Impact of Data-driven Methods for Causal Inference in Observational Studies: Simulations in Three Real-world Datasets

Thibaut Pressat-Laffouilhère (✉ t.pressat@chu-rouen.fr)

CHU Rouen: Centre Hospitalier Universitaire de Rouen <https://orcid.org/0000-0002-0378-175X>

Clément Massonnaud

CHU Rouen: Centre Hospitalier Universitaire de Rouen

Hélène Bréard

CHU Rouen: Centre Hospitalier Universitaire de Rouen

Margaux Lefebvre

CHU Rouen: Centre Hospitalier Universitaire de Rouen

Bruno Falissard

CESP: Centre de recherche en Epidemiologie et Sante des Populations

André Gillibert

CHU Rouen: Centre Hospitalier Universitaire de Rouen

Research article

Keywords: variable selection, causal diagram, observational study, causal inference

Posted Date: June 9th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-591022/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: Although data-driven methods for selecting covariates in multivariable models ignore confusion, mediation and collision, they are still used in causal inference. This study, through three real-world datasets, shows the impact of data-driven methods on causal inference.

Methods: A research question leading to multivariate model was raised for each of three real-world datasets. Three covariate selection methods were compared on their performances to correctly answer the question: Augmented Backward Elimination with BIC criterion and “change-in-estimate” threshold set at 0.05, Backward Elimination with BIC criterion and a knowledge-based method relying on causal diagrams. The covariates were classified as indispensable, prohibited and optional, considering the potential bias they could cause on the estimate. For each dataset and sample size ($N=75, 300$ and $3,000$), 10,000 Monte Carlo samples were drawn. Percentages of inclusion of each covariate in models were computed. Coverages of Wald’s 95% confidence interval of exposure effects were computed with two different theoretical values (the analysed method, the knowledge-based method).

Results: Even with the largest sample size ($n=3,000$), data-driven methods were not reproducible, with 8.6% to 53% of covariates included in 20% to 80% of experiences. Prohibited covariates could be included in more than 80% of experiences and indispensable covariates missed in more than 80% of experiences even with $n=3,000$. With the largest sample sizes, coverages of the theoretical knowledge-based value by data-driven methods ranged from 0% to 83.7%; coverages of the theoretical value of the same data-driven method ranged from 73.2% to 91.1% and were asymmetrical.

Conclusion: In conclusion, data-driven methods should not be used in causal inference.

Background

Many approaches are used to select covariates to be included in multivariable regression models in the context of observational studies [1]. Schematically, modeling can be performed for two different aims: to predict or to explain [2]. The first is to create a model by choosing variables that can best predict an outcome (e.g. diagnosis of a disease with imaging and clinical variables). The second is to estimate the strength of the causal link between two variables (e.g., exposure of interest: tobacco, with the outcome lung cancer) adjusted for any confounding variables. A clear assessment of the study aim is a necessary step to determine optimal methods for selecting covariates in the model.

When the aim is to predict, data-driven methods are widely used. There are two classical approaches [3]. Older methods are based on maximum-likelihood model fit and rely on the P-value, Bayesian information criterion (BIC) and Akaike information criterion (AIC). This includes Stepwise, Backward elimination and Forward methods [4]. Newer methods are based on penalization of regression coefficients. This includes, for the most known, the Least Absolute Shrinkage and Selection Operator (LASSO) [5] and the elastic net [6]. Contrary to ridge regression which penalizes regression coefficients without exclusion, these methods

apply penalties forcing the value of some coefficients to zero, leading to exclusion of variables. Users try to obtain the “optimal” penalty parameter using cross-validation, AIC or BIC.

When the aim is to explain (exposure of interest), data-driven methods are also used. One of the most reported methods is the preselection of variables by univariable analysis, selecting adjustment variables that are associated with the outcome with a p-value cut-off [7,8]. The « change in estimate » selection method keeps an adjustment variable in the model depending on the change in the estimate of the coefficient of the exposure of interest (e.g., > 10%) that is induced by the removal of the adjustment variable, allowing conservation of “strong” potential confounders [9]. The purposeful selection method [10] is a hybrid method between a first univariable preselection (e.g., p-value ≤ 0.25), then backward elimination of variables with non-significant association (e.g., p-value > 0.1) and a small “change in estimate” (e.g., $\leq 15\text{--}20\%$) and finally the entry of excluded variables from the initial univariable analysis (e.g., univariable p-value > 0.25) if significant (e.g., multivariable p-value < 0.10). However, in causal inference prior knowledge is primordial according to Witte J *et al.* [11]: “all selection methods rely on assumptions that can only be justified with subject-matter knowledge”. Part of this knowledge can be represented with causal diagrams such as causal directed acyclic graphs (DAGs) [12]. In a causal DAG, variables are represented by nodes and causal relationships between variables by directed edges. These representations help the identification of confusion, mediation, or collision relations between variables. A collider is a variable caused jointly by the exposure and outcome of interest while a mediator is caused by the exposure and itself causes the outcome. Contrary to the causal DAG approach, the data-driven methods presented above do not take into account the difference between confounders, mediators or colliders and may lead to inappropriate adjustment on mediators and colliders unless variables are thoroughly preselected.

