Variable selection methods were poorly reported but rarely misused in major medical journals: Literature review

Thibaut Pressat-Laffouilhère, Romain Jouffroy, A. Leguillou, Gaëtan Kerdelhué, Jacques Bénichou, André Gilibert

Version of Record: https://www.sciencedirect.com/science/article/pii/S0895435621002146
Manuscript_7b00a74b519009b8c98574a2be37f680

Variable selection methods were poorly reported but rarely misused in major medical journals: literature review

T Pressat-Laffouilhère[1,2,3*], R Jouffroy[4], A Leguillou[5], G Kerdelhue[2], J Benichou[1,6], A Gillibert[1].

1- CHU Rouen, Department of Biostatistics, F-76000 Rouen, France
2- CHU Rouen, Department of Biomedical Informatics, F-76000 Rouen, France
3- Normandie Univ, UNIROUEN, LITIS EA 4108, F-76000 Rouen, France
4- Intensive care unit, Anaesthesiology Department and SAMU of Paris, Necker University Hospital, Assistance Publique Hôpitaux de Paris and Paris Descartes university, 75015 Paris, France
5- Reims University Hospital, Reims, France
6- Inserm U 1018, University of Rouen and University Paris-Saclay, France
*:t.pressat@chu-rouen.fr; Rouen University Hospital 37 Boulevard Gambetta, 76000 Rouen, France

Abstract

Objective: This work presents a review of the literature on reporting, practice and misuse of *knowledge-based* and data-driven variable selection methods, in five highly cited medical journals, considering recoding and interaction unlike previous reviews.

Study Design and Setting: Original observational studies with a predictive or explicative research question with multivariable analyses published in NEJM, Lancet, JAMA, BMJ and AIM between 2017 and 2019 were searched. Article screening was performed by a single reader, data extraction was performed by two readers and a third reader participated in case of disagreement. The use of data-driven variable selection methods in *causal explicative* questions was considered as misuse.

Results: 488 articles were included. The variable selection method was unclear in 234 (48%) articles, data-driven in 78 (16%) articles and *knowledge-based* in 176 (36%) articles. The most common *data-driven* methods were: Univariate selection (n=22, 4.5%) and model comparisons or testing for interaction (n=17, 3.5%). Data-driven methods were misused in 51 (10.5%) of articles.

Conclusion: Overall reporting of variable selection methods is insufficient. *Data-driven* methods seem to be used only in a minority of articles of the big five medical journals.

Keywords: Review, Observational study, covariate selection, variable selection, reporting, data-driven
Running title: Variable selection methods: reporting, practice, misuse.

Article Wordcount = 2937

Introduction

Variable selection in multivariate analysis is a wide-ranging and still controversial issue affecting all quantitative areas of biomedical research (e.g. medecine, epidemiology, social sciences, etc.). Relevant variable selection is critical in observational studies to address confounding issues in explanatory models and reach high predictive performance in predictive models.

There are two general approaches to variable selection, (i) knowledge-based approaches based on published knowledge or expert opinion that may be synthesized in causal diagrams [1], and (ii) data-driven methods. There is a wide range of data-driven methods, including backward elimination, forward selection [2] and stepwise, Least Absolute Shrinkage and Selection Operator (LASSO) [3], Elasticnet [4], the "change in estimate", augmented backward elimination [5], which combines backward elimination and "change in estimate" (more details in supplementary). These methods have been recently reviewed by Desboulet and al [6], Witte J and al [7]and Heinze G and al [8].

Data-driven methods are generally appropriate for building *predictive models* of a prognostic or diagnostic nature but are more questionable for *explanatory models* in etiologic research (*e.g.* models pertaining to the assessment of causal risk factors or treatment effects). Indeed, the latter requires making inference on the exposure (or treatment) coefficient of the model, with statistical tests or confidence intervals that are affected by bias when data-driven methods are used. Indeed, usual Wald, Rao's score and likelihood ratio have been shown to be biased with too narrow confidence intervals and too high estimates in case of data-driven pre-selection of variables [9]. Moreover, data-driven methods may inappropriately adjust on mediation variables or omit relevant confounders, biasing inference on the main causal effect in explanatory models. In case of predictive models these adjustments are less deleterious

because the aim is to obtain high predictive performance whatever variable is used in the model.

