

RESEARCH ARTICLE

Review of guidance papers on regression modeling in statistical series of medical journals

Christine Wallisch^{1,2*}, Paul Bach^{1,3}, Lorena Hafermann¹, Nadja Klein³, Willi Sauerbrei⁴, Ewout W. Steyerberg⁵, Georg Heinze², Geraldine Rauch^{1*}, on behalf of topic group 2 of the STRATOS initiative[¶]

1 Institute of Biometry and Clinical Epidemiology, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Charité—Universitätsmedizin Berlin, Berlin, Germany, **2** Center for Medical Statistics, Informatics and Intelligent Systems, Section for Clinical Biometrics, Medical University of Vienna, Vienna, Austria, **3** School of Business and Economics, Emmy Noether Group in Statistics and Data Science, Humboldt-Universität zu Berlin, Berlin, Germany, **4** Faculty of Medicine and Medical Center, Institute of Medical Biometry and Statistics, University of Freiburg, Freiburg, Germany, **5** Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

¶ Membership of the topic group 2 of the STRATOS initiative is listed in the Acknowledgments.

* christine.wallisch@meduniwien.ac.at (CW); geraldine.rauch@charite.de (GR)



OPEN ACCESS

Citation: Wallisch C, Bach P, Hafermann L, Klein N, Sauerbrei W, Steyerberg EW, et al. (2022) Review of guidance papers on regression modeling in statistical series of medical journals. PLoS ONE 17(1): e0262918. <https://doi.org/10.1371/journal.pone.0262918>

Editor: Tim Mathes, Witten/Herdecke University, GERMANY

Received: June 28, 2021

Accepted: January 8, 2022

Published: January 24, 2022

Copyright: © 2022 Wallisch et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The data was collected within the review and is available as [supporting information S6](#).

Funding: CW: I-4739-B Austrian Science Fund, <https://www.fwf.ac.at/en/>; LH: RA 2347/8-1, German Research Foundation, <https://www.dfg.de/en/>; WS: SA 580/10-1, German Research Foundation, <https://www.dfg.de/en/>. All funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Although regression models play a central role in the analysis of medical research projects, there still exist many misconceptions on various aspects of modeling leading to faulty analyses. Indeed, the rapidly developing statistical methodology and its recent advances in regression modeling do not seem to be adequately reflected in many medical publications. This problem of knowledge transfer from statistical research to application was identified by some medical journals, which have published series of statistical tutorials and (shorter) papers mainly addressing medical researchers. The aim of this review was to assess the current level of knowledge with regard to regression modeling contained in such statistical papers. We searched for target series by a request to international statistical experts. We identified 23 series including 57 topic-relevant articles. Within each article, two independent raters analyzed the content by investigating 44 predefined aspects on regression modeling. We assessed to what extent the aspects were explained and if examples, software advices, and recommendations for or against specific methods were given. Most series (21/23) included at least one article on multivariable regression. Logistic regression was the most frequently described regression type (19/23), followed by linear regression (18/23), Cox regression and survival models (12/23) and Poisson regression (3/23). Most general aspects on regression modeling, e.g. model assumptions, reporting and interpretation of regression results, were covered. We did not find many misconceptions or misleading recommendations, but we identified relevant gaps, in particular with respect to addressing nonlinear effects of continuous predictors, model specification and variable selection. Specific recommendations on software were rarely given. Statistical guidance should be developed for nonlinear effects, model specification and variable selection to better support medical researchers who perform or interpret regression analyses.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Rationale

Knowledge transfer from the rapidly growing body of methodological research in statistics to application in medical research does not always work as it should [1]. Possible reasons for this problem are the lack of guidance and that not all statistical analyses are conducted by statistical experts but often by medical researchers who may or may not have a solid statistical background. Applied researchers cannot be aware of all statistical pitfalls and the most recent developments in statistical methodology. Keeping up is already challenging for a professional biostatistical researcher, who is often restricted to an area of main interest. Moreover, articles on statistical methodology are often written in a rather technical style making knowledge transfer even more difficult. Therefore, there is a need for statistical guidance documents and tutorials written in more informal language, explaining difficult concepts intuitively and with illustrative educative examples. The international STRENGTHENING Analytical Thinking for Observational Studies (STRATOS) initiative (<http://stratos-initiative.org>) aims to provide accessible and accurate guidance documents for relevant topics in the design and analysis of observational studies [1]. Guidance is intended for applied statisticians and other medical researchers with varying levels of statistical education, experience and interest. Some medical journals are aware of this situation and regularly publish isolated statistical tutorials and shorter articles or even whole series of articles with the intention to provide some methodological guidance to their readership. Such articles and series can have a high visibility among medical researchers. Although some of the articles are short notes or rather introductory texts, we will use the phrase ‘statistical tutorial’ for all articles in our review.

Regression modeling plays a central role in the analysis of many medical studies, in particular, of observational studies. More specifically, regression model building involves aspects such as selection of a model type that matches the type of outcome variable, selection of explanatory variables to include in a model, choosing an adequate coding of the variables, deciding on how flexibly the association of continuous variables with the outcome should be modeled, planning and performing model diagnostics, model validation and model revision, reporting of a model and describing how well differences in the outcome can be explained by differences in the covariates. Some of the choices made during model building will strongly depend on the aim of modeling. Shmueli (2010) [2] distinguished between three conceptual modeling approaches: descriptive, predictive and explanatory modeling. In practice these aims are still often not well clarified, leading to confusion about which specific approach is useful in a modeling problem at hand. This confusion, and an ever-growing body of literature in regression modeling may explain why a common state-of-the-art is still difficult to define [3]. However, not all studies require an analysis with the most advanced techniques and there is the need for guidance for researchers without a strong background in statistical methodology, who might be “medical students or residents, or epidemiologists who completed only a few basic courses in applied statistics” according to the definition of level-1 researchers by the STRATOS initiative [1].

If suitable guidance for level-1 researchers in peer-reviewed journals was available, many misconceptions about regression model building could be avoided [4–6]. The researchers need to be informed about methods that are easily implemented, and they need to know about strengths and weaknesses of common approaches [3]. Suitable guidance should also point to possible pitfalls, elaborate on dos and don'ts in regression analyses, and provide software recommendations and understandable code for different methods and aspects. In this review, we focused on low-dimensional regression models where the sample size exceeds the number of candidate predictors. Moreover, we will not specifically address the field of causal inference, which goes beyond classical regression modeling.

So far, it is unclear what aspects of regression modeling have already been well-covered by related tutorials and where gaps still exist. Furthermore, suitable tutorial papers may be published but they are unknown to (nearly all) clinicians and therefore widely ignored in their analyses.

Objectives

The objective of this review was to provide an evidence-based information basis assessing the extent to which regression modeling has been covered by series of statistical tutorials published in medical journals. Specifically, we sought to define a catalogue of important aspects on regression modeling, to identify series of statistical tutorials in medical journals, and to evaluate which aspects were treated in the identified articles and at which level of sophistication. Thereby, we put an intended focus on the choice of the regression model type, on variable selection and for continuous variables on the functional form. Furthermore, this paper will provide an overview, which helps to inform a broad audience of medical researchers about the availability of suitable papers written in English.

The remainder of this review is organized as follows: In the next section, the review protocol is described. Subsequently, we summarize the results of the review by means of descriptive measures. Finally, we discuss implications of our results suggesting potential topics for future tutorials or entire series.

Material and methods

The protocol of this review describing the detailed design was already published by Bach et al. (2020) [7]. In here, we summarize its main characteristics.

