Recent methodological trends in Epidemiology: No need for data-driven variable selection?

Christian Staerk[1] Alliyah Byrd[2] Andreas Mayr[1]

[1]Department of Medical Biometry, Informatics and Epidemiology,
University Hospital Bonn, Bonn, Germany

[2]Surgical Sciences, Department of Surgery, Duke University, Durham, USA

Abstract

Variable selection in regression models is a particularly important issue in epidemiology, where one usually encounters observational studies. In contrast to randomized trials or experiments, confounding is often not controlled by the study design, but has to be accounted for by suitable statistical methods. For instance, when risk factors should be identified with unconfounded effect estimates, multivariable regression techniques can help to adjust for confounders. We investigated the current practice of variable selection in four major epidemiological journals in 2019 and found that the majority of articles used subject-matter knowledge to determine a priori the set of included variables. In comparison with previous reviews from 2008 and 2015, fewer articles applied data-driven variable selection. Furthermore, for most articles the main aim of analysis was hypothesis-driven effect estimation in rather low-dimensional data situations (i.e., large sample size compared to the number of variables). Based on our results we discuss the role of data-driven variable selection in epidemiology.

Keywords: Confounding; Epidemiologic methods; Modelling; Regression; Variable selection.

1 Introduction

Variable selection in the context of regression models is particularly important in the field of epidemiology. While randomized controlled trials (RCTs) are regarded as the "gold standard" to control for confounding in clinical medicine [1], the nature of epidemiological studies is usually observational as they are concerned with the determinants of public health. In observational studies, the issue of confounding needs to be addressed by suitable data analysis strategies. For instance, when the effect of a potential risk factor for a certain disease or health status should be examined, it is desirable to estimate this effect via a multivariable regression model including those covariables that may confound the effect estimate for the risk factor of interest. Thus, the issue of variable selection is central to explicative research in epidemiology to control for confounders (e.g., [2, 3]). But also in the context of predictive research questions, variable selection in regression models plays an important role to facilitate the applicability and interpretability of prediction models (e.g., [4, 5]).

From a methodological perspective, strategies for variable selection can be broadly classified into two main categories: *data-driven* variable selection techniques and *a priori* variable selection based on subject-matter knowledge by the researchers. In data-driven approaches, variables are selected based on the observed data by using various methods, including univariate selection [6], classical stepwise selection [7], change-in-estimate methods [8], and regularization methods such as the Lasso [9], the elastic net [10], and the adaptive Lasso [11]. On the other hand, in a priori selection approaches, the inclusion of variables is decided prior to the data analysis based on subject-matter knowledge. In particular, methods from causal inference, including directed acyclic graphs (DAGs), can be used to make the modelling assumptions of the researchers explicit and select a priori sets of confounders accordingly [3, 12, 13].

The discussion of which variable selection method is most appropriate for a research question has to be placed in the context of the main aim of the respective epidemiological study, which may be of a primarily hypothesis-driven, hypothesis-generating, descriptive or predictive nature (cf., [14, 15, 16]). Particularly in settings where the main aim of analysis is hypothesis-driven effect estimation, data-driven variable selection may even distort statistical inference when confidence intervals and p-values are computed without taking the previous variable selection into account (cf., [4, 17, 18]). Thus, for such classical epidemiological estimation settings, a priori variable selection based on subject-matter knowledge or information from the literature is often preferred to data-driven selection approaches (e.g., [3, 4]). Nevertheless, Hafermann et al. (2021, [19]) illustrate that variable selection based on background "knowledge" can also lead to model misspecification when covariates are wrongly included or excluded based on previous studies, calling for a critical assessment of the source of this prior knowledge. Another important practical aspect is the dimensionality of the research data. While data-driven variable selection may often not be beneficial in classical settings with relatively large sample sizes compared to numbers of available covariates ($n \gg p$), in high-dimensional situations with large numbers of candidate variables (particularly if $p \gg n$), some form of variable selection may be inevitable to allow for the estimation of regression coefficients (cf., [20]).

