Appendix 1 –statistics

General aspects

The overall aim was to compare ADL with the best possible model containing customary risk factors.
To achieve this, two models were developed and compared, the “model without ADL” and the “full model”.
Data was originally stored in an SPSS file. All data analysis was performed in R [1]. Code is provided in appendix 2.

1. Outcome

Generally, it is important to describe the quantity of cases with missing outcome and to determine if there are any underlying patterns. Otherwise, simple exclusion may affect representativity [2].

In the study
Survival status was determined using the local region’s electronic registry on the 6th February 2014. The time variable was defined as days from discharge to death or censoring at study endpoint, whichever came first. Those surviving at endpoint had been followed for a median of 1428 days (range 1312-1548). The baseline survival function is shown in figure e1.

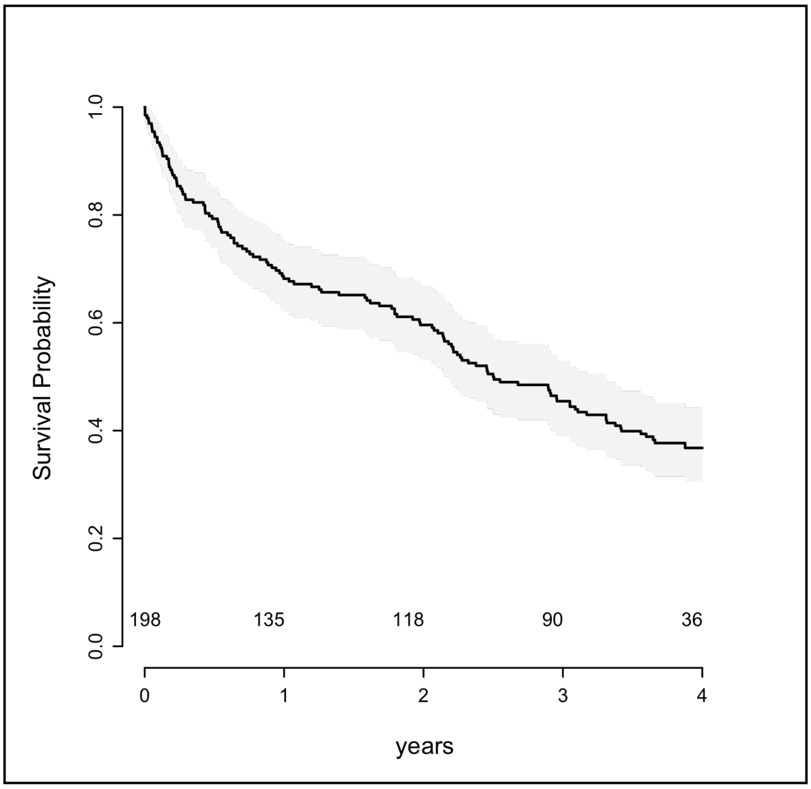

figure e1. Baseline survival function.

In our study, two cases were missing survival status due to having moved abroad (no longer in the region's registry) Hypothetically, these cases could be assumed to be in better health (severely diseased patients are unlikely to move abroad?). However, they were considered too few to affect representativity and were discarded from further analysis. Thus, the number of cases decreased from 200 to 198.

2. Crude analysis

Before any modifications are done to a variable, a crude analysis for the intented outcome could be of interest, to obtain an initial estimate of the effect of the predictor

In the study
Bivariate Cox proportional hazards regressions were carried out separately for all variables, including only outcome and the variable. All variables were treated in their original form, on their original scale. Observations with missing values were excluded from crude analysis. Data is presented with β coefficients, Standard errors, Wald χ^2, p value and hazard ratios in table 2 in the article.

In the crude analysis, all variables/potential predictors were statistically significant except sex, control/intervention status in the original study and the ADL item "food intake". Regarding the latter, the distribution was severely skewed, with only 18 cases (9%) having a non-zero value. To obtain a preliminary ranking of importance, the variables were sorted by decreasing Wald χ^2 in the table in the article.

In crude analysis, all separate GBS-ADL items were included but in further multivariate analysis, only the total GBS-ADL score was used, to avoid fitting too many variables and multicolinearity (the ADL items were intercorrelated at r = 0.8-0.9)

3. Missing data

In general, it is important to analyse missing data patterns in predictors. The first step is to determine the quantity of missing data. The second step is whether data is missing completely at random or if there are underlying patterns. When these prerequisites have been fulfilled, there are several approaches to missing data:

1. Listwise deletion, discarding all observations with any missing data points. The advantage of this approach is that no "manipulation" is done. Therefore, this method may seem intuitively most correct. The obvious disadvantage is that sample size could be substantially diminished. In addition, representativity could be affected, if missing a variable is systematically associated with other characteristics.

2. Using simple imputation. This technique substitutes missing values with the mean, mode or median value. This could be acceptable only if the variable is missing completely at random and the percentage of missing values small.

3. Using a more complex imputational technique. This approach uses customised regression models including all other covariates to obtain a stable prediction of the missing values. This method has been described and emphasized in several publications [3-7]. When using complex imputations, single or multiple imputations could be chosen. In the latter case, a separate

dataset is analysed for each imputational iteration, leading to a much larger complexity in the analysis.

In the study
When analysing the quantity of missing data, eGFR was missing in one case, BMI in three cases. Albumin was missing in 17 (9%) cases. BNP was missing in 113 (56%) observations.

The BNP variable was discarded from further analysis, as it had more than 50% missing. BMI and eGFR were considered missing completely at random. However, we found that cases with missing albumin were predominantly female (15 female vs. 2 male, $\chi^2 = 3.32$, p = 0.056) and had lower score on Charlson comorbidity index (1.47 vs. 2.33, F = 11.3, p = 0.002). Thus, excluding cases with missing albumin would affect representativity. Discarding the albumin variable would affect the overall aim, to compare ADL with the best possible traditional model. Therefore, the missing values in BMI, eGFR and Albumin were imputed using a single conditional imputation method (with the `transcan` function in R). In total, the effect of imputations was very small on the variable properties, as shown below.