Eligibility criteria

First, we identified series of statistical tutorials and papers published in medical journals with a target audience mainly or exclusively consisting of medical researchers or practitioners. Second, we searched for topic-relevant articles on regression modeling within these series. Journals with a target audience of pure theoretical, methodological or statistical focus were not considered. We included medical journals if they were available in English language since this implies high international impact and broad visibility. Moreover, the series had to comprise at least five or more articles including at least one topic-relevant article. We focused on statistical series only since we believed that entire series have higher impact and visibility than isolated articles.

Sources of information & search strategy

After conducting a pilot study for a systematic search for series of statistical tutorials, we had to adapt our search strategy since sensitive keywords to identify statistical series could not be found. Therefore, we consulted more than 20 members of the STRATOS initiative via email in spring 2018 for suggestions on statistical series addressing medical researchers. We also asked them to forward this request to colleagues, which resembles snowball sampling [8, 9]. This call was repeated at two international STRATOS meetings in summer 2018 and in 2019. The search was closed on June 30st, 2019. Our approach also included elements of respondent-driven sampling [10] by offering collaboration and co-authorship in case of relevant contribution to the review. In addition, we included several series that were additionally proposed by a reviewer during the peer-review process of this manuscript, and which were published by the end of June, 2019 to be consistent with the original request.

Data management & selection process

The list of all resulting statistical series suggested is available as [S1 File](#).

Two independent raters selected relevant statistical series from the pool of candidate series by applying the inclusion criteria outlined above.

An article within a series was considered to be topic-relevant if the title included one of the following keywords: *regression, linear, logistic, Cox, survival, Poisson, multivariable, multivariate*, or if the title suggested that the main topic of the article was *statistical regression modeling*. Both raters decided on the topic-relevance of an article independently and resolved discrepancies by discussion. To facilitate the selection of relevant statistical series, we designed a report form called *inclusion form* ([S2 File](#)).

Data collection process

After the identification of relevant series and topic-relevant articles, a content analysis was performed on all topic-relevant articles using an *article content form* ([S3 File](#)). The article content form was filled-in for every identified topic-relevant article by the two raters independently and again discrepancies were resolved by discussion. The results of completed article content forms were copied into a data base for further quantitative analysis.

Data items

In total 44 aspects of regression modeling were examined in the article content form ([S3 File](#)), which were related to four areas: *type of regression model, general aspects of regression modeling, functional form of continuous predictors, and selection of variables*. The 44 aspects cover topics of different complexity. Some aspects can be considered basic, others are more advanced. This was also commented in the [S3 File](#) for orientation. We mainly focused on predictive and descriptive models and did not consider particular aspects attributed to ethological models.

For each aspect, we evaluated whether it was mentioned at all, and if yes, the extent of explanation (short = one sentence only / medium = more than one sentence to one paragraph / long = more than one paragraph) [7]. We recorded whether examples and software commands were provided, and if recommendations or warnings were given with respect to each aspect. A box for comments provided space to note recommendations, warnings and other issues. In the article content form, it was also possible to add further aspects to each area. A manual for raters was created to support an objective evaluation of the aspects ([S4 File](#)).

Summary measures & synthesis of results

This review was designed as an explorative study and uses descriptive statistics to summarize results. We calculated absolute and relative frequencies to analyze the 44 statistical aspects. We used stacked bar charts to describe the ordinal variable *extent of explanation* for each aspect. To structure the analysis, we grouped the aspects into the afore mentioned areas: *type of regression model, general aspects of regression modeling, determination of functional form for continuous predictors and selection of variables*.

We conducted the above analyses both article-wise and series-wise. In the article-wise analysis, each article was considered individually. For the series-wise analysis, the results from all articles in a series were pooled and each series was considered as the unit of observation. This means, if an aspect was explained in at least one article, this also counted for the entire series.

Risk of bias

The risk of bias by missing a series was addressed extensively in the protocol of this study [7, 11, 12]. Moreover, bias could result from the inclusion criterion of series, which was the requirement of at least five articles in a series. This may have led to a less representative set of series. We set this inclusion criterion to identify highly visible series. Bias could also result from the specific choice of aspects of regression modeling to be screened. We tried to minimize this bias by the possibility for free text entries that could later be combined into additional aspects.

This review has been written according to the PRISMA reporting guideline [13, 14], compare [S1 Checklist](#). This review does not include patients or humans. The data that were collected within the review are available in [S1 Data](#).

Results

Selection of series and articles

The initial query revealed 47 series of statistical tutorials ([Fig 1](#) and [S1 File](#)). Out of these 47 series, two series were not published in a medical journal and five series did not target an

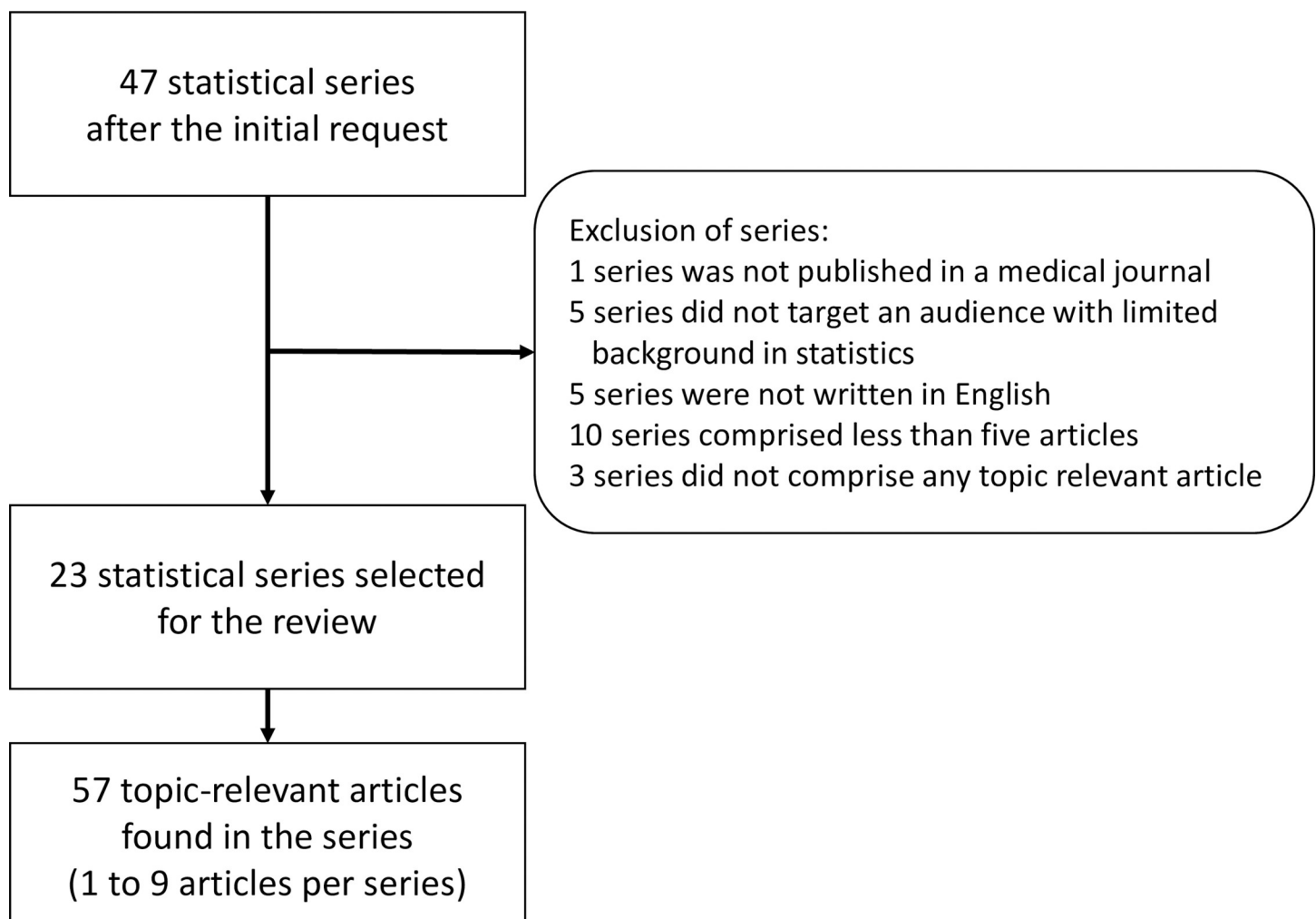


Fig 1. Flowchart of selection of statistical series and topic-relevant articles.