In this study, we investigated the current practice of variable selection in the context of regression models by analyzing articles published in four major epidemiological journals in 2019. We found that the majority of articles used subject-matter knowledge to determine a priori the set of included variables. When compared to two previous reviews for the years 2008 and 2015 [21, 22], a decreasing and relatively small number of articles applied data-driven variable selection. In line with this, for most considered articles, the main aim of analysis was hypothesis-driven effect estimation in rather low-dimensional data situations. We additionally collected data on the reporting of p-values and sensitivity analyses in the context of variable selection. Based on our findings we discuss the current role of data-driven variable selection methods in epidemiology.

2 Methods

We reviewed the variable selection methodology of articles published in 2019 in the following four journals: the American Journal of Epidemiology, Epidemiology, the European Journal of Epidemiology, and the International Journal of Epidemiology. To analyze trends regarding the usage of different variable selection methods over time, we employed a similar methodology as two previous descriptive reviews for the years 2008 and 2015 [21, 22], focusing on the same four journals.

Initially, each article was screened and evaluated to determine if it was eligible. Only original research articles that analyzed individual data were included. Similar to previous reviews [21, 22], we only included explicative and predictive studies and excluded commentaries, essays, correspondences, editorials, opinions, corrections, book reviews, data resources, study design reports, cohort profiles, software application profiles, reviews and meta-analyses, descriptive studies, validation studies, methodological studies, simulation and modelling studies, case-crossover studies, randomized controlled trials, genome-wide, epigenome-wide and similar association studies, as well as instrumental variable and Mendelian randomization studies. Each article was assessed for eligibility and classified by one of the authors. Another author reviewed the inclusion and initial classification of each article and, after discussion, articles were reclassified if necessary. Additional information on the exclusion of articles can be found in Web Appendix 1 (see Web Figure 1 for a PRISMA-type flowchart [23] and Web Table 1 for numbers of articles identified, excluded, and reasons for exclusion by journal). Furthermore, Web Table 2 provides detailed data on the exclusion and classification of all individual articles considered in our review.

Articles were classified regarding the usage of variable selection techniques, including the following categories: “prior knowledge”, “stepwise selection”, “change-in-estimate”, “univariate selection”, “regularization methods”, “other methods”, and “not described”. In particular, articles were classified as “prior knowledge” if they contained any justification for the inclusion of variables based on previous studies or subject-matter knowledge. Articles classified as “prior knowledge” were also checked for the use of causal graphs to specify the assumptions of the researchers and justify the selection of variables (cf., [12]); such articles were additionally classified in the subcategory “prior knowledge and causal graphs”. Articles that employed forward selection, backward selection, or a combination of forward and backward selection using various selection criteria such as the statistical significance or information criteria (cf., [7]) were classified as “stepwise selection”. The category “change-in-estimate” consists of articles where the selection of variables was based on changes in estimated regression coefficients, event risks, or correlations (cf., [8]). “Univariate selection” refers to articles that based the variable selection on univariate correlations or p-values for associations between individual covariates and the outcome of interest (cf., [6]). Articles using regularized regression such as the

Lasso [9] were classified as “regularization methods”. The category “other methods” summarizes articles that used techniques that did not fit into the classification groups previously listed. Finally, articles were classified as “not described” if they did not report the use of any data-driven variable selection method and also gave no explicit information regarding the inclusion of variables in a regression context based on prior knowledge. Note that articles classified as “not described” may also implicitly have used prior knowledge for variable selection, but did not include any explicit explanations or references for the inclusion of variables (cf., [21]).

In addition to collecting data on the use of variable selection methods in regression models, we also categorized articles based on the primary aim of the analysis as well as the reporting of p-values and sensitivity analyses. In particular, the primary aim of an article was determined to be “hypothesis-driven”, when it examined one or more formulated hypotheses based on previous research or prior knowledge. On the other hand, the aim of an article was classified as “hypothesis-generating”, when it conducted analyses for patterns and associations without stating any prespecified hypothesis. The category “prediction” was for articles where the main aim was to develop or evaluate prediction models for particular outcome variables. Furthermore, the results section and all tables of eligible articles were scanned for any reporting of p-values. Finally, we examined whether articles employed sensitivity analyses using different sets of included covariables to report multiple adjusted or unadjusted effect estimates in the context of regression.

3 Results

Table 1 summarizes the results of our review of articles published in the four epidemiological journals in 2019 and also provides comparisons with two previous reviews of articles published in 2008 and 2015 [21, 22]. Note that articles using two or more variable selection techniques were classified into multiple categories accordingly.