Variable	β	S.E	Wald X2	p value	HR (95% CI)
Albumin, g/L, n = 181	-0.064	0.018	13.1	<0.001	0.94 (0.90 - 0.97)
- with imputation, transcan	-0.066	0.017	14.7	<0.001	0.94 (0.91 - 0.97)
eGFR, ml/min, n = 197	-0.029	0.005	29.3	<0.001	0.97 (0.96 - 0.98)
- with imputation, transcan	-0.029	0.005	29.2	<0.001	0.97 (0.96 - 0.98)
BMI, kg/m^2 , n = 195	-0.053	0.020	7.4	0.007	0.95 (0.91 - 0.99)
- with imputation, transcan	-0.053	0.020	7.4	0.006	0.95 (0.91 - 0.99)

Table e1. Effect of imputation on variable properties.

4. Variable considerations

Extreme outliers
In regression, outliers may be defined as observations with more than 3 interquartile ranges over the third quartile or below the 1st quartile. Such extreme values may affect a regression model significantly. First data entry errors should be considered and pursued. Then the biological plausibility should be considered. If plausible, we may consider a truncation at the 99th or 1st percentile [8].

In the study
In our study, data screening revealed, that for eGFR there was one extreme outlier with an estimated value of 198 ml /min (> 6 IQR over 3rd quartile), see boxplot.

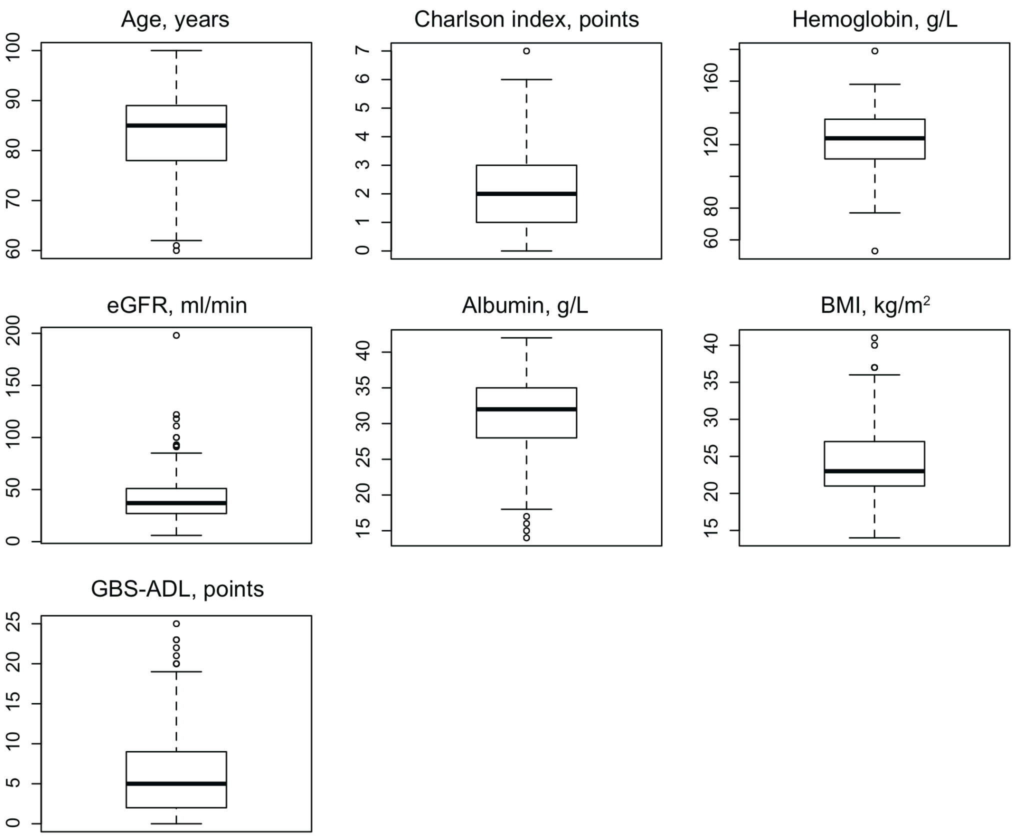

Figure e2. Boxplots of the continuous predictors. eGFR = Glomerular filtration rate, BMI = Body mass index, ADL = Activities of daily living.

This case was screened for data entry errors but none were found. Regarding biological plausibility, eGFR was measured with the Cockcroft-Gault formula ((140-age) * weight * constant)/Serum Creatinine in µmol/L, where the constant is 1.23 for men and 1.04 for women. Thus, GFR was not measured directly, but estimated and sensitive to extreme values in both serum creatinine, age and body weight. With this reservation, we considered the value to be biologically plausible. However, we did not consider it clinically important to compare one elderly patient with 198 ml/min in eGFR with another with 120 ml/min with regard to mortality. Therefore, eGFR was winsorized at the 99th percentile (118 ml/min). This led to a slightly improved fit in univariate performance.

Variable	β	S.E	Wald X2	p value	HR (95% CI)
eGFR, ml/min	-0.029	0.005	29.2	<0.001	0.97 (0.96 - 0.98)
- winsorized at 99th percentile	-0.029	0.005	29.8	<0.001	0.97 (0.96 - 0.98)

Table e2. effect of winsorization on variable properties.

Non-linearity

Most regression model assume that the predictors are linearly related to the outcome. However, non-linear relationships, such as U-shapes, for continuous variables are common.

There are several ways to address non-linearity:

First, assuming that the variable is linear. The advantage of this approach is that it results in an easily interpreted main effect, for example the Hazard Ratio in survival analysis. This is the approach used in our crude comparisons. However, the approach is potentially problematic. For hemoglobin, this would mean that the risk difference between two individuals with 170 and 130 g/L would be the same as between two with 90 and 50, respectively. In addition, this approach cannot handle U-shaped risks, it is likely that someone with 200 g/L in Haemoglobin with dehydration or polycytemia does not have better survival than someone with 140 g/L

Second, to dichotomize the variable, using a previously established cut-off, is another frequently used approach. However it is not recommended as it ignores a lot of information [9]. In our example, applying the WHO cut-off for anemia (120 g/L for women, 130 for men) would attribute the same risk for an individual with Hb of 119 g/L as for one with 53 g/L (the lowest in our material).

Third, to categorise the variable into categories that are clinically important, creating dummy variables. This approach could handle U-shaped risks. However, previously defined clinically important categories are needed and several degrees of freedom is spent in the analysis. As with a dichotomous transformation, all cases within a category are attributed the same risk.