For several years now, stepwise methods have been criticized, as they lead to underestimated standard errors, non-normal sampling fluctuations and erroneous p-values when using usual confidence interval and hypothesis test procedures: Wald, Score and likelihood ratio. Moreover, they amplify collinearity problems [13–15]. Peter L. Flom [15] proposed to use other methods such as the LASSO or the least-angle regression (LARS) which are supposed to be better alternatives. However, the stepwise method, and more widely, data-driven methods (including the penalized method proposed above), have well-known issues associated with p-values (even if AIC or BIC are used instead): risk of false discovery increased by the multiplicity of tests [16] and inappropriate acceptance of the null hypothesis when an effect is not significant [17]. Despite these warnings, researchers continue to use data-driven methods [7,8,18], and the development of data-driven methods for covariate selection in explanatory models is ongoing in 2019 with the article Confounder Selection via Support Intersection by Lee S *et al.*

But some researchers advocate the use of prior knowledge in explanatory models: VanderWeele TJ *et al.* [19] presented theoretical examples with DAGs. Heinz *et al.* [1] emitted recommendations. Witte J *et al.* [11] demonstrated the strengths and weaknesses of data-driven methods in causal inference using simulation with important sample size (minimum sample at 500), few variables, prevalence treatment and outcome around 0.5. However, these studies are too far from real-life settings.

There is a need for examples based on real datasets with prevalences far from 0.5, smaller datasets reflecting the low statistical power of some current studies and larger number of covariables. In this article, knowledge-based and data-driven covariate selection methods were tested in three causal research questions along with three real-life datasets. The objectives were to assess whether data-driven covariate selection methods were fragile without prior knowledge and whether they could lead to erroneous inference.

Methods

Data are derived from three different studies, published or submitted to peer-reviewed journals. Each one is described briefly in terms of size, available variables, and target population. Then, one causal research question is raised for each study. Covariate selection methods are compared on their propensity to select the appropriate set of variables and 95% confidence interval coverage of the true value that they are supposed to estimate.

Datasets and research questions

Dataset #1 is from a retrospective obstetrical study conducted in 200 patients with failure of medical abortion defined as absent or incomplete expulsion requiring suction. The routinely collected clinical variables were: age (years), body mass index (BMI), number of vaginal deliveries, abortions, pregnancies, miscarriages, cesarean sections, children born alive, use of analgesic level 2 (yes or no), antiemetic (yes or no), weeks of gestation, length of hospitalization (hours), treatment route (oral or vaginal), additional dose of misoprostol, number of spontaneous fetal expulsions, doubt of expulsion (as declared by the physician: yes, don't know, or no), resort to suction (yes or no). Question #1 was: How much does failure of medical abortion increase the length of hospitalization?

Dataset #2 is from a prospective study conducted in 665 adult women who underwent medical abortion. The variables collected a priori with an auto-questionnaire and a hetero-questionnaire were: Age (18–20, 20–25, 25–35, >35 years), medical history of abortion (medical, surgical, both or no) and pregnancy (yes or no), weeks of gestation, initial preference for medical abortion (yes or no), fear of hospitalization (yes or no), total misoprostol dose (200–400, 600–800, 1000–2000 mg), treatment route, pain (from 0 to 10), nausea/vomiting (yes or no), fever (yes or no), diarrhea (yes or no), bleeding estimate (as usual during menstruation, more, less than menstruation), misoprostol side effects intensity (from 0 to 3), and the question “If I were to have an abortion again, would I prefer the medical or surgical method?” (yes or no). Question #2 was: Does nausea increase the rate of preference for the surgical method of abortion for a future abortion?

Dataset #3 is from a retrospective study conducted in 265 patients with stroke who underwent a combination of thrombolysis and thrombectomy or thrombectomy alone between 2011 and 2019. The variables collected, based on routine patient care, were: age (< 60, 60–70, 70–80, >80 years), sex, anticoagulant use, hypertension, diabetes mellitus, anti-platelet use, alcohol consumption, pre-treatment

volume of cerebral necrosis (< 1, 1–10, 10–20, 20–30, > 30 cm³), time to imaging after stroke (0–90, 90–150, 150–270, > 270 min), post-treatment volume of cerebral necrosis (> 70cm³), wake-up stroke, pre-hospitalization Rankin score (normal or abnormal), thrombolysis treatment, pre-treatment National institute of health stroke score (NIHSS), post-treatment cerebral hemorrhage, pre-treatment Alberta Stroke Program Early CT Score (ASPECTS) (1–7, 8–10), tobacco use, pre-treatment clot burden score (CBS) (0–6, 7–10), overweight, hypercholesterolemia, sleep apnea syndrome, thrombus localization (middle cerebral artery, internal carotid artery, basilar artery, tandem), medical history of stroke, medical history of myocardial infarction, collateral circulation, atrial fibrillation. Question #3 was: Does combination therapy (thrombectomy plus thrombolysis) decrease, compared to thrombectomy alone, the risk of massive necrosis, defined as more than 70 cm³ of necrosis?

Statistical analysis

Initial variable subset for data-driven methods

Considering that minimal knowledge about the chronology of variable measurement was available, the starting pool of variables could differ depending on questions. All variables measured after the exposure of interest were removed to avoid inclusion of mediation variables [20] except for the first dataset where all pre-suction variables were kept. Indeed, resort to suction was timestamped, unlike abortion failure in a context where suction is a late proxy of abortion failure.

Initial datasets

To use data-driven methods incomplete observations in the originals datasets #1 and #2 were excluded before any simulation was performed. In dataset #3, missing data were managed by simple imputation by chained equations before any simulation was performed. The predictive matrix included all pre-exposure covariates from the initial subset.