Although STROBE guidelines [10] encourage authors to clearly define which variables are potential confounders and to thoroughly describe statistical methods in their papers, this is not always done [11] and many statistical choices are left unexplained, sometimes because of insufficient space.

Walter *and Timeier* reported frequencies of variable selection methods in four major epidemiological journals in 2008 and found widespread use of stepwise selection methods and frequent insufficient or missing reporting of methods used [12]. Talbot *et al* updated the review in 2015 and found a lower use of stepwise [13]. Of note, these reviews did not report on practices regarding assessment and inclusion of interaction terms, recoding of variables (e.g. the variable 'age' may be included as a continuous, categorical [18-25][26-35][36-45], or polynomial variable with age+age^2+age^3), or variable selection in sensitivity analyses.

Practices have evolved and may be different in medical journals with the highest impact factor, considered more influential as they are often seen as the cream of the crop. The goal of this study was therefore to review articles published from 2017 to 2019 in the five medical journals with the highest impact factor in the "Medicine, general & internal" category according to the journal citation reports of 2016 [14], often referred to as the big-five medical journals (i.e., New England Journal of Medicine, The Lancet, Journal of American Medical Association, British Medical Journal, and Annals of Internal Medicine) in order to assess reporting, to describe current practice and to estimate misuse of variable selection methods. Because of their bearing on variable selection [15], other factors were also examined, *i.e.*, recoding and interaction, in the primary and sensitivity analyses.

Methods

Inclusion criteria

Screening was performed by the first author (T.P.-L.) using the online tables of contents of the big-five medical journals for original articles published between January 2017 and December 2019 and reporting observational studies with multivariate statistical models. Articles were screened on the basis of title, abstract and full text assessment.

Only observational health studies on the human subject were considered; *i.e.,* studies including human individuals (patients, healthy volunteers, or healthcare practitioners) from whom health variables were measured without any forced intervention. Design included cross-sectional, case-control, prospective and retrospective cohort studies, and some quasi-experimental studies where adjustments are needed to correct bias, with or without a control group. Only studies reporting estimates from at least one model addressing the main study objective and requiring the selection of a set of covariates were included. Machine learning models were also considered except in cases where the variables could not be selected by a human (e.g. pixel array). Economic, genetic, descriptive epidemiological studies, meta-analyses (or pooled cohorts) and systematic reviews were excluded. Finally, only research articles having a *predictive* or *explicative* main research objective (as defined below) were included.

Data extraction

Types of research question

The *type of research question*, in the abstract, at the end of the introduction section, in the methods section and the discussion, was categorized as either *predictive* or *explicative; explicative* questions were further divided into *causal* and *risk factor/association*. A research question was considered as *predictive* if authors specified that they aimed to build a "predictive model" or a "prognostic model". A research question was considered as *causal*

explicative in the following situations: (i) the exposure is a treatment, environment (e.g. pollution) or health behavior (e.g. tobacco consumption) that may be controlled, (ii) the authors used keywords related to causality such as 'mediation', 'causal path', 'causal association' (we did not consider reverse causality bias as it can be present in a predictive setting, or confounder as it can be used in risk factor/association studies), (iii) the authors considered that the exposure has an 'impact' or may 'affect' the outcome, (iv) use of propensity score, (v) the authors discussed or concluded that modifying the exposure might modify the outcome or suggested a policy controlling theexposure. For non modifiable variables such as race, distinction between causal or risk factor/association was made on a case-by-case basis. A research question was considered as *risk factor/association explicative* in the following situations: (i) authors used "risk factor" without further specification, (ii) none of the previous definitions (predictive or *causal explicative*) enabled to categorize the research question. No articles were simultaneously assigned to the two categories.

Variable selection methods

The variable selection method was searched in the multivariate model of each article. If there were several analyses, only the primary analysis corresponding to the primary aim was retained at first. The variable selection method was categorized according to three exclusive groups: -"Knowledge-based" including "knowledge-based without citation": the article provides bibliographic references supporting the process of selecting covariates (at least for one covariate) and " knowledge-based with citation": the article uses terms suggesting that the variables are selected based on knowledge or hypothesis prior to the analysis as suggested by "**known** to be confounders/potential confounder", "previously found to be associated" and '*a priori*', or reports that the adjustment variables have been chosen with an explicit thought of causal pathways, as well as an analysis labelled as a "mediation analysis".

