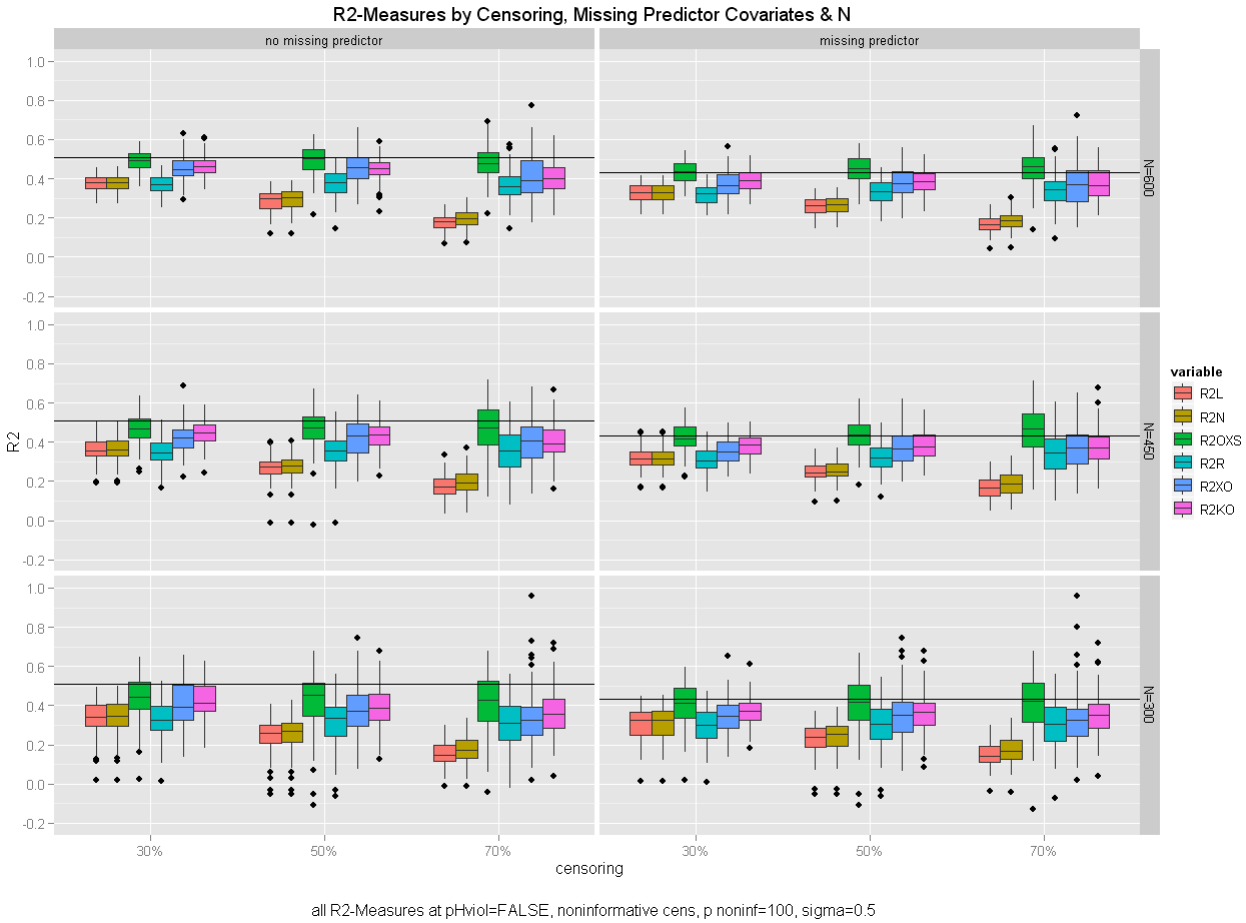

Figure 3: R^2-measures by Censoring, Missing Predictor Covariates & N

- Darstellung aller 3 Diargramme mit "Missing Covariates" (N, Noise,Sigma)

