

Only Time Will Tell? - Identification of predictors for binary outcomes in longitudinal designs

Knut Mueller, UCB Biosciences, Monheim, Germany

Abstract

Longitudinal observational data (i.e. real world data) is increasingly used to identify factors that may predict a certain outcome due to the large number of participants under study. There are several appropriate methods to tackle such kind of analyses. Although there is often more than one possible statistical method, depending on the circumstances, one method may be preferable to another depending upon the research hypothesis and available data. Especially in designs where the focus is on a comparison of groups that reach or do not reach a certain endpoint, a responder analysis may be the preferred option, when the time component is of minor interest. However, it may be beneficial to include the available time information to increase the strength of the analysis, especially when missing or censored data is expected. Using the example of a predictive model, this paper compares the Proportional Hazard regression with methods like conditional logistic regression to show the pros and cons of each of the approaches using simulated data.

Methods

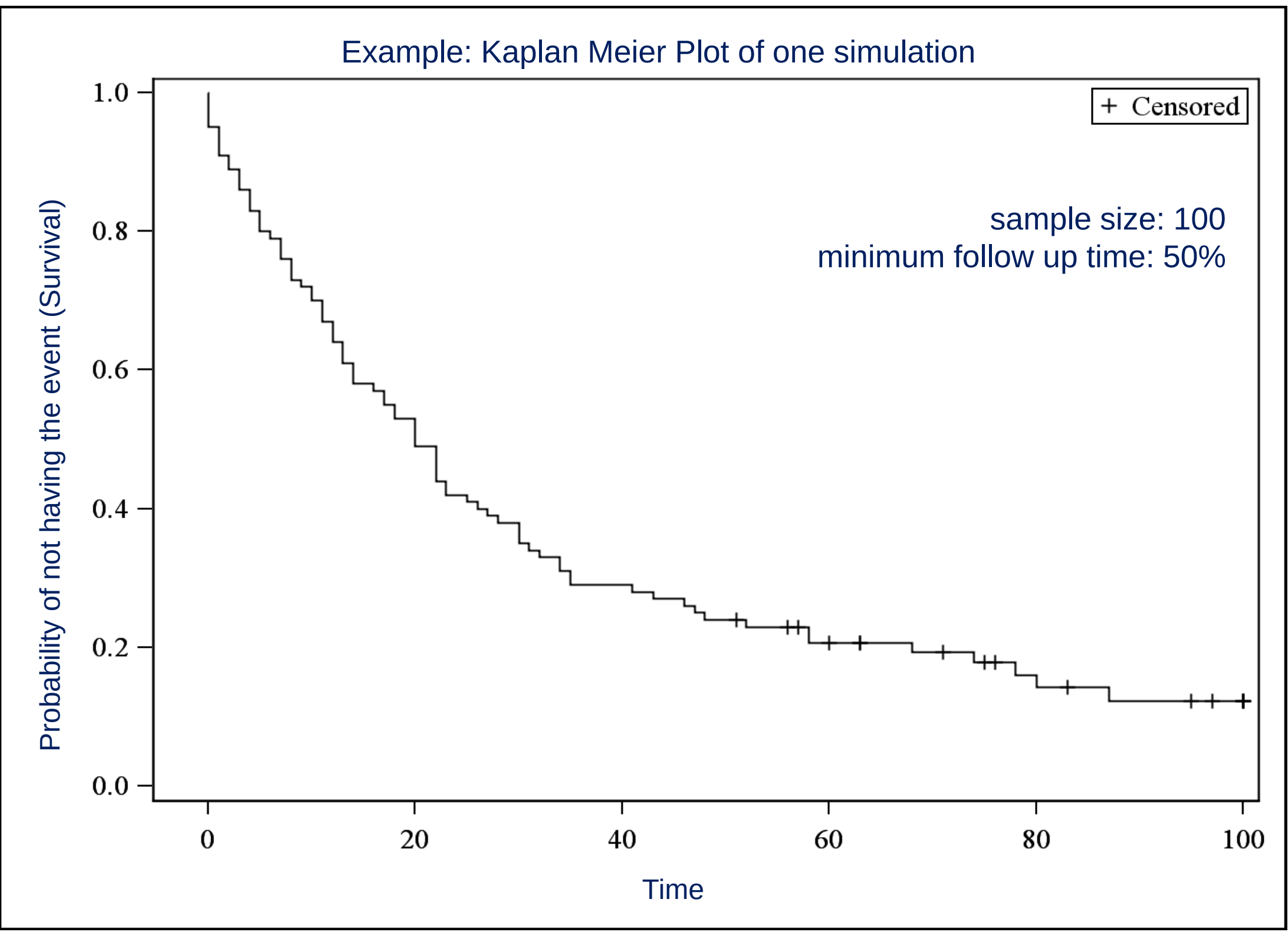
We simulated time to event data based on a Weibull distribution and a proportional hazard regression model (1). The simulated data introduce 4 predicting factors, of which one is unobserved, here referred to as a confounder. The data also contain 3 observed random variables that were not contributing to the model, here referred to as distractors. In the simulations we varied the incidence and beta of the predictors, the incidence and beta of the confounder, the incidence of the distractors, the sample size, and the proportion of patients lost to follow up. We then analyzed the data using a Cox regression Model and a conditional logistic regression and compared the two methods regarding their capability of correctly identifying the predictor variables (see box "outcomes"). The evaluation was done using a repeated measurement ANCOVA with PROC MIXED. All parameters including interactions were analyzed.