- "data-driven method": the article specifies that at least one data-driven method was used for automatically selecting covariates;

- "*unclear*": the article is unclear, where reporting was insufficient to allow categorization into data-driven or knowledge-based methods.

In case an article used a combination of knowledge-based and data-driven methods, it was classified as using data-driven methods. The choice between alternative coding (continuous age versus age groups) and the selection of interaction terms were considered as part of the variable selection process. Hence, if authors statistically tested which recoding or interaction term induced the best model fit in order to include or not an interaction term or modify the coding of variables in the final model, the article was considered as using a data-driven method. The only exception being tests of interaction between covariates and time, usually tested in Cox regression models since they primarily serve a purpose of validating model assumptions rather than including additional terms in the model.

Studies categorized as having used "data-driven" methods, were further categorized according to 11 non-exclusive non-limitative method types defined by a prior literature search [6,7,13]: Backward elimination, Forward selection, Stepwise selection, Univariate selection, LASSO, Elasticnet, Change-in-estimate criterion, Purposeful method [16], High dimensional propensity score [17] considered as a "*data-driven method*" because the list of covariates included is data-driven, inclusion of an interaction term in the final model depending on the result of interaction test, variable coding depending on a test of linear fit.

Then, we searched variable selection in the sensitivity analyses section concerning the primary aim because information about variable selection may be managed in a sensitivity analysis. Sensitivity analyses were recorded in three non-exclusive categories depending on their objective and category "none" as follows: (i) alternative variable adjustment; (ii) alternative recoding; (iii) other sensitivity analyses. In case of data-driven alternative variable adjustment or recoding, the name of the method was recorded. Sensitivity analyses were considered only if authors employed the term 'sensitivity analyses'.

Excerpts

Selected excerpts were collected by the first author (T.P.-L.) when they were found to be particularly illustrative of the reporting of a method.

Source of information

Information on variable selection methods and excerpts was only collected from the methods section of each article but information on sensitivity analyses was searched in the entire article. Appendices were not considered. All references were reviewed blindly by two authors of this work: T.P.-L. and R.J. In case of disagreement, the article was read by a third author, A.L, and decisions were based on consensus or majority vote.

Statistical analysis

As data-driven methods could bias inference in case of causal questions they were considered as a misuse. The proportions of the variable selection methods are presented separately for the primary and sensitivity analyses. To estimate the misuse of data-driven methods, the distribution of the type of research question is described in this subgroup, including articles where data-driven methods were used in primary analyses or sensitivity analyses.

We conducted two post-hoc sensitivity analyses to explore the results about unclear reporting. First the definition of “knowledge based” was extended considering the articles that present lists of covariates right after sentences such as ‘these are (potential) confounders/mediators’. Second, 25 articles were randomly selected with variable selection methods that were still rated as “unclear” (from the method section alone, see above) after the extension of the “knowledge-based” definition. Additional information on variable selection methods (and its location) was searched in all 25 articles, appendix included. Confidence intervals at 95% were computed with Clopper-Pearson method. Data-Management and statistical analyses were

performed with R statistical software (version 3.5, The R Foundation for Statistical Computing, Vienna, Austria).