<https://doi.org/10.1371/journal.pone.0262918.g001>

audience with low statistical knowledge. Therefore, these series were excluded. Five and ten series were excluded because they were not written in English or they did not comprise at least five articles, respectively. Further, we excluded three series because they did not contain any topic-relevant article. The list of the series and the reason for each excluded series is found in [S1 File](#). Finally, we included 23 series with 57 topic-relevant articles.

Characteristics of the series

Each series contained between one to nine topic-relevant articles (two on average, [Table 1](#)). The variability of the average number of article pages per series illustrates that the extent of the articles was very different (1 to 10.3 pages). Whereas the series *Statistics Notes* in the *BMJ* typically used a single page to discuss a topic, hence pointing only to the most relevant issues, there were longer papers with a length of up to 16 pages [[15](#), [16](#)]. The series in the *BMJ* is also the one spanning over the longest time period (1994–2018). Beside of the series in the *BMJ*, only the *Archives of Disease in Childhood* and the *Nutrition* series started publishing papers already in the last century. [Fig 2](#) shows that most series were published only during a short period, perhaps paralleling terms of office of an Editor.

The most informative series with respect to our pre-specified list of aspects was published in *Revista Española de Cardiología*, which mentioned 35 aspects in two articles on regression modeling ([Table 1](#)). Similarly, *Circulation* and *Archives of Disease in Childhood* covered 31 and 30 aspects in three articles each. The number of articles and the years of publication varied across the series ([Fig 2](#)). Some series comprised only five articles whereas *Statistics Notes* of the *BMJ* published 68 short articles, which was very successful with some articles that were cited about 2000 times. Almost all series covered multivariable regression in at least one article. The range of regression types varied across series. Most statistical series were published with the intention to improve the knowledge of their readership about how to apply appropriate methodology in data analyses and how to critically appraise published research [[17–19](#)].

Characteristics of articles

The top three articles that covered the highest number of aspects (27 to 34 out of 44 aspects) on six to seven pages were published in *Revista Española de Cardiología*, *Deutsches Ärzteblatt International*, and in *European Journal of Cardio-Thoracic Surgery* [[20–22](#)]. The article of Nuñez et al. [[22](#)] published in *Revista Española de Cardiología* covered the most popular regression types (linear, logistic and Cox regression) and explained not only general aspect but also gave insights into non-linear modeling and variable selection. Schneider et al. [[20](#)] covered all regression types that we considered in our review in their publication in *Deutsches Ärzteblatt International*. The top-ranked article in *European Journal of Cardio-Thoracic Surgery* [[21](#)] particularly focused on the development and validation of prediction models.

Explanation of aspects in the series

Almost all statistical series included at least one article that mentioned or explained *multivariable regression* ([Table 1](#)). *Logistic regression* was the most frequently described regression type in 19 out of 23 series (83%), followed by *linear regression* (78%). *Cox regression/survival model* (including proportional hazards regression) was mentioned in twelve series (52%) and was less extensively described than linear and logistic regression. *Poisson regression* was covered by three series (13%). Each of the considered general aspects of regression modeling were mentioned in at least four series (17%) ([Fig 3](#)) except for *random effect models*, which were treated in only one series (4%). *Interpretation of regression coefficients*, *model assumptions*, and *different purposes of regression mode* were covered in 19 series (83%). The aspect *different purposes*

Table 1. Characteristics of included statistical series ranked by number of covered aspects. We considered 44 aspects, see [S3 File](#).

Rank	Journal	Statistical series	Years of publications	No. of articles	No. of topic-relevant articles	Average no. of pages in topic-relevant articles (range)	No. of aspects covered	Covered regression types	Covered multi-variable models	References
1	Revista Española de Cardiología	Contemporary methods in biostatistics	2011	6	2	7.0 (7)	35	Linear, logistic and Cox	Yes	[22, 23]
2	Circulation	Statistical primer for cardiovascular research	2006–2009	23	3	6.0 (6)	31	Linear, logistic and Cox	Yes	[24–26]
3	Archives of Disease in Childhood	Statistics from the inside	1991–1995	16	3	4.7 (4–5)	30	Linear, and logistic	Yes	[27–29]
4	Deutsches Ärzteblatt International	Series of evaluation of scientific publications	2009–2017	24	1	7.0 (7)	27	Linear, logistic, Cox, and Poisson	Yes	[20]
4	European Journal of Cardio-Thoracic Surgery	Statistical primer	2018	5	1	6.0 (6)	27	Linear, logistic and Cox	Yes	[21]
6	American Journal of Roentgenology	Fundamentals of clinical research for radiologists	2000–2005	22	3	10.3 (4–16)	25	Linear, logistic and Cox	Yes	[15, 16, 30]
6	European Heart Journal	Statistical tutorials	2011–2014	5	1	8.0 (8)	25	Logistic and Cox	Yes	[31]
8	Nephrology Dialysis Transplantation	Clinical research in kidney diseases	2017	9	2	8.5 (8–9)	23	Linear, logistic, Cox, and Poisson	Yes	[32, 33]
8	Nature Methods	Points of significance	2013–2019	42	9	2.0 (2)	23	Linear and logistic	Yes	[34–42]
8	Journal of Thoracic Disease	Statistics corner	2015–2018	33	5	4.8 (2–6)	23	Linear, logistic and Cox	Yes	[43–47]
11	Critical Care	Statistics review		14	3	7.3 (6–9)	23		Yes	[48–50]
12	Radiology	Statistical concepts series	2002–2004	17	2	6.0 (6)	22	Linear and logistic	Yes	[51, 52]
13	Journal of Clinical Psychopharmacology	Statistics commentary series	2014–2019	32	2	3.0 (3)	19	Linear	Yes	[53, 54]
14	Kidney International	abc of epidemiology	2007–2009	17	3	5.0 (5)	15	Linear, logistic and Cox	Yes	[55–57]
14	Advances in Physiology Education	Explorations in statistics	2010–2018	13	1	6.0 (6)	15	Linear	No	[58]
14	Journal of the American Medical Association	JAMA guide to statistics and methods	2014–2017	20	2	2.0 (2)	15	Logistic	Yes	[59, 60]
17	The Medical Journal of Australia	Accessible series on statistics for clinicians	2016–2018	10	1	3.0 (3)	13	Linear and logistic	Yes	[61]
17	Nephron Clinical Practice	Kidney disease and population health	2009–2011	17	1	6.0 (6)	13	Cox	Yes	[62]
17	American Journal of Ophthalmology	Series on statistics	2008–2010	14	3	2.0 (2)	13	Linear, logistic, Cox and Poisson	Yes	[63–65]
20	British Medical Journal	Statistics notes	1994–2018	68	4	1.0 (1)	12	Linear	Yes	[66–69]
21	Nephrology Dialysis Transplantation	Clinical epidemiology in nephrology	2010–2013	9	2	5.5 (5–6)	11	Logistic	Yes	[70, 71]

