

The Problem with Univariable Selection in Regression Modelling — and What to Do Instead

Theresa Ullmann, Georg Heinze, Franziska Kappenberg, Marc Henrion, Willi Sauerbrei, Gary Collins, Carl-Stephan Leonhardt, Mariana Nold, Daniela Dunkler

for Topic Group 2 of the STRATOS initiative

ISCB, October 1st, 2026, Freiburg

Multivariable modeling

- Multivariable regression modeling relates multiple variables to an outcome of interest
- E.g., linear, logistic, or Cox regression
- For different purposes, such as:
 - Description: e.g., what are the most important factors associated with high blood pressure?
 - Prediction: e.g., what is a patient's risk of 30-day mortality after surgery?

Multivariable modeling

- Multivariable modeling begins with set of candidate variables (based on background knowledge, data availability,...)
- In many studies, large models are not desirable
 - If several competing models describe the observed outcomes equally well, simpler models are preferred
 - Simpler prediction models are often easier to communicate and implement in practice

→ data-driven variable selection approaches

Univariable selection

Among data-driven selection approaches, **univariable selection** is one of the **most widely used** in health research

(Vacheron et al. 2021, Su et al. 2026)

Procedure:

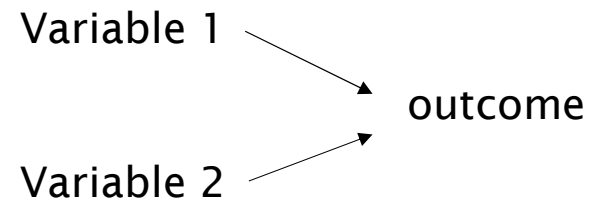
- Each variable is analyzed separately in a univariable model
- Only those showing a statistically significant association with the outcome (typically $\alpha = 0.05$) are selected for the multivariable model

But: univariable selection lacks a sound statistical basis!

Why univariable selection can fail

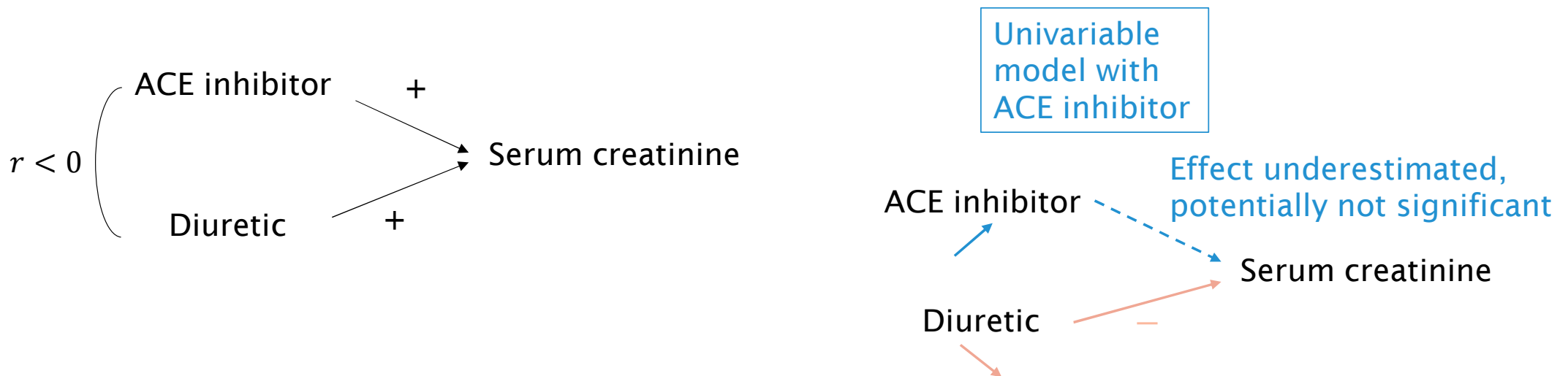
Three mechanisms:

- Masking
- Amplification
- Loss of precision



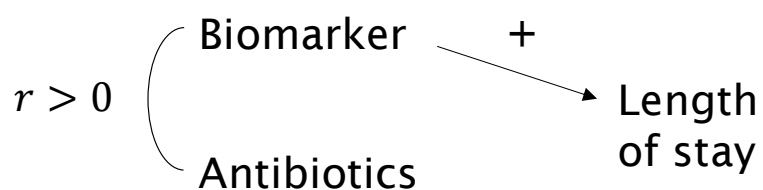
Masking

- Variables: doses of two antihypertensive drugs (e.g., diuretic and angiotensin-converting enzyme inhibitor)
- Outcome: serum creatinine (continuous measure of kidney function)
- Higher doses of either drug increase serum creatinine
- But: drug doses negatively correlated ($r < 0$) → effects obscured in univariable models