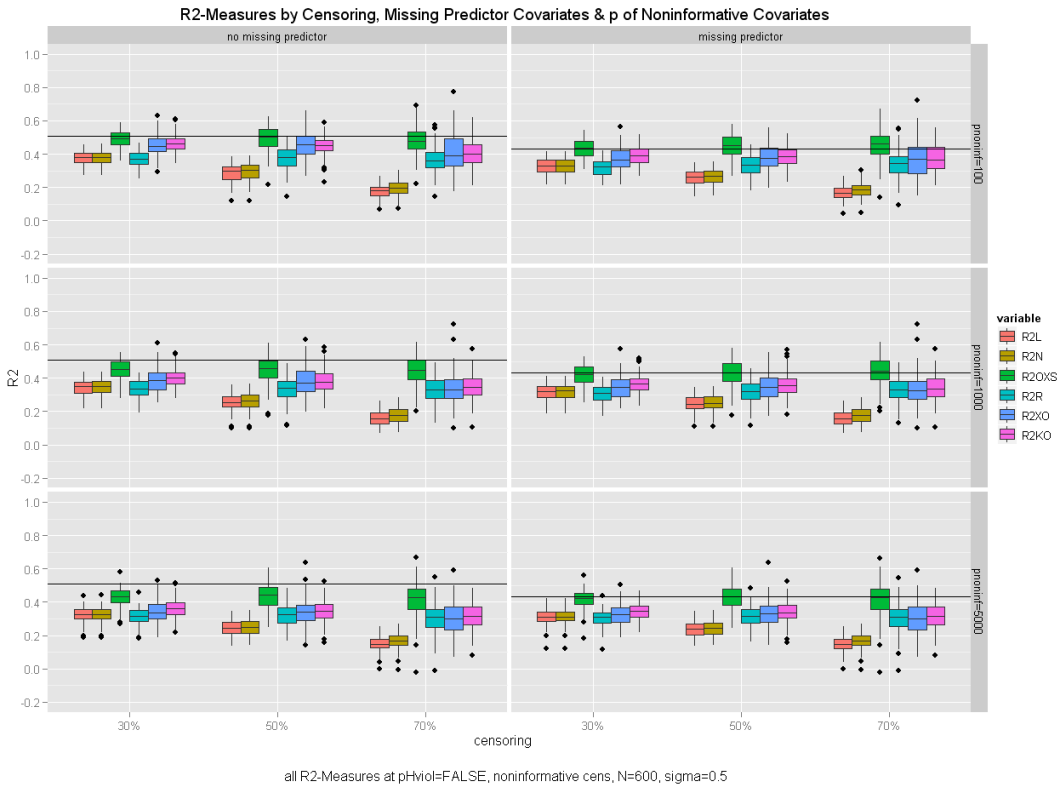

Figure 4: R^2-measures by Censoring, Missing Predictor Covariates & p of Noninformative Covariates