Covariate selection methods

The following covariate selection methods were used in order to build the explanatory models: (1) Knowledge-based with causal diagram using bidirectional edges to represent association between two variables that share ancestors (bidirectional edges were already used for this purpose in causal diagram for epidemiology [12]); (2) Augmented Backward Elimination (ABE) [21] with a “change in estimate” threshold set at 0.05 and BIC criterion; (3) Backward elimination (BE) with BIC criterion. Selection of the variable representing the exposure of interest was forced in the modeling process. For each data set, causal diagrams were designed by the author T.PL based on previous discussion with one medical expert of the field (scientific literature, guidelines and opinions), then validated by the same medical expert.

Knowledge-based approach

The “disjunctive criterion” [22] was used to select potential confounders from the initial set of candidate variables. In contrast with one of the classical definitions of confounder which causes both the outcome

and exposure, the “disjunctive criterion” selects any covariate that causes outcome or exposure or both. The disjunctive criterion method has been used more sparingly:

- For a given potential variable, it was excluded if the existence of a selection bias leading to an imbalance between groups did not seem possible AND if this variable did not seem to be a proxy of an unmeasured confounder or an unknown confounder.
- Any potential variable considered as redundant with any selected confounder [23], exposure or outcome was excluded.

For each dataset, variable adjustments were categorized as either *optional* when it would not add or remove any important bias, *prohibited* when it might introduce an important bias or *indispensable* when it would remove an important bias. The lists of variables selected by the knowledge-based approach appear in the result tables. These lists of variables were not adapted to the sample size so as not to disadvantage the automatic methods. Here are some non-exhaustive explanations about the choice of variables:

The research question associated with dataset #1 was the analysis of the causal link between “failure” of abortion and “length of hospitalization”. “Doubt of expulsion” was considered as mediation variables between “failure” and “length of hospitalization” and so was *prohibited* adjustment. “Doubt of expulsion” and “additional dose of misoprostol” precede “resort to suction” and are completely or partly due to actual failure of abortion. The variables “anti-emetic” and “use of analgesic level 2” were caused by side effect caused by “additional dose of misoprostol” and so, were *prohibited* adjustments too.

The research question associated with dataset #2 was the analysis of the causal link between “nausea” and “future preference for surgical abortion compared to medical abortion after past medical abortion”, the adjustment on “misoprostol side effect intensity” was *prohibited* because it seemed to be redundant with the exposure “nausea”. That would be comparable to selecting “pain” (yes, no) and “pain” (0 to 10) in the same model. The adjustment on “total misoprostol dose” was *optional* because redundant with the adjustment on side effects on the one hand and “weeks of gestation” on the other. “Weeks of gestation” indicates misoprostol dose and may lead to “failure” or to “pain” which is one misoprostol side effect leading to “future preference for surgical abortion compared to medical abortion after past medical abortion”. Side effects (“pain”, “bleeding estimate”) adjustments were indispensable because they are concomitant with “nausea” hence they may confound the real effect.

The research question associated with dataset #3 was the analysis of the negative causal link between “thrombolysis” and “massive brain necrosis”. As this was a therapeutic question in a retrospective study, the indication bias could be important. This bias was corrected mostly by adjustment on “anticoagulant use”, “time to imaging after stroke”, “wake-up stroke”, “medical history of stroke” which were classed as *indispensable* adjustments. There were no *prohibited* variable adjustments in this dataset.

Monte Carlo sampling and criteria of assessment

A total of 10,000 samples (5,000 for dataset #3) each of three sizes ($N = 75, 300$ and 3000) of independent and identically distributed observations were randomly drawn with replacement from each initial dataset. The ordinary least squares general linear model was used for the first research question while maximum likelihood logistic regressions were used for the second and third questions. Samples with constant outcome or exposure were excluded.

Criteria of assessment

The *inclusion rate* of a variable was defined as the proportion of Monte Carlo experiences in which the variable was selected in the model by the data-driven method. We defined *inconstant inclusion* of variables as any inclusion rate between 20 and 80%. The actual coverage of Wald's 95% confidence interval of the effect (odds ratio or adjusted difference of means) of the exposure, was assessed. Two different coverages were computed, with two different theoretical values of the real effect of the exposure on the outcome. The theoretical value was set to be the effect of the exposure on the outcome in the original dataset (duplicated to reach at least 1,000,000 observations in order to reach the asymptotic effect); first in a model fitted by the analyzed method, second in a model fitted by the knowledge-based method. Overestimation and underestimation percentages were separately computed to describe coverage faults.

All analyses were performed with R statistical software (version 4.0.3, The R Foundation for Statistical Computing, Vienna, Austria) with 'abe', 'dagitty' [24], 'ggdag', 'mice', and data.table packages. The computations were performed on the GenOuest cloud computing platform, leveraging parallelization on a cluster of 50 processor cores.

Results

General characteristics of samples

The numbers of patients excluded in dataset #1 and dataset #2 were respectively 24 (12%) and 67 (10%). In dataset #3, 95 (35%) of the 265 patients had incomplete observations that were kept using a simple imputation by chained equation. Primary exposures in datasets #1, #2 and #3, respectively "failure", "nausea" and "thrombolysis" had frequency rates respectively equal to 8.6%, 52% and 48%. The primary outcomes in datasets #1, #2 and #3 were respectively "length of hospitalization", "preference for medical abortion" and "massive necrosis". In dataset #1 the mean (standard deviation) "length of hospitalization" in the suction group was 27 (10) hours versus 11 (6) hours in the group without suction. In datasets #2 and #3, the frequencies of the primary outcomes were respectively 50% and 34%. The sets of eligible covariates, excluding the primary exposure, in datasets #1, #2 and #3 contained respectively 15, 13 and 23 potential covariates.