Evaluation Methods

- Repeated Measurement ANCOVA
- Each simulation was analysed using a Cox regression model compared to a conditional logistic regression model
- Factors in the evaluation model see parameters of the simulations

Parameters of the simulations

- Event time distribution: Weibull
- Distribution of Censoring time: Uniform
- Proportional Hazard Regression Model
- 3 predictors, that contribute to the model
- 1 unobserved confounder that may contribute to the model
- 3 distractors, that do not contribute to the model
- Predictor beta: 0.5 / 1.0 / 2.0
- Predictor incidence: 0.1 / 0.5
- Confounder beta: 0 / 0.5 / 1.0 / 2.0
- Confounder incidence: 0.1 / 0.5
- Distractor incidence: 0.1 / 0.5
- Sample size: 100 / 500
- Minimal follow -up time: 25% / 50% / 100%



Compared Analysis Methods:

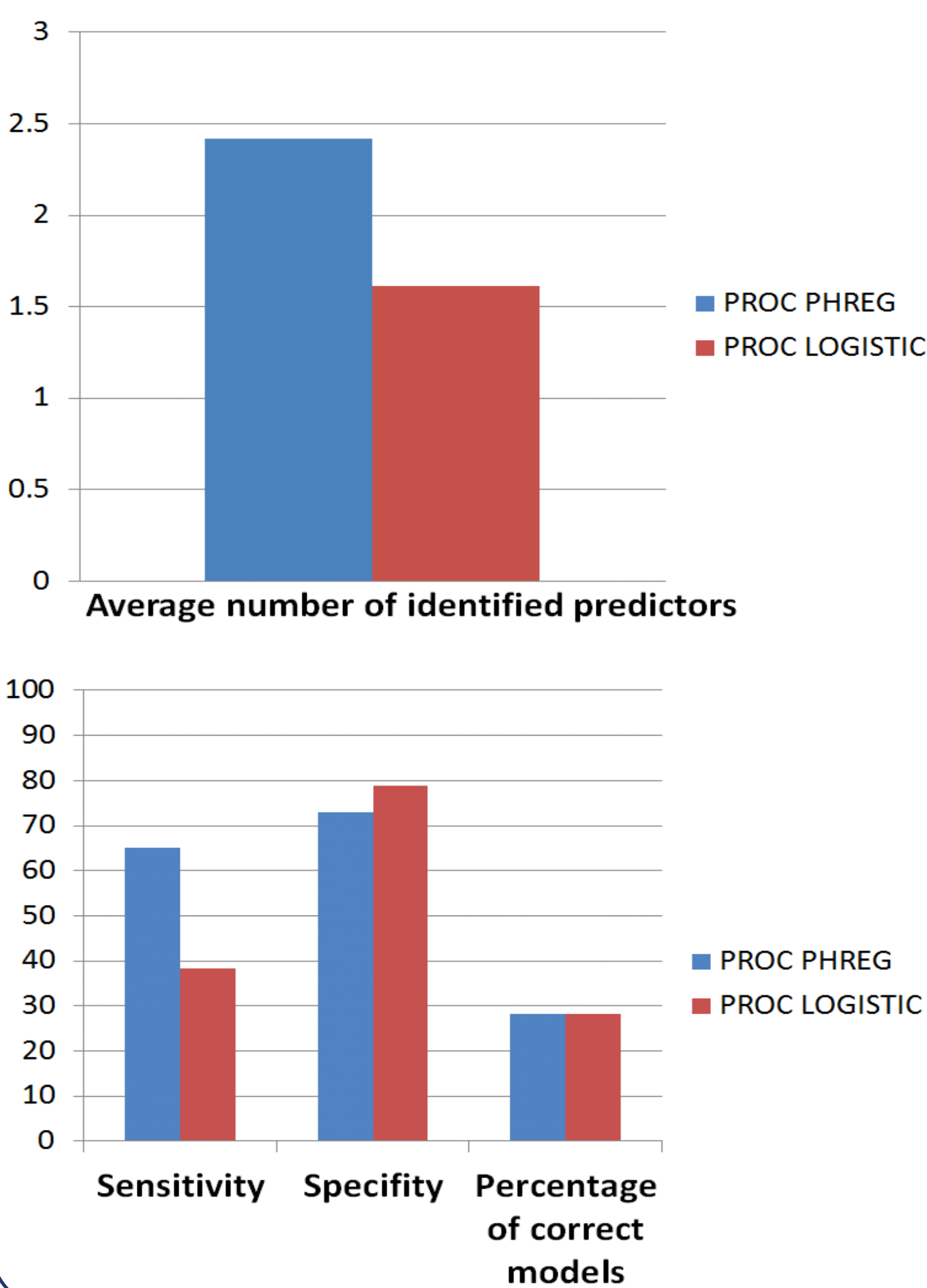
Proportional Hazards Regression
(Cox Regression) using PROC PHREG

Conditional Logistic Regression
with matched pairs (matched by follow up time) using PROC LOGISTIC