(Continued)

Table 1. (Continued)

Rank	Journal	Statistical series	Years of publications	No. of articles	No. of topic-relevant articles	Average no. of pages in topic-relevant articles (range)	No. of aspects covered	Covered regression types	Covered multi-variable models	References
22	Annals of Thoracic Surgery	The Statistician's page	2001–2016	24	1	1.0 (1)	8	Logistic	Yes	[72]
23	Nutrition	Random bytes	1996–1997	8	2	1.5 (1–2)	3	Linear and logistic	No	[73, 74]

<https://doi.org/10.1371/journal.pone.0262918.t001>

of regression models comprised at least one statement in an article concerning purposes of regression models, which could be identified by keywords like prediction, description, explanation, etiology, or confounding. More than one sentence was used for the explanation of different purposes in 15 series (65%). In 18 series (78%), reporting of regression results and regression diagnostics were described, which was done extensively in most series. Aspects like treatment of binary covariates, missing values, measurement error, and adjusted coefficient of determination were rather infrequently mentioned and found in four to seven series each (25–30%).

At least one aspect of functional forms of continuous predictors, was mentioned in 17 series (74%), but details were hardly ever given (Fig 4). The possibility of non-linear relation and non-linear transformations were raised in 16 (70%) and eleven series (48%), respectively. Dichotomization of continuous covariates was found in eight series (35%) and it was extensively discussed in two (9%). More advanced techniques like the use of splines or fractional polynomials were mentioned in some series but detailed information for splines was not provided. Generalized additive models were never mentioned.



Fig 2. Publication years and number of articles in statistical series from highest to lowest.

<https://doi.org/10.1371/journal.pone.0262918.g002>

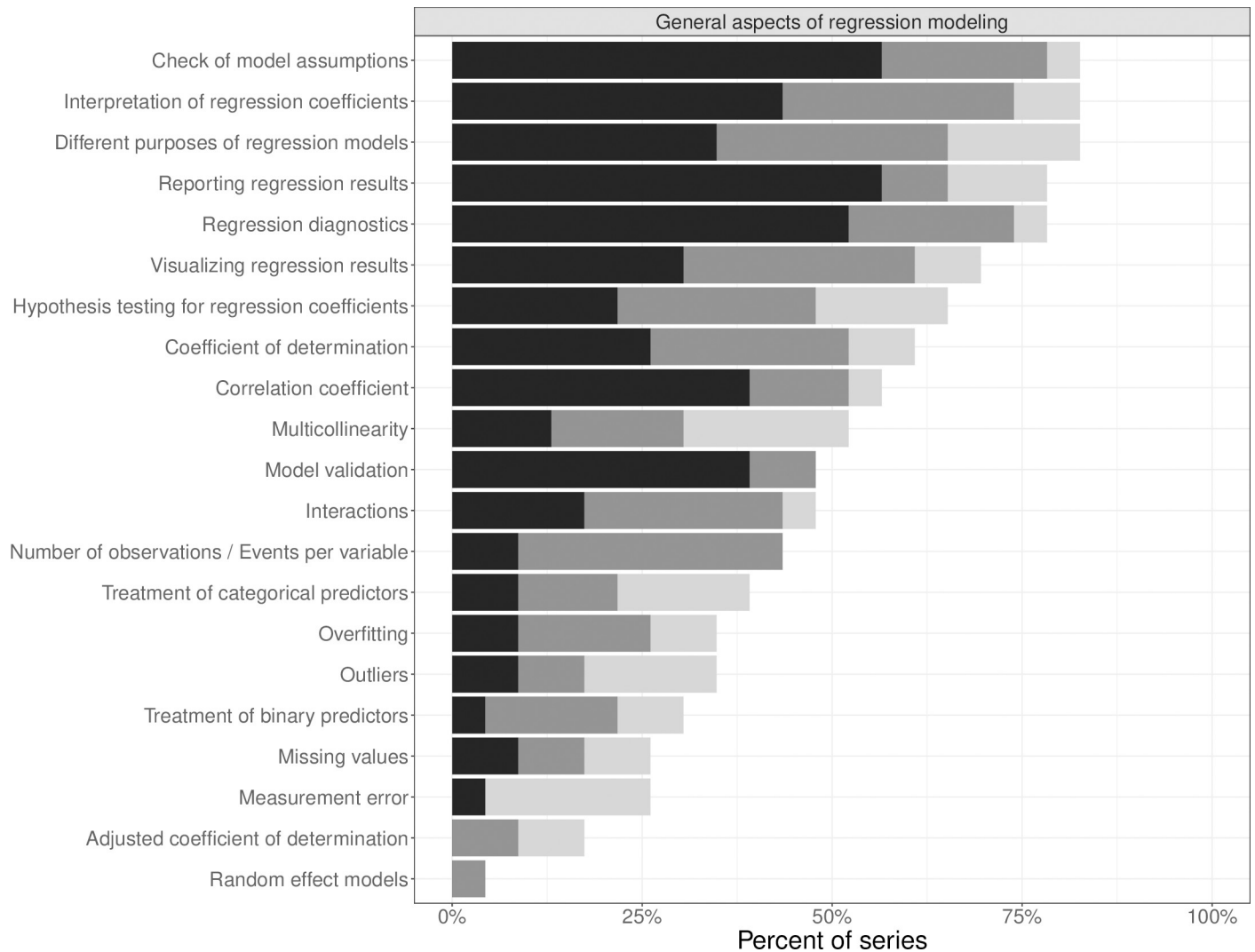


Fig 3. Extent of explanation of general aspects of regression modeling in statistical series: One sentence only (light grey), more than one sentence to one paragraph (grey) and more than one paragraph (black).

<https://doi.org/10.1371/journal.pone.0262918.g003>

Selection of variables was mentioned in 15 series (65%) and described extensively in ten series (43%) (Fig 5). However, specific variable selection methods were rarely described in detail. *Backward elimination*, *selection based on background knowledge*, *forward selection*, and *stepwise selection* were the most frequently described selection methods in seven to eleven series (30–48%). *Univariate screening*, which is still popular in medical research, was only described in three series (13%) in up to one paragraph. Other aspects of variable selection were hardly ever mentioned. *Selection based on AIC/BIC*, relating to best subset selection or stepwise selection based on these information criteria, and the *choice of the significance level* were found in 2 series only (9%). Relative frequencies of aspects mentioned in articles are detailed in Figs 1–3 in S5 File.

