Author(s): Reinikainen, Jaakko; Karvanen, Juha; Tolonen, Hanna

Title: How many longitudinal covariate measurements are needed for risk prediction?

Year: 2016

Version:

Please cite the original version:

Reinikainen, J., Karvanen, J., & Tolonen, H. (2016). How many longitudinal covariate measurements are needed for risk prediction?. Journal of Clinical Epidemiology, 69, 114-124. https://doi.org/10.1016/j.jclinepi.2015.06.022

How many longitudinal covariate measurements are needed for risk prediction?

Jaakko Reinikainen[a,*], Juha Karvanen[a], Hanna Tolonen[b]

[a]*Department of Mathematics and Statistics, University of Jyvaskyla, P.O. Box 35 (MaD), FI-40014 University of Jyvaskyla, Finland*
[b]*Department of Chronic Disease Prevention, National Institute for Health and Welfare, P.O. Box 30, FI-00271 Helsinki, Finland*

Abstract

Objective: In epidemiological follow-up studies, many key covariates, such as smoking, use of medication, blood pressure and cholesterol, are time-varying. Because of practical and financial limitations, time-varying covariates cannot be measured continuously, but only at certain prespecified time points. We study how the number of these longitudinal measurements can be chosen cost-efficiently by evaluating the usefulness of the measurements for risk prediction.

Study Design and Setting: The usefulness is addressed by measuring the improvement in model discrimination between models using different amounts of longitudinal information. We use simulated follow-up data and the data from the Finnish East–West study, a follow-up study, with eight longitudinal covariate measurements carried out between 1959 and 1999.

Results: In a simulation study, we show how the variability and the hazard ratio of a time-varying covariate are connected to the importance of re-measurements. In the East–West study, it is seen that for older people, the risk predictions obtained using only every other measurement are almost equivalent to the predictions obtained using all eight measurements.

Conclusion: Decisions about the study design have significant effects on the costs. The cost-efficiency can be improved by applying the measures of model discrimination to data from previous studies and simulations.

Keywords: study design; longitudinal measurements; model discrimination; risk prediction

*Corresponding author. Tel.: +358 440 366 896
Email address: jaakko.o.reinikainen@jyu.fi (Jaakko Reinikainen)

What is new?

- The usefulness of longitudinal measurements can be systematically summarized by measures of model discrimination.
- The cost-efficiency of the follow-up study design can be improved in both ongoing and completely new follow-up studies by considering model discrimination comparisons based on simulations and data from previous studies.
- The results from the East–West study with over 50 years of follow-up suggest that carrying out covariate measurements every ten years may by sufficient for older people when all-cause mortality is considered as an end point.

1. Introduction

Epidemiological follow-up studies usually include time-varying covariates, such as smoking, use of medication, blood pressure, cholesterol and body mass index. Especially in long follow-up studies, these kinds of covariates may lose their predictive power over time, if only the baseline measurements are used. This can be seen as one form of the regression dilution problem (1). An ideal solution would be to measure these covariates continuously, but this is usually impossible because of practical and financial limitations. Here we use the term 'covariate' to mean both variables of direct interest and control variables measured on continuous or categorical scale. Longitudinal measurements carried out at prespecified time points are often used, when the speed of change in the covariates is relatively slow. Planning longitudinal measurements, however, raises many questions related to the costs and efficiency of the study. Which individuals should be measured and how frequently? Often available resources and traditions guide these decisions.

According to our knowledge, the question presented in the title has not previously been formulated as a statistical problem. In addition to the practical importance in designing an epidemiological study, the question has also wider theoretical interest. In the general form, the question is about estimation or approximation of a stochastic continuous-time process on the basis of a small number of discrete-time observations. In the context of causal

inference, the problem can be formulated as a question on the relationship between continuous-time processes and causal directed acyclic graphs (2; 3). We do not aim to solve the general problem, but to present tools that can be used to support fact-based decision making on the study design in practical situations.

Cost-efficiency of a follow-up study can be considered from different viewpoints. We assume a follow-up study with time-varying covariates and a continuously observed survival outcome. Our objective is to study the determination of the reasonable number of longitudinal measurements needed for risk prediction. Another aim is to study whether a new measurement is worth carrying out in an ongoing follow-up study. We approach these questions by using simulation studies and empirical evidence from previous studies. The combination of these is also discussed.

Some other aspects of cost-efficiency of follow-up studies have already been explored. Our previous work (4) considered the optimal selection of a subset of individuals for a new measurement when we cannot afford to remeasure the entire cohort. The timing of follow-up visits has been analyzed in a case where an examination is needed to determine if an event of interest has occurred (5; 6). Optimal design of follow-up studies has also been investigated when a subset of individuals is selected for expensive genotyping (7). In the case of longitudinal response, the optimal number of repeated measurements has been studied (8) and so called triggered sampling design has been proposed to improve cost-efficiency (9).

Risk prediction is motivated by the need to assign interventions on the basis of the individual-level risk profiles. Before carrying out a re-examination of the covariates in an ongoing follow-up study, researchers may want to know how valuable this would be for risk prediction. This can be addressed by simulating the unknown covariate measurements and survival times and comparing the predictive abilities of a model using new measurements and a model fitted without new measurements. If the incremental benefit would be small or negligible, the re-examination could be considered to be conducted later. When we are planning a completely new follow-up study, we could utilize similar studies conducted earlier to learn about the importance of longitudinal measurements. By analyzing data from similar studies, we may understand better the role of re-examinations in the new study.

To evaluate the usefulness of longitudinal covariate measurements, we use measures of model discrimination (10; 11; 12) to compare models using different amounts of longitudinal information. These measures have al-

ready been applied to specific cases to show that using longitudinal covariate measurements improves model performance compared to using only baseline measurements (13; 14). In this article, we present the concepts on a general level and, in addition, study a practical example based on data from the East–West study, the Finnish part of an international follow-up study called the Seven Countries Study (15; 16). These data suit our purposes well, because the Finnish cohorts have eight longitudinal measurements carried out between 1959 and 1999 and the information on mortality is available until the end of 2011.

2. Risk models and measures of model performance

2.1. Models for risk prediction

The usefulness of longitudinal covariate measurements for risk prediction depends on the risk prediction model used. Therefore, we have to define our models of interest and design the study with respect to them. Two main approaches for modeling survival time with time-varying covariates are time-dependent Cox model (17) and so called joint modeling (18; 19). In the time-dependent Cox model covariate values are updated at measurement times, whereas a joint model includes models for the covariate process and survival times and allows them to be associated.