Fourth, to use a more complex fitting function, such as a restricted cubic spline [10, 11]. This approach uses so called knots, point estimates where the risk is determined. A cubic function is used to fit the function between knots. Near the ends the risk is modelled linear.

In the study

We prespecified Hemoglobin to be non-linear and tried the approaches above, see figure e4. We decided to use the 4-knot restricted cubic spline as both the best performance and was most appropriate from a clinical perspective. The knots were placed at the 5th, 35th, 65th and 95th percentiles where Hb was 92.25, 118, 130 and 148.15, respectively. The resulting function to fit Hb was:

```
2.2712251-0.017758194* hb-6.2295666e-06*pmax(hb-
92.25,0)^3+5.8240197e-05*pmax(hb-118,0)^3-7.7559735e-
05*pmax(hb-130,0)^3+2.5549104e-05*pmax(hb-148.15,0)^3
```

As opposed to the easily interpreted hazard ratio from the linear function, this is not easy to interpret without a graph, the graphic display of the four approaches is presented in figure e3.

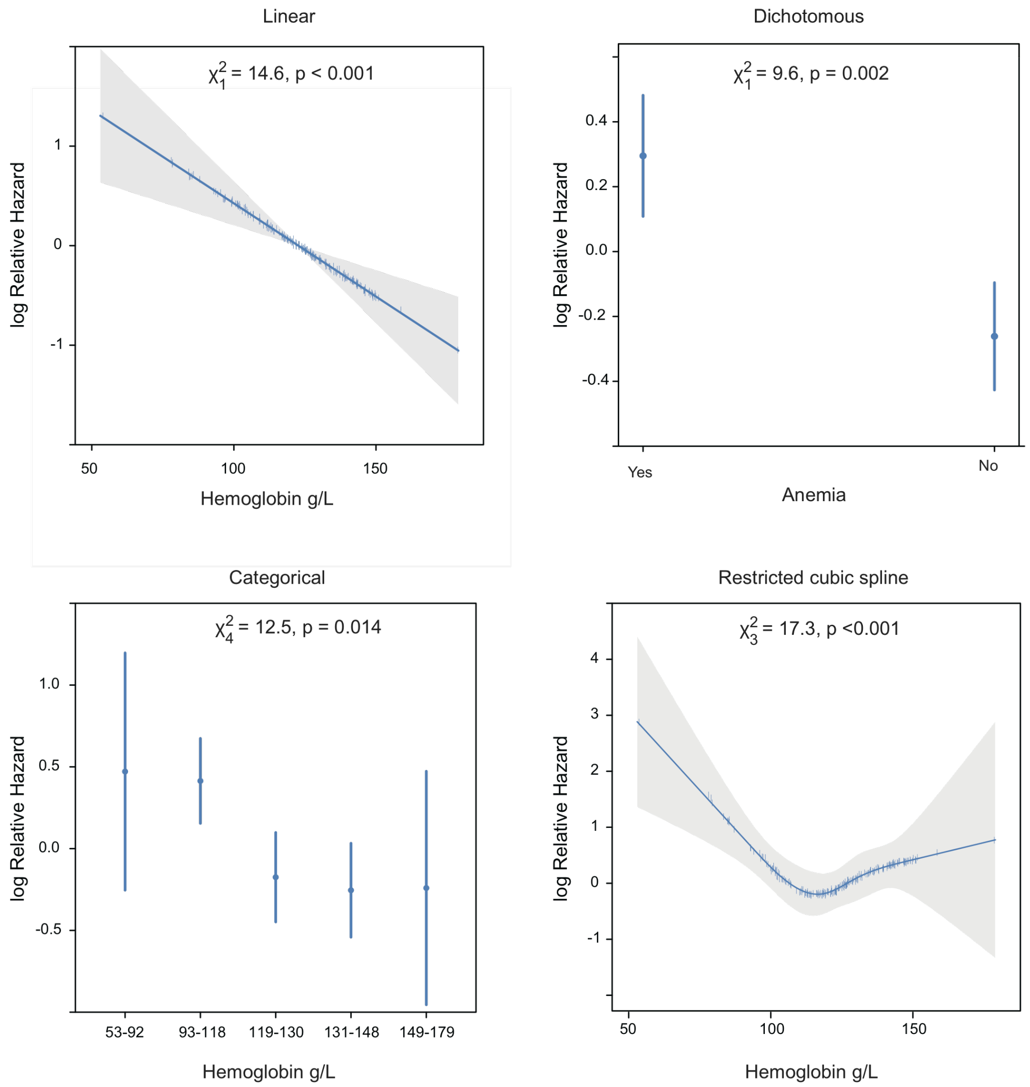

Figure e3. Different transformations of Hemoglobin. For dichotomous, the WHO definition of anemia is used. For categorical, the 5th, 35th, 65 and 95th percentiles were used, for easier comparison with the spline fit.

Apart from Hemoglogin, all other variables were bivariately tested for non-linearity by using 4-knot splines followed by ANOVA tests to determine if there was a significant non-linear component. GBS-ADL showed significant non-linearity and different codings were tested. We tested dichotomizing at the median and categorizing at the quartiles. A polynomial showed good fit but was not clinically plausible, with decreasing risks at the higher end of ADL impairment. The restricted cubic spline resembled a log fit and indeed the log fit was chosen, with fewer degrees of freedom spent, see figure e4. No other variables showed significant non-linear effects.