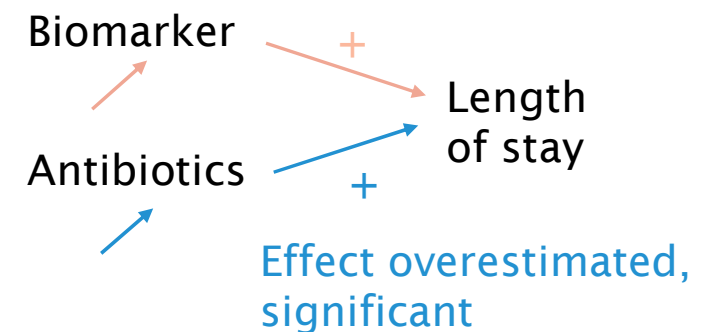


Amplification

- Variables: 1) inflammatory biomarker of disease severity, 2) prescription of broad-spectrum antibiotics
- Outcome: length of hospital stay
- Higher biomarker levels associated with longer hospital stay, antibiotic itself has no direct effect on the outcome
- Variables positively correlated ($r > 0$) → antibiotic use may appear to be associated with longer stay in univariable model

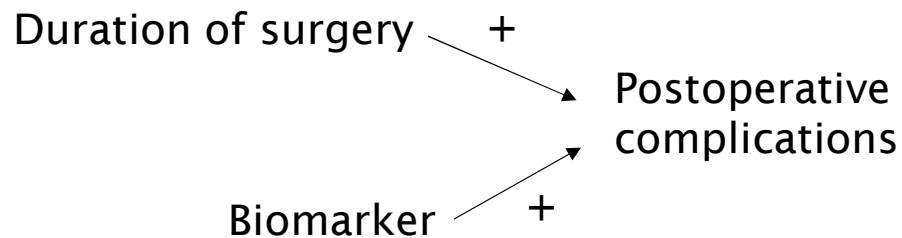


Univariable model
with antibiotics

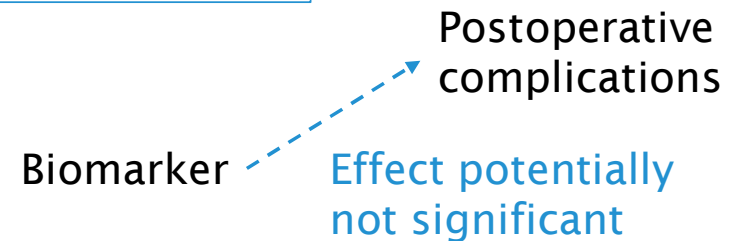


Loss of precision

- Variables: 1) duration of surgery, 2) preoperative biomarker
- Outcome: postoperative complications
- Even if both variables are truly associated with the outcome, a univariable analysis may fail to detect them
- Statistical significance depends not only on effect size, but also on how precisely it is estimated!

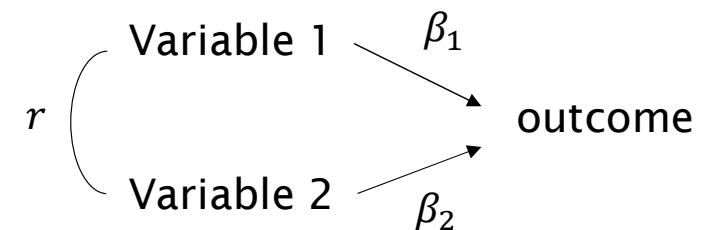


Univariable model
with biomarker



A simple simulation study

- Three mechanisms can be studied more systematically via simulation
- Simulation studies use artificially generated data in which true relationships between variables are known
- Allows us to assess whether univariable selection can recover these relationships, e.g.,
 - how often variables are wrongly included or excluded
 - whether estimated effects in selected model are biased
 - whether confidence intervals have appropriate coverage
- We can vary parameters such as correlation, effect sizes, sample size,...