Joint modeling is often preferred because time-dependent Cox models may provide biased estimates of the regression coefficients if the longitudinal process is measured with error or includes random variation that is not captured by the measurements (20). Bias is a less serious concern in risk prediction because the calibration of the model can be checked and if necessary, the model can be recalibrated. There are also cases where time-dependent Cox models are appropriate (21). Further, although some specialized methods have been proposed for joint modeling with multiple longitudinal covariates, including conditional score estimator (22), latent class approach (23) and Bayesian methods (24; 25), the computational methods and software for multivariate joint modeling are not fully developed. For these reasons, the time-dependent Cox model was used in this work.

The choice of the type of the model has to be study-specific to obtain reasonable estimates of predicted probabilities. It is also worth noticing that there are several different ways to use longitudinal measurement information in risk prediction models. New time-dependent covariates derived from the original measurements may be, for example, average of the most recent and

all the previous measurements (26), standard deviation or maximum value of the measurements (27) or change in the latest two measurements (28).

2.2. Measures of model performance

Several measures of model performance have been developed to give insight into the usefulness of predictors or models (29; 11; 30; 12). These measures typically fall into one of the two main categories: model calibration or model discrimination. Calibration quantifies how well the predicted risk agrees with the actual observed risk. Discrimination is a measure of how well the model can separate events and non-events.

The area under the receiver operating characteristic curve (AUC) (29) is a widespread measure of model discrimination. The AUC is the probability that risk prediction of a randomly selected pair of individuals, one with the primary outcome of interest and one without, is properly ordered (31). However, many researchers have observed that the improvement in AUC may be small even if the addition of the predictor to the model can be otherwise justified (32; 33). Recently, a decision analytic approach has been proposed to interpret small changes in AUC (34).

The net reclassification improvement (NRI) and the integrated discrimination improvement (IDI) indices (11) have gained popularity as alternatives for the improvement in AUC although these measures have also been criticized (35; 36). The NRI is estimated as the proportion of correct minus incorrect reclassifications among events, plus the proportion of correct minus incorrect reclassifications among non-events. A reclassification is considered to be correct if an event is classified to a higher risk category by the new model than by the old model or if a non-event is classified to a lower risk category. In our applications, we do not have any established risk categories, so we use the continuous (or category-less) version of the NRI (37), where upward and downward 'reclassifications' are defined as any upward or downward change in predicted probabilities. We denote the continuous NRI simply by NRI. The IDI is estimated as the average risk of events minus the average risk of non-events obtained from the new model, minus the average risk of events minus the average risk of non-events obtained from the old model.

Decision-analytic measures quantify the clinical usefulness of models by incorporating relative consequences of false positives and negatives. A risk threshold T is defined as a risk of the outcome at which one is indifferent about whether to classify an individual to the high or low risk category. The odds of T is then the ratio of harm to benefit. Net benefit (38) is a

decision-analytic measure, which is given as follows:

$$\mathrm{NB} = \frac{\mathrm{TP}}{N} - w\frac{\mathrm{FP}}{N},$$

where TP is the number of true positives, FP the number of false positives, N is the size of the data set and $w = T/(1-T)$. When net benefit is used in comparing two models, the difference in NB is interpreted as the difference in the proportion of true positives at the same level of false positives.

It is not always clear whether model calibration or discrimination should be preferred and recently there has been some controversy over the performance and use of risk prediction models (39; 40). Discrimination could be preferred, because recalibration is always possible, but poor discrimination cannot be corrected afterwards (10). On the other hand, calibration should not be overlooked, when the goal is to stratify individuals accurately into risk categories (41). In addition, it has been noted that using the NRI, IDI and net benefit requires calibrated models (42; 43). In this paper, we restrict our considerations into model discrimination, but if the particular aim of a study was to construct risk assessment algorithms, measures of calibration, e.g the Hosmer-Lemeshow statistic (44), calibration-in-the-large (45) and calibration slope (45), should also be applied when planning a study design.

3. Study design based on simulations

A reasonable simulation study to investigate the incremental benefit of new measurements provides the distributions of covariates and survival times and their relations. Particularly, understanding of the covariate processes and their effects on survival is needed. This information can be obtained from the data already collected in an ongoing study or from similar studies conducted earlier.

In general, the simulation of proportional hazards models (46; 47) requires several distributions and parameters to be determined. In the context of this paper, these include the sample size, the distributions of covariates, the correlations between longitudinal measurements, the distribution of survival times, the effect of the covariates on survival and the length of the follow-up. It is worth considering the sensitivity of these choices by trying different distributions and parameter values.

3.1. Description of the simulation study

Simulation examples are presented in different settings to demonstrate how the variability of a time-varying covariate and the hazard ratio of this covariate affect the usefulness of the new measurements. A ten-year follow-up study with 10 000 individuals of age 60 years at baseline, is considered. The baseline measurement is carried out for the entire cohort. It is evaluated, whether a new measurement five years after the baseline significantly improves the predictions.

The baseline measurement x_{0i}, for individual i, of a time-varying piecewise constant covariate is generated from the normal distribution with mean $\mu_0 = 0$ and variance $\sigma^2 = 1$. Second measurement value x_{1i} five years after baseline is simulated from the normal distribution conditioned on the baseline measurement with mean $\mu_{1i} = \gamma x_{0i}$ and variance $\sigma^2_\varepsilon = \sigma^2 - \gamma^2\sigma^2$ resulting in the same variance for the baseline and the second measurements, when $-1 \leq \gamma \leq 1$. As a matter of fact, the parameter γ represents also the correlation between x_0 and x_1, because these variables have the same variance. Values $0, 0.2, 0.4, 0.6$ and 0.8 are used separately for γ to see the effect of correlation between longitudinal measurements on the importance of re-measurement. We denote the covariate shortly by $x(t)$, which takes the baseline value x_0 in the time interval $(t_0, t_1]$ and x_1 in the interval $(t_1, t_2]$, where t_0, t_1 and t_2 are times of the baseline measurement, the second measurement and the end of the follow-up, respectively.

