Replication study of Vittinghof and McCulloch’s (2007) simulation study.

Rick Nijman Jolien Ketelaar [1], Lieke Hesen [2]

[1] ~

[2] ~

June 17, 2021

Abstract

`<a summary of the replication effort>`

Correspondence concerning this replication report should be addressed to:

...

1 Introduction

This replication report documents the replication attempt of the simulation study of Vittinghoff and McCulloch (2007). Following the definition of Rougier et al. (2017) we understand the replication of a published study as writing and running new code based on the description provided in the original publication with the aim of obtaining the same results.

2 Method

2.1 Information basis

The original article is solely used as an information basis.

2.2 Data Generating Mechanism

Information provided in the above mentioned sources indicated that the following simulation factors were systematically varied in generating the artificial data.

Simulation factor	No. levels	Levels
General		
Events per variable	15	2,3,4,5,6,7,8,9,10,11,12,13,14,15,16
Number of predictors	4	2,4,8,16
Sample sizes	4	128, 256, 512, 1024
B1 (coefficient primary predictor)	4	0,log(1.5),log(2),log(4)
Pairwise correlation (with binary predictor of other predictors)	1 (fixed)	0.25
Specific for binary primary predictor		
Expected prevalence of primary predictor	3	0.1, 0.25, 0.5
Multiple correlation of primary predictor with covariates	4	0,0.25,0.5,0.75
Specific for continuous primary predictor		
Variance primary predictor	1 (fixed)	0.16
Multiple correlation of primary predictor with covariates	5	0,0.1,0.25, 0.5, 0.9

2.2.1 Events per variable

“We considered values for the EPV from two to 16[…]” (p. 710) althus the authors.

2.2.2 Number of predictors

Four different levels are used for the number of predictors, namely “[…] models with a total of two, four, eight and 16 predictor variables […]” (p. 710).

2.2.3 Sample sizes

The authors describe the sample sizes as follows: “[…] sample sizes of 128, 256, 512 and 1,024 […]” (p.710).

2.2.4 B1 (coefficient primary predictor)

The regression coefficients of B1, from the primary predictor on the outcome variable range from 0 to 0.75, as described by: “[…] values of B1, the regression coefficient for the primary predictor, of 0, 0.25, 0.5, or 0.75 […]” (p.710).

2.2.5 Pairwise correlation (with binary predictor of other predictors)

The authors state that: “With a binary primary predictor, the other predictors were multivariate normal with pairwise correlation of 0.25.” (p. 710). Later on the authors state that: “With the continuous primary predictor, all predictors were multivariate normal and equally intercorrelated.” (p. 711). Therefore, we assumed that the pairwise correlation of 0.25 is also used for the continuous predictor and is not varied over scenarios.

2.2.6 Expected prevalence

“The binary primary predictor was generated with expected prevalence of 0.1, 0.25, 0.5, 0.75.” (p. 710).

2.2.7 Multiple correlation of primary predictor with covariates

The authors state that: “The binary predictor was generated with […], and multiple correlation with the covariates of 0, 0.25, 0.5 and 0.75” (p. 711). Later on the authors state that: ”[…] for comparability with the binary predictors, and the multiple correlation between the primary predictor and adjustment variables was set to 0, 0.1, 0.25, 0.5 or 0.9 (p.711- p.712). The authors are first talking about covariates and later about adjustment variables. We interpreted this as the same concept: the correlation between primary predictor and covariance matrix.

2.2.8 Variance primary predictor and covariates

Vittinghof and McCulloch (2007) describe “The variance of the continuous primary predictor was set to 0.16” (p. 710). This was constant over all conditions with the continuous primary predictor. Neither the variance for the binary primary predictor, nor the variance of the covariates was described.

2.2.9 Omit extreme cases

Vittinghof and McCulloch (2007) state that: “The factorial omitted extreme cases with outcome prevalence of greater than 50 percent” (p.710).

We assumed that the authors did this in the following way. In each factor the number of predictors and the number of events per variable is known. The number of predictors multiplied by the events per variable is the number of events per factor. Also, the sample size is determined beforehand. The number of events divided by the sample size is the outcome prevalence. Each factor has its own outcome prevalence, when the outcome prevalence was higher than 0.5, we filtered out the cases.

This method leads to 10176 for the binary predictor and 4240 for the continuous, where the authors examined 9328 and 3392 scenarios (for details on the calculation the code is included on github).

2.3 Data generating mechanism

2.3.1 The aggregate effect

Vittinghof and McCulloch (2007) state that: “The aggregate effect of the covariates is held constant across models with two, four, eight and 16 predictors” (p. 711). However, the aggregate effect is not specified.

2.3.2 Binary outcome data

2.3.3 Retrospective sampling

Vittinghof and McCulloch (2007) state that: “In logistic models, we kept the first “cases” and “controls” generated up to the required numbers of each, taking advantage of the fact that under the logistic model only the intercept is affected by such retrospective sampling” (p. 710-711). It is unclear how this procedure is conducted. We assumed that the authors oversampled cases and controls and kept the required first cases and controls. Unfortunately, it was not defined how many cases and controls were oversampled.

2.3.4 Logistic model with binary predictor

Vittinghof and McCulloch (2007) are not giving information about how they generated the data for the logistic model with the binary predictor. It is unclear how the binary primary predictor with the predefined prevalence is generated while keeping the predefined correlation structure intact.