Masking scenario

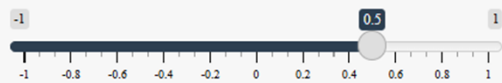
One variable acts as a suppressor of the other because the two effect coefficients (β_1 , β_2) and the correlation (r) are arranged so that their product is *negative* [$\text{sign}(\beta_1 \times \beta_2 \times r) < 0$]. At least one univariable p-value is inflated relative to the multivariable model. Univariable preselection is likely to drop it even if it truly belongs in the model. See Introduction for details.

True model parameters

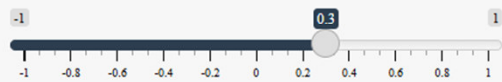
Linear model:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

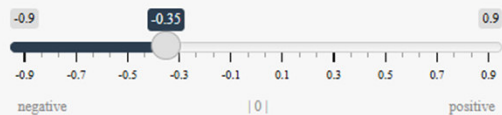
β_1 (X1)



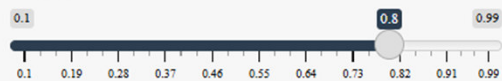
β_2 (X2)



Correlation of X1 and X2 (r)



R^2 (population)



Study & simulation settings

Sample size (n)

$n = 100$

100

α (selection threshold)

Introduction

1. Selection Frequency

2. Coefficient Estimates

3. p-Value Distributions

4. Bias & Variance

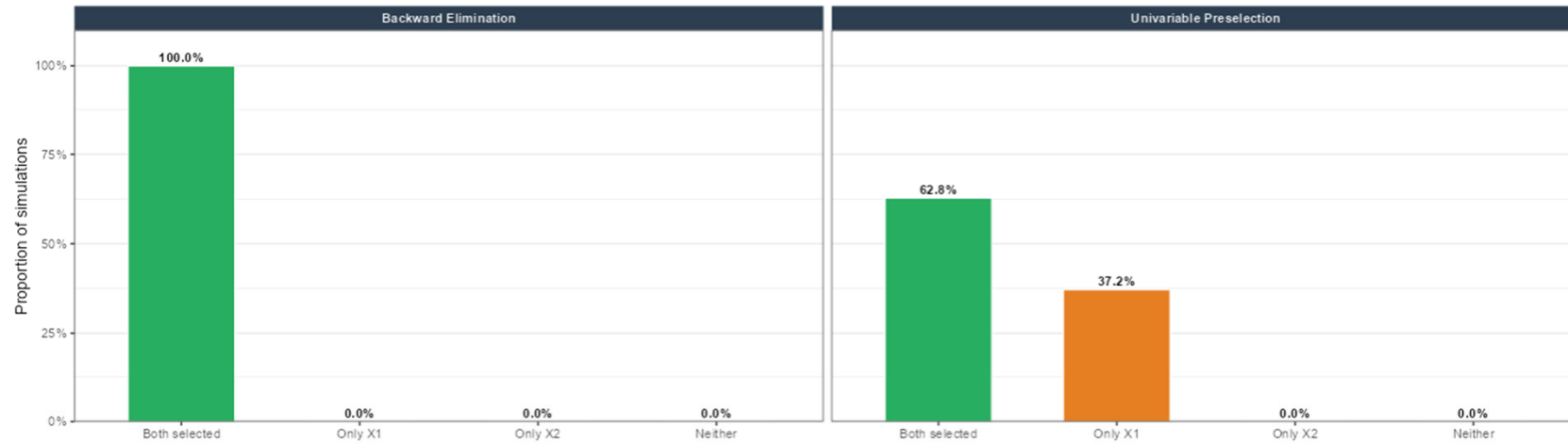
5. Scatterplot of Estimates

6. Sensitivity Analysis

7. Single Dataset

Variable Selection Frequency

nsim = 500 | n = 100 | beta1 = 0.50 | beta2 = 0.30 | r = -0.35 | R2 = 0.80 | alpha = 0.05



Interpretation

- **Both selected** corresponds to the correct decision. A higher proportion in the simulation is better.
- **Only X₁ / Only X₂** means that one true predictor was missed (false negative).
- **Neither** corresponds to both predictors dropped by the variable selection method. This is the worst outcome.
- This is a **masking scenario** [$\text{sign}(\beta_1 \times \beta_2 \times r) < 0$]: one variable masks the effect of the other in a univariable test, inflating its p-value. Univariable preselection is particularly likely to drop a variable that genuinely belongs in the model.
- $R^2 = 0.8$. A higher R^2 increases statistical power, so both methods are more likely to select both X_1 and X_2 correctly. A lower R^2 makes correct selection harder for both methods, but the advantage of backward elimination over univariable preselection is typically maintained.