In total, 199 (73%) of the 272 included articles for the year 2019 used prior knowledge to select variables, among which 35 articles (13%) also employed causal graphs to summarize and justify the selection of variables based on subject-matter knowledge. In all of these 35 articles, directed acyclic graphs (DAGs) were used; specifically, 9 articles (26%) provided at least one DAG in the main body of the article, 19 articles (54%) provided at least one DAG in supplementary materials, while 7 articles (20%) only mentioned the use of DAGs without providing any explicit graphs. Notably, only 43 (16%) of the included articles reported the use of data-driven variable selection methods. The most commonly applied data-driven method was based on change-in-estimate considerations (18 articles, 7% of all included articles), followed by stepwise selection methods (11 articles, 4%), univariate selection methods (9 articles, 3%), and other methods (7 articles, 3%). Only two of the included articles in 2019 reported the use of regularization methods [24, 25]. Finally, 43 articles (16%) did not provide any information regarding the employed variable selection techniques and also did not provide justifications for the included variables. While results were generally similar across the four different journals, some differences could also be observed. For example, articles published in the European Journal of Epidemiology (EJE) in 2019 used prior knowledge for variable selection more frequently (88%) compared with the other three journals (65% – 73%), while articles in EJE were less frequently (2%) classified into the category “not described” compared with the other three journals (13% – 26%).

Among the seven articles classified as “other methods” for variable selection, one article [26] used high-dimensional propensity scores [27]; one study [28] employed prior knowledge in combination with model averaging based on the Akaike information criterion (AIC) [29]; one study [30] used several variable selection methods including random forests [31]; one study [32] used the AIC (without model averaging); one study [33] used the Bayesian information criterion (BIC) for variable selection; one study [34] employed a Bayesian kernel machine regression approach [35] using posterior inclusion probabilities (PIPs) for variable selection; and one study [36] used a data-driven Bayesian variable selection approach based on DAGs [37].

When compared with two previous reviews for articles in the same four journals published in the years 2008 and 2015 [21, 22], an increasing trend for the use of prior knowledge to determine the included variables is apparent (28% in 2008, 50% in 2015, and 73% in 2019). On the other hand, the use of data-driven variable selection has been steadily declining (37% in 2008, 24% in 2015, and 16% in 2019). In particular, in comparison to previous reviews, we find that a decreasing fraction of studies employed change-in-estimate (15% in 2008, 12% in 2015, and 7% in 2019), stepwise selection (20% in 2008, 5% in 2015, and 4% in 2019), and univariate selection methods (9% in 2015, 3% in 2019). While no articles in 2008 or 2015 reported the use of regularization methods for variable selection, in 2019 one article employed the Lasso [9] to select metabolic correlates of habitual sleep quality [24] and one article applied the adaptive elastic net [38] to classify suicidal behavior. Furthermore, we found that the fraction of articles in 2019 that did not provide any information or reasoning regarding the variable selection was smaller compared to previous reviews (35% in 2008, 37% in 2015, and 16% in 2019). Note that while this finding indicates more detailed reporting in 2019, it may also be partly due to potential differences between reviews regarding the classification of articles into the categories "prior knowledge" or "not described". In particular, while previous reviews may have focused on the methods sections of the articles to identify justifications for the prior inclusion of variables, in our review we considered the full research articles (including also the introduction, results and discussion sections).

Regarding the main aims of analysis, most included articles in our 2019 review focused on hypothesis-driven research questions (255 articles, 94%), especially the estimation of effects of particular variables (e.g., risk factors) on outcomes, adjusting for potential confounders in rather low-dimensional data situations. Only a minority of the included articles were of a primarily hypothesis-generating (11 articles, 4%) or predictive nature (6 articles, 2%). A majority of included articles (190 articles, 70%) used sensitivity analyses to report effect estimates based on different sets of included variables in the models, also considering the case where both unadjusted estimates (with no additional covariates in the model) and adjusted estimates were reported. Furthermore, a majority of included articles (167 articles, 61%) reported p-values in their results. Among those articles that reported p-values, 31 articles (11% of all included articles) had used data-driven variable selection, without indications of adjustments for variable selection via post-selection inference.