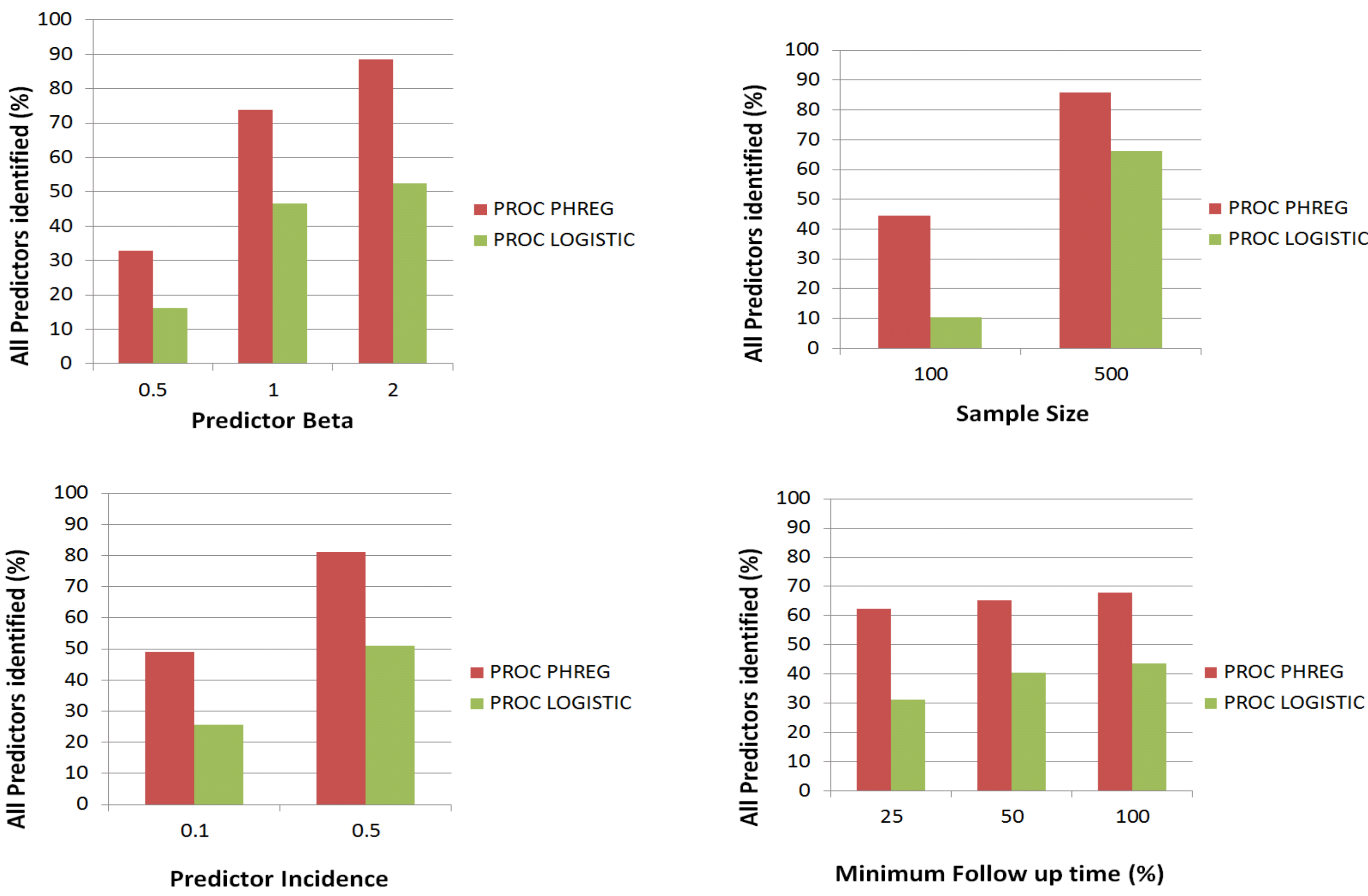
Outcomes

- Percentage of Simulations with correct model identified
- Percentage of simulations with all predictors identified (sensitivity)
- Percentage of simulations with all distractors identified (specificity)
- Average number of correctly identified predictors