Inclusion rates of variables

With BE, in datasets #1, #2 and #3, respectively 3 (20%), 4 (33%) and 10 (45%) covariates were inconstantly included (i.e., included in 20 to 80% of experiences) with a sample size of 300. With ABE, these figures were respectively 8 (53%), 10 (83%) and 11 (50%).

The set of included variables did not completely stabilize as the sample size grew. With BE in datasets #1, #2 and #3, respectively 8 (53%), 3 (23%) and 2 (9%) variables were inconstantly included with the largest sample size ($n = 3000$). With ABE, these figures were respectively 4 (27%), 3 (23%) and 3 (14%). Although that was expected for highly correlated variables such as “number of vaginal deliveries” and “number of children born alive”, it was less expected for variables such as “body mass index” (Table 1).

In large samples ($n = 3000$), several prohibited variables were frequently ($\geq 80\%$) included: 3/4 for both BE and ABE in dataset #1 and 1/1 for ABE in dataset #2 (see Tables 1 and 2). With BE in large samples ($n = 3000$), 1/3 indispensable variables were frequently missed ($\geq 80\%$) for dataset #1, 1/7 for dataset #2 and 4/8 for dataset #3. With ABE in large samples, these figures were respectively 1/3, 1/7 and 3/8.

With BE, some variables, such as “doubt of expulsion” in dataset #1 had an inclusion rate growing with sample size and tending towards 100%, while some other variables such as “use of analgesic level 2” had an inclusion rate that was constantly low (see Table 1). Strong U-shaped associations between the sample size and inclusion rate were found for “sleep apnea syndrome”, “time to imaging after stroke” and “thrombus localization” in dataset #2 (see Table 2). Eventually, strong decreasing associations exist between sample size and inclusion rate, such as that of the variables “anticoagulant use”, “diabetes mellitus” and “hypertension” in dataset #3 (see Table 3).

With ABE in small samples ($n = 75$), a high frequency of inclusions ($\geq 80\%$) was found for 7 (46%) variables of dataset #1, 12 (100%) of dataset #2 and 22 (100%) of dataset #3 (Tables 1 to 3).

Coverage

Compared to knowledge-based theoretical value

The coverage of the knowledge-based theoretical value by BE and ABE decreased with the sample size and, for the largest sample size ($n = 3000$), reached 0% for both BE and ABE in dataset #1, 58.9% and 57.6% for BE and ABE in dataset #2, and 77 and 88.7% for BE and ABE in dataset #3 (Tables 4 to 6).

Compared to the method itself

The coverages of the asymptotic BE value for BE on large sample sizes ($n = 3000$) in datasets #1, #2 and #3 were respectively 73.2%, 88.8% and 77%. For ABE, these figures were respectively 73.6%, 91.1% and 88.7%; for the knowledge-based method, they were 85%, 94.7% and 94.9%.

Discussion

In the three datasets, data-driven confidence intervals had poor coverage, even when compared to their own asymptotic effect. This may be due to the invalidity of Wald's confidence intervals for test-based data-driven methods (BE, ABE) without shrinkage [25]. Bias in point estimates makes the coverage tend towards zero when the sample size tends towards infinity as the width of the confidence interval tends towards zero, centered around the wrong value.

The inclusion of prohibited variables leads to bias in estimations to assess the causal effect of exposure on the outcome. Judd and McClelland wrote: "It seems unwise to let an automatic algorithm determine the questions we do and do not ask of our data" [26]. Since different models answer different research questions (e.g., adjusting or not on a mediation variable depending on whether the direct or total effect is to be analyzed), data-driven methods with unstable inclusions of covariables lead to random models, answering random research questions. In dataset #1, on large sample sizes, data-driven methods tended to answer the question "How much does suction increase the length of hospitalization?" instead of the initial question "How much does failure of medical abortion increase the length of hospitalization?", because they adjusted on "doubt of expulsion", "additional dose of misoprostol", "antiemetic".

In their article presenting ABE [21] authors assumed that the initial subset of variables respected the "disjunctive criterion". Hence, they did not use it as a blind method but as a complementary method of prior knowledge based on a DAG. They consider ABE as a way to improve reproducibility where DAGs are not unanimous or when knowledge is poor. As we saw in this article ABE and BE provide non-reproducible results, with random inclusion (inclusion rates between 20 and 80%) of many variables, even on quite large sample sizes, and some U-shape associations between sample size and inclusion rates. Moreover, ABE and BE include mediators and colliders without warning. Furthermore, the numerous options available in ABE function (p-value and "change-in-estimate" thresholds, AIC, BIC) may further increase non-reproducibility.