- Text

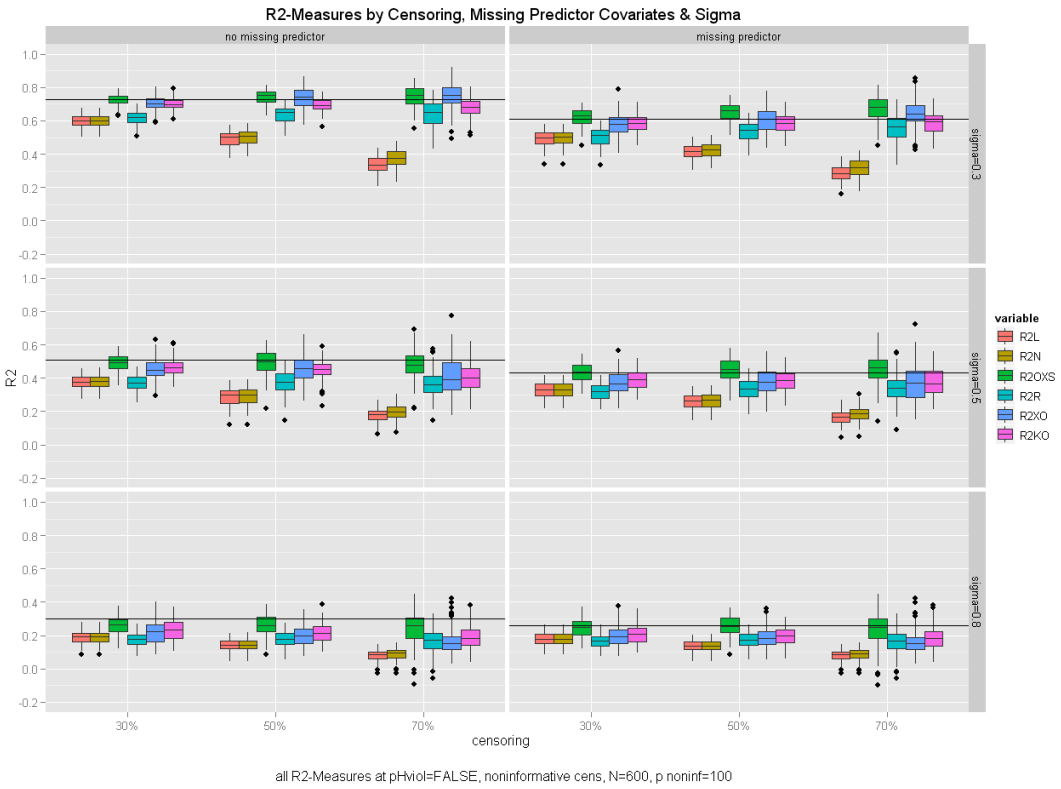

Figure 5: R^2-measures by Censoring, Missing Predictor Covariates & Sigma

Informativ Censoring

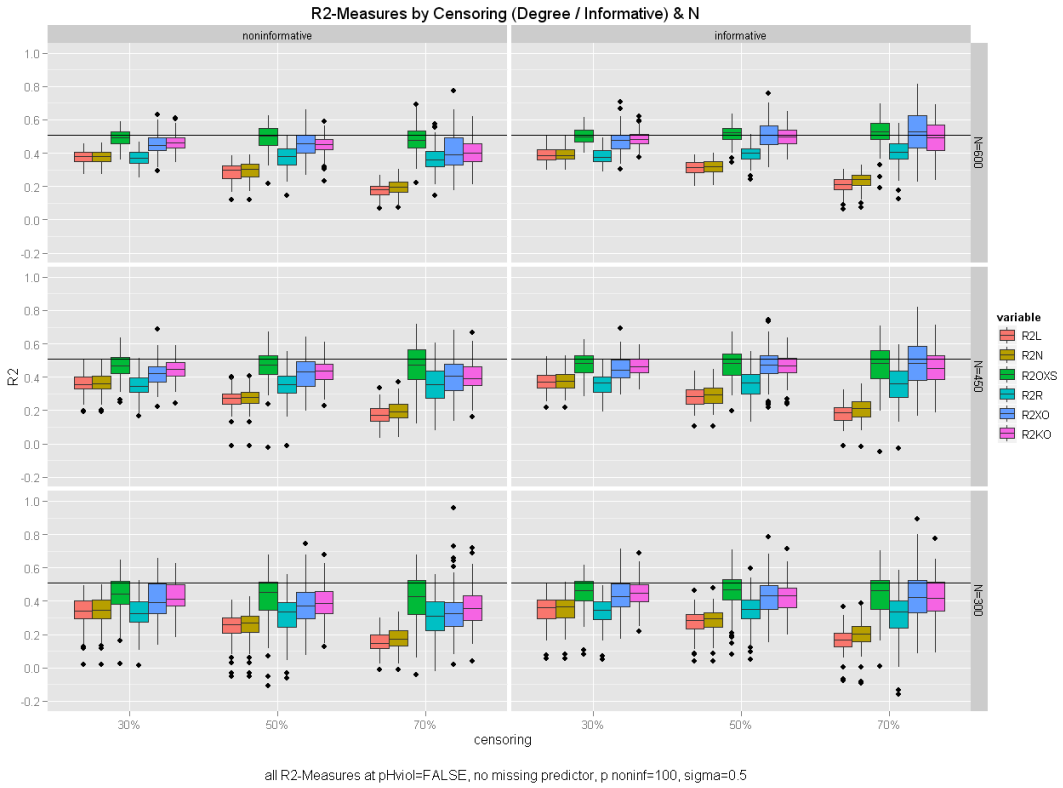

Figure 6: R^2-measures by Censoring (Degree / Informative) and N

- Darstellung aller 3 Diargramme mit "Informative Censoring" (N, Noise,Sigma)