4 Discussion

In comparison with previous descriptive reviews for the years 2008 and 2015 [21, 22], our current review of articles from 2019 in four major epidemiological journals revealed a clear trend towards less data-driven variable selection and towards more a priori variable selection based on subject-matter knowledge. Furthermore, we found that a considerable fraction of articles (13%) also explicitly justified the prior inclusion of variables using directed acyclic graphs (DAGs) and that a majority of articles conducted sensitivity analyses to report differently adjusted (or unadjusted) effect estimates based on various sets of included variables.

In light of these trends, a central question arises: where is the need for data-driven variable selection in modern epidemiological research? To put this question into context it is important to consider the main aim or type of a particular data analysis that involves regression. In our view, the main targets of epidemiological research in this context can be grouped as follows (cf., [14, 16]): (1) identification of risk factors or other relevant variables (*hypothesis-generating*), (2) estimation of the effects of particular risk factors on health outcomes, adjusting for confounders (*hypothesis-driven*), and (3) prediction of health outcomes based on several risk factors and other variables (*prediction*). Variable selection can play an important but different role in each of the three research settings. In hypothesis-generating research (1), data-driven variable selection methods are particularly suited to identify sets of relevant risk factors in an exploratory way. On the other hand, in hypothesis-driven research (2), variable selection is primarily concerned with confounder selection, which is often based on subject-matter knowledge. Finally, in predictive research settings (3), both a priori and data-driven variable selection methods can be incorporated into the construction of prediction models, while the main focus is on the predictive performance and not on the interpretation of the effects of particular risk factors.

From a general statistical perspective, "descriptive modelling" can be considered as a fourth target of variable selection, where the expectation of a particular outcome should be represented parsimoniously by a set of the most relevant variables [15]. While descriptive studies have an important role in epidemiology [39], the adjustment for covariates is usually a different issue in this context compared to the parsimonious descriptive modelling in statistics (cf., [40]). In particular, epidemiological studies on describing a population of interest do not necessarily require control for covariates and covariate adjustment can even be diametrical to the descriptive research question [41]. Therefore, as also in previous reviews on variable selection in epidemiology [21, 22], we excluded descriptive studies from our review and focused on (hypothesis-generating or hypothesis-driven) explicative and predictive studies, where variable selection is typically a crucial issue in the context of regression models.

Modern statistical learning methods including regularization methods like the Lasso [9, 10, 11] and boosting methods (inducing implicit regularization via early stopping of the algorithm) [42, 43, 44] are primarily designed for predictive settings but encourage sparse models (including only smaller subsets of all considered variables) to enhance the applicability and interpretability of the models [45]. Since most of the articles included in our review were concerned with research focusing on hypothesis-driven estimation, it is not a big surprise that, as in previous reviews [21, 22], very few articles employed modern regularization methods like the Lasso for variable selection. Furthermore, we believe that, from a methodological perspective, the trend towards less data-driven and more a priori variable selection should be viewed generally positive in the context of "classical" hypothesis-driven research questions, as data-driven variable selection complicates post-selection statistical inference (cf., [4, 17, 18]). Despite recent methodological advancements in the field of post-selection inference (e.g., [46, 47, 48, 49, 50]), in the reviewed articles we could not identify any explicit applications of modern post-selection inference methods to adjust confidence intervals and p-values for data-driven variable selection.

In the context of hypothesis-driven research, causal directed acyclic graphs (DAGs) are particularly flexible tools for the specification of prior knowledge and assumptions to justify the a priori selection of confounders. Among the articles in our review which used prior knowledge for variable selection, 35 articles (18%) incorporated DAGs for variable selection, while 164 articles (82%) employed a priori variable selection without mentioning the use of causal graphs. Furthermore, 7 articles which mentioned DAGs did not make any causal graphs available. Thus, while our review indicates an increasing prevalence of a priori variable selection in hypothesis-driven epidemiological research, we believe that a wider adoption of causal graphs and further improvements on the reporting would be desirable (cf., [51] for a recent review with recommendations on the reporting of DAGs).