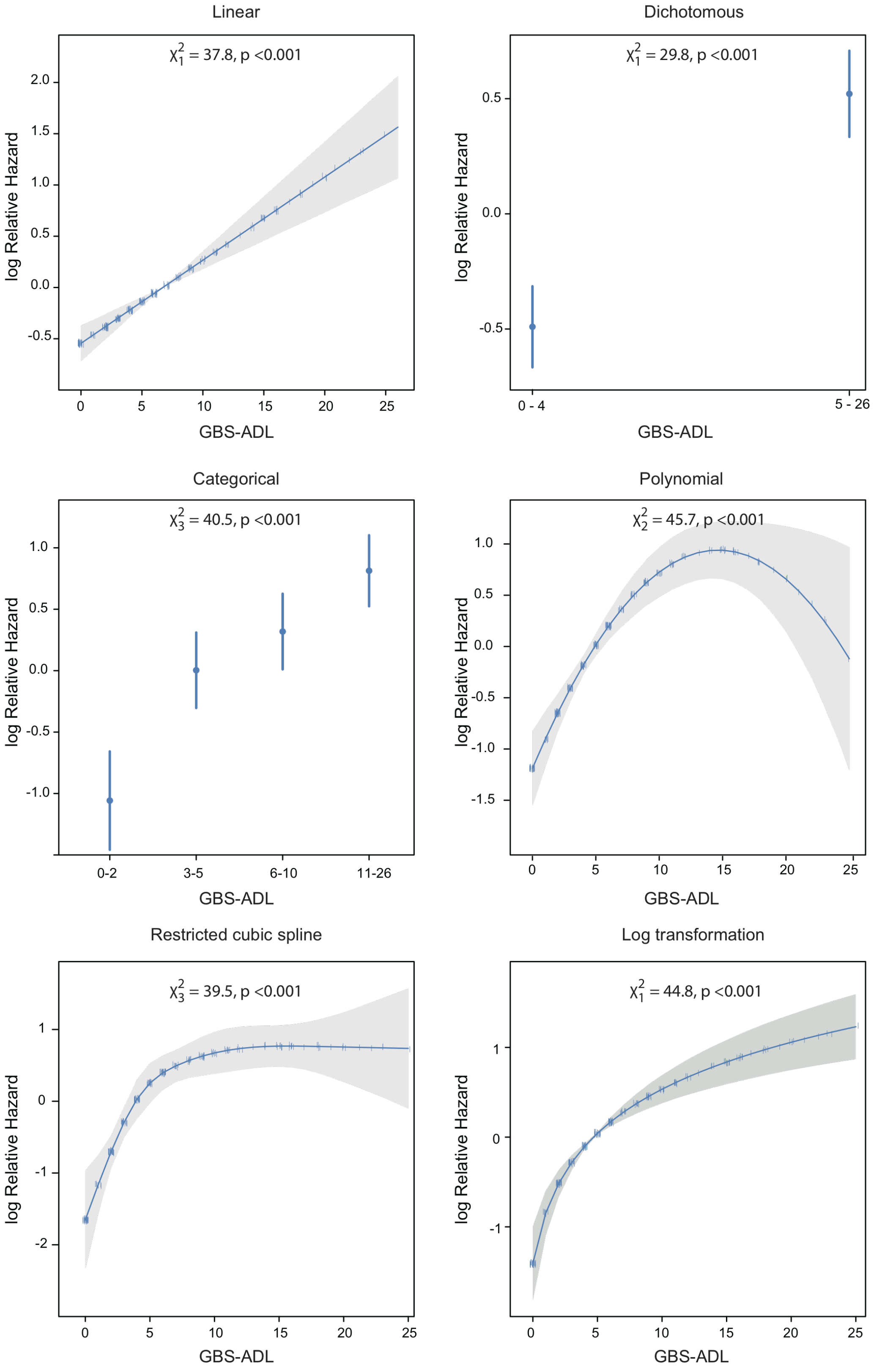

Figure e4. Different transformations tested for GBS-ADL.

5. Fitting the multivariate models

In the study
The two models were fitted, using the imputations and transformations above. The "model without ADL" used the covariates age, sex, charlson comorbidity index, albumin, BMI, eGFR, control/intervention status, and hemoglobin fitted as a restricted cubic spline The "full model" also included log(GBS-ADL).

6. Multivariate Diagnostics - Multicolinearity

Predictors with strong intercorrelations could cause interpreting problems, this is tested using the variance inflation factor (VIF). The interpretation of VIF has been disputed, a rule of thumb saying that VIF > 4 or > 10 signals a problematic multicolinearity problem have been suggested. However, these cut-offs may be too low, as a VIF over 10 could be acceptable [12]. To address multicolinearity, clustering of variables or data reduction could be applied.

In the study
In our models, all variables were simultaneously tested for colinearity. VIF Values were ranging between 1.02 and 1.47 in the "model without ADL" and between 1.10 and 1.52 in the "full model". The strongest bivariate correlation was between age and eGFR (r = -0.49). Thus, no apparent multicolinearity was present and no further action was taken.

7. Interactions – additivity assumption

A two-way interaction occurs when the effect of one predictor is dependending on the value of one other predictor. There are several recommendations regarding the number of interactions to test for. Only clinically plausible interactions could be tested, however, this requires prior knowledge. Another strategy is to test for all possible interactions, this requires a very large sample, to avoid overfitting. A compromise is to do a pooled interaction test for each variable and if the test is significant, the specific interactions are pursued [11].

In the study
We did not have prespecified interactions for ADL and the sample size did not permit testing for all possible interactions. Therefore we opted for a global test approach. As we did not want to give ADL any advantages compared to the other variables, we also performed global tests for the other variables, one at a time. In the "model without ADL", the global test was significant for sex and BMI and an interaction term of sex * BMI was found (low BMI was a risk factor in men, not in women). This interaction was included in the model. In the "full model" another interaction, GBS-ADL*eGFR, was also found (the effect of impaired GBS-ADL was higher when eGFR was less imparied and vice versa). One interpretation of this interaction could be that impaired GBS-ADL is associated with weight loss and thus lower eGFR. To test properly for this we would need to apply three-way interactions (such as GBS-ADL*BMI*eGFR), which was beyond the scope of this paper.

8. Assumption of proportional hazards

The assumption of proportional hazards is the assumption that hazards from predictors do not vary over time. Proportional hazards can be tested in several different ways. Graphically, schoenfeld residuals are often plottet against time, then a straight line at zero is ideal. There are also different approaches to compensate for non-proportional hazards, the most common being adding an interaction term with time.

In the study
The proportional hazards assumption was first tested using a global test (`cox.zph` in R) as well as specific tests for all variables. In the "model without ADL", the global test gave a p value of 0.72 and in the "full model" a p value of 0.70, signalling no violations of the PH assumption. The variable closest was eGFR, with a p value of 0.14. For eGFR, a schoenfeld residual plot is shown in figure e5. No further action was taken.

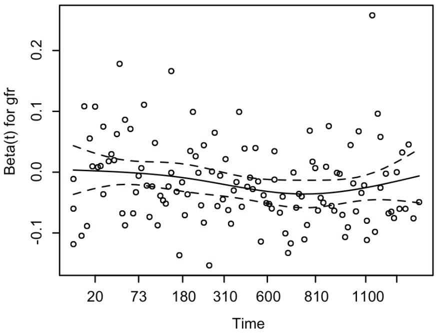

figure e5. Schoenfeld residual plot for eGFR.