Numerical results (500 simulations)

Selection outcome	Backward Elimination	Univariable Preselection
Both selected	100.0%	62.8%
Only X ₁	0.0%	37.2%
Only X ₂	0.0%	0.0%
Neither	0.0%	0.0%



MEDICAL UNIVERSITY
OF VIENNA

STRATOS
INITIATIVE

Theresa Ullmann
Center for Medical Data Science - Institute of Clinical Biometrics



Masking scenario

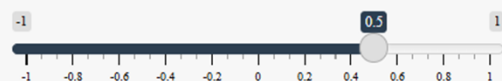
One variable acts as a suppressor of the other because the two effect coefficients (β_1 , β_2) and the correlation (r) are arranged so that their product is *negative* [$\text{sign}(\beta_1 \times \beta_2 \times r) < 0$]. At least one univariable p-value is inflated relative to the multivariable model. Univariable preselection is likely to drop it even if it truly belongs in the model. See Introduction for details.

True model parameters

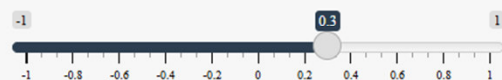
Linear model:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

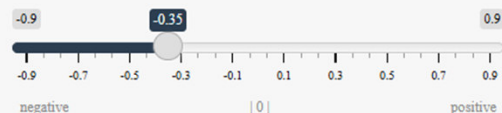
β_1 (X1)



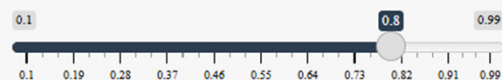
β_2 (X2)



Correlation of X1 and X2 (r)



R^2 (population)



Study & simulation settings

Sample size (n)

$n = 50$

50

α (selection threshold)

Introduction

1. Selection Frequency

2. Coefficient Estimates

3. p-Value Distributions

4. Bias & Variance

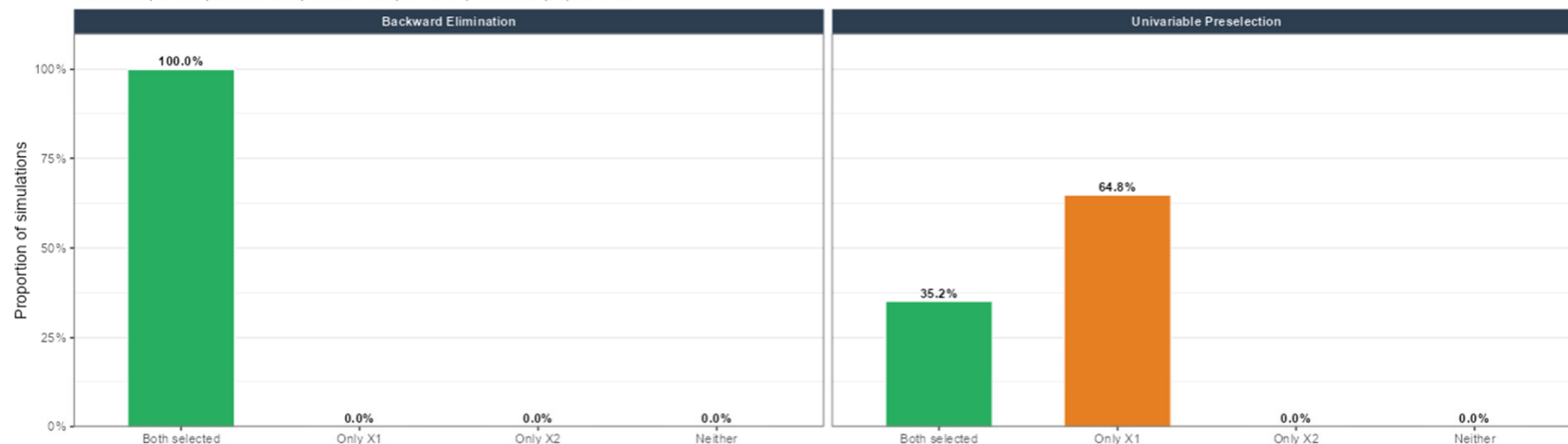
5. Scatterplot of Estimates

6. Sensitivity Analysis

7. Single Dataset

Variable Selection Frequency

nsim = 500 | n = 50 | beta1 = 0.50 | beta2 = 0.30 | r = -0.35 | R2 = 0.80 | alpha = 0.05