Other data-driven methods based on shrinkage such as Glider with group LASSO [27] or outcome adaptive LASSO [28] have been promoted as better alternatives, but they are as blind as test-based methods about relations (mediation, collision, confusion) between variables. On small-to-medium sample sizes, data-driven methods can omit confounders (e.g., "weeks of gestation" in dataset #1) leading to under-adjustment unless the sample size is very large besides the fact that most of the time under-powered studies lead to non-reproducible results [29]. A previous study, comparing forward, backward, stepwise and purposeful with simulation including six variables of which three had an effect on the outcome, showed the low proportion of experiences where the correct model was selected (even for $n = 600$ with 70% of correct models) [10]. However, in our study, even if some indispensable variables were not included, they were sometimes replaced by optional variables that could be proxies (e.g., "weeks of gestation" and "total misoprostol dose" in dataset #2 or "atrial fibrillation" and "anticoagulant use" in dataset #3). Indeed, in the causal DAG theory, an adjustment is sufficient to control confounding if it blocks all backdoor paths. If several variables are on this path it is sufficient to adjust on one or all of them to obtain the desired effect [30].

The variability of relations of inclusion rates to sample size (U-shape, decrease or increase) may be explained by contradictory forces growing in different directions with the sample size. First, the statistical power grows with the sample size for all effects, tending to an increase of the inclusion rate with the sample size. Second, the sensitivity of the “change in estimate” threshold is inversely proportional to sample size, leading to a decrease of inclusion rate with sample size. Third, two correlated variables compete for inclusion in the sample size. On small samples, the variable having the largest effect on the outcome and the highest statistical power may win while, on large samples, both variables have a statistical power tending towards 100% and other factors come into play, such as the change in the estimate of the primary exposure or the novel inclusion of a third variable that attenuate or increase the effect of one of the two concurrent variables.

All these chaotic inclusions and biased estimates could have an impact on clinical conclusions of articles [31].

Strengths & limits

Causal diagrams were validated by only one clinical practitioner who has studied the concerned field, hence they are not universally consensual. The use of three datasets is a way to present diverse situations but is in no way exhaustive. Although real-life datasets were used, their small sample sizes made them imperfect, with idiosyncrasies that do not exist in the actual population. Data-driven methods used to recode variables or add interaction terms were not analyzed. Shrinkage-based methods were not analyzed, because they are rarely used with explanatory models. Although there are some methods to correct the bias of point estimate and confidence intervals after data-driven selection of variables, none were analyzed [32,33]. However, these methods provide inference that is only valid conditional on the selected variables, and so, fail to solve the problem of randomness of the research question due to methods randomly selecting variables. Moreover, these methods are not implemented in major statistical software: Stata, SPSS and SAS provide Wald’s p-values and confidence intervals without warning when stepwise, backward or forward selection is performed.

Conclusion

This work shows that variables selected by data-driven methods (backward elimination and augmented backward elimination) are inconstant for a given sample size, vary with the sample size and are often inappropriate for causal inference (e.g., random adjustment on mediator). It also shows that coverage of the primary effect by 95% confidence interval estimated by data-driven methods can tend towards zero as the sample size grows. Selection of variables is not an easy task, relying on a complex context. Data-driven methods, blind to this context, should not be used to build an explanatory model.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

Data are not available because of individual privacy.

R script is available at <https://github.com/ClementMassonnaud/dataDrivR>.

Conflicts of interest/Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Funding

No funding was received to assist with the preparation of this manuscript.

Authors' contributions

T.PL and A.G designed the study, interpreted the results and drafted the article. CR.M corrected, rewrote the R code to allow parallelization for simulations and rerun the analysis. H.B and M.L provided data and helped construct the causal diagram(s). B.F interpreted the results. All of the authors reviewed the article and approved the final version.

Acknowledgments

The authors are grateful to Nikki Sabourin-Gibbs, Rouen University Hospital, for her help in editing the manuscript.

References

1. Heinze G, Wallisch C, Dunkler D. Variable selection - A review and recommendations for the practicing statistician. *Biom J*. 2018 May;60(3):431–49. <https://doi.org/10.1002/bimj.201700067>
2. Shmueli G. To Explain or to Predict? *Stat Sci*. 2010 Aug;25(3):289 – 310. <https://doi.org/10.1214/10-STS330>