9. Influential observations

With small sample size, a few influential observations could affect a model significantly. One way to screen for influential observations is by using what is called dfBeta, that shows to what extent the regression coefficient would change, if that case should be removed. Every case is designated a dfBeta value for each variable. Often, standardised dfBetas, with a cutoff of 0.20 is used to signify an influential observation. Thus, if deleting one observation led to a change in a predictor's β coefficient of more than 0.2 standard error, that observation was noted. For variables of specific interest, a sensitivity analysis could be performed without the observations with dfBeta > 0.2 to determine whether the effect is mainly due to a few highly influential observations.

In the study
In the "model without ADL", a total of 23 (12%) observations had any DfBeta > ±0.20. The lowest dfBeta was -0.39 and the highest 0.32. In the "full model", 21 observations were considered influential. DfBetas ranged from -0.46 to 0.48. Nine cases had a dfBeta > ± 0.20 for GBS-ADL and/or its interaction with eGFR. A sensitivity analysis was done, with these nine observations excluded. In that model the overall χ^2 increased from 123 to 124 and the GBS-ADL χ^2 from 32 to 37. Thus, the effects of GBS-ADL in the "full model" were not due to a few influential observations. In all further analysis the influential observations were kept in the model.

10. Relative contribution of variables

Describing the main effects of predictors including non-linear terms and interaction terms is not as intuitive as for simpler models, using Hazard Ratios. This is especially true if the model contains continuous-by-continuous interactions.

In the study
To obtain an estimate of the relative importance of the different predictors, we used the anova approach, developed by Harrell (`anova.rms` in R)[11]. Simple anova plots were included in

the article as figure 1. Plots of the variable effects are shown below in figure e6. In these plots, interaction terms have been incorporated into the variables' relative importance.

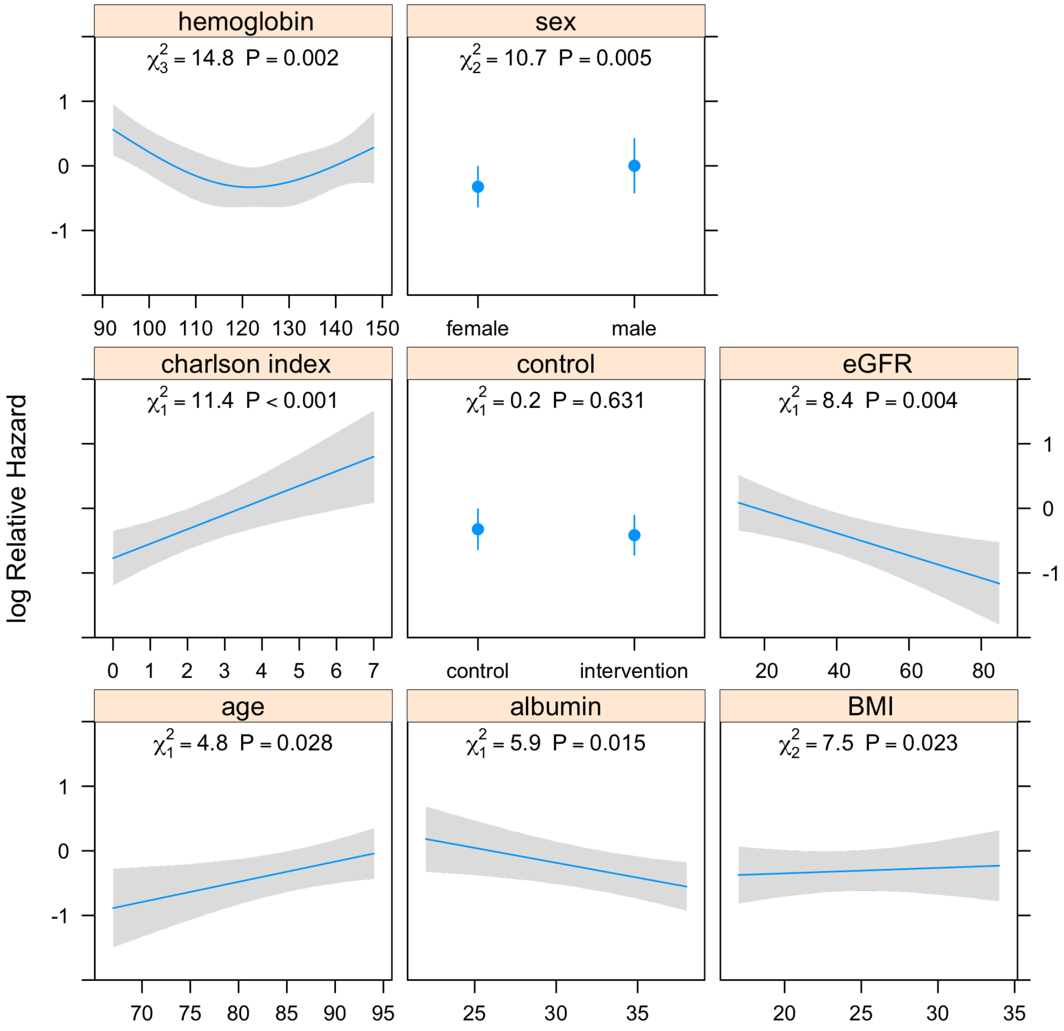

Figure e6a. Plot of variable effects in the "model without ADL"