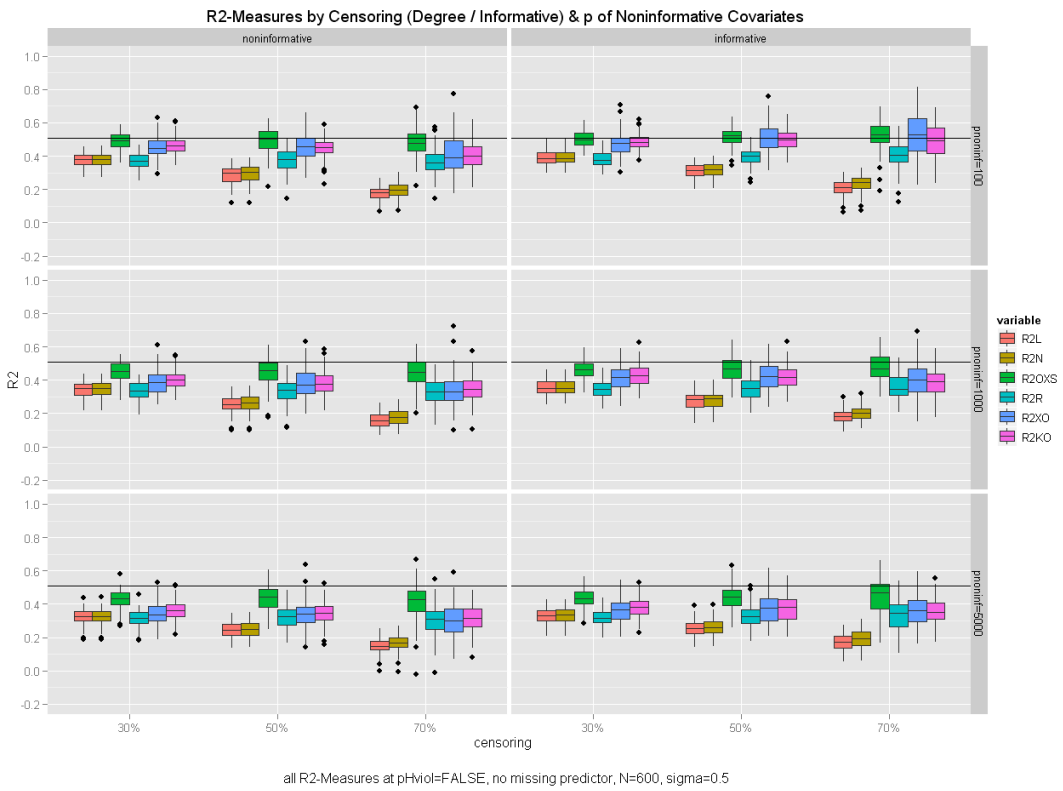

Figure 7: R^2-measures by Censoring (Degree / Informative) & p of Noninformative Covariates