Software

We found general recommendations for software in nine articles of nine different series. Authors mentioned R, Nanostat, GLIM package, SAS and SPSS [75–78]. SAS as well as R were

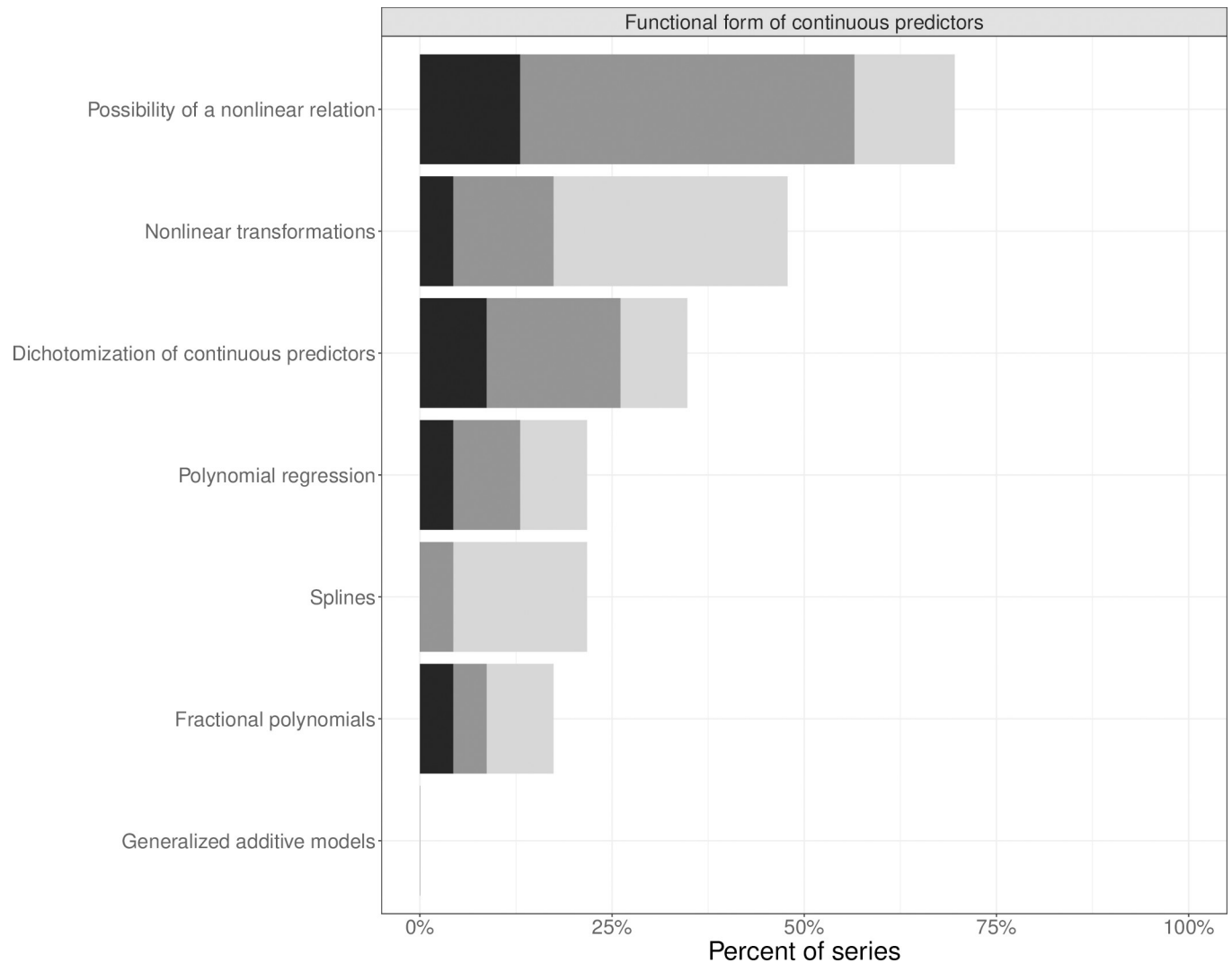


Fig 4. Extent of explanation of aspects of functional forms of continuous predictors in statistical series: One sentence only (light grey), more than one sentence to one paragraph (grey) and more than one paragraph (black).

<https://doi.org/10.1371/journal.pone.0262918.g004>

recommended in three articles. In only one article the authors referred to a specific package in R. Detailed code examples were provided in two articles only [16, 58]. In the article of Curran-Everett [58], the R script file was provided as appendix and in the article of Obuchowski [16], code chunks were included throughout the text directly showing how to derive the reported results. In all, software recommendations were rare and mostly not detailed.

Recommendations and warnings in the series

Recommendations and warnings were given on many aspects of our list. All statements are listed in [S5 File: Table 1](#) and some frequent statements across articles are summarized below.

Statements on general aspects

We found numerous recommendations and warnings on general aspects as described in the following. Concerning data preparation, some authors recommended to impute missing

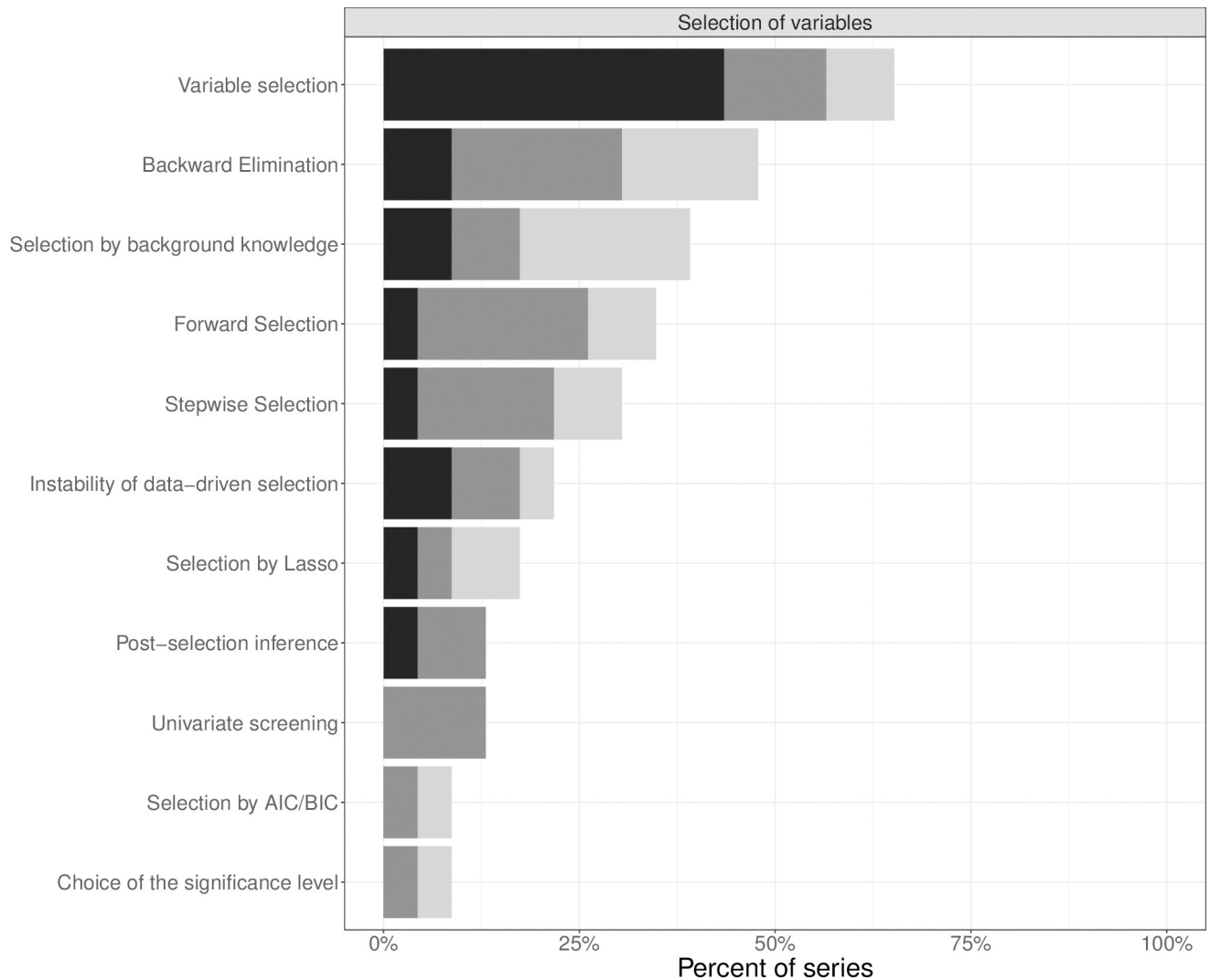


Fig 5. Extent of explanation of aspects of selection of variables in statistical series: One sentence only (light grey), more than one sentence to one paragraph (grey) and more than one paragraph (black).