Survival times are drawn from the Weibull distribution conditioned on the time-varying covariate through the proportional hazards model

$$\lambda(t|x(t)) = \lambda_0(t)e^{\beta x(t)}. \tag{1}$$

In the Weibull distribution, using parameterization where the baseline hazard function is expressed as

$$\lambda_0(t) = \frac{a}{b}\left(\frac{t}{b}\right)^{a-1},$$

we use the shape parameter $a = 6.1$ and scale (in days) parameter $b = 28\ 000$, which roughly equal those estimated from the real data used in Section 4. The regression parameter is set separately to $\beta = 0.1$ and $\beta = 0.2$ to illustrate how the hazard ratio of the covariate affects the importance of using the second measurements in the risk prediction. When the regression coefficient in the Weibull proportional hazards model is $\beta = 0.1$ or $\beta = 0.2$, the hazard ratio is $e^\beta = 1.11$ or $e^\beta = 1.22$ per an increase of the size of the

standard deviation, respectively. The survival time is censored at the end of the follow-up t_2, if the event has not occurred before that time.

In these simulations, we first fit a Weibull proportional hazards model to survival information from the interval $(t_0, t_1]$ using the baseline measurements. Then, this model is used to calculate predicted event probabilities for the interval $(t_1, t_2]$ in two ways: using the baseline measurement or using both the baseline and the second measurement. The discrimination ability of these predictions is compared using 1000 simulation runs in each setting.

3.2. Simulation results

Figure 1 shows the estimates of (continuous) NRI, IDI and incremental AUC (iAUC) for the added predictive ability of using new covariate measurements instead of the baseline measurements. As expected, the use of the new measurements becomes more important when the hazard ratio increases. However, the figure reveals that IDI reacts more strongly to the difference in the hazard ratio than NRI and iAUC.

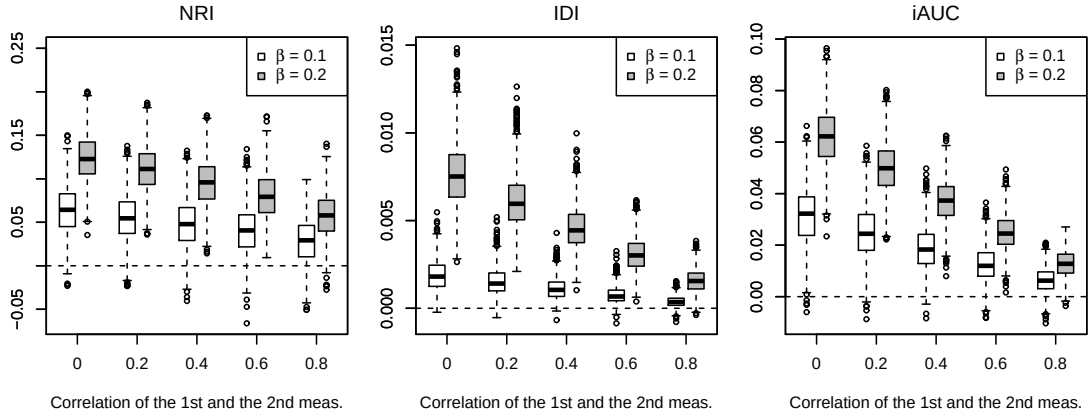

Figure 1: Estimates of NRI, IDI and incremental AUC using 1000 simulations in each setting. The horizontal dashed line is at the level of zero.

The results also illustrate to what extent the correlation of longitudinal measurements affects the usefulness of the second measurement. Naturally, the greater the correlation is, the lesser is the additional information obtained by carrying out a re-examination. If the correlation was one, the models to be compared would be exactly the same. These kinds of simulations do not give a direct answer to the question whether the re-measurement should be

carried out, but they help researchers in making fact-based decision about the study design.

4. Analysis of historical data

Planning of completely new follow-up studies should be based on experiences from similar studies conducted earlier, if such are available. In this section we propose an approach for evaluating the usefulness of longitudinal covariate measurements in the risk prediction using data from a completed follow-up study. This guides researchers in understanding the role of re-examinations and in allocating the study resources efficiently.

4.1. The East–West study

We use data from the East–West study, the Finnish part of the Seven Countries Study, which is one of the first international follow-up studies in the field of cardiovascular epidemiology. The Seven Countries Study was initiated in the late 1950s to study cardiovascular disease and their risk factors in different countries (16). The Finnish data consist of one cohort from Eastern and one from South-Western Finland, from which comes the name the East–West study. All the men born between 1900 and 1919 and living in these geographically defined areas were included in the cohorts ($N = 1711$). The baseline measurements were carried out in 1959 and re-examinations in 1964, 1969, 1974, 1984, 1989, 1994 and 1999. The cohorts were followed up for mortality until the end of 2011.

Some characteristics of the data are presented in Table 1. Smoking is a binary variable describing the current smoking status. Systolic blood pressure (SBP), total cholesterol, body mass index (BMI) and resting heart rate (HR) are continuous variables. Table 1 presents only those variables from the East–West study, which were used in our analyses. All-cause mortality was used as the response variable. Correlations between longitudinal covariate measurements are shown to enable the reflection on the effect of the variability of the covariates over time to the importance of the re-measurements. Our analysis of these data did not include all the relevant covariates known nowadays as the baseline measurement of the East–West study was carried out in 1959.

4.2. Model comparisons

To learn about the importance of the longitudinal measurements in the East–West study, we compared four different ways of using the data:

Table 1: Characteristics of the East–West data by the examination years.

	Year							
	1959	1964	1969	1974	1984	1989	1994	1999
Number of individuals alive	1711	1594	1428	1225	766	525	317	189
Participation rate (%)	98	97	96	96	92	86	87	68
Average age of individuals alive	49.8	54.7	59.4	64.0	73.0	76.9	80.9	84.1
SBP: corr. with the previous meas.		0.70	0.66	0.55	0.56	0.48	0.26	0.25
Chol.: corr. with the previous meas.		0.70	0.73	0.63	0.54	0.69	0.67	0.66
BMI: corr. with the previous meas.		0.90	0.91	0.88	0.81	0.87	0.83	0.78
HR: corr. with the previous meas.		0.54	0.50	0.53	0.37	0.45	0.26	0.29
Smoking: corr. with the previous meas.		0.71	0.72	0.65	0.63	0.81	0.81	0.56

Participation rate was calculated as the proportion of individuals with SBP, Chol, BMI, HR or Smoking measured among all individuals alive.
Spearman correlation is used for smoking, Pearson correlation for other variables.
Abbreviations: SBP = systolic blood pressure, Chol. = total cholesterol, BMI = body mass index, HR = resting heart rate

(a) The null model, which does not use covariate data at all, so the predictions are based only on the information about individuals' ages.

(b) Only the baseline measurements are used.

(c) Every other measurement is used, that is, the measurements from the years 1959, 1969, 1984 and 1994 are used.