- Text

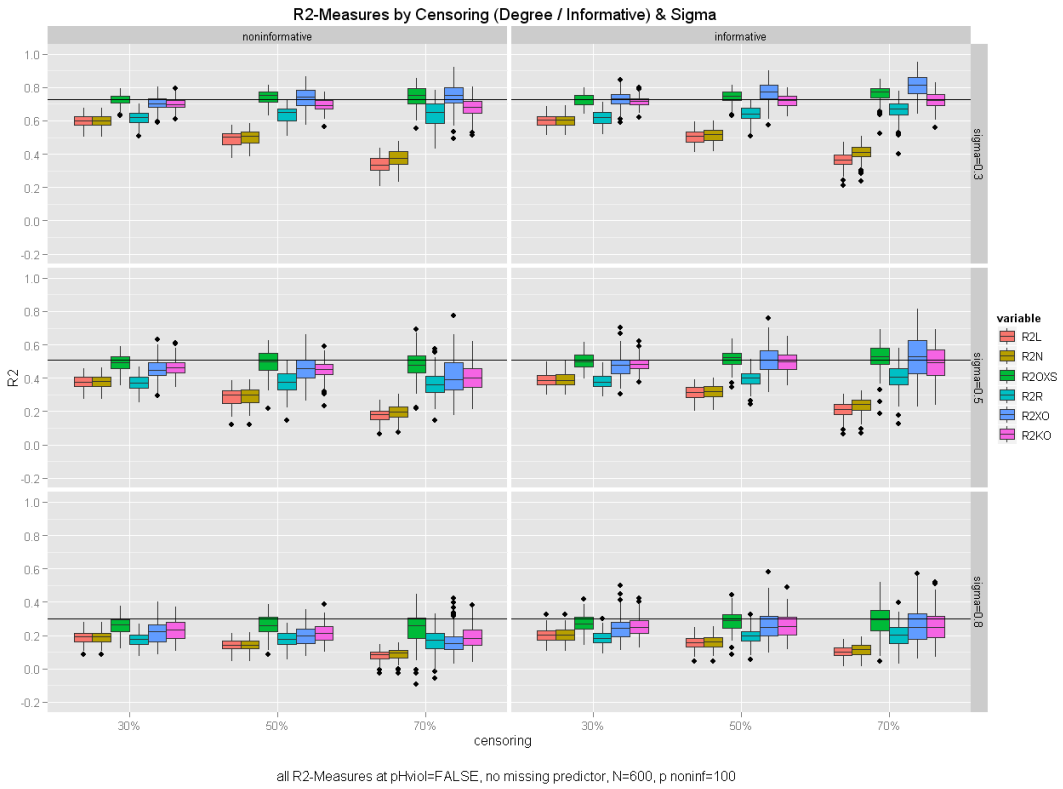

Figure 8: R^2-measures by Censoring (Degree / Informative) & Sigma

Violation Of PH-Assumption

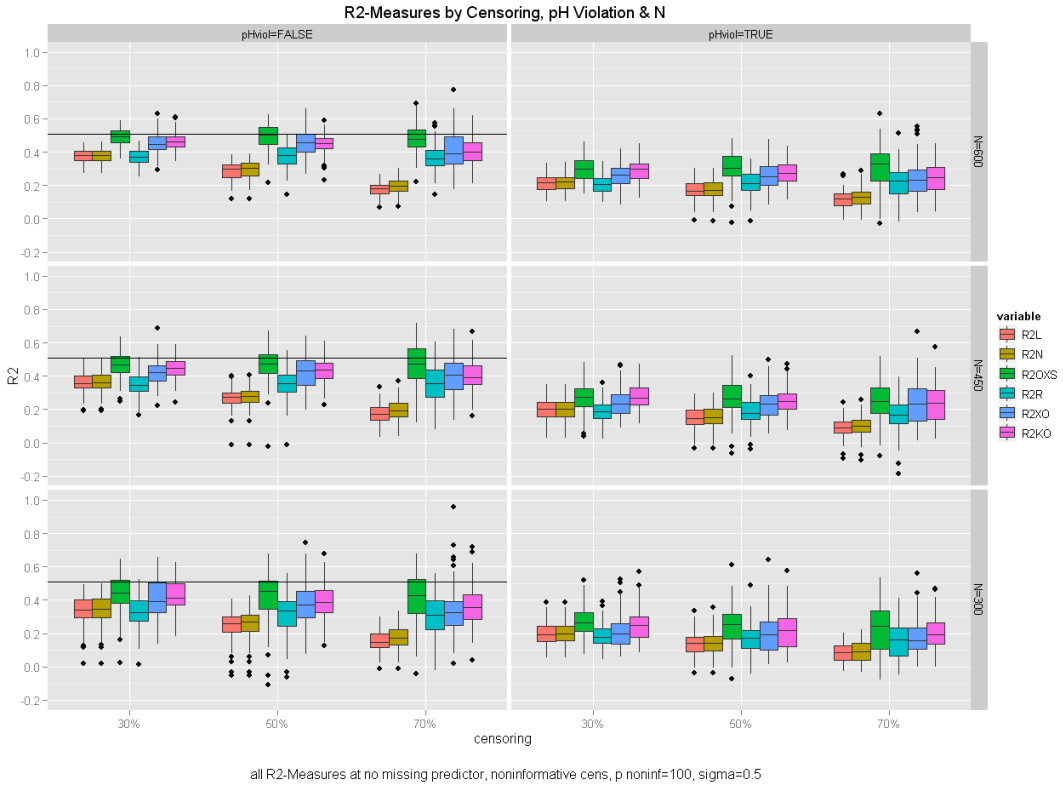

Figure 9: R^2-measures by Censoring, PH-Violation & N

- Darstellung aller 3 Diargramme mit "Violation of PH-Assumption" (N, Noise,Sigma)