<https://doi.org/10.1371/journal.pone.0262918.g005>

values in multivariable models, e.g. by multiple imputation [20–22, 31]. Steyerberg et al. [31] and Grant et al. [21] discouraged from using a complete case analysis to handle missing values. As an aspect of model development, *number of observations/events per variable* was a disputed topic in several articles [79–81]. In seven articles, we found explicit recommendations for the number of observations (in linear models) or the events per variable (in logistic and Cox/survival models), varying between at least ten to 20 observations/events per variable [16, 20, 22, 25, 31, 33, 55]. Several recommendations and warnings were given on *model assumptions* and *model diagnostics*. Many series authors recommended to check assumptions graphically [24, 27, 44, 58, 72] and they warned that models may be inappropriate if the assumptions are not met [20, 24, 31, 33, 52, 55, 56, 62]. In the context of Cox proportional hazards model, authors especially mentioned the proportional hazards assumption [24, 44, 49, 56, 62]. Concerning reporting of results, some authors warned to not confuse odds ratios with relative risks or hazard ratios [25, 44, 59]. Several warnings could also be found on reporting performance of a model. Most authors did not recommend to report the coefficient of determination R^2 [20, 27,

[51, 61] and indicated that the pitfall of R^2 is that its value increases with increasing number of covariates in the model [15]. Schneider et al. [20] and Richardson et al. [61] recommended to use the adjusted coefficient of determination instead. We also found many recommendations and statements about *model validation* for prediction models. Authors of the evaluated articles recommended cross-validation or bootstrap validation instead of split sample validation if internal validation is performed [21, 22, 31, 70, 72]. It was also suggested that internal validation is not sufficient for the model to be used in clinical practice and an external validation should be executed as well [21]. In several articles, we found that authors warned about applying the Hosmer-Lemeshow test because of potential pitfalls [31, 60, 61]. For *reporting regression results*, in two articles the guideline for Transparent Reporting of multivariable prediction models for Individual Prognosis or Diagnosis (TRIPOD) was mentioned [21, 71, 82].

Statements on functional form of continuous predictors

Dichotomization of continuous predictors is an aspect of *functional forms of continuous predictors* that was frequently discussed. Many authors argued against categorization of continuous variables because it may lead to loss of power, to increased risk of false positive results, to underestimation of variation, and to concealment of non-linearities [21, 26, 31, 69]. However, other authors advised to categorize continuous variables if the relation to the outcome is non-linear [24, 25, 59].

Statements on variable selection

We also found recommendations in favor of or against specific variable selection methods. Four articles explicitly recommended to take advantage of background knowledge to select variables [15, 20, 48, 59]. Univariate screening was advised against by one article [19]. Comparing stepwise selection methods, Grant et al. [21] preferred backward elimination over forward selection. Authors warned about consequences of stepwise methods such as unstable selection and overfitting [21, 31]. It was also pointed out that selected models must be interpreted with greatest caution and implications should be checked on new data [28, 53].

Methodological gaps in the series

This descriptive analysis of contents gives rise to some observations on important gaps and possibly misleading recommendations. First, we found that one general type of regression models, Poisson regression, was not treated in most series. This omission is probably due to the fact that Poisson regression is less frequently applied in medical research because most outcomes are binary or time-to-event and, therefore, logistic and Cox regression are more frequent. Second, several series introduced the possibility of non-linear relations of continuous covariates with the outcome. However, only few statements on how to deal with non-linearities by specifying flexible functional forms in multiple regression were available. Third, we did not find very detailed information on advantages and disadvantages of data-driven variable selection methods in any of the series. Finally, tutorials on statistical software and on specific code examples were hardly found in the reviewed series.

Misleading recommendations in the series

Quality assessment of recommendations would have been controversial and we did not intend doing it. Nevertheless, here we mention two issues that we consider as severely misleading. Although univariate screening as a method for variable selection was never recommended in any of the series, one article showed an example with the application of this procedure to pre-

filter the explanatory variables based on their associations with the outcome variable [47]. It is known since long that univariate screening should be avoided because it has the potential to wrongly reject important variables [83]. In another article it was suggested that a model can be considered robust if results from both backward elimination and forward selection agree [20]. Such agreement does not support robustness of stepwise methods: relying on agreement is a poor strategy [84, 85].

Series and articles recommended to read

Depending on the aim of the planned study, as well as the focus and knowledge level of the reader, different series and articles might be recommended. The series in *Circulation* comprised three papers about multiple linear and logistic regression [24–26], which provide basics and describe many essential aspects of univariable and multivariable regression modeling. For more advanced researchers, we recommend the article of Nuñez et al. in *Revista Española de Cardiología* [22], which gives a quick overview of aspects and existing methods including functional forms and variable selection. The *Nature Methods* series published short articles focusing on few, specific aspects of regression modeling [34–42]. This series might be of interest if one likes to spent more time on learning about regression modeling. If someone is especially interested in prediction models, we recommend a concise publication in the *European Heart Journal* [31], which provides details on model development and validation for predictive purposes. For the same topic we can also recommend the paper by Grant et al. [21]. We consider all series and articles recommended in this paragraph as suitable reading for medical researchers but this does not imply that we agree to all explanations, statements and aspects discussed.

Discussion

Summary and consequences for future work

This review summarizes the knowledge about regression modeling that is transferred through statistical tutorials published in medical journals. A total of 23 series with 57 topic-relevant articles were identified and evaluated for coverage of 44 aspects of regression modeling. We found that almost all aspects of regression modeling were at least mentioned in any of the series. Several aspects of regression modeling, in particular most general aspects, were covered. However, detailed descriptions and explanations of non-linear relations and variable selection in multivariable models were lacking. Only few papers provided suitable methods and software guidance for analysts with a relatively weak statistical background and limited practical experience as recommended by the STRATOS initiative [1]. However, we confess that currently there is no agreement on state of the art methodology [3].

Nevertheless, readers of statistical tutorials should not only be informed about the possibility of non-linear relations of continuous predictors with the outcome but they should also be given a brief overview about which methods are generally available and may be suitable. This could be achieved by tutorials that introduce readers to methods like fractional polynomials or splines, explaining similarities and differences between these approaches, e.g., by comparative, commented analyses of realistic data sets. Such documents could also show how alternative analyses (considering/ignoring potential non-linearities) may result in conflicting results and explain the reasons for such discrepancies.

Detailed tutorials on variable selection could aim at describing the mechanism of different variable selection methods, which can easily be applied with standard statistical software, and should state in what situations variable selection methods are needed and could be used. For example, if sufficient background knowledge is available, prefiltering or even the selection of variables should be based on this information rather than using data-driven methods on the

entire data set. Such tutorials should provide comparisons and interpretation of the results of various variable selection methods and suggest adequate methods for different data settings.