Aside from the main aim of the epidemiological analysis, another crucial aspect in the context of variable selection is the dimensionality of the research data. While in classical epidemiological research the available datasets are often low-dimensional with a considerably smaller number of observed variables in relation to the sample size ($p \ll n$), in genetic epidemiology one is usually confronted with high-dimensional ($p \gg n$) or large-scale data (large number of variables p and large sample size n). In high-dimensional and large-scale situations, some form of variable selection or regularization is typically required to fit multivariable regression models. Even when the main aim is hypothesis-driven estimation, adjusting for all potential confounders in a classical regression model may not be feasible or inefficient if the number of candidate genes or genetic variants (SNPs) is very large. Furthermore, variable selection based on prior knowledge and DAGs tends to be very challenging in such settings, although recent methods have been proposed for data-driven confounder selection in high-dimensional settings using graphical models when the underlying causal structure is unknown [52, 53].

While articles on genetic epidemiology were excluded from our review to provide comparability with previous reviews [21, 22], in an additional exploratory analysis we also investigated the variable selection methodology in original research articles concerned with genetic epidemiology published in the four epidemiological journals in 2019. Based on a limited number of published articles on genetic epidemiology in the four journals in 2019, comprising 23 Mendelian randomization studies and 6 (epi)-genome-wide association studies, we found that genetic variables were selected independently from each other in a data-driven way or were included based on univariate associations with the considered outcome from previous studies (e.g., p-values from univariate summary statistics of genome-wide association studies). While the latter may also be viewed as a priori variable selection, it entails a large risk of model misspecification in subsequent studies resulting from the wrong inclusion or exclusion of genetic variables (cf., [19]). In Mendelian randomization studies, for example, it is important to select valid genetic instruments to identify and estimate causal effects (e.g., [54, 55, 56]). Statistical fine-mapping approaches are refinements to univariate selection approaches, which aim to select candidate causal variants by applying modern multivariable regression methods on individual-level data instead of relying only on univariate summary statistics (e.g., [57]). Scalable multivariable statistical learning methods such as extensions of the Lasso or boosting may also play an increasingly important role in the development of polygenic risk models for the prediction of complex traits (e.g., [58, 59]).

We conclude that a priori variable selection based on subject-matter knowledge is increasingly predominant in epidemiological research with a focus on hypothesis-driven estimation in data situations with a limited number of variables. However, there is a particular need for data-driven variable selection in high-dimensional and large-scale data situations, which are typically encountered in genetic epidemiology. While further methodological research on scalable variable selection and post-selection inference methods is warranted, a special focus should also be on the transfer of new methodological developments from statistics towards epidemiological research and practice.

References

[1] David S Jones and Scott H Podolsky. The history and fate of the gold standard. *Lancet*, 385(9977):1502–1503, 2015.

[2] Sander Greenland. Invited commentary: variable selection versus shrinkage in the control of multiple confounders. *Am J Epidemiol*, 167(5):523–529, 2008.

[3] Tyler J VanderWeele. Principles of confounder selection. *Eur J Epidemiol*, 34 (3):211–219, 2019.

[4] Georg Heinze and Daniela Dunkler. Five myths about variable selection. *Transpl Int*, 30(1):6–10, 2017.

[5] Mohammad Ziaul Islam Chowdhury and Tanvir C Turin. Variable selection strategies and its importance in clinical prediction modelling. *Fam Med Community Health*, 8(1):e000262, 2020.

[6] Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *J R Stat Soc Series B Stat Methodol*, 70(5):849–911, 2008.

[7] Ronald R Hocking. A Biometrics invited paper. The analysis and selection of variables in linear regression. *Biometrics*, 32:1–49, 1976.

[8] Denis Talbot, Awa Diop, Mathilde Lavigne-Robichaud, and Chantal Brisson. The change in estimate method for selecting confounders: A simulation study. *Stat Methods Med Res*, 30(9):2032–2044, 2021.

[9] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol*, 58(1):267–288, 1996.

[10] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol*, 67(2):301–320, 2005.

[11] Hui Zou. The adaptive lasso and its oracle properties. *J Am Stat Assoc*, 101 (476):1418–1429, 2006.

[12] Sander Greenland, Judea Pearl, and James M Robins. Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48, 1999.

[13] Ian Shrier and Robert W Platt. Reducing bias through directed acyclic graphs. *BMC Med Res Methodol*, 8(1):1–15, 2008.