Interpretation

- **Both selected** corresponds to the correct decision. A higher proportion in the simulation is better.
- **Only X₁ / Only X₂** means that one true predictor was missed (false negative).
- **Neither** corresponds to both predictors dropped by the variable selection method. This is the worst outcome.
- This is a **masking scenario** [$\text{sign}(\beta_1 \times \beta_2 \times r) < 0$]: one variable masks the effect of the other in a univariable test, inflating its p-value. Univariable preselection is particularly likely to drop a variable that genuinely belongs in the model.
- $R^2 = 0.8$. A higher R^2 increases statistical power, so both methods are more likely to select both X_1 and X_2 correctly. A lower R^2 makes correct selection harder for both methods, but the advantage of backward elimination over univariable preselection is typically maintained.

Numerical results (500 simulations)

Selection outcome	Backward Elimination	Univariable Preselection
Both selected	100.0%	35.2%



Univariable selection also fails in more realistic scenarios

- So far: two variables
- We also simulated realistic data with 20 variables (ten true predictors, ten noise variables)
- Univariable selection often made selection errors
- Downstream consequences for selected model:
 - biased and imprecise estimates of regression coefficients
 - confidence intervals often did not achieve their nominal coverage
 - predictive performance was also often inferior
- Most pronounced in smaller samples and in settings with weaker signal-to-noise

Real-world case study: biomarker for cancer prognosis

- Tong et al. (2010) assessed the prognostic relevance of the gene expression of peripheral myelin protein 22 (PMP22) in breast cancer
- Retrospective multicenter study in Austria, 249 primary breast tumours
- PMP22 expression was measured alongside established clinical factors such as nodal status, estrogen receptor (ER) status, tumor size, and differentiation grade
- Outcome: disease-free survival, i.e., time between diagnosis of disease and recurrence or distant metastasis (109 events during a median observation time of 8.89 years)

Real-world case study: biomarker for cancer prognosis

- In *univariable* Cox regression: PMP22 expression not significant at $\alpha = 0.05$ level
→ would have been excluded by univariable selection
- In *multivariable* Cox regression: patients with high (above-median) PMP22 expression had a 1.77-fold increased risk of relapse compared to those with lower expression

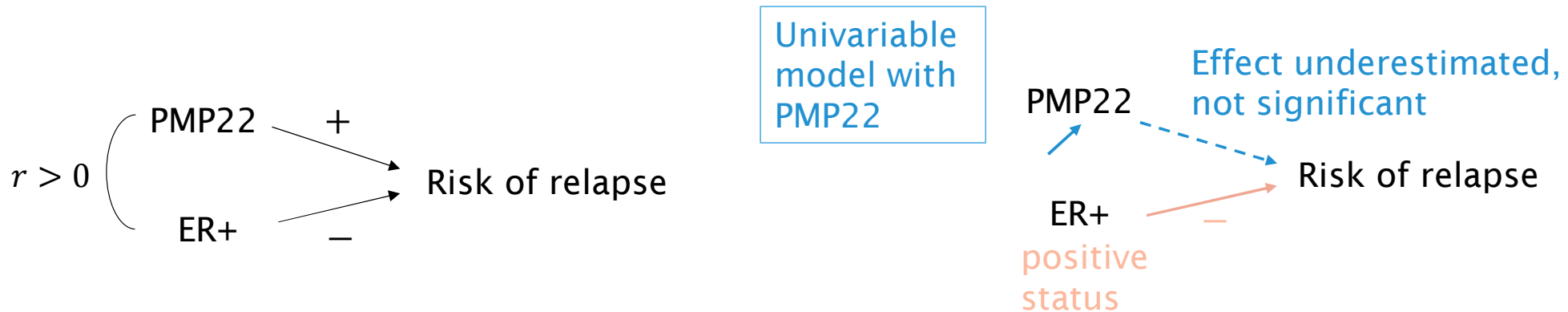
Variable	5 univariable models		Multivariable model	
	Hazard ratio (95%-CI)	<i>p</i>	Hazard ratio (95%-CI)	<i>p</i>
PMP22	1.01 (0.69 - 1.46)	0.978	1.77 (1.09 - 2.87)	0.020
nodal status	2.72 (1.73 - 4.27)	<0.001	2.72 (1.69 - 4.38)	<0.001
ER	0.63 (0.43 - 0.92)	0.017	0.49 (0.31 - 0.78)	0.003
tumor size	1.32 (0.99 - 1.78)	0.061	1.27 (0.93 - 1.74)	0.137
grade	1.06 (0.78 - 1.44)	0.703	0.98 (0.70 - 1.38)	0.913