3. Desboulets L. A Review on Variable Selection in Regression Analysis. *Econometrics*. 2018 Nov 23;6(4):45. <https://doi.org/10.3390/econometrics6040045>
4. Hamaker HC. On multiple regression analysis. *Stat Neerlandica*. 1962 Mar;16(1):31–56. <https://doi.org/10.1111/j.1467-9574.1962.tb01184.x>
5. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. *J R Stat Soc Ser B Methodol*. 1996 Jan;58(1):267–88. <https://www.jstor.org/stable/2346178>
6. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B Stat Methodol*. 2005 Apr;67(2):301–20. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>
7. Walter S, Tiemeier H. Variable selection: current practice in epidemiological studies. *Eur J Epidemiol*. 2009;24(12):733–6. <https://doi.org/10.1007/s10654-009-9411-2>
8. Talbot D, Massamba VK. A descriptive review of variable selection methods in four epidemiologic journals: there is still room for improvement. *Eur J Epidemiol*. 2019 Aug;34(8):725–30. <https://doi.org/10.1007/s10654-019-00529-y>
9. Lee PH. Is a cutoff of 10% appropriate for the change-in-estimate criterion of confounder identification? *J Epidemiol*. 2014;24(2):161–7. <https://doi.org/10.2188/jea.JE20130062>
10. Bursac Z, Gauss CH, Williams DK, Hosmer DW. Purposeful selection of variables in logistic regression. *Source Code Biol Med*. 2008 Dec 16;3:17. <https://doi.org/10.1186/1751-0473-3-17>
11. Witte J, Didelez V. Covariate selection strategies for causal inference: Classification and comparison. *Biom J*. 2019 Sep;61(5):1270–89. <https://doi.org/10.1002/bimj.201700294>
12. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiol Camb Mass*. 1999 Jan;10(1):37–48. <https://doi.org/10.1097/00001648-199901000-00008>
13. Thompson B. Why Won't Stepwise Methods Die? *Meas Eval Couns Dev*. 1989 Jan;21(4):146–8. <https://doi.org/10.1080/07481756.1989.12022899>
14. Whittingham MJ, Stephens PA, Bradbury RB, Freckleton RP. Why do we still use stepwise modelling in ecology and behaviour?: Stepwise modelling in ecology and behaviour. *J Anim Ecol*. 2006 Sep;75(5):1182–9. <https://doi.org/10.1111/j.1365-2656.2006.01141.x>
15. Flom PL, Development N, Institutes R, York N. Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. 2007;7.
16. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p -values. *R Soc Open Sci*. 2014 Nov;1(3):140216. <https://doi.org/10.1098/rsos.140216>
17. Goodman S. A dirty dozen: twelve p -value misconceptions. *Semin Hematol*. 2008 Jul;45(3):135–40. <https://doi.org/10.1053/j.seminhematol.2008.04.003>
18. Liao H, Lynn HS. A survey of variable selection methods in two Chinese epidemiology journals. *BMC Med Res Methodol*. 2010 Sep 29;10:87. <https://doi.org/10.1186/1471-2288-10-87>
19. VanderWeele TJ. Principles of confounder selection. *Eur J Epidemiol*. 2019 Mar 15;34(3):211–9. <https://doi.org/10.1007/s10654-019-00494-6>

20. VanderWeele TJ. Mediation and mechanism. *Eur J Epidemiol*. 2009 May;24(5):217–24.
<https://doi.org/10.1007/s10654-009-9331-1>
21. Dunkler D, Plischke M, Leffondré K, Heinze G. Augmented Backward Elimination: A Pragmatic and Purposeful Way to Develop Statistical Models. Olivier J, editor. *PLoS ONE*. 2014 Nov 21;9(11):e113677. <https://doi.org/10.1371/journal.pone.0113677>
22. VanderWeele TJ, Shpitser I. A new criterion for confounder selection. *Biometrics*. 2011 Dec;67(4):1406–13. <https://doi.org/10.1111/j.1541-0420.2011.01619.x>
23. Ikram MA. The disjunctive cause criterion by VanderWeele: An easy solution to a complex problem? *Eur J Epidemiol*. 2019 Mar;34(3):223–4. <https://doi.org/10.1007/s10654-019-00501-w>
24. Textor J, van der Zander B, Gilthorpe MS, Liśkiewicz M, Ellison GTH. Robust causal inference using directed acyclic graphs: the R package ‘dagitty’. *Int J Epidemiol*. 2017 Jan 15;dyw341.
<https://doi.org/10.1093/ije/dyw341>
25. Morozova O, Levina O, Uusküla A, Heimer R. Comparison of subset selection methods in linear regression in the context of health-related quality of life and substance abuse in Russia. *BMC Med Res Methodol*. 2015 Aug 30;15:71. <https://doi.org/10.1186/s12874-015-0066-2>
26. Judd CM, McClelland GH, Ryan CS. *Data analysis: a model comparison approach*. 2nd ed. New York; Hove: Routledge; 2009. 328 p.
27. Koch B, Vock DM, Wolfson J. Covariate selection with group lasso and doubly robust estimation of causal effects: GLiDeR. *Biometrics*. 2018 Mar;74(1):8–17. <https://doi.org/10.1111/biom.12736>
28. Shortreed SM, Ertefaie A. Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*. 2017 Dec;73(4):1111–22. <https://doi.org/10.1111/biom.12679>
29. Ioannidis JPA. Why Most Published Research Findings Are False. *PLoS Med*. 2005 Aug 30;2(8):e124.
<https://doi.org/10.1371/journal.pmed.0020124>
30. Pearl J. [Bayesian Analysis in Expert Systems]: Comment: Graphical Models, Causality and Intervention. *Stat Sci*. 1993 Aug 1 ;8(3). <http://dx.doi.org/10.1214/ss/1177010891>
31. Stavseth MR, Clausen T, Røislien J. The clinical consequences of variable selection in multiple regression models: a case study of the Norwegian Opioid Maintenance Treatment program. *Am J Drug Alcohol Abuse*. 2020;46(1):13–21. <https://doi.org/10.1080/00952990.2019.1648484>
32. Berk R, Brown L, Buja A, Zhang K, Zhao L. Valid post-selection inference. *Ann Stat*. 2013 Apr 1;41(2).
<http://dx.doi.org/10.1214/12-AOS1077>
33. Lee JD, Sun DL, Sun Y, Taylor JE. Exact post-selection inference, with application to the lasso. *Ann Stat*. 2016 Jun 1;44(3). <http://dx.doi.org/10.1214/15-AOS1371>

Tables

Table 1

percentages of variable inclusions in the model depending on covariate selection method (BE, ABE) and sample size (75, 300, 3000) for dataset #1