Results

Screening

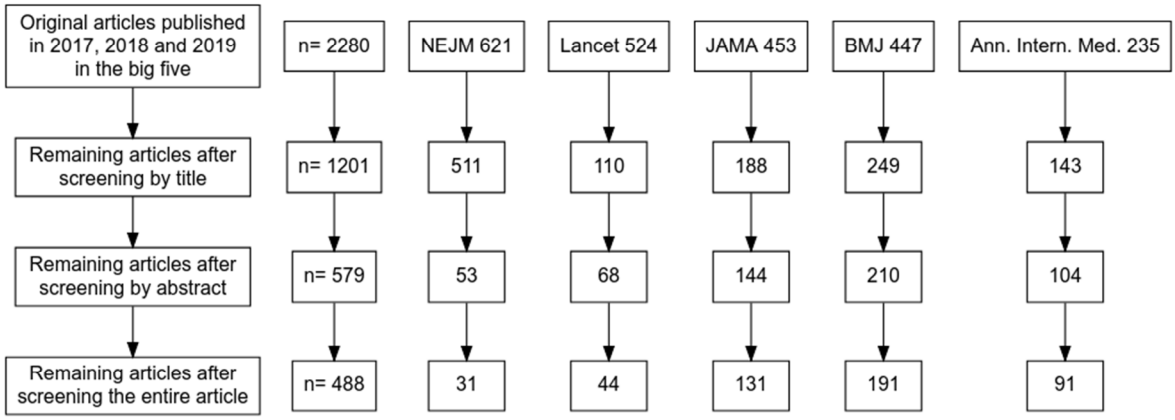

Figure 1: Flowchart review

Overall, 488 articles fulfilled inclusion criteria and were selected. Articles from the BMJ represented 39% of all included articles whereas the Lancet and the NEJM respectively represented 9% and 6.4% (Figure 1 and supplementary figure S1 for details about excluded articles).

Practice and reporting:

The variable selection method was unclear in 234 (48%) of the 488 articles. Data-driven methods were used in 78 (16%) articles in their primary analyses and in 83 (17%) if data-driven methods were searched in both the primary and sensitivity analyses rather than in the primary analysis alone. Among the 83 articles which used data-driven methods the majority (43 articles) used test-based methods (univariate selection, backward, stepwise, forward, purposeful, double univariate analysis and unclear data-driven methods). Data-driven methods were not systematically explained 3 (3.6%). Both change-in-estimate and univariate selection methods used in 29 articles were not named properly but were explicit enough. Alternative variable adjustment and alternative recoding performed in sensitivity analyses concerned respectively 101 (20.6%) and 30 (6.2%) of the articles (Table 1).

Table 1: distribution of variable selection methods in the main analysis, then in the sensitivity analyses of 488 articles

	Total n(%)
Variable selection method in main analysis	
Unclear	234 (48%)
Knowledge-based	176 (36%)
Knowledge-based with citation	82 (16.8%)
Knowledge-based without citation	94 (19.2%)
Directed acyclic graph	9 (1.8%)
Data-driven methods (detailed category non-exclusive)	78 (16%)
Univariate selection	18 (3.7%)
Interaction test	15 (3.1%)
Linearity test	13 (2.7%)
Change-in-estimate	8 (1.6%)
Stepwise	7 (1.4%)
Backward	6 (1.2%)
High dimensional propensity score	4 (0.8%)
Elastic Net	2 (0.4%)
Forward	1 (0.2%)
LASSO	1 (0.2%)
Purposeful	1 (0.2%)
Principal Component Analysis	1 (0.2%)
Deletion, substitution, and addition [18]	1 (0.2%)
Regression tree [19]	1 (0.2%)
Net reclassification index [20]	1 (0.2%)
Kernel regularized least squares [21]	1 (0.2%)
Colinearity test	1 (0.2%)
Double univariate analysis*	1 (0.2%)
Unclear data-driven method	3 (0.6%)
Variable selection method in Sensitivity analyses (detailed category non-exclusive)	
Alternative variable adjustment	101(20.7%)
Data-driven methods	9(1.8%)
Univariate selection	4(0.9%)
Forward	1(0.2%)
High dimensional propensity score	1(0.2%)
Interaction test	2(0.4%)
Purposeful	1(0.2%)
Alternative recoding	30(6.1%)
Others	245(50.2%)
None	120(24.6%)

* select all variables that significantly correlate with both the exposure and the outcome

Reporting post-hoc sensitivity analysis

After applying a slightly different definition of knowledge-based method, the variable selection method of 176 (36.1%) articles remained unclear.

The variable selection method remained unclear in 15 (60% CI: 39%; 79%) articles out of 25, nine were knowledge-based with citation, information found in introduction (n=1), discussion(n=3), appendix (n=3) or abstract (n=2) and one was univariate selection post matching (information in discussion)

Excerpts

A total of ten excerpts were selected. One explicitly did not use data-driven methods: "...rather than deferring to statistical criteria.". Three mentioned prior knowledge: " ...existing literature…", "...a priori assumptions...", "... reviewing the literature and consulting clinical experts.". Four defined relation between variables: "... identified potential confounders…", " ...potential mediators.", "... roles as either confounders or mediators", "... risk factors may be in the causal pathway.". Data-driven methods are presented for recoding and interaction. (Table 2).