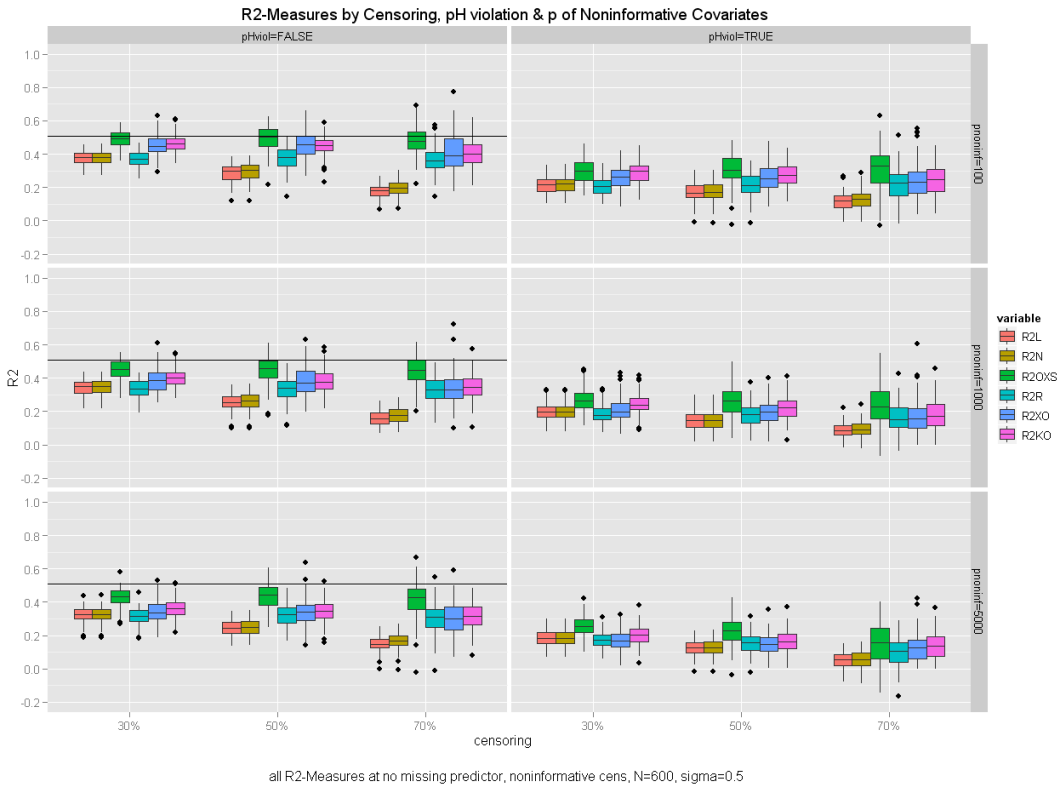

Figure 10: R^2-measures by Censoring, PH-Violation & p of Noninformative Covariates

- Text

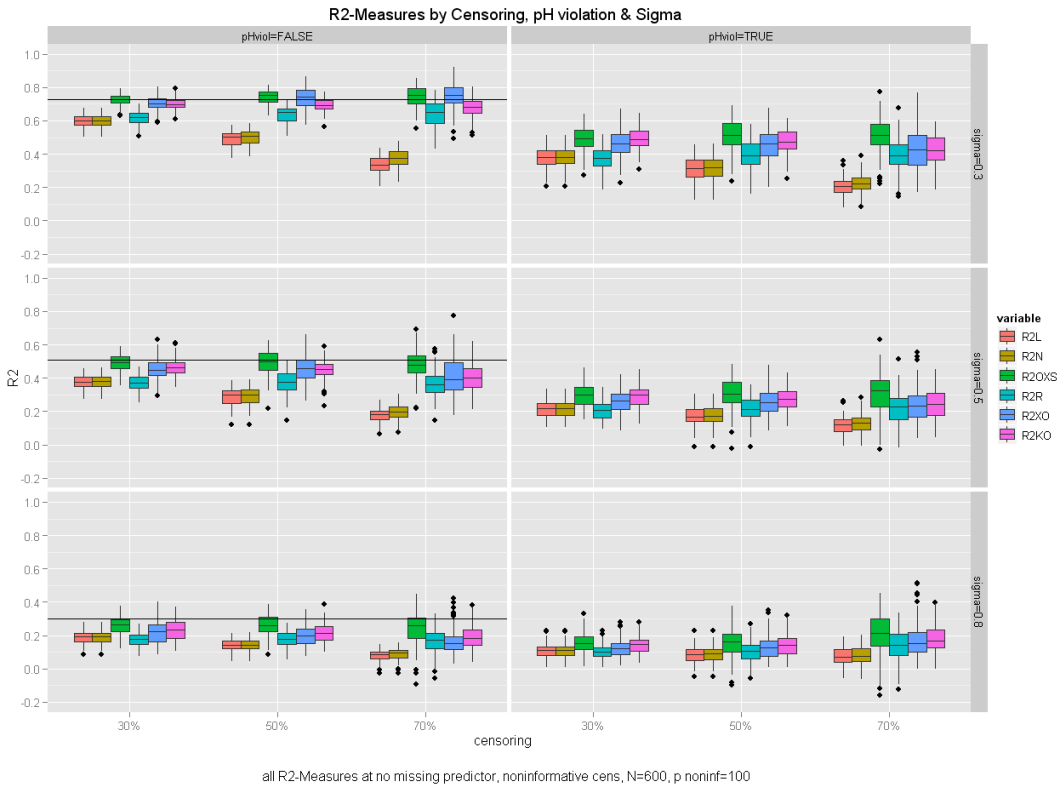

Figure 11: R^2-measures by Censoring, PH-Violation & Sigma

(Selected) Combinations / Interactions

Summary and Discussion

Convenient properties for performance measures in survival analysis were defined e.g. by Schemper & Stare [8], Royston and Sauerbrei [4] or Royston [25]. Summarized they include (a) that R^2 increases with strength of association between covariates and outcome and provides the nesting property, i. e. in nested models M1 $\subset$ M2 ($'\subset$' denoting nesting) follows $R^2(M1) < R^2(M2)$, (b) approximate independence from the amount of censoring in terms of unbiasedness, although lower precision would be expected, (c) invariance for monotonic transformation of time scale, (d) close relationship or "numerical consistency" to ordinary R^2 that would be obtained in a fictive equivalent linear regression analysis with the same data set, (e) an intuitively clear interpretation, (f) the availability of confidence intervals and (g) robustness against incorrectly specified models. Subsidiary computational simplicity may have advantages, if all other properties are equally satisfied.