	Augmented backward elimination (ABE) inclusion rate (%)			Backward elimination (BE) inclusion rate (%)		
	n = 75	n = 300	n = 3000	n = 75	n = 300	n = 3000
Sample size	n = 75	n = 300	n = 3000	n = 75	n = 300	n = 3000
Covariates						
Body Mass Index*	39.2%	1.6%	54.3%	5.3%	1.3%	56.5%
Antiemetic [†]	59.5%	34.2%	98.8%	19.2%	27.2%	98.8%
Number of vaginal deliveries [‡]	90.6%	79.0%	40.7%	11.1%	5.4%	20.6%
Number of cesarean sections*	82.1%	72.5%	85.3%	8.8%	7.8%	80.8%
Doubt of expulsion [†]	90.4%	96.0%	100.0%	38.8%	78.1%	100.0%
Number of children born alive*	90.5%	83.3%	96.3%	9.0%	4.8%	48.3%
Number of spontaneous fetal expulsions*	74.0%	48.1%	43.6%	7.9%	10.6%	41.5%
Number of miscarriages*	79.0%	71.1%	93.1%	9.1%	9.6%	49.5%
Number of pregnancies*	86.6%	81.0%	99.4%	13.3%	9.6%	68.9%
Number of abortions*	85.4%	76.5%	99.3%	10.6%	6.7%	63.1%
Treatment route [‡]	72.0%	35.6%	3.0%	25.5%	16.9%	2.5%
Additional dose of misoprostol [†]	100.0%	100.0%	100.0%	99.5%	100.0%	100.0%
Use of analgesic level 2 [†]	51.9%	5.7%	1.9%	6.6%	1.7%	2.3%
Weeks of gestation [‡]	75.4%	87.8%	100.0%	34.9%	76.4%	100.0%
Age*	63.3%	27.4%	77.5%	8.4%	8.3%	73.3%
n = sample size, [‡] indispensable covariate, * optional covariate, [†] prohibited covariate.						

Table 2

percentages of variable inclusions in the model depending on covariate selection method (BE, ABE) and sample size (75, 300, 3000) for dataset #2

	Augmented backward elimination (ABE) inclusion rate (%)			Backward elimination (BE) inclusion rate (%)		
	n = 75	n = 300	n = 3000	n = 75	n = 300	n = 3000
Sample size						
Covariates						
Age*	99.9%	78.5%	18.2%	14.9%	0.2%	0.6%
Medical history of abortion [‡]	99.6%	90.2%	100.0%	23.6%	25.4%	100.0%
Medical history of pregnancy [‡]	94.0%	70.2%	79.1%	17.9%	7.8%	60.3%
Initial preference for medical abortion [‡]	94.0%	62.4%	97.4%	23.6%	26.5%	97.7%
Fear of hospitalization [‡]	93.2%	63.1%	99.8%	25.1%	31.8%	99.8%
Diarrhea*	92.1%	42.8%	4.1%	17.1%	2.9%	3.1%
Pain [‡]	95.0%	89.9%	100.0%	32.5%	55.2%	100.0%
Bleeding estimate [‡]	99.2%	72.0%	67.8%	18.1%	4.8%	64.2%
Fever*	91.4%	29.1%	5.9%	16.3%	3.4%	5.8%
Misoprostol side effect intensity [†]	99.3%	79.3%	85.1%	19.1%	9.5%	73.4%
Total misoprostol dose*	99.4%	84.2%	98.2%	18.2%	14.0%	96.4%
Weeks of gestation [‡]	99.6%	73.9%	18.6%	14.5%	1.5%	6.2%
Treatment route*	93.6%	59.7%	22.0%	17.1%	3.0%	9.7%
n = sample size, [‡] indispensable covariate, * optional covariate, [†] prohibited covariate						

Table 3

percentages of variable inclusions in the model depending on covariate selection method (BE, ABE) and sample size (75, 300, 3000) for dataset #3

	Augmented backward elimination (ABE) inclusion rate (%)			Backward elimination (BE) inclusion rate (%)		
	n = 75	n = 300	n = 3000	n = 75	n = 300	n = 3000
Sample size						
Covariates						
Age [‡]	100.0%	100.0%	100.0%	60.4%	74.3%	100.0%
ASPECTS*	100.0%	96.7%	100.0%	73.3%	74.8%	100.0%
Medical history of stroke [‡]	99.9%	65.2%	32.6%	52.0%	3.3%	12.2%
Medical history of myocardial infarction*	99.9%	83.3%	99.7%	53.9%	32.2%	99.7%
Wake-up stroke [‡]	99.9%	66.1%	14.4%	55.8%	6.7%	12.0%
Collateral circulation [‡]	100.0%	84.2%	99.0%	71.0%	61.4%	91.7%
Diabetes Mellitus*	100.0%	59.1%	2.5%	57.4%	6.8%	4.9%
Time to imaging after stroke [‡]	100.0%	96.8%	99.1%	42.0%	8.2%	80.1%
Atrial fibrillation*	100.0%	96.5%	100.0%	66.8%	60.4%	100.0%
Hypertension*	100.0%	61.0%	3.2%	60.0%	8.3%	4.1%
Hypercholesterolemia*	100.0%	90.6%	100.0%	64.5%	36.8%	99.9%
Thrombus localization*	100.0%	98.3%	99.4%	38.5%	7.4%	37.7%
Pre-treatment necrosis volume [‡]	100.0%	100.0%	100.0%	15.8%	57.1%	100.0%
Pre-treatment NIHSS [‡]	100.0%	97.8%	100.0%	74.0%	79.9%	100.0%
Anti-platelet use*	99.9%	69.4%	30.7%	59.3%	10.7%	18.7%
Anticoagulant use [‡]	99.9%	64.9%	15.5%	56.5%	8.4%	2.6%
Pre-hospitalization Rankin*	99.9%	67.9%	93.7%	59.8%	21.4%	92.4%
Sleep apnea syndrome*	99.2%	59.8%	28.2%	42.0%	2.9%	33.4%
Sex*	100.0%	98.0%	100.0%	70.9%	70.7%	100.0%