(d) All the measurements carried out in the study are used.

We can imagine that these represent four different scenarios of carrying out the measurements and we refer to the models based on them as Models (a)–(d). Cox proportional hazards models of the type of formula (1), with all-cause mortality as the end-point and age as the time-scale, were employed in each case. Models (a) and (b) were time-fixed and Models (c) and (d) were time-dependent models. All the variables presented in Table 1 were used as covariates in Models (b)–(d).

An attempt was made to take full advantage of the measurements available for each model. Smoothing splines (48) were applied to take into account the nonlinear effects of the covariates. Models (c) and (d) use covariate data as time-dependent averages and latest changes. The proportional hazards assumption of the models was checked using Schoenfeld residuals (48). The

model equations and selection of degrees of freedom for the splines are presented in Appendix. As seen in Table 1, the participation rate was high in the East–West study, so missing data was not a substantial problem here. The missing longitudinal measurements were imputed simply by carrying forward the last observation. This imputation method is similar in treating covariates, and has also the same drawbacks, as the time-dependent Cox model (21). We point out that the purpose of these analyses is to illustrate the concepts of this article rather than being a thorough investigation of all-cause mortality.

Figure 2 shows AUC estimates calculated from five-year predictions. These estimates, as well as NRIs, IDIs and net benefits in this section, are calculated for a binary outcome (death/survival during the appropriate five-year period). The Cox models are fitted to the entire follow-up period and the predictions are computed using these models and the covariate data collected at the beginning of a prediction interval and before that. Although the follow-up started from the year 1959, the first AUC for the model using all the measurements is for the prediction interval of 1964–1969 (from autumn 1964 to autumn 1969), because two longitudinal measurements are needed to calculate the changes in the risk factors. Respectively, the first AUC for the model, which uses every other measurement, is for 1969–1974.

It has been stated that the improvement in AUC may not be a very good tool for comparing predictiveness of models (32; 33), but plotting the AUCs in this way illustrates how the predictive ability of each model changes over time, which could not be seen by using NRI or IDI. For example, Figure 2 shows that the AUC of the model using only the baseline measurement approaches the AUC of the null model as time goes on. As we can see, the order of the AUCs between the models indicates that predictions obtained using every other measurements are virtually as good as those obtained using all the measurements. However, the predictions are improved by carrying out longitudinal measurements after the baseline measurement, except for the latter part of the follow-up, where age is still predictive but the risk factors start to lose their predictive power. This may be due to an increasing role of frailty as a predictor in elderly (49; 50). As seen in Table 1, the average age of individuals alive is already over 80 in the last two measurement times. The exceptionally high AUCs in the prediction interval of 1984–1989 are independent of the risk factors, because this spike is observed also for the null model. For some reason, individuals' age is correlated with the survival status more strongly in this interval than in other five-year intervals considered.

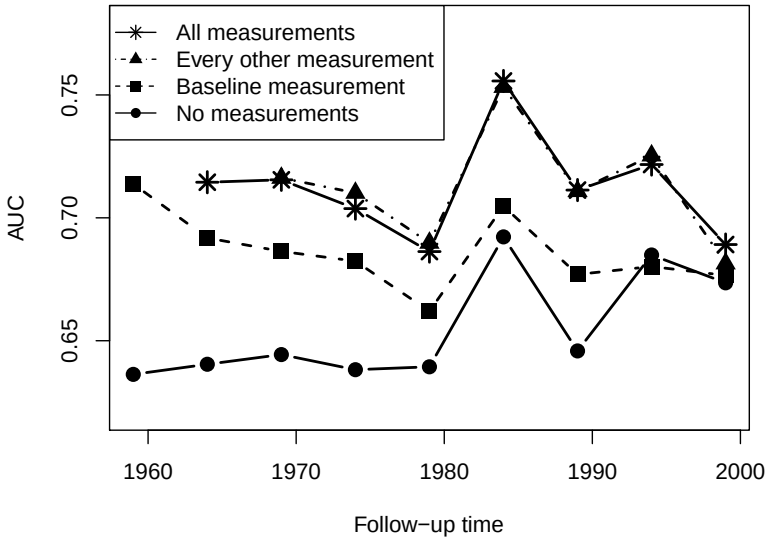

Figure 2: AUC estimates for the five-year predictions from the models using different amounts of longitudinal measurements in the East–West study. The estimates are plotted on the horizontal axis at the starting point of the prediction intervals.

In general, the increase in the predictive ability of the null model over time can be explained by the usual shape of a survival curve: when the cohort grows older, the difference in the survival rates between younger and older individuals becomes larger.

To evaluate whether increasing the number of longitudinal measurements improves the predictive ability significantly, we may compare the models using NRI and IDI indices. Pairwise model comparisons of the five-year predictions using NRI are presented in Figure 3. In most of the time intervals, the predictions are significantly improved by adding measurements. Every other measurement seems to be enough compared to carrying out all the measurements. However, it is clear that the predictions can be improved by carrying out some longitudinal measurements after the baseline measurement. In the last time interval, the covariate measurements do not seem to be very useful in predicting. Reasons for this may be the decreasing cohort size or a possibility of the diminishing predictive power of risk factors in elderly. It is also possible that the same prediction model does not fit at the

oldest ages.