In our view the more frequently case in practical application of Cox model is the one of a not correctly specified model, because nearly never all informative covariates are known.

- cardinally /general properties and advantages of likelihood based approaches (see Nagelkerke, Magee, Schemper & Stare)
  - general liability of likelihood-based approaches to overfiitting, because they are calculated from models likelihood

- transferability of properties of R^2-measures to high dimensional data with LASSO-estimation
- not considered misspecifications / situations:
  - nonlinear explained / unexplained effects ?
- generizability of results
- summarizing table of affectations (e.g. 3 categories) of each measure by certain misspecification
- ev. table of compliance to defined measure-properties (a-g) +
  - possibility for time dependent covariates
  - possibility of general application in other than survival models
- x

References

[1] Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA: **Evaluating the yield of medical tests'**. *Journal of the American Medical Association* 1982, **247**:2543–2546.

[2] Harrell FE, Lee KL, Mark DB: **Tutorial in Biostatistics. Multivariable Prognostic Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors**. *Statistics in Medicine* 1996, **15**:361–387.

[3] Kattan MW: **Judging new markers by their ability to improve predictive accuracy**. *Journal of the National Cancer Institute* 2003, **95**:634–635.

[4] Royston P, Sauerbrei W: **A new measure of prognostic separation in survival data**. *Statistics in Medicine* 2004, **23**:723–748.

[5] Steyerberg E, Vickers A, Cook N, Gerds T, Gonen M, Obuchowski N, Pencina M, Kattan W: **Assessing the Performance of Prediction Models. A Framework for Traditional and Novel Measures**. *Epidemiology* 2010, **21**:128–138.

[6] Wyatt JC, Altman DG: **Prognostic Models: clinically useful or quickly forgotten?** *British Medical Journal* 1995, **311**:1539–1541.

[7] Muers MF, Shevlin P, Brown J: **Prognosis in lung cancer: Physicians' opinion compared with outcome and a predictive model**. *Thorax* 1996, **51**:894–902.

[8] Schemper M, Stare J: **Explained variation in survival analysis**. *Statistics in Medicine* 1996, **15**:1999–2012.

[9] Henderson R, Jones M, Stare J: **Accuracy of point predictions in survival analysis**. *Statistics in Medicine* 2001, **20**:3083–3096.

[10] Witten D, Tibshirani R: **Survival analysis with high-dimensional covariates**. *Statistical Methods in Medical Research* 2010, **19**:29–51.

[11] Henderson R: **Problems and prediction in survival-data analysis**. *Statistics in Medicine* 1995, **14**:161–184.

[12] Boulesteix AL, Sauerbrei W: **Added predictive high-throughput molecular data to clinical data and its validation**. *Briefings in Bioinformatics* 2001, **bbq085v1-bbq085**:1–15.

[13] Pepe S, Janes H, Longton G, Leisenring W, Newcomb P: **Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker**. *American Journal of Epidemiology* 2004, **159**(9):882–890.

[14] Simon R, Radmacher M, Dobbin K, McShane L: **Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification**. *Journal of the National Cancer Institute* 2003, **95**:14–18.

[15] Bovelstad HM, Nygard S, Storvold HL, Aldrin M, Borgan O, Frigessi A, Lingjarde OC: **Prediction survival from microarray data - a comparative study**. *Bioinformatics* 2007, **23**:2080–2087.

[16] Hielscher T, Zucknick M, Werft W, Benner A: **On the prognostic value of survival models with application to gene expression signatures**. *Statistics in Medicine* 2010, **29**:818–829.

[17] Schmid M, Hielscher T, Augustin T, Gefeller O: **A Robust Alternative to Shemper-Henserson Estimator of Prediction Error**. *Biometrics* 2011, **no. doi: 10.1111/j.1541-0420.2010.01459.x**.

[18] Bovelstad HM, Ornulf B: **Assessment of evaluation criteria for survival prediction from genomic data**. *Biometrical Journal* 2011, **53**(2):202–216.

[19] Cox D, Snell E: *The Analysis of Binary Data*. Chapman & Hall, London, second ed edition 1989.

[20] Kent J, O'Quigley J: **Measures of dependence for censored survival data**. *Biometrica* 1988, **75**:525–534.

[21] Magee L: **R2 measures based on Wald and likelihood ratio joint significance tests**. *American Statistican* 1990, **44**:250–253.

[22] Nagelkerke N: **A note on a general definition of the coefficient of determination**. *Biometrica* 1991, **78**:691–692.