PMP22 (> median gene expression) versus ($\leq$ median gene expression); nodal status 1 versus 0; ER + versus -; tumor size (1-4); grade (1-3)

Real-world case study: biomarker for cancer prognosis

Explanation: **masking**

Variable	5 univariable models		Multivariable model	
	Hazard ratio (95%-CI)	p	Hazard ratio (95%-CI)	p
PMP22	1.01 (0.69 - 1.46)	0.978	1.77 (1.09 - 2.87)	0.020
nodal status	2.72 (1.73 - 4.27)	<0.001	2.72 (1.69 - 4.38)	<0.001
ER	0.63 (0.43 - 0.92)	0.017	0.49 (0.31 - 0.78)	0.003
tumor size	1.32 (0.99 - 1.78)	0.061	1.27 (0.93 - 1.74)	0.137
grade	1.06 (0.78 - 1.44)	0.703	0.98 (0.70 - 1.38)	0.913



When new prognostic biomarkers are correlated with established clinical factors, univariable selection may falsely discard promising biomarker candidates

Conclusion & what to do instead

- Do not use univariable selection, neither on its own nor in combination with other methods (such as stepwise selection)
- Instead:
 - Clearly define your modeling aim
 - Define the set of candidate variables using subject-matter knowledge
 - Decide if data-driven variable selection is appropriate in this setting (depends, e.g., on sample size/events per variable, signal strength)
 - If data-driven selection is used:
 - Use methods that operate within a multivariable framework
 - Perform stability investigations
 - Interpret regression coefficients, confidence intervals, and p -values with caution
 - Report everything transparently



Journal of Clinical Epidemiology

Volume 199, November 2026, 112423



Original Research

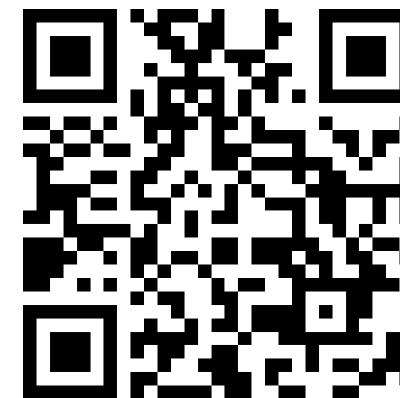
A preregistered simulation study provided evidence on the appropriate use of data-driven variable selection in descriptive multivariable regression modeling

Theresa Ullmann ^a, Georg Heinze ^a, Michael Kammer ^{a,b}, Daniela Dunkler ^a ,
TG2 of the STRATOS Initiative

Show more

Add to Mendeley Share Cite

App with univariable selection simulation:
<https://biometrician.shinyapps.io/UnivarSelection/>



REVIEW ARTICLE

Open Access



Variable selection – A review and recommendations for the practicing statistician

[Georg Heinze](#) , [Christine Wallisch](#), [Daniela Dunkler](#)

First published: 02 January 2018 | <https://doi.org/10.1002/bimj.201700067> |

VIEW METRICS

References

- Su, X., Ambler, G., Green, N., & Pavlou, M. (2026). The use of variable selection in clinical prediction modelling for binary outcomes: a systematic review. *Journal of Clinical Epidemiology*, 197, 112371.
- Tong, D., Heinze, G., Pils, D., Wolf, A., Singer, C. F., Concini, N., Hofstetter, G., Schiebel, I., Rudas, M. & Zeillinger, R. (2010). Gene expression of PMP22 is an independent prognostic factor for disease-free and overall survival in breast cancer patients. *BMC Cancer*, 10(1), 682.
- Vacheron, C. H., Friggeri, A., Allaouchiche, B., Maucourt-Boulch, D., & Coz, E. (2021). Quiet scandal: variable selection in three major intensive care medicine journals. *Intensive Care Medicine*, 47(12), 1487-1489.