[14] Sean Carroll and David Goodstein. Defining the scientific method. *Nat Methods*, 6(237), 2009.

[15] Galit Shmueli. To Explain or to Predict? *Stat Sci*, 25(3):289 – 310, 2010.

[16] Bradley Efron. Prediction, estimation, and attribution. *J Am Stat Assoc*, 115 (530):636–655, 2020.

[17] Georg Heinze, Christine Wallisch, and Daniela Dunkler. Variable selection – A review and recommendations for the practicing statistician. *Biom J*, 60(3): 431–449, 2018.

[18] Willi Sauerbrei, Aris Perperoglou, Matthias Schmid, Michal Abrahamowicz, Heiko Becher, Harald Binder, Daniela Dunkler, Frank E Harrell, Patrick Royston, and Georg Heinze. State of the art in selection of variables and functional

forms in multivariable analysis—–outstanding issues. *Diagnostic and prognostic research*, 4(1):1–18, 2020.

[19] Lorena Hafermann, Heiko Becher, Carolin Herrmann, Nadja Klein, Georg Heinze, and Geraldine Rauch. Statistical model building: Background “knowledge” based on inappropriate preselection causes misspecification. *BMC Med Res Methodol*, 21(1):1–12, 2021.

[20] Yoav Benjamini. Selective inference: The silent killer of replicability. *Harv Data Sci Rev*, 2(4), 2020.

[21] Stefan Walter and Henning Tiemeier. Variable selection: Current practice in epidemiological studies. *Eur J Epidemiol*, 24(12):733–736, 2009.

[22] Denis Talbot and Victoria K Massamba. A descriptive review of variable selection methods in four epidemiologic journals: There is still room for improvement. *Eur J Epidemiol*, 34(8):725–730, 2019.

[23] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, and David Moher. Updating guidance for reporting systematic reviews: development of the PRISMA 2020 statement. *J Clin Epidemiol*, 134:103–112, 2021.

[24] Tianyi Huang, Oana A Zeleznik, Elizabeth M Poole, Clary B Clish, Amy A Deik, Justin M Scott, Céline Vetter, Eva S Schernhammer, Robert Brunner, Lauren Hale, et al. Habitual sleep quality, plasma metabolites and risk of coronary heart disease in post-menopausal women. *Int J Epidemiol*, 48(4): 1262–1274, 2019.

[25] Qiu-Yue Zhong, Leena P Mittal, Margo D Nathan, Kara M Brown, Deborah Knudson González, Tianrun Cai, Sean Finan, Bizu Gelaye, Paul Avillach, Jordan W Smoller, et al. Use of natural language processing in electronic medical records to identify pregnant women with suicidal behavior: towards a solution to the complex classification problem. *Eur J Epidemiol*, 34:153–162, 2019.

[26] Rachel P Ogilvie, Richard F MacLehose, Alvaro Alonso, Faye L Norby, Kamakshi Lakshminarayan, Conrad Iber, Lin Y Chen, and Pamela L Lutsey. Diagnosed sleep apnea and cardiovascular disease in atrial fibrillation patients: The role of measurement error from administrative data. *Epidemiology*, 30(6): 885–892, 2019.

[27] Sebastian Schneeweiss, Jeremy A Rassen, Robert J Glynn, Jerry Avorn, Helen Mogun, and M Alan Brookhart. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, 20 (4):512–522, 2009.

[28] Daniel M Weinberger, Virginia E Pitzer, Gili Regev-Yochay, Noga Givon-Lavi, and Ron Dagan. Association between the decline in pneumococcal disease in unimmunized adults and vaccine-derived protection against colonization in toddlers and preschool-aged children. *Am J Epidemiol*, 188(1):160–168, 2019.

[29] Kenneth P Burnham and David R Anderson. Multimodel inference: Understanding AIC and BIC in model selection. *Sociol Methods Res*, 33(2):261–304, 2004.

[30] Marisa A Hast, Mike Chaponda, Mbanga Muleba, Jean-Bertin Kabuya, James Lupiya, Tamaki Kobayashi, Timothy Shields, Justin Lessler, Modest Mulenga, Jennifer C Stevenson, et al. The impact of 3 years of targeted indoor residual spraying with pirimiphos-methyl on malaria parasite prevalence in a high-

transmission area of northern Zambia. *Am J Epidemiol*, 188(12):2120–2130, 2019.