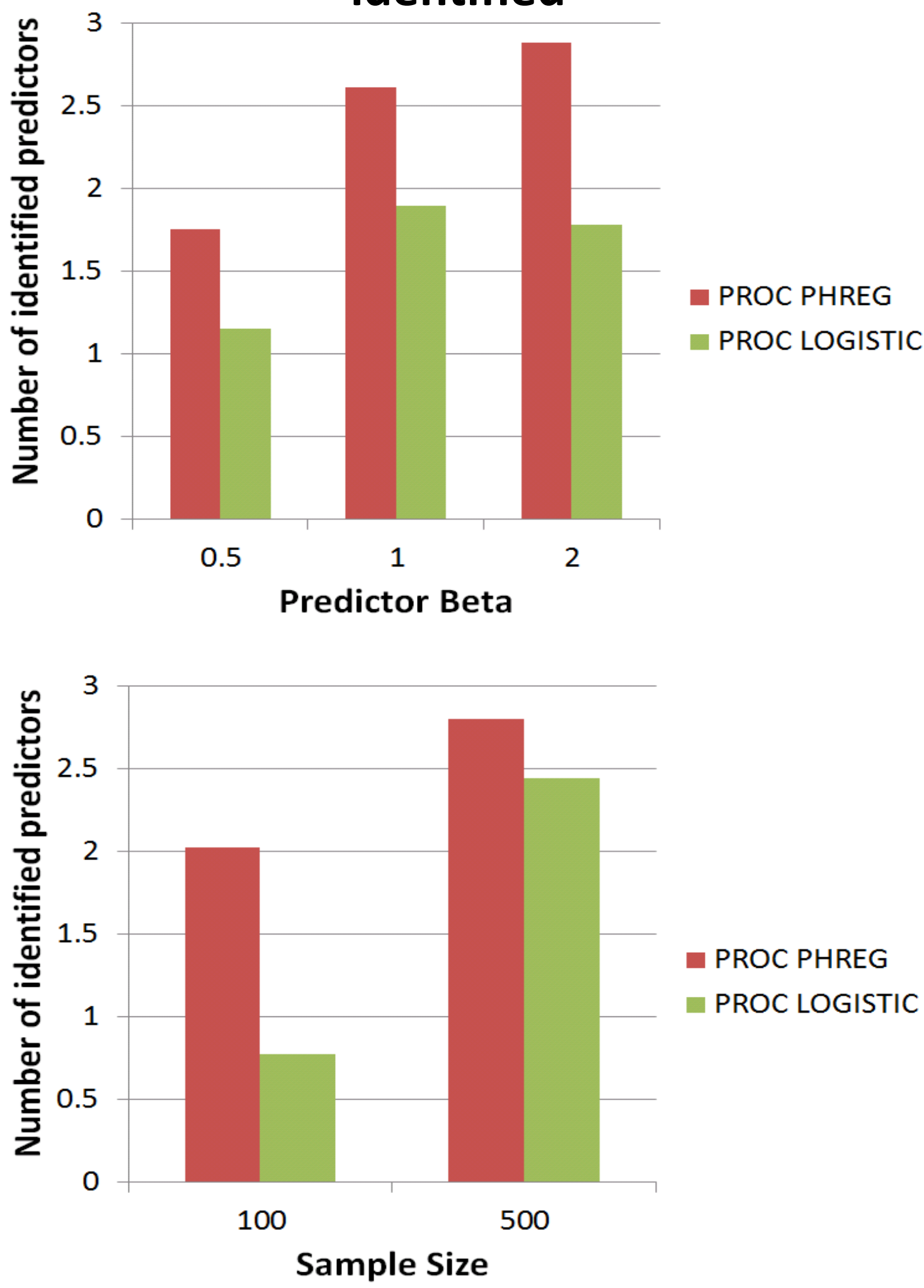
Overall Comparison of Methods



Sensitivity



Average number of predictors identified



Results

The Cox Regression Model performed much better with regard to correctly identifying predictor variables for reaching the event. On average 2.4 predictors were correctly identified. In 60 % of the simulations all 3 predictors were identified. The Conditional Logistic Regression model was less capable of identifying the predictors correctly. However, compared to the conditional logistic regression models, the Cox regression models more often incorrectly identified distractor variables as predictors for the event of interest. The sample sized turned out to be the most obvious predictor for the performance of both methods, but even more so for the logistic regression. With a sample size of 100, the logistic regression identified less than 1 predictor correctly. Other modifying factors were the number of patients lost to follow up and the incidences of the prediction variables. All other factors were not modifying the performance of the two methods.

Conclusions

- Cox Regression Models perform better with low sample sizes compared to Conditional Logistic Regression Models
- The much better performance with regard to identifying predictors correctly (sensitivity) goes at the cost of a slightly lower specificity (falsely identified predictors)
- Additional analyses are necessary to explore the robustness of both methods when the proportional hazards assumption is violated.

References

(1) Kleinman K (2010): Simulate censored survival data. <http://sas-and-r.blogspot.de/2010/03/example-730-simulate-censored-survival.html>