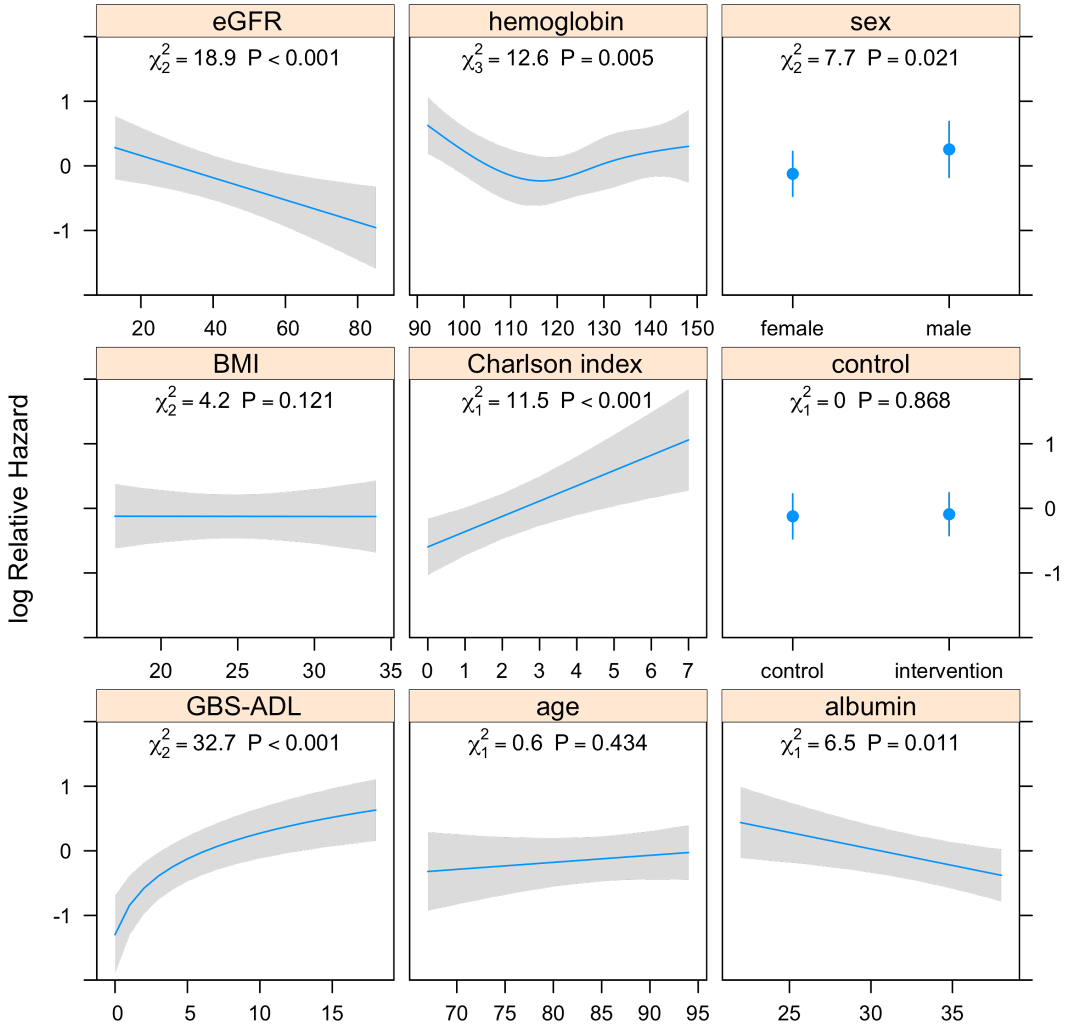

Figure e6b. Plots of the variable effects in the “full model”.

11. Added value of an added variable

There are several ways to determine the added value of a variable in a regression model.

a. Likelihood ratio test. With two nested models (where the smaller model is also a part of the full model) a Likelihood ratio test could be performed as a χ^2 test over df = number of additional independent variables in the new model.

In the study
The results are shown in table 3 in the article. For the “model without ADL”, LR χ^2 was 78.4 and for the “full model” 121.0. The degrees of freedom were 11 and 13, respectively. Therefore the LR test resulted in a χ^2 (df = 2, N = 198) = 42.5, p < 0.001. Thus the “full model” had a significantly better fit.

b. Discrimination, measured with the C, or concordance, statistic. The C statistic is the probability that, in a case-control pair, the case (deceased) will be given a higher predicted risk from the model than the control (survivor). C statistics range from 0.5 (coin toss, useless) to 1.0 (perfect discrimination). In logistic regression (without time-to-event data), the c statistic is the same as ROC. For survival analysis, time is incorporated, so a case at time t is compared with a survivor at time t, albeit this survivor could be dead at time t+1 (the next day). C statistics in survival analysis are often lower than ROC in logistic analysis. In addition, there are several different ways to calculate c statistic for time-to-event data.

In the study
We chose the method by Uno, to be able to compare between models. The "model without ADL" had a c statistic of 0.72 and the "full model" of 0.78. We set the follow-up time to 1428 days, as this was our median follow-up time of survivors. C statistics from the two models were compared using the method described by Uno et al. in the SurvC1 package [13]. Difference in c statistic between the model without ADL and the full model was 0.058 (95% CI = 0.022 - 0.094, p value 0.002).

c. NRI >0. Continuous net reclassification index[14, 15]. This index determines to what extent adding a new variable to a model leads to a change in the correct direction in predicted risk for each observation at time t (towards higher risk for deceased, towards lower for survivors). NRI>0 ranges from 0 (no increased value, useless) to 1(all observations reclassified in the right direction). NRI>0 has been shown to be more sensitive than change in C index, especially when the baseline model has a good performance. NRI>0 only describes the share of observations that have been reclassified, it does not quantify the amount of change in risk. Thus, it cannot distinguish between adding a variable that increases the predicted mortality risk for all cases with 1% or one that increases it with 50%.

For interpretation, the original NRI > 0 has been compared to the effect size of the added variable, where NRI>0 of 0.6 should be considered strong, 0.4 intermediate and below 0.2 weak [16]. However, after the initial development, Pencina et al. have suggested that ½ NRI>0 shoud be reported, as an average[15]. This is also what is given by the `IDI.INF` function in the `SurvIDINRI` package in R.

In the study
In our study ½ NRI>0 (95%CI) was 0.42 (0.22-0.58) with a p value <0.001. Again the follow-up time was set to 1428 days, to avoid extensive censoring. By doubling the point estimate of ½ NRI>0, the original NRI>0 would be 0.84, indicating a substantial effect size of adding ADL.

IDI. Integrated discrimination improvement. Originally developed by Pencina et al. for logistic models, IDI has been extended to time-to-event data [14, 17]. While NRI>0 displays the percentage of observtations being reclassified in the desirable direction, IDI is related to mean change in predicted probabilities within cases and controls. IDI is similar to testing the difference in R2, or discrimination slope, in logistic regrssion. IDI and NRI with confidence intervals were calculated (using the **`IDI.INF`** function) with the method by Uno et al. [18].