[31] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[32] Michael G Walsh, Siobhan M Mor, Hindol Maity, and Shah Hossain. Forest loss shapes the landscape suitability of Kyasanur Forest disease in the biodiversity hotspots of the Western Ghats, India. *Int J Epidemiol*, 48(6):1804–1814, 2019.

[33] Jiamin Yin, Camille Lassale, Andrew Steptoe, and Dorina Cadar. Exploring the bidirectional associations between loneliness and cognitive functioning over 10 years: the English longitudinal study of ageing. *Int J Epidemiol*, 48(6): 1937–1948, 2019.

[34] Arce Domingo-Relloso, Maria Grau-Perez, Laisa Briongos-Figuero, Jose L Gomez-Ariza, Tamara Garcia-Barrera, Antonio Duenas-Laita, Jennifer F Bobb, F Javier Chaves, Marianthi-Anna Kioumourtzoglou, Ana Navas-Acien, et al. The association of urine metals and metal mixtures with cardiovascular incidence in an adult population from Spain: the Hortega Follow-Up Study. *Int J Epidemiol*, 48(6):1839–1849, 2019.

[35] Jennifer F Bobb, Linda Valeri, Birgit Claus Henn, David C Christiani, Robert O Wright, Maitreyi Mazumdar, John J Godleski, and Brent A Coull. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics*, 16(3):493–508, 2015.

[36] Yi-Chia Lee, Chen-Yang Hsu, Sam Li-Sheng Chen, Amy Ming-Fang Yen, Sherry Yueh-Hsia Chiu, Jean Ching-Yuan Fann, Shu-Lin Chuang, Wen-Feng Hsu, Tsung-Hsien Chiang, Han-Mo Chiu, et al. Effects of screening and universal healthcare on long-term colorectal cancer mortality. *Int J Epidemiol*, 48 (2):538–548, 2019.

[37] Isabelle Bray. Application of Markov chain Monte Carlo methods to projecting cancer incidence and mortality. *J R Stat Soc Ser C Appl Stat*, 51(2):151–164, 2002.

[38] Hui Zou and Hao Helen Zhang. On the adaptive elastic-net with a diverging number of parameters. *Ann Stat*, 37(4):1733–1751, 2009.

[39] Matthew P Fox, Eleanor J Murray, Catherine R Lesko, and Shawnita Sealy-Jefferson. On the need to revitalize descriptive epidemiology. *Am J Epidemiol*, 191(7):1174–1179, 2022.

[40] Catherine R Lesko, Matthew P Fox, and Jessie K Edwards. A framework for descriptive epidemiology. *Am J Epidemiol*, 191(12):2063–2070, 2022.

[41] Sara Conroy and Eleanor J Murray. Let the question determine the methods: descriptive epidemiology done right. *Br J Cancer*, 123(9):1351–1352, 2020.

[42] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Ann Stat*, 29(5):1189–1232, 2001.

[43] Peter Bühlmann and Torsten Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Stat Sci*, 22(4):477–505, 2007.

[44] Andreas Mayr, Harald Binder, Olaf Gefeller, and Matthias Schmid. The evolution of boosting algorithms. *Methods Inf Med*, 53(06):419–427, 2014.

[45] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proc Natl Acad Sci USA*, 116(44):22071–22080, 2019.

[46] Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *Ann Stat*, 41(2):802–837, 2013.

[47] Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference, with application to the lasso. *Ann Stat*, 44(3):907–927, 2016.

[48] Ryan J Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact post-selection inference for sequential regression procedures. *J Am Stat Assoc*, 111(514):600–620, 2016.

[49] Rina Foygel Barber and Emmanuel J Candès. A knockoff filter for high-dimensional selective inference. *Ann Stat*, 47(5):2504–2537, 2019.

[50] David Rügamer, Philipp F M Baumann, and Sonja Greven. Selective inference for additive and linear mixed models. *Comput Stat Data Anal*, 167:107350, 2022.

[51] Peter WG Tennant, Eleanor J Murray, Kellyn F Arnold, Laurie Berrie, Matthew P Fox, Sarah C Gadd, Wendy J Harrison, Claire Keeble, Lynsie R Ranker, Johannes Textor, et al. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *Int J Epidemiol*, 50(2):620–632, 2021.