Table 2: selected excerpts from articles

	Excerpts
Variable selection	As recommended, we identified potential confounders based on existing literature, rather than deferring to statistical criteria. [22] †
Variable selection	Model 2 was the primary model because model 3 risk factors may be in the causal pathway. [23] †
Variable selection	...adjusting for such variables is known to result in reduced precision and potential amplification of bias. [24] †
Variable selection	We included covariates on the basis of a priori assumptions about their roles as either confounders or mediators. [25] †
Sensitivity analysis	In a sensitivity analysis we additionally adjusted for the following potential mediators. [26] †
Interaction	The regression model was supplemented by adding interactions of covariates one at a time and selecting the model with superior balance. [27] ‡
Interaction	Where there was statistical significance, we included the interaction term in the final model and expressed the results using the interaction. [28]‡
Recoding	We defined [X1] categories after reviewing the literature and consulting clinical experts. [29] †
Recoding	...using cubic spline models to account for possible non-linear relations with the outcome. [30]
Recoding	...Akaike information criterion had been considered as the most reliable, flexible criterion for fitting penalised splines in Cox Models. [31]‡

† categorized in knowledge-based, ‡ categorized in data-driven

Misuse of data-driven methods:

Misuse of data-driven methods as defined in the method section was found in 51 (10.5%) of the 488 articles. Methods such as machine learning, shrinkage methods or high dimensional propensity score were used to compute propensity score in 8 cases. In 5/8 cases, univariate selection was used after propensity score matching for adjusting on unbalanced covariates (standardized difference > X). Principal component analysis was used because of genetic variables comprised in the list of covariates. In 3 cases, data-driven methods were only used in sensitivity analyses. Interaction or linear testing was the only data-driven method used in

respectively 7 cases and 8 other cases. Linear test concerning exposure in 5/9 cases and huge sample dataset (>100 000) in 4/9 cases. (Table 3).

Table 3: distribution of the type of research questions and distribution of variable selection methods for articles with data-driven methods in a primary or sensitivity analysis

	n (%)	
Articles with data-driven methods	83 (100%)	
Type of research question		
Predictive	14 (16.9%)	
Explicative	69 (83.1%)	
Risk/Association	18 (21.7%)	
Causal	51 (61.4%)	
Variable selection methods in the "causal" subgroup	**Primary analysis n**	**Sensitivity analysis n**
Total	48	7
Univariate selection	5	3
Interaction test	9	1
Linearity test	9	0
Change-in-estimate	8	0
Stepwise	3	0
Backward	3	0
High dimensional propensity score	4	1
Forward	0	1
LASSO	1	0
Purposeful	1	1
Principal Component Analysis	1	0
Deletion, substitution, and addition [18]	1	0
Regression tree [19]	1	0
Double univariate analysis*	1	0
Unclear data-driven method	3	0

* select all variables that significantly correlate with both the exposure and the outcome

Discussion

The percentage of unclear variable selection methods was very high, accounting for nearly 50% of all articles. However, our post-hoc analysis lowered the percentage to 36% and revealed that among articles graded as unclear 40% [21%;61%] of variable selection methods were not in the method section.The data-driven methods used are heterogeneous (test based, shrinkage, machine learning) and are not widely used (17%) in observational studies of the

big five medical journals. Furthermore, more recent variable selection methods for causal inference have been published such as augmented backward elimination, or group lasso and doubly robust estimation of causal effect [32], or outcome adaptive lasso [33] but none were found in our review. The time between the publication of the method and the creation of a corresponding package, and the statisticians' habits may delay their use.

Misuse of data-driven methods was low (10.5%), corresponding to use in *causal explicative* studies, and was rare if only primary analyses were considered and interaction, linear tests were not accounted (6.8%).

Comparison with the literature

Previous reviews reported 35% of unclear variable selection methods [12,13]. There are some explanations to this difference with our review (48%). First, some associations may be considered so well known, e.g., association between smoking and cardiovascular outcomes, that authors may think they do not need to provide any explanation or citation. Second, associations could be explained and cited in other sections than the Methods section (40% in our post-hoc analysis). Third, implicit justification of selection and poor reporting may be a consequence of lack of space in the methods section but authors could add supplemental data. Fourth, the level of reporting required for an article to be categorized as "knowledge based" may impact the percentage of unclear variable selection methods (36% with a slightly different definition) but we cannot compare with previous reviews that did not detailed their “knowledge based” definition.