[23] Xu R, O'Quigley J: **A measure of dependence for propotional hazards models**. *Journal of Nonparametric Statistics* 1999, **12**:83–107.

[24] O'Quigley J, Xu R, Stare J: **Explained randomness in proportional hazards models**. *Statistics in Medicine* 2005, **24**:479–489.

[25] Royston P: **Explained variation for survival models**. *The Stata Journal* 2006, **6**:83–96.

[26] Heagerty PJ, Zheng Y: **Survival model predictive accuracy and ROC curves**. *Biometrics* 2005, **61**:92–105.

[27] Pepe M, Zheng Y, Jin Y, Huang Y, Parikh C, Levy W: **Evaluating the ROC performance of markers for future events**. *Lifetime Data Analysis* 2008, **14**:86–113.

[28] Graf E, Schmoor C, Sauerbrei W, Schumacher M: **Assessment and comparision of prognostic classification schemes for survival data**. *Statistics in Medicine* 1999, **18**:2529–2545.

[29] Schemper M, Henderson R: **Predictive accuracy and explained variation in cox regression**. *Biometrics* 2000, **56**:249–255.

[30] Gerds T, Schumacher M: **Consistent estimation of the expected Brier score in general survival models with right-censored event times**. *Biometrical Journal* 2006, **48**:1029–1040.

[31] Korn E, Simon R: **Measures of explained variation for survival data**. *Statistics in Medicine* 1990, **9**:487–503.

[32] Cox D: **Regression models and life tables (with discussion)**. *Journal of the Royal Statistical Society, Series B* 1972, **74**:187–220.

[33] Therneau T, Grambsch P: *Modeling Survival Data. Extending the Cox Model.* Springer, New York, Berlin, Heidelberg 2000.

[34] Collett D: *Modelling Survival Data in Medical Research.* CRC: Chapman & Hall, 2. edition 2003.

[35] Tibshirani R: **Regression shrinkage and selection via the LASSO**. *Journal of the Royal Statistical Society, Series B* 1996, **58**:267–288.

[36] Tibshirani R: **The LASSO method for variable selection in the Cox model**. *Statistics in Medicine* 1997, **16**:385–395.

[37] Hastie T, Tibshirani R, Friedman J: *The Elements of Statistical Learning.* Springer, New York, 2. ed edition 2009.

[38] Datta S, Le-Rademacher J, Datta S: **Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and LASSO**. *Biometrics* 2007, **63**:259–271.

[39] Tibshirani R: **Univariate shrinkage in the Cox model for high dimensional data**. *Statistical Applications in Genetics and Molecular Biology* 2009, **8**:Article 21.

[40] Goeman J: **L1 penalized estimation in the cox proportional hazards model**. *Biometrical Journal* 2010, **52**:70–84.

[41] Verweij P, Houwelingen H: **Crossvalidation in survival analysis**. *Statistics in Medicine* 1993, **12**:2305–2314.

[42] Efron B: **Regression and ANOVA with zero-one data: measures of residual variations'**. *Journal of the American Statistical Association* 1978, **73**:113–121.

[43] Maddala G: *Limited-dependent and Qualitative Variables in Econometrics.* Cambridge University Press 1983.

[44] Kullback S, Leibler R: **On information and sufficiency**. *Annals of Mathematical Statistics* 1951, **22**:79–86.

[45] Kent J: **Information gain and general measure of correlation**. *Biometrica* 1983, **70**:163–173.

[46] Heinzl H: **Using SAS to calculate the Kent and O'Quigley measure of dependence for Cox proportional hazards regression model**. *Computer Methods and Programs in Biomedicine* 2000, **63**:71–76.

[47] Heinzl H, Stare J, Mittlböck M: **A Measure of dependence for the stratified Cox Proportional Hazards Regression Model**. *Biometrical Journal* 2002, **44**:671–682.

[48] Schoenfeld D: **Partial residuals for the proportional hazards regression model**. *Biomeirika* 1982, **69**:239–241.

[49] R Development Core Team: **R: A Language and Environment for Statistical Computing** 2009.

[50] Bender R, Augustin T, Blettner M: **Generating survival times to simulate Cox proportional hazards model**. *Statistics in Medicine* 2005, (24):1713–1723.

[51] Lin L: **A concordance correlation coefficient to evaluate reproducibility**. *Biometrics* 1989, **45**:255–268.

[52] Lin L: **Assay validation using the concordance correlation coefficient**. *Biometrics* 1992, **48**:599–604.

[53] Lin L: **Correction: A note on concordance correlation coefficient**. *Biometrics* 2000, **56**:324–325.