	Augmented backward elimination (ABE) inclusion rate (%)			Backward elimination (BE) inclusion rate (%)		
Clot burden score [‡]	100.0%	60.7%	3.4%	59.4%	6.6%	2.3%
Overweight*	100.0%	59.1%	7.9%	59.1%	4.4%	11.9%
Tobacco*	100.0%	86.9%	99.9%	62.0%	30.0%	97.9%
Alcohol consumption*	99.9%	59.3%	13.6%	50.3%	2.2%	4.3%
n = sample size, [‡] indispensable covariate, * optional covariate, [†] prohibited covariate						

Table 4

percentages of overestimation, underestimation and correct coverage of the effect of the exposure on the outcome depending on variable selection method (ABE, BE, KB) and sample size (75, 300, 3000) for dataset #1.

Method	ABE			ABE		
Theoretical value	asymptotic ABE			asymptotic KB		
Sample size	75	300	3000	75	300	3000
95% CI overestimation	19.9	15.9	14.0	3.1	0.1	0.0
95% CI underestimation	14.8	12.3	12.4	53.3	91.9	100
95% CI coverage	65.2	71.8	73.6	43.6	8.1	0.0
Method	BE			BE		
Theoretical value	asymptotic BE			asymptotic KB		
Sample size	75	300	3000	75	300	3000
95% CI overestimation	35.4	25.6	13.1	3.6	0.2	0.0
95% CI underestimation	14.2	12.9	13.7	47.7	86.1	100
95% CI coverage	50.4	61.5	73.2	48.7	13.7	0.0
Method	KB			KB		
Theoretical value	asymptotic KB			asymptotic KB		
Sample size	75	300	3000	75	300	3000
95% CI overestimation	9.1	8.2	7.4	9.1	8.2	7.4
95% CI underestimation	7.1	6.5	7.6	7.1	6.5	7.6
95% CI coverage	83.8	85.2	85.0	83.8	85.2	85.0
BE = backward elimination, ABE = augmented backward elimination, KB = Knowledge-based, 95% CI = 95% confidence interval						

Table 5

percentages of overestimation, underestimation and correct coverage of the effect of the exposure on the outcome depending on variable selection method (ABE, BE, KB) and sample size (75, 300, 3000) for dataset #2.

Method	ABE			ABE		
Theoretical value	asymptotic ABE			asymptotic KB		
Sample size	75	300	3000	75	300	3000
95% CI overestimation	4.6	4.8	6.5	3.4	1.5	0.1
95% CI underestimation	2.8	2.7	2.4	3.9	8.0	42.3
95% CI coverage	92.7	92.5	91.1	92.7	90.5	57.6
Method	BE			BE		
Theoretical value	asymptotic BE			asymptotic KB		
Sample size	75	300	3000	75	300	3000
95% CI overestimation	5.0	7.9	8.7	2.8	2.5	0.1
95% CI underestimation	2.3	2.1	2.5	4.1	6.8	41.0
95% CI coverage	92.7	90.0	88.8	93.1	90.8	58.9
Method	KB			KB		
Theoretical value	asymptotic KB			asymptotic KB		
Sample size	75	300	3000	75	300	3000
95% CI overestimation	4.0	3.1	2.7	4.0	3.1	2.7
95% CI underestimation	2.5	2.2	2.5	2.5	2.2	2.5
95% CI coverage	93.4	94.7	94.7	93.4	94.7	94.7
BE = backward elimination, ABE = augmented backward elimination, KB = Knowledge-based, 95% CI = 95% confidence interval						

Table 6

percentages of overestimation, underestimation and correct coverage of the effect of the exposure on the outcome depending on variable selection method (ABE, BE, KB) and sample size (75, 300, 3000) for dataset #3.

Method	ABE			ABE		
Theoretical value	asymptotic ABE			asymptotic KB		
Sample size	75	300	3000	75	300	3000
95% CI overestimation	0.4	3.7	7.8	0.4	2.1	0.8
95% CI underestimation	1.1	8.1	3.5	1.1	12.7	19.3
95% CI coverage	98.5	88.1	88.7	98.5	85.2	79.9
Method	BE			BE		
Theoretical value	asymptotic BE			asymptotic KB		
Sample size	75	300	3000	75	300	3000
95% CI overestimation	0.5	12.6	21.5	0.5	7.6	5.2
95% CI underestimation	1.1	4.4	1.6	1.1	7.5	11.1
95% CI coverage	98.4	83.0	77.0	98.4	84.9	83.7
Method	KB			KB		
Theoretical value	asymptotic KB			asymptotic KB		
Sample size	75	300	3000	75	300	3000
95% CI overestimation	2	2.0	2.1	2	2.0	2.1
95% CI underestimation	5	3.5	3.0	5	3.5	3.0
95% CI coverage	93	94.6	94.9	93.5	93.7	94.9
BE = backward elimination, ABE = augmented backward elimination, KB = Knowledge-based, 95% CI = 95% confidence interval						

Figures

Figure 1

Causal diagram built with prior knowledge and medical practitioner expertise for dataset #1