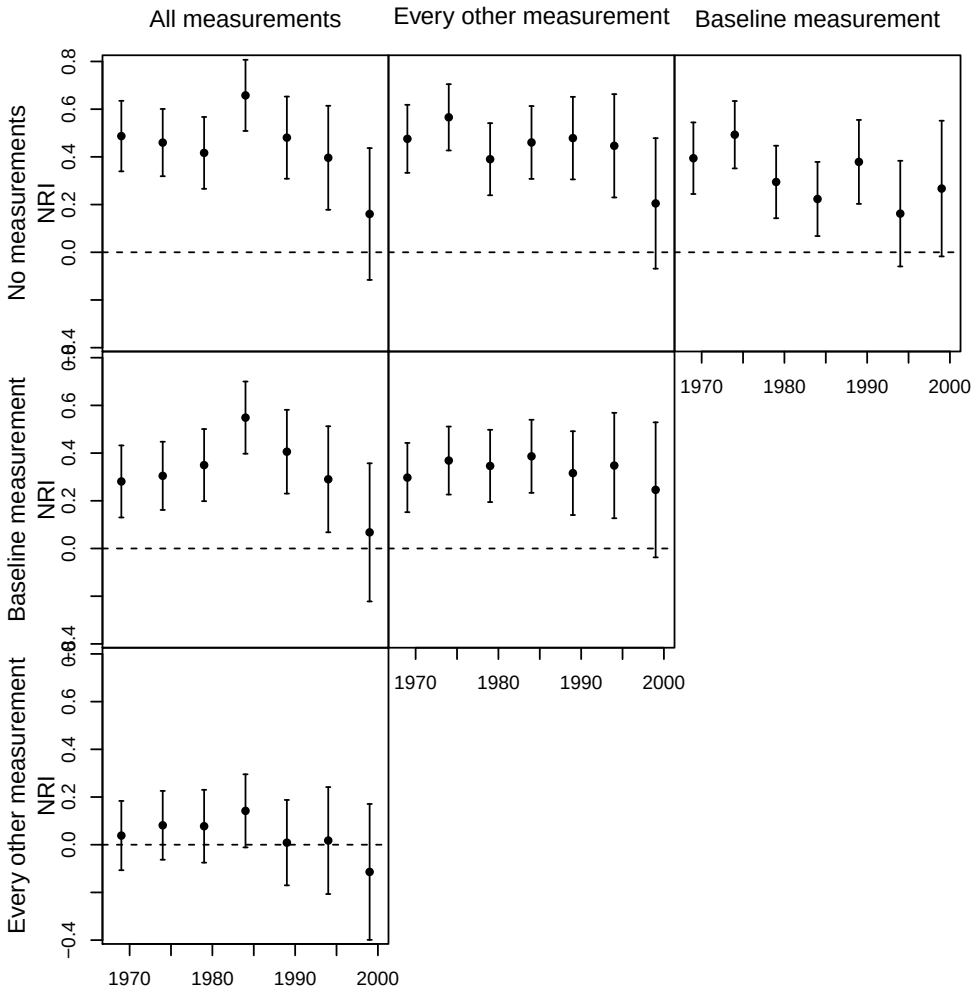

Figure 3: NRI estimates with 95% confidence intervals for pairwise comparisons of the five-year predictions from models using different amounts of longitudinal measurements in the East–West study. The estimates are plotted on the horizontal axis at the starting point of the prediction intervals. A positive value of an estimate means that the model named above the figure is better than the model named on the left-hand side of the figure.

We observe mostly the same patterns in the model comparisons, when IDI is used instead of NRI to measure the difference in predictive abilities (Figure 4). Increasing the number of measurements improves the predictions,

but carrying out all the measurements may bring negligible improvement compared to carrying out every other measurement.

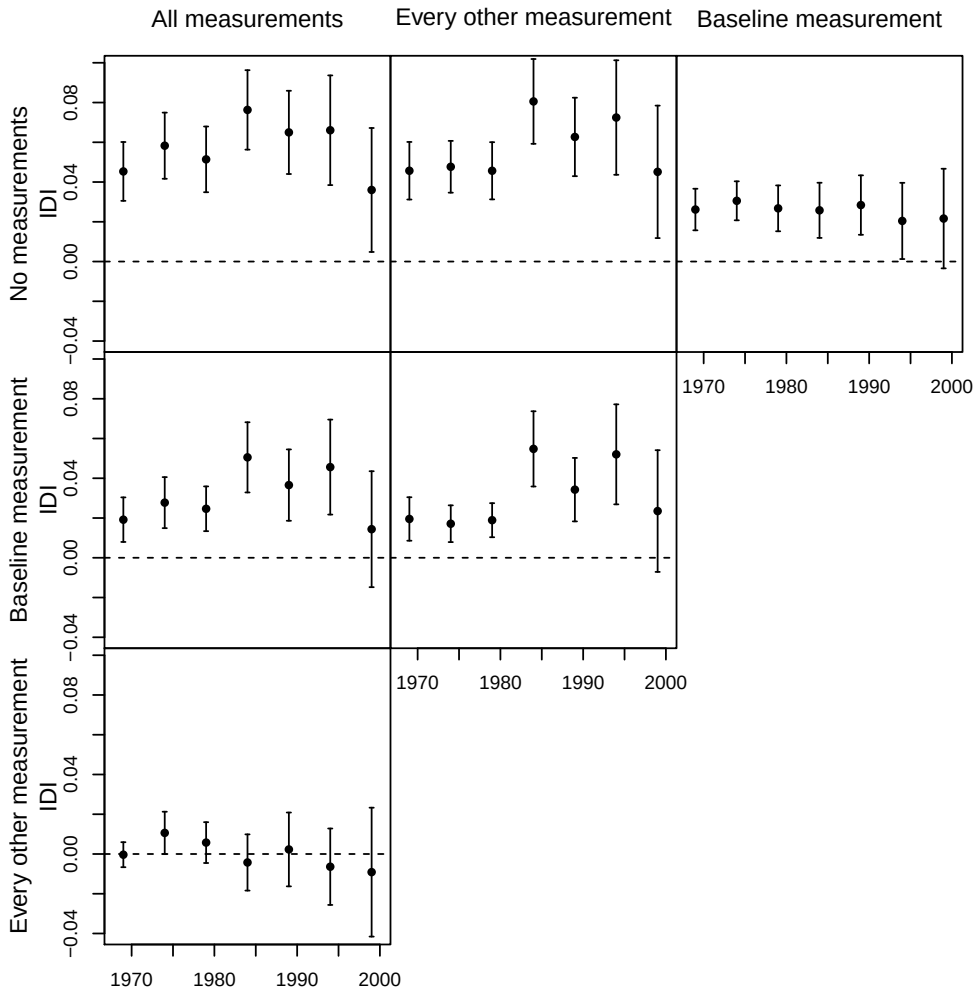

Figure 4: IDI estimates with 95% confidence intervals for pairwise comparisons of the five-year predictions from models using different amounts of longitudinal measurements in the East–West study. The estimates are plotted on the horizontal axis at the starting point of the prediction intervals. A positive value of an estimate means that the model named above the figure is better than the model named on the left-hand side of the figure.

NRI and IDI do not measure clinical consequences of using different models. In order to develop models to assist clinical decisions, the follow-up study

can be designed using decision-analytic measures. Here, we demonstrate the use of net benefit, where relative misclassification costs are given by a risk threshold. Figure 5 shows the differences in net benefit when other models are compared to Model (a). If we assume that further examinations are conducted for individuals with high risk but want to save costs by avoiding unnecessary examinations, a risk threshold of 10% could be used to assume that one true positive is worth nine false positives.