2.3.5 Time-to-event data

Vittinghof & McCulloch (2007) provide rather limited data on how the data for the cox model was generated. It is unclear which distribution underlies the survival analysis. One can guess by looking back at earlier mentioned articles, such as Peduzzi et al. (1995). However Vittinghof & McCulloch do not state they follow a similar procedure. Not knowing which distribution (e.g. Weibull) makes the simulation impossible to replicate.

2.3.6 Additional simulations

The original authors describe that they also “examined models with all binary predictors” (p.717), again with a logistic and Cox model. They did not provide any parameters they used in these additional simulations.

2.4 Compared Methods

NA. We did not compare methods since we were unable to generate the data.

2.5 Performance measures

NA. We did not evaluate performance measures since we were unable to generate the data.

2.6 Technical implementation

NA. We did not implement the simulation study since we were unable to generate the data.

3 Results

NA. No results were obtained since we were unable to generate the data. We did find it striking that in Table 1 of the original article, the percentage for the problem of a relative bias from above 15 percent was higher than the percentage of any of the three problems for more than half of the scenarios. This probably has to do with the exclusion of some scenarios for which the decision criteria were not described. Furthermore, Table 2 of the original article shows percentages of higher than 100%, e.g. 260.1 for 2-4 events per variable for the problem of maximum relative bias. Here, an explanation is missing.

3.1 Replicability

The paper of Vittinghof and McCulloch (2007) is not replicable. We cannot recreate the factorial as described in the paper, because the data generating parameters are insufficiently described. We attempted to guess what the intended set-up of the simulation was. Nevertheless, we attempted to recreate the intended set-up by making a best guess of the factorial. This resulted in 10176 and 4240 scenarios for the binary and continuous primary predictors, whereas the authors of the original paper mention 9328 and 3392 scenarios. Even when making this best guess, many subsequent factors were still unknown, such as the underlying model of the Cox regression. In conclusion, the provided information was so scarce that the replicator degrees of freedom were too large to perform any of the subsequent steps of the analysis.

3.2 Replicator degrees of freedom

We were not able to replicate this study because the replicator degrees of freedom tended to go to infinity. The table below specifies the issue, our decision and justification if applicable.

Issue	*Replicator decision*	*Justification*
Variance binary primary predictor and covariates unknown	We decided to use a variance of 1.	A variance of 1 seemed the most convenient choice, because this means the covariance matrix is equivalent to the correlation matrix. This choice, however, resulted in some covariance matrices that were non-positive definite
Generation of correlation binary primary predictor with covariates while keeping the multivariate normal covariance structure	We used a logistic regression model, where we tried to find beta values such that the correlation would resemble the described values.	To our knowledge, this was the only way to keep the overall covariance structure. We didn't know a way to simulate both binary and continuous variables given a certain covariance matrix. We were not able to find a beta value to get to a correlation of 0.75 between the primary predictor and covariates.
The aggregate effect of the covariates was unknown	This unknown was not resolved	-
The underlying distribution of the survival analysis is unknown	The decision was made to stop the replication study	We could have looked back at older papers, however the authors do not specify they followed the same procedure. The amount of degrees of freedom lead to the decision to stop the replication

Issue	*Replicator decision*	*Justification*
The pairwise correlation of the continuous primary predictor was not explicitly specified	We assume the same pairwise correlation as was specified for the binary predictor	The authors state that: “With the continuous primary predictor, all predictors were multivariate normal and equally intercorrelated.” (p. 711). Therefore, we assumed that the pairwise correlation of 0.25 is also used for the continuous predictor and is not varied over scenarios.
The authors use both the term covariates and the term adjustment variables	We interpreted this as the same concept: the correlation between primary predictor and covariance matrix.	The authors do not specify which is meant by adjustment variables otherwise
The way the factorial omitted extreme cases is unclear	We assumed that the authors did this in the following way. In each factor the number of predictors and the number of events per variable is known. The number of predictors multiplied by the events per variable is the number of events per factor. Also, the sample size is determined beforehand. The number of events divided by the sample size is the outcome prevalence. Each factor has its own outcome prevalence, when the outcome prevalence was higher than 0.5, we filtered out the cases.	This made the most sense

3.3 Equivalence of results

NA. Results were not obtained since we were unable to generate the data.

4 Acknowledgments

We want to acknowledge Kim Luijken and David Koops for their help during the difficult steps of this replication study. They provided us with insights and tips, which inspired us to try and replicate the study. Unfortunately their enthusiasm did not help us do the impossible: replicate the study.

5 References

Vittinghoff, E., & McCulloch, C. E. (2007). Relaxing the rule of ten events per variable in logistic and Cox regression. American journal of epidemiology, 165(6), 710-718.
Peduzzi P, Concato J, Feinstein AR, et al. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol 1995;48:1503–10
Rougier, Nicolas P., Konrad Hinsen, Frédéric Alexandre, Thomas Arildsen, Lorena A. Barba, Fabien C. Y. Benureau, C. Titus Brown, et al. 2017. "Sustainable Computational Science: The ReScience Initiative." *PeerJ Computer Science* 3 (December): e142. https://doi.org/10.7717/peerj-cs.142.