Generally, the articles also lacked details on software to perform statistical analysis and usually did not provide code chunks, descriptions of specific functions, an appendix with commented code or references to software packages. Future work should also focus on filling this gap by recommendations of software as well as providing well commented and documented code for different statistical methods in a format that is accessible by non-experts. We recommend that software, packages and functions therein to apply certain methods should be reported in every statistical tutorial article. The respective code to derive analysis results could be provided in an appendix or directly in the manuscript text, if not too lengthy. Any provided code in the appendix should be well-structured and lavishly commented referring to the particular method and describing all defined parameter settings. This will encourage medical researchers to increase the reproducibility of their research by also publishing their statistical code, e.g., in electronic appendices to their publications. For example, worked examples with openly accessible data sets and commented code allowing fully reproducible results have a high potential to guide researchers in their own statistical tasks. On the contrary, we discourage from using point-and-click software programs, which sometimes output far more analysis results than requested. Users may pick inadequate methods or report wrong results inadvertently, which could debilitate their research work.

Generally, our review may stimulate the development of targeted gap-filling guidance and tutorial papers in the field of regression modeling, which should support medical researchers in several ways: 1) by explaining how to interpret published results correctly, 2) by guiding them how to critically appraise the methodology used in a published article, 3) by enabling them to plan, perform basic statistical analyses and report results in a proper way and 4) by helping them to identify situations in which the advice of a statistical expert is required. In [S3 File](#): CRF article screening we commented which aspects should usually be addressed by an expert and which aspects are considered basic.

Strengths and limitations

According to our knowledge this is the first review on series of statistical tutorials in the medical field with the focus on regression modeling. Our review followed a pre-specified and published protocol to which many experienced researchers in the field of applied regression modeling contributed. One aspect of this contribution was the collection of series of statistical tutorials that could not be identified by common keyword searches.

We standardized the selection process by designing an inclusion checklist for series of statistical tutorials and by providing a manual for the content form with which we extracted the actual information of the article and series. Another strength is that the data collection process was performed objectively since each article was analyzed by two out of three independent raters. Discrepancies were discussed among all three of them to find a consent. This procedure avoided that single opinions were transferred to the output of this review. This review is informative for many clinical colleagues who are interested in statistical issues in regression modeling and search for suitable literature.

This review also has limitations. An automated, systematic search was not possible because series could not be identified by common keywords neither on the series' title level nor on the article's title level. Thus, not all available series may have been found. To enrich our initial query, we also searched on certain journals' webpages and requested our expert panel from the STRATOS initiative to complement our list with other series they were aware of. We also included series that were suggested by one reviewer during the peer-review procedure of this manuscript. This selection strategy may impose a bias towards higher-quality journals since

series of less prestigious journals might not be known to the experts. However, the higher-quality journals can be considered as the primary source of information for researchers seeking advice on statistical methodology.

We considered only series with at least five articles. This boundary is of course to a certain extent arbitrary. It was motivated by the fact that we intended to do analyses on the series level, which is only reasonable if a series covers an adequate number of articles. We also assumed that larger series are more visible and well-known to researchers.

We also might have missed or excluded some important aspects of regression modeling in our catalogue. The catalogue of aspects was developed and discussed by several experienced researchers of the STRATOS initiative working in the field of regression modeling. After submission of the protocol paper some more aspects were added on request of its reviewers [7]. However, further important aspects such as meta-regression, diagnostic models, causal inference, reproducibility or open data and open software code were not addressed. We encourage researchers to repeat similar reviews on these related fields.

A third limitation is that we only searched for series whereas there might be other educational papers on regression modeling that were published as single articles. However, we believe that the average visibility of an entire series and thereby its educational impact is much higher than for isolated articles. This does not negate that there could be excellent isolated articles, which can have a high impact for training medical researchers. While working on the final version of this paper we became aware of the series *Big-data Clinical Trial Column* in the *Annals of Translational Medicine*. Until 1 January 2019 they had published 36 papers and the series would have been eligible for our review. Obviously, we might have overseen further series, but it is unlikely that it has a larger effect on the results of our review.

Moreover, there are many introductory textbooks, educational workshops and online video tutorials, some of them with excellent quality, which were not considered here. A detailed review of such sources clearly was out of our scope.

Conclusion

Despite many series of statistical tutorials being available to guide medical researchers on various aspects of regression modeling, several methodological gaps still persist, specifically on addressing nonlinear effects, model specification and variable selection. Furthermore, papers are published in a large number of different journals and are therefore likely unknown to many medical researchers. This review fills the latter gap, but many more steps are needed to improve the quality and interpretation of medical research. More detailed statistical guidance and tutorials with a low technical level on regression modeling and other topics are needed to better support medical researchers who perform or interpret regression analyses.

Supporting information

S1 Checklist. PRISMA reporting guideline.

(PDF)

S1 File. List of candidate series for potential inclusion in the review.

(PDF)

S2 File. Case report form–series inclusion.

(PDF)

S3 File. Case report form–article screening.

(PDF)

S4 File. Manual for the article screening sheet.
(PDF)

S5 File. Supplementary figures and tables.
(PDF)

S1 Data. Collected data.
(XLSX)

Acknowledgments

When this article was written, topic group 2 of STRATOS consisted of the following members: Georg Heinze (Co-chair, georg.heinze@meduniwien.ac.at), Medical University of Vienna, Austria; Willi Sauerbrei (wfs@imbi.uni-freiburg.de), University of Freiburg, Germany; Aris Perperoglou (co-chair, aris.perperoglou@astrazeneca.com), AstraZeneca, London, Great Britain; Michal Abrahamowicz, Royal Victoria Hospital, Montreal, Canada; Heiko Becher, Medical University Center Hamburg, Eppendorf, Hamburg, Germany; Harald Binder, University of Freiburg, Germany; Daniela Dunkler, Medical University of Vienna, Austria; Rolf Groenwold, Leiden University, Leiden, Netherlands; Frank Harrell, Vanderbilt University School of Medicine, Nashville TN, USA; Nadja Klein, Humboldt Universität, Berlin, Germany; Geraldine Rauch, Charité–Universitätsmedizin Berlin, Germany; Patrick Royston, University College London, Great Britain; Matthias Schmid, University of Bonn, Germany.

We thank Edith Motschall (Freiburg) for her important support in the pilot study where we tried to define keywords for identifying statistical series within medical journals. We thank several members of the STRATOS initiative for proposing a high number of candidate series and we thank Frank Konietschke for English language editing in our protocol.

Author Contributions

Conceptualization: Willi Sauerbrei, Georg Heinze, Geraldine Rauch.

Data curation: Christine Wallisch, Paul Bach, Lorena Hafermann.

Formal analysis: Christine Wallisch, Paul Bach, Lorena Hafermann.

Funding acquisition: Willi Sauerbrei, Georg Heinze, Geraldine Rauch.

Investigation: Christine Wallisch, Paul Bach, Lorena Hafermann, Georg Heinze, Geraldine Rauch.

Methodology: Christine Wallisch, Paul Bach, Geraldine Rauch.

Project administration: Willi Sauerbrei, Geraldine Rauch.

Resources: Willi Sauerbrei, Geraldine Rauch.

Software: Christine Wallisch.

Supervision: Willi Sauerbrei, Georg Heinze, Geraldine Rauch.

Validation: Nadja Klein, Ewout W. Steyerberg, Georg Heinze, Geraldine Rauch.

Visualization: Christine Wallisch.

Writing – original draft: Christine Wallisch, Geraldine Rauch.

Writing – review & editing: Christine Wallisch, Paul Bach, Lorena Hafermann, Nadja Klein, Willi Sauerbrei, Ewout W. Steyerberg, Georg Heinze, Geraldine Rauch.