In the study
IDI was 0.15 (95%CI 0.07-0.27, p < 0.001), indicating that the mean change in the correct direction was 15% (cases (deceased) were given 15% higher mortality risk by adding ADL while controls (survivors) were given 15% lower).

12. Overfitting - internal validation

Overfitting occurs when sample size is small. Then the fitted model becomes too optimistic and dependent on the present dataset. Thus, the findings will neither be reproducable nor valid in other populations. Ideally a model could be tested in another population at another location and setting, what is known as external validation. If that is not possible, there are several ways to accomplish internal validation. The recommended approach is via bootstrapping. In bootstrapping a new dataset, of the same size as the original, is constructed from the original dataset by resampling with replacement (an observation could be selected several times). This dataset is then used to develop the model, which is then tested on the full original population. The difference in apparent performance and resampled performance is called optimism. The procedure is repeated 200-1000 times and the mean optimism is subtracted from the apparent performance estimates. This way an optimism-corrected estimate is obtained. In a future external population, this corrected estimate should be considered the best estimate possible[2, 8, 11, 19, 20].

In the study
Our aim was not to develop a valid prediction model, as this would have called for a larger sample size. Instead, we aimed to determine the relative imporance and added value of ADL when compared to the best possible traditional model. In the trade-off between overfitting and a well-performing traditional model, we empahsised the latter. A heuristic estimate of shrinkage would be χ^2 - d.f. / χ^2 = 123 - 14/123 = 0.89 for the “full model”. However, this d.f. is falsely low as we tested many more interactions and transformations.

To better determine, the extent of overfitting, we carried out an internal validation with 1000 bootstraps. For the “model without ADL” and the “full model”, the calibration slopes were 0.84 and 0.83, respectively, indicating a substantial amount of overfitting. The optimism-corrected R^2 was 0.26 vs. 0.39. Optimism-corrected c statistics were 0.69 and 0.76.

13. Updating the model - final nomogram and Kaplan-Meier curves

An overfitted model could be updated using a model that shrinks the regression coefficients. One such method is the LASSO (least absolute shrinkage and selection operator) model [21, 22]. LASSO could be used to both shrink factors as well as to eliminate variables.

In the study
Even if our aim was not to develop a valid prediction model, the amount of overfitting suggested that we should try to update the “full model”. In a LASSO model, the interaction terms and non-linear terms were combined into single terms. We used the `coxpath` function in R. We considered the model with the lowest AIC (Akaike Information Criterion). In this model, the variable “control” was shrunk to zero, all other variables remained. The mean shrinkage was 0.84. The lasso path and AIC is shown in figure e7.

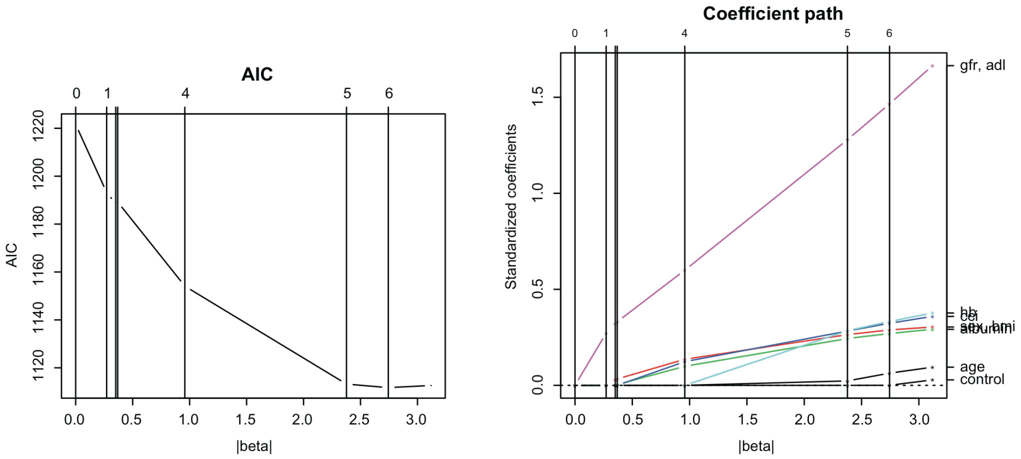

Figure e7. Lasso plots. AIC (Akaike information criterion) is lowest when control is set to zero.

A final nomogram, using the shrunk Lasso coefficients, was built, see figure e8. The cases were divided into four equally sized risk groups, by the quartiles of the linear predictor. To display the discrimination of the model, a kaplan-Meier curve was built, included in the article as figure 2.

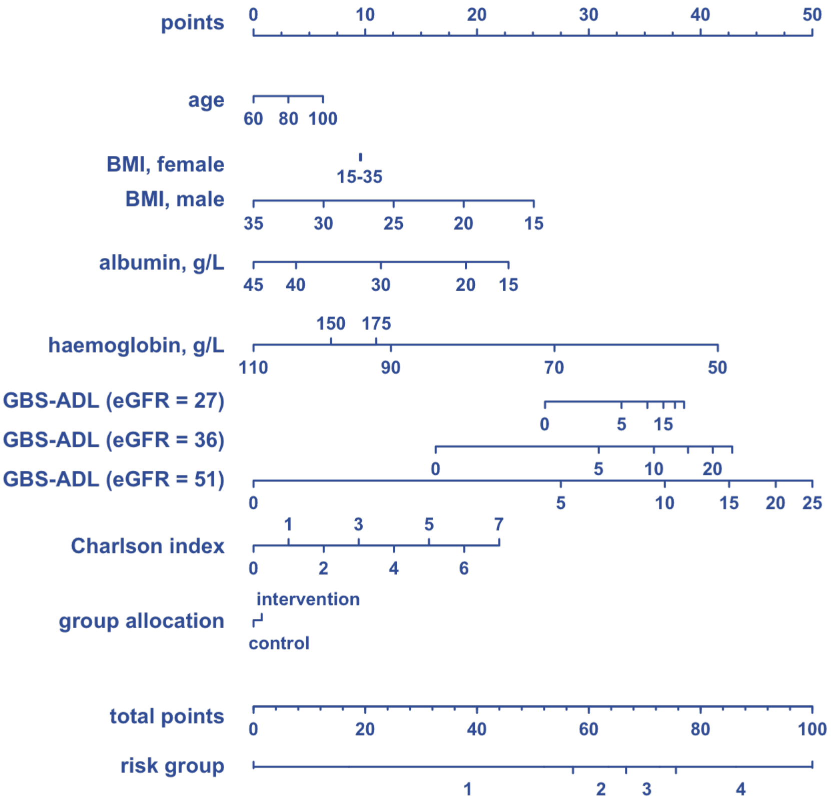

Figure e8a. Nomogram. Interpretation: For an individual, the variables are compared with the upper ”points” line, one at a time. These scores are then added for a total score that is plotted at the ”total points” line at the bottom. This could then be used to designate the person to a ”risk group” Notice the effect of interactions, low BMI is only a risk factor in men and the risk of GBS-ADL is moderated by eGFR, which is presented by median and quartiles. The cutoffs in the nomogram for the risk groups are completely arbitrary here, created to obtain 4 equally sized groups. In another scenario, cutoffs could be established to obtain for example a group with 90% chance of 3-year survival.