[52] Jenny Häggström. Data-driven confounder selection via Markov and Bayesian networks. *Biometrics*, 74(2):389–398, 2018.

[53] Janine Witte and Vanessa Didelez. Covariate selection strategies for causal inference: Classification and comparison. *Biom J*, 61(5):1270–1289, 2019.

[54] Tyler J VanderWeele, Eric J Tchetgen Tchetgen, Marilyn Cornelis, and Peter Kraft. Methodological challenges in Mendelian randomization. *Epidemiology*, 25(3):427–435, 2014.

[55] Jack Bowden, George Davey Smith, and Stephen Burgess. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol*, 44(2):512–525, 2015.

[56] Stephen Burgess and Simon G Thompson. Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur J Epidemiol*, 32 (5):377–389, 2017.

[57] Daniel J Schaid, Wenan Chen, and Nicholas B Larson. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet*, 19(8):491–504, 2018.

[58] Junyang Qian, Yosuke Tanigawa, Wenfei Du, Matthew Aguirre, Chris Chang, Robert Tibshirani, Manuel A Rivas, and Trevor Hastie. A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. *PLoS Genet*, 16(10):e1009141, 2020.

[59] Carlo Maj, Christian Staerk, Oleg Borisov, Hannah Klinkhammer, Ming Wai Yeung, Peter Krawitz, and Andreas Mayr. Statistical learning for sparser fine-mapped polygenic models: The prediction of LDL-cholesterol. *Genet Epidemiol*, 46(8):589–603, 2022.

Table 1: Results of our review for included articles published in four major epidemiological journals in 2019.

	Am J Epi		Epidem		Eur J Epi		Int J Epi		Total 2019		Total 2015[a]		Total 2008[a]	
	($n = 109$)		($n = 43$)		($n = 42$)		($n = 78$)		($n = 272$)		($n = 292$)		($n = 300$)	
Classification category	**No.**	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**
Variable selection methods														
Prior knowledge	77	71	28	65	37	88	57	73	199	73	146	50	83	28
Prior knowl. & causal graphs	16	15	9	21	4	10	6	8	35	13	—[b]	—[b]	—[b]	—[b]
Data-driven variable selection	16	15	4	9	5	12	18	23	43	16	69	24	112	37
Change-in-estimate	9	8	1	2	3	7	5	6	18	7	34	12	44	15
Stepwise selection	4	4	1	2	1	2	5	6	11	4	16	5	59	20
Univariate selection	4	4	1	2	0	0	4	5	9	3	26	9	—[b]	—[b]
Regularization methods	0	0	0	0	1	2	1	1	2	1	0	0	0	0
Other methods	2	2	1	2	0	0	4	5	7	3	5	2	9	3
Not described	21	19	11	26	1	2	10	13	43	16	107	37	105	35
Main aim of analysis														
Hypothesis-driven	105	96	43	100	36	86	71	91	255	94	—[b]	—[b]	—[b]	—[b]
Hypothesis-generating	4	4	0	0	3	7	4	5	11	4	—[b]	—[b]	—[b]	—[b]
Prediction	0	0	0	0	3	7	3	4	6	2	—[b]	—[b]	—[b]	—[b]
Reporting of results														
Sensitivity analyses[c]	73	67	28	65	36	86	53	68	190	70	—[b]	—[b]	—[b]	—[b]
P-values	75	69	11	26	25	60	56	72	167	61	—[b]	—[b]	—[b]	—[b]
P-values & data-driven sel.	12	11	1	2	3	7	15	19	31	11	—[b]	—[b]	—[b]	—[b]

Abbreviations: Am J Epi, American Journal of Epidemiology; data-driven sel., data-driven variable selection; Epidem, Epidemiology (journal); Eur J Epi, European Journal of Epidemiology; Int J Epi, International Journal of Epidemiology; knowl., knowledge; mult. models, multiple models; No., number.

[a] For comparisons, total results of two previous reviews on the use of variable selection methods for the years 2008 [21] and 2015 [22] are also shown.
[b] Classification categories were not considered in previous reviews for the years 2008 and 2015.
[c] Sensitivity analyses with different sets of included covariables.