Comparisons of net benefit are presented for two different time intervals in Figure 5: years 1969–1974 and 1989–1994. For the latter interval, 10% risk threshold does not discriminate individuals well because increased age has increased also the risk of death. Hence, a higher threshold could be used for an older cohort. For instance, a 25% threshold would mean that one true positive is worth three false positives. The figure also shows that conclusions based on comparisons with respect to a single risk threshold may be very sensitive to the choice of the threshold.

To summarize our findings on the East–West study, we can say that measuring the covariates in ten-year intervals might be sufficient. By contrasting these results with correlations between longitudinal measurements (Table 1), we do not find similar relation as in the results of the simulation study of Section 3.2. Correlations mainly decrease over time, so one could expect that re-measurements would become more important. On the other hand, smoking status, which is a very strong predictor of all-cause mortality, has high correlations in the latter part of the follow-up.

5. Combining simulations and real data

The concepts of Sections 3 and 4 would perhaps be the most advantageous in a situation where the extension of a follow-up study having already at least two measurement times is to be planned. In this case we have some information on the correlations of the covariate measurements, which describes the variability of the covariates in time. We also have learned about the effects of the covariates on survival and at least two measurements allow us to investigate how the changes in the covariates affect survival, for instance. If we have earlier measured some expensive potential risk factor, which does not seem to have predictive power, this could be omitted from the upcoming examinations.

An ongoing follow-up study with at least two measurement times offers the information required for a reliable simulation. First, the next covariate

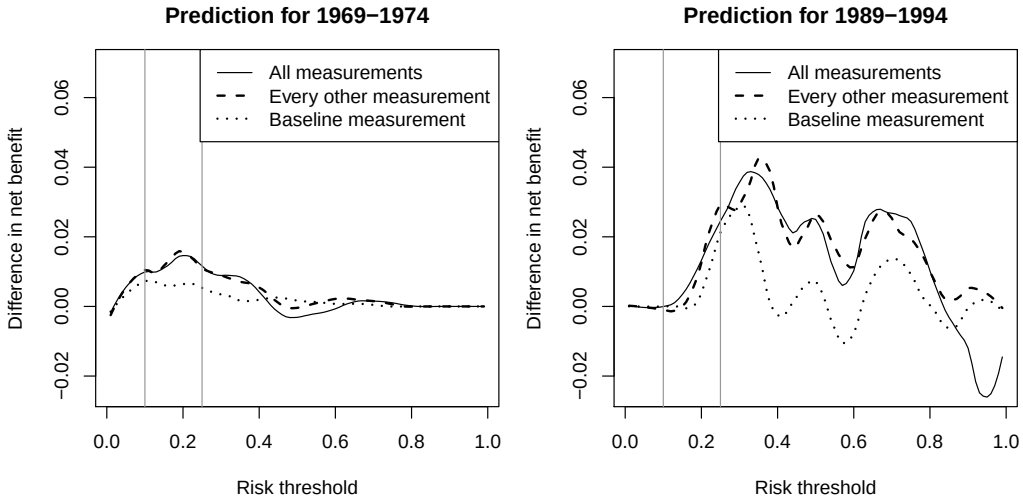

Figure 5: Smoothed curves showing the differences in net benefit compared to the model without any measurement information (Model (a)) for different risk thresholds on the horizontal axis. Vertical lines indicate the risk thresholds of 10% and 25%.

values are generated so that they correlate appropriately with the previous measurements. Then, forthcoming survival times are simulated conditioned on the new covariate values. Of course, the simulation of survival times can also be conditioned, for example, on the average of new simulated covariate values and previous observed values. After this, the predictions are calculated for the desired time interval with and without new simulated covariate values and finally these predictions are compared using measures of model performance. The procedure is replicated in the same simulation setting. The sensitivity of the choices made can be addressed by trying different simulation parameters and risk prediction models.

One way to obtain more accurate information about the covariate processes, is to carry out re-examinations for a small subcohort more frequently than for the entire cohort. This increases our knowledge about correlations between measurements in different time intervals and thus improves the simulations. Moreover, the subset of individuals can be selected optimally for longitudinal measurements to result in more precise estimates of the regression parameters than a simple random sample (4).

6. Discussion

Conducting a follow-up study requires large financial resources. To allocate the resources efficiently, the study has to be designed well. We considered methods to evaluate the number of longitudinal covariate measurements needed for the risk prediction. Simulations and data from a previous study was used to evaluate the importance of longitudinal measurements. Measures of model discrimination, namely AUC, NRI, IDI and net benefit, were used to compare models using different amounts of longitudinal information. It should be noted, however, that the application of the concepts presented in this paper does not require the use of these specific measures, but other metrics could be used as well.

As an example of using historical data, we evaluated the usefulness of longitudinal measurements in a long follow-up study, the East–West study. These data were suitable for our purposes, because the study contains eight longitudinal measurements and a follow-up period of more than 50 years. To our knowledge, these kinds of evaluations have not been presented before. By performing analyses considering the usefulness of longitudinal measurements, we can say to what extent we could decrease the number of the measurements without significantly losing the precision of the predictions.

The tools presented here seem to be helpful in designing follow-up studies, although they alone may not give exact answers to questions about the number of measurements needed. Because the decisions about study design have significant effects on costs, researchers should have some insight into the usefulness of longitudinal measurements and, particularly, effects of not carrying out longitudinal measurements should be understood. The collected data may be used for several purposes, and the usefulness of longitudinal measurements could be evaluated using several different models and variables to obtain a comprehensive understanding of the phenomenon to be studied.

Although the presented methodology is developed for follow-up studies in general, certain limitations should be noted. The tools may not be useful for completely new research questions for which no previous data or prior knowledge exists. As always in statistical modeling, the conclusions may depend on the modeling assumptions. With East–West data, splines were used to allow more flexible modeling but, for instance, interactions of the covariates were not present in the models and time-dependent Cox models cannot fully take into account the true variability of the covariates. The discussion on the

merits and demerits of different measures of model performance is expected to continue. Therefore, we do not provide recommendations of the measures to be preferred. In the East-West example the main conclusions were similar with all the measures considered.

The role of missing data was not studied in this article, although missingness of survival and especially longitudinal data is a common problem in follow-up studies. An individual may be lost during the follow-up, refuse to continue participation or miss a visit for some reason and then return to the study. Even if the missingness would be non-informative, it might reduce the power of the study, and more seriously, if the missingness would be informative, the estimates might also be considerably biased. The topic of missing longitudinal data is treated in more detail by, for example, Engels and Diehr (51) and Twisk (52).