In previous reviews, the use of data-driven methods was more frequent: 84% for explicative and predictive studies combined in a review in two Chinese epidemiologic journals published between 2004 and 2008 [34], 35% in four major epidemiologic journals published in 2008 [12], and finally 23% in the same epidemiologic journals (explicative studies) published in 2015 [13]. For the last two reviews, stepwise selection was respectively

used in 20% and 5% of the articles whereas the "change in estimate" method was used in 15% and 12% of the articles. These differences may be due to journal requirements or reporting, and publication year. However, we had a very broad definition of data-driven methods that increased their frequency compared to previous articles. Indeed, in our article interactions and recoding (even in sensitivity analyses), both data-driven were considered. Our definition of explicative studies was dichotomized in 'risk factor/association' and 'causal' that enabled to estimate misuse of data-driven methods contrary to these previous studies.

Strengths and limitations

Only five journals were searched but these "big five" are the most widely read medical journals worldwide. The New England Journal of Medicine and the Lancet contributed few articles to this study, because most of their articles report randomised or non-randomised interventional studies.

The type of research question that we defined as *explicative causal* may be controversial because there are several definitions of causality. Moreover, we based our extraction of the *type of research question* on authors' reporting, and our interpretation.

Variable selection methods were only screened in the methods section as reporting methods outside of the methods section is considered as poor reporting but a sensitivity analysis was carried out to measure the degree of missing information in the entire article, appendix included.

We performed an up to date review with double data extraction, making interpretation of the current practice reliable, but providing no insights on temporal trends since articles from only three consecutive years were assessed. Many authors listed covariates, with citations, or defining as confounder or potential confounder to justify their choices, but did not explicitly specify that all choices were made *a priori*, or without the use of data-driven methods. Hence, some models that we considered as being built by a knowledge-based method could have

been partly built with data-driven methods, leading to an underestimation of data-driven methods in our study.

Conclusion

Stepwise and data-driven variable selection methods do not seem to be widely used in the "big five" medical journals. Unfortunately, the variable selection method is not clearly reported in many articles, and the actual proportion of data-driven variable selection may be higher. Poor specification of the variable selection and coding scheme as well as the absence of published protocol leaves room for p-Hacking. Authors should clearly indicate that they did not use or rely on any of the data-driven methods such as those presented in the excerpts. Therefore, we recommend that the STROBE recommendation to authors to "Describe **all** statistical methods […]" be taken to the letter.

Conflict of interest statement

Declarations of interest: none

Acknowledgments

The authors are grateful to Nikki Sabourin-Gibbs, Rouen University Hospital, for her help in editing the manuscript and Emeline Lejeune for helping to extract the list of articles from the online tables of contents.

References

1. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiol Camb Mass. 1999 Jan;10(1):37–48.

2. Hamaker HC. On multiple regression analysis. Stat Neerlandica. 1962 Mar;16(1):31–56.

3. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. J R Stat Soc Ser B Methodol. 1996 Jan;58(1):267–88.

4. 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.

5. 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.

6. Desboulets L. A Review on Variable Selection in Regression Analysis. Econometrics. 2018 Nov 23;6(4):45.

7. Witte J, Didelez V. Covariate selection strategies for causal inference: Classification and comparison. Biom J Biom Z. 2019 Sep;61(5):1270–89.

8. Heinze G, Wallisch C, Dunkler D. Variable selection - A review and recommendations for the practicing statistician. Biom J Biom Z. 2018 May;60(3):431–49.

9. Harrell FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Second edition. Cham Heidelberg New York: Springer; 2015. 582 p. (Springer series in statistics).

10. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Ann Intern Med. 2007 Oct 16;147(8):W163-194.

11. Sharp MK, Bertizzolo L, Rius R, Wager E, Gómez G, Hren D. Using the STROBE statement: survey findings emphasized the role of journals in enforcing reporting guidelines. J Clin Epidemiol. 2019 Dec;116:26–35.