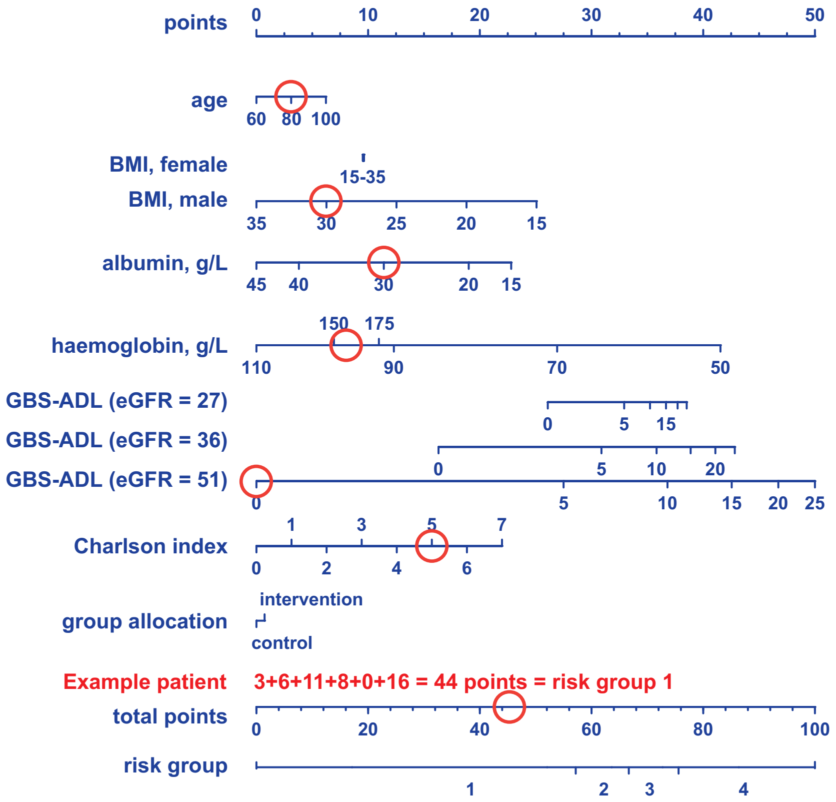

Figure e8b. Example of a scoring: This patient is 80 years old (3 points), male with BMI 30 (6 points), has an albumin of 30 (11 points) and a hemoglobin of 98 (8 points), normal kidney function and GBS-ADL (0 points) and a Charlson index score of 5 (16 points). The total score would be 3+6+11+8+0+16 = 44 points, placing this patient well within risk group 1. If this patient had all other variables constant but a functional decline, with a GBS-ADL score of 7, this would result in a total score of 44+30 = 74, placing the patient in risk group 3. The risk attributed to the functional decline would be equivalent to a hemoglobin drop from 98 to 55 g/L. Would it infer the same sense of urgency to the clinician?

References - statistical appendix

1. *R: A language and environment for statistical computing.* [computer program]. Version. VIenna; 2014.

2. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med.* Feb 28 1996;15(4):361-387.

3. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol.* Oct 2006;59(10):1087-1091.

4. Horton NJ, Kleinman KP. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Am Stat.* Feb 2007;61(1):79-90.

5. Janssen KJ, Donders AR, Harrell FE, Jr., et al. Missing covariate data in medical research: to impute is better than to ignore. *J Clin Epidemiol.* Jul 2010;63(7):721-727.

6. Marshall A, Altman DG, Royston P, Holder RL. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC Med Res Methodol.* 2010;10:7.

7. Moons KG, Donders RA, Stijnen T, Harrell FE, Jr. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol.* Oct 2006;59(10):1092-1101.

8. Steyerberg EW. *Clinical Prediction Models - A Practical Approach to Development, Validation, and Updating*. New York: Springer; 2009.

9. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med.* Jan 15 2006;25(1):127-141.

10. Harrell FE, Jr., Lee KL, Pollock BG. Regression models in clinical studies: determining relationships between predictors and response. *J Natl Cancer Inst.* Oct 5 1988;80(15):1198-1202.

11. Harrell Jr FE. *Regression Modeling Strategies - With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer-Verlag; 2001.

12. O'Brien RM. A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity* 2007-10-01 2007;41(5):673-690.

13. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med.* May 10 2011;30(10):1105-1117.

14. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med.* Jan 30 2008;27(2):157-172; discussion 207-112.

15. Pencina MJ, D'Agostino RB, Sr., Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med.* Jan 15 2011;30(1):11-21.

16. Pencina MJ, D'Agostino RB, Pencina KM, Janssens AC, Greenland P. Interpreting incremental value of markers added to risk prediction models. *Am J Epidemiol.* Sep 15 2012;176(6):473-481.

17. Chambless LE, Cummiskey CP, Cui G. Several methods to assess improvement in risk prediction models: extension to survival analysis. *Stat Med.* Jan 15 2011;30(1):22-38.

18. Uno H, Tian L, Cai T, Kohane IS, Wei LJ. A unified inference procedure for a class of measures to assess improvement in risk prediction systems with survival data. *Stat Med.* Jun 30 2013;32(14):2430-2442.

19. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med.* Aug 30 2004;23(16):2567-2586.

20. Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol.* Aug 2001;54(8):774-781.

21. Tibshirani R. Regression shrinkage and selection via the Lasso. *J. Royal. Statist. Soc B.* 1996;58(1):267-288.

22. Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med.* Feb 28 1997;16(4):385-395.