This paper considered only one aspect of cost-efficiency in follow-up studies. More work is needed to combine the different viewpoints of cost-efficient follow-up designs. Future research will consider the optimal selection of individuals for longitudinal measurements (4) together with the determination of the reasonable number of measurements.

Acknowledgements

The research of the first author was supported by the Emil Aaltonen Foundation. Authors thank Antti Penttinen, Sara Taskinen and Satu Helske for helpful comments and suggestions.

Appendix

The model equations of Section 4.2 are as follows:

(a) $\lambda(t_i) = \lambda_0(t_i)$

(b) $\lambda(t_i|x(t_0)) = \lambda_0(t_i)\exp[\text{spline}(SBP(t_0), \text{df} = 4) + \text{spline}(CHOL(t_0), \text{df} = 2) + \text{spline}(BMI(t_0), \text{df} = 3) + \beta_1 HR(t_0) + \beta_2 Smoking(t_0)]$

(c) $\lambda(t_i|x(t)) = \lambda_0(t_i)\exp[\beta_1 SBP_{mean}(t) + \text{spline}(CHOL_{mean}(t), \text{df} = 2) + \text{spline}(BMI_{mean}(t), \text{df} = 5) + \beta_2 HR_{mean}(t) + \beta_3 Smoking_{mean}(t) + \text{spline}(SBP_{change}(t), \text{df} = 2) + \text{spline}(CHOL_{change}(t), \text{df} = 3) + \text{spline}(BMI_{change}(t), \text{df} = 3) + \text{spline}(HR_{change}(t), \text{df} = 5)]$

(d) $\lambda(t_i|x(t)) = \lambda_0(t_i)\exp[\text{spline}(SBP_{mean}(t), \text{df} = 2) + \text{spline}(CHOL_{mean}(t), \text{df} = 2) + \text{spline}(BMI_{mean}(t), \text{df} = 5) + \beta_1 HR_{mean}(t) + \beta_2 Smoking_{mean}(t) + \text{spline}(SBP_{change}(t), \text{df} = 5) + \text{spline}(CHOL_{change}(t), \text{df} = 4) + \text{spline}(BMI_{change}(t), \text{df} = 2) + \beta_3 HR_{change}(t)]$

where t_0 refers to the time of the baseline measurements and i to an individual. The time-dependent mean and change variables in Models (c) and (d) are calculated using the measurements available, i.e. every other in Model (c) and all the measurements in Model (d). The function 'spline()' means a smoothing spline (48). The degrees of freedom (df) controlling the amount of smoothing are selected from integers between 1 and 5, where df = 1 means a linear effect, using Akaike information criterion.

References

[1] R. Clarke, M. Shipley, S. Lewington, L. Youngman, R. Collins, M. Marmot, R. Peto, Underestimation of risk associations due to regression dilution in long-term follow-up of prospective studies, American Journal of Epidemiology 150 (4) (1999) 341–353.

[2] O. Aalen, K. Røysland, J. Gran, R. Kouyos, T. Lange, Can we believe the DAGs? A comment on the relationship between causal DAGs and mechanisms, Statistical Methods in Medical Research (January 23, 2014) doi: 10.1177/0962280213520436.

[3] A. Sokol, N. R. Hansen, Causal interpretation of stochastic differential equations, Electronic Journal of Probability 19 (100) (2014) 1–24.

[4] J. Reinikainen, J. Karvanen, H. Tolonen, Optimal selection of individuals for repeated covariate measurements in follow-up studies, Statistical Methods in Medical Research (February, 24, 2014) doi: 10.1177/0962280214523952.

[5] L. Y. Inoue, G. Parmigiani, Designing follow-up times, Journal of the American Statistical Association 97 (459) (2002) 847–858.

[6] G. M. Raab, J. Davies, A. B. Salter, Designing follow-up intervals, Statistics in Medicine 23 (20) (2004) 3125–3137.

[7] J. Karvanen, S. Kulathinal, D. Gasbarra, Optimal designs to select individuals for genotyping conditional on observed binary or survival outcomes and non-genetic covariates, Computational Statistics & Data Analysis 53 (5) (2009) 1782–1793.

[8] F. B. Tekle, F. E. Tan, M. P. Berger, Too many cohorts and repeated measurements are a waste of resources, Journal of Clinical Epidemiology 64 (12) (2011) 1383–1390.

[9] J. A. Dubin, L. Han, T. R. Fried, Triggered sampling could help improve longitudinal studies of persons with elevated mortality risk, Journal of Clinical Epidemiology 60 (3) (2007) 288–293.

[10] F. Harrell, K. L. Lee, D. B. Mark, Tutorial in biostatistics: multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, Statistics in Medicine 15 (1996) 361–387.

[11] M. J. Pencina, R. B. D'Agostino, R. S. Vasan, Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond, Statistics in Medicine 27 (2) (2008) 157–172.

[12] B. Van Calster, A. J. Vickers, M. J. Pencina, S. G. Baker, D. Timmerman, E. W. Steyerberg, Evaluation of markers and risk prediction models overview of relationships between NRI and decision-analytic measures, Medical Decision Making 33 (4) (2013) 490–501.

[13] J. Wong, M. Taljaard, A. J. Forster, G. J. Escobar, C. van Walraven, Addition of time-dependent covariates to a survival model significantly improved predictions for daily risk of hospital death, Journal of Evaluation in Clinical Practice 19 (2) (2013) 351–357.

[14] R. Sutradhar, C. Atzema, H. Seow, C. Earle, J. Porter, L. Barbera, Repeated assessments of symptom severity improve predictions for risk of death among patients with cancer, Journal of Pain and Symptom Management 48 (6) (2014) 1041–1049.

[15] M. J. Karvonen, G. Blomqvist, V. Kallio, E. Orma, S. Punsar, P. Rautaharju, J. Takkunen, A. Keys, Men in rural East and West Finland, Acta Medica Scandinavica 180 (1966) 169–190.

[16] A. Keys, Coronary heart disease in seven countries., Circulation 41 (1) (1970) 186–195.

[17] L. D. Fisher, D. Y. Lin, Time-dependent covariates in the Cox proportional-hazards regression model, Annual Review of Public Health 20 (1) (1999) 145–157.

[18] R. Henderson, P. Diggle, A. Dobson, Joint modelling of longitudinal measurements and event time data, Biostatistics 1 (4) (2000) 465–480.

[19] D. Rizopoulos, Joint models for longitudinal and time-to-event data: With applications in R, CRC Press, 2012.