12. Walter S, Tiemeier H. Variable selection: current practice in epidemiological studies. Eur J Epidemiol. 2009;24(12):733–6.

13. 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.

14. 2016 Journal Impact Factor, Journal Citation Reports (Clarivate Analytics, 2020)

15. for TG2 of the STRATOS initiative, Sauerbrei W, Perperoglou A, Schmid M, Abrahamowicz M, Becher H, et al. State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues. Diagn Progn Res. 2020 Dec;4(1):3, s41512-020-00074–3.

16. Bursac Z, Gauss CH, Williams DK, Hosmer DW. Purposeful selection of variables in logistic regression. Source Code Biol Med. 2008 Dec 16;3:17.

17. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiol Camb Mass. 2009 Jul;20(4):512–22.

18. Sinisi SE, van der Laan MJ. Deletion/substitution/addition algorithm in learning with applications in genomics. Stat Appl Genet Mol Biol. 2004;3:Article18.

19. Loh W. Classification and regression trees. WIREs Data Min Knowl Discov. 2011 Jan;1(1):14–23.

20. Pencina MJ, D'Agostino RB, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. 2008 Jan 30;27(2):157–72; discussion 207-212.

21. Hainmueller J, Hazlett C. Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach. Polit Anal. 2014;22(2):143–68.

22. Fiolet T, Srour B, Sellem L, Kesse-Guyot E, Allès B, Méjean C, et al. Consumption of ultra-processed foods and cancer risk: results from NutriNet-Santé prospective cohort. BMJ. 2018 Feb 14;k322.

23. Zhong VW, Van Horn L, Cornelis MC, Wilkins JT, Ning H, Carnethon MR, et al. Associations of Dietary Cholesterol or Egg Consumption With Incident Cardiovascular Disease and Mortality. JAMA. 2019 Mar 19;321(11):1081.

24. Desai RJ, Bateman BT, Huybrechts KF, Patorno E, Hernandez-Diaz S, Park Y, et al. Risk of serious infections associated with use of immunosuppressive agents in pregnant women with autoimmune inflammatory conditions: cohort study. BMJ. 2017 Mar 6;j895.

25. Timpka S, Stuart JJ, Tanz LJ, Rimm EB, Franks PW, Rich-Edwards JW. Lifestyle in progression from hypertensive disorders of pregnancy to chronic hypertension in Nurses' Health Study II: observational cohort study. BMJ. 2017 Jul 12;j3024.

26. Nelson SM, Haig C, McConnachie A, Sattar N, Ring SM, Smith GD, et al. Maternal thyroid function and child educational attainment: prospective cohort study. BMJ. 2018 Feb 20;k452.

27. Helenius K, Longford N, Lehtonen L, Modi N, Gale C. Association of early postnatal transfer and birth outside a tertiary hospital with mortality and severe brain injury in extremely preterm infants: observational cohort study with propensity score matching BMJ 2019; 367 :l5678

28. Wallis CJD, Juvet T, Lee Y, et al. Association Between Use of Antithrombotic Medication and Hematuria-Related Complications. JAMA. 2017;318(13):1260–1271.

29. Thayakaran R, Adderley NJ, Sainsbury C, Torlinska B, Boelaert K, Šumilo D, et al. Thyroid replacement therapy, thyroid stimulating hormone concentrations, and long term health outcomes in patients with hypothyroidism: longitudinal study. BMJ. 2019 Sep 3;l4892.

30. Abrahami D, Douros A, Yin H, Yu OHY, Renoux C, Bitton A, et al. Dipeptidyl peptidase-4 inhibitors and incidence of inflammatory bowel disease among patients with type 2 diabetes: population based cohort study. BMJ. 2018 Mar 21;k872.

31. Lv Y-B, Gao X, Yin Z-X, Chen H-S, Luo J-S, Brasher MS, et al. Revisiting the association of blood pressure with mortality in oldest old people in China: community based, longitudinal prospective study. BMJ. 2018 Jun 5;k2158.

32. 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.

33. Shortreed SM, Ertefaie A. Outcome‐adaptive lasso: Variable selection for causal inference. Biometrics. 2017 Dec;73(4):1111–22.

34. Liao H, Lynn HS. A survey of variable selection methods in two Chinese epidemiology journals. BMC Med Res Methodol. 2010 Dec;10(1):87.