[20] R. Prentice, Covariate measurement errors and parameter estimation in a failure time regression model, Biometrika 69 (2) (1982) 331–342.

[21] T. E. Hanson, A. J. Branscum, W. O. Johnson, Predictive comparison of joint longitudinal-survival modeling: a case study illustrating competing approaches, Lifetime Data Analysis 17 (1) (2011) 3–28.

[22] X. Song, M. Davidian, A. A. Tsiatis, A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data, Biometrics 58 (4) (2002) 742–753.

[23] C. Proust-Lima, P. Joly, J.-F. Dartigues, H. Jacqmin-Gadda, Joint modelling of multivariate longitudinal outcomes and a time-to-event: a nonlinear latent class approach, Computational Statistics & Data Analysis 53 (4) (2009) 1142–1154.

[24] E. R. Brown, J. G. Ibrahim, V. DeGruttola, A flexible B-spline model for multiple longitudinal biomarkers and survival, Biometrics 61 (1) (2005) 64–73.

[25] D. Rizopoulos, P. Ghosh, A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event, Statistics in Medicine 30 (12) (2011) 1366–1380.

[26] J. Reinikainen, T. Laatikainen, J. Karvanen, H. Tolonen, Lifetime cumulative risk factors explain cardiovascular disease mortality in a 50-year follow-up study in Finland, International Journal of Epidemiology 44 (1) (2015) 108–116.

[27] P. M. Rothwell, S. C. Howard, E. Dolan, E. O'Brien, J. E. Dobson, B. Dahlöf, P. S. Sever, N. R. Poulter, Prognostic significance of visit-to-visit variability, maximum systolic blood pressure, and episodic hypertension, The Lancet 375 (9718) (2010) 895–905.

[28] H. D. Sesso, M. J. Stampfer, B. Rosner, J. M. Gaziano, C. H. Hennekens, Two-year changes in blood pressure and subsequent risk of cardiovascular disease in men, Circulation 102 (3) (2000) 307–312.

[29] J. A. Hanley, B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1982) 29–36.

[30] L. E. Chambless, C. P. Cummiskey, G. Cui, Several methods to assess improvement in risk prediction models: extension to survival analysis, Statistics in Medicine 30 (1) (2011) 22–38.

[31] M. S. Pepe, The statistical evaluation of medical tests for classification and prediction, Oxford University Press, 2003.

[32] T. J. Wang, P. Gona, M. G. Larson, G. H. Tofler, D. Levy, C. Newton-Cheh, P. F. Jacques, N. Rifai, J. Selhub, S. J. Robins, et al., Multiple biomarkers for the prediction of first major cardiovascular events and death, New England Journal of Medicine 355 (25) (2006) 2631–2639.

[33] M. R. Spitz, C. Amos, A. D'Amelio Jr, Q. Dong, C. Etzel, Re: Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk, JNCI Journal of the National Cancer Institute 101 (24) (2009) 1731.

[34] S. G. Baker, E. Schuit, E. W. Steyerberg, M. J. Pencina, A. Vickers, K. G. Moons, B. W. Mol, K. S. Lindeman, How to interpret a small increase in AUC with an additional risk prediction marker: decision analysis comes through, Statistics in Medicine 33 (22) (2014) 3946–3959.

[35] A. J. Vickers, M. Pepe, Does the net reclassification improvement help us evaluate models and markers?, Annals of Internal Medicine 160 (2) (2014) 136–137.

[36] J. Hilden, Commentary: On NRI, IDI, and "good-looking" statistics with nothing underneath, Epidemiology 25 (2) (2014) 265–267.

[37] M. J. Pencina, R. B. D'Agostino, E. W. Steyerberg, Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers, Statistics in Medicine 30 (1) (2011) 11–21.

[38] A. J. Vickers, E. B. Elkin, Decision curve analysis: a novel method for evaluating prediction models, Medical Decision Making 26 (6) (2006) 565–574.

[39] P. M. Ridker, N. R. Cook, Statins: new American guidelines for prevention of cardiovascular disease, The Lancet 382 (9907) (2013) 1762–1765.

[40] D. M. Lloyd-Jones, D. Goff, N. J. Stone, Statins, risk assessment, and the new American prevention guidelines, The Lancet 383 (9917) (2014) 600–602.

[41] N. R. Cook, Use and misuse of the receiver operating characteristic curve in risk prediction, Circulation 115 (7) (2007) 928–935.

[42] M. J. Leening, E. W. Steyerberg, B. Van Calster, R. B. D'Agostino, M. J. Pencina, Net reclassification improvement and integrated discrimination improvement require calibrated models: relevance from a marker and model perspective, Statistics in Medicine 33 (19) (2014) 3415–3418.

[43] B. Van Calster, A. J. Vickers, Calibration of risk prediction models impact on decision-analytic performance, Medical Decision Making 35 (2) (2015) 162–169.

[44] D. W. Hosmer, S. Lemeshow, Goodness of fit tests for the multiple logistic regression model, Communications in Statistics – Theory and Methods 9 (10) (1980) 1043–1069.

[45] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, M. W. Kattan, Assessing the performance of prediction models: a framework for some traditional and novel measures, Epidemiology 21 (1) (2010) 128.

[46] R. Bender, T. Augustin, M. Blettner, Generating survival times to simulate Cox proportional hazards models, Statistics in Medicine 24 (11) (2005) 1713–1723.

[47] P. C. Austin, Generating survival times to simulate Cox proportional hazards models with time-varying covariates, Statistics in Medicine 31 (29) (2012) 3946–3958.

[48] T. M. Therneau, P. M. Grambsch, Modeling survival data: extending the Cox model, Springer, 2000.

[49] J. Afilalo, S. Karunananthan, M. J. Eisenberg, K. P. Alexander, H. Bergman, Role of frailty in patients with cardiovascular disease, The American Journal of Cardiology 103 (11) (2009) 1616–1621.

[50] J. S. Farhat, V. Velanovich, A. J. Falvo, H. M. Horst, A. Swartz, J. H. Patton Jr, I. S. Rubinfeld, Are the frail destined to fail? Frailty index as predictor of surgical morbidity and mortality in the elderly, The Journal of Trauma and Acute Care Surgery 72 (6) (2012) 1526–1531.

[51] J. M. Engels, P. Diehr, Imputation of missing longitudinal data: a comparison of methods, Journal of Clinical Epidemiology 56 (10) (2003) 968–976.

[52] J. W. Twisk, Applied longitudinal data analysis for epidemiology: a practical guide, Cambridge University Press, 2013.