References

1. Sauerbrei W, Abrahamowicz M, Altman DG, le Cessie S, Carpenter J, initiative S. STRENGTHENING analytical thinking for observational studies: the STRATOS initiative. *Stat Med*. 2014; 33(30):5413–32. <https://doi.org/10.1002/sim.6265> PMID: 25074480
2. Shmueli G. To explain or to predict? *Stat Sci*. 2010; 25(3):289–310. <https://doi.org/10.1214/10-Sts330>
3. Sauerbrei W, Perperoglou A, Schmid M, Abrahamowicz M, Becher H, Binder H, et al. State of the art in selection of variables and functional forms in multivariable analysis-outstanding issues. *Diagn Progn Res*. 2020; 4:3. <https://doi.org/10.1186/s41512-020-00074-3> PMID: 32266321
4. Wynants L, van Smeden M, McLernon DJ, Timmerman D, Steyerberg EW, Van Calster B, et al. Three myths about risk thresholds for prediction models. *Bmc Med*. 2019; 17(1). <https://doi.org/10.1186/s12916-019-1425-3> PMID: 31651317
5. Heinze G, Dunkler D. Five myths about variable selection. *Transpl Int*. 2017; 30(1):6–10. <https://doi.org/10.1111/tri.12895> PMID: 27896874
6. van Smeden M, Lash TL, Groenwold RHH. Reflection on modern methods: five myths about measurement error in epidemiological research. *Int J Epidemiol*. 2020; 49(1):338–47. <https://doi.org/10.1093/ije/dyz251> PMID: 31821469
7. Bach P, Wallisch C, Klein N, Hafermann L, Sauerbrei W, Steyerberg EW, et al. Systematic review of education and practical guidance on regression modeling for medical researchers who lack a strong statistical background: Study protocol. *PLoS One*. 2020; 15(12):e0241427. <https://doi.org/10.1371/journal.pone.0241427> PMID: 33347441
8. Goodman LA. Snowball sampling. *Ann Math Stat*. 1961; 32(1):148–70. <https://doi.org/10.1214/aoms/1177705148>
9. Faugier J, Sargeant M. Sampling hard to reach populations. *J Adv Nurs*. 1997; 26(4):790–7. <https://doi.org/10.1046/j.1365-2648.1997.00371.x> PMID: 9354993
10. Heckathorn DD. Respondent-driven sampling: A new approach to the study of hidden populations. *Soc Probl*. 1997; 44(2):174–99. <https://doi.org/10.1525/sp.1997.44.2.03x0221m>
11. Sterne JAC, Savovic J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *Bmj-Brit Med J*. 2019; 366. <https://doi.org/10.1136/bmj.l4898> PMID: 31462531
12. Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. *Ann Intern Med*. 2019; 170(1):W1–W33. <https://doi.org/10.7326/M18-1377> PMID: 30596876
13. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev*. 2015; 4. <https://doi.org/10.1186/2046-4053-4-1> PMID: 25554246
14. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JPA, et al. The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and elaboration. *Ann Intern Med*. 2009; 151(4):W65–W94. <https://doi.org/10.7326/0003-4819-151-4-200908180-00136> PMID: 19622512
15. Dendukuri N, Reinhold C. Correlation and regression. *Am J Roentgenol*. 2005; 185(1):3–18. <https://doi.org/10.2214/ajr.185.1.01850003> PMID: 15972391
16. Obuchowski NA. Multivariate statistical methods. *Am J Roentgenol*. 2005; 185(2):299–309. <https://doi.org/10.2214/ajr.185.2.01850299> PMID: 16037496
17. Proto AV. Radiology 2002—Statistical concepts series. *Radiology*. 2002; 225(2):317. <https://doi.org/10.1148/radiol.2252020996> PMID: 12409559
18. Prel JD, Rohrig B, Blettner M. Statistical methods in medical research. *Dtsch Arztebl Int*. 2009; 106(7):99. <https://doi.org/10.3238/arztebl.2009.0099>
19. Attia JR, Jones MP. Introducing an accessible series on statistics for clinicians. *Med J Aust*. 2016; 205(9):392. <https://doi.org/10.5694/mja16.00981> PMID: 27809732
20. Schneider A, Hommel G, Blettner M. Linear regression analysis. Part 14 of a series on evaluation of scientific publications. *Dtsch Arztebl Int*. 2010; 107(44):776–82. <https://doi.org/10.3238/arztebl.2010.0776> PMID: 21116397
21. Grant SW, Collins GS, Nashef SAM. Statistical Primer: developing and validating a risk prediction model. *Eur J Cardio-Thorac*. 2018; 54(2):203–8. <https://doi.org/10.1093/ejcts/ezy180> PMID: 29741602
22. Nuñez E, Steyerberg EW, Nuñez J. Regression modeling strategies. *Rev Esp Cardiol*. 2011; 64(6):501–7. <https://doi.org/10.1016/j.recsep.2011.01.019> PMID: 21531065

23. Steyerberg EW, Van Calster B, Pencina MJ. Performance measures for prediction models and markers: evaluation of predictions and classifications. *Rev Esp Cardiol*. 2011; 64(9):788–94. <https://doi.org/10.1016/j.recesp.2011.04.017> PMID: 21763052
24. Slinker BK, Glantz SA. Multiple linear regression—Accounting for multiple simultaneous determinants of a continuous dependent variable. *Circulation*. 2008; 117(13):1732–7. <https://doi.org/10.1161/CIRCULATIONAHA.106.654376> PMID: 18378626
25. LaValley MP. Logistic regression. *Circulation*. 2008; 117(18):2395–9. <https://doi.org/10.1161/CIRCULATIONAHA.106.682658> PMID: 18458181
26. Crawford SL. Correlation and regression. *Circulation*. 2006; 114(19):2083–8. <https://doi.org/10.1161/CIRCULATIONAHA.105.586495> PMID: 17088476
27. Healy MJR. 15. Multiple regression. *Arch Dis Child*. 1995; 73(2):177–81. <https://doi.org/10.1136/adc.73.2.177> PMID: 7574870
28. Healy MJR. 16. Multiple regression (2). *Arch Dis Child*. 1995; 73(3):270–4. <https://doi.org/10.1136/adc.73.3.270> PMID: 7492177
29. Healy MJR. 7. Regression and correlation. *Arch Dis Child*. 1992; 67(10):1306–9. <https://doi.org/10.1136/adc.67.10.1306> PMID: 1444537
30. Stolberg HO, Norman G, Trop I. Survival analysis. *Am J Roentgenol*. 2005; 185(1):19–22. <https://doi.org/10.2214/ajr.185.1.01850019> PMID: 15972392
31. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J*. 2014; 35(29):1925. <https://doi.org/10.1093/eurheartj/ehu207> PMID: 24898551
32. Ravani P, Parfrey P, Murphy S, Gadag V, Barrett B. Clinical research of kidney diseases IV: standard regression models. *Nephrol Dial Transpl*. 2008; 23(2):475–82. <https://doi.org/10.1093/ndt/gfm880> PMID: 18182407
33. Ravani P, Parfrey P, Gadag V, Malberti F, Barrett B. Clinical research of kidney diseases III: Principles of regression and modelling. *Nephrol Dial Transpl*. 2007; 22(12):3422–30. <https://doi.org/10.1093/ndt/gfm777> PMID: 18029371
34. Altman N, Krzywinski M. Simple linear regression. *Nat Methods*. 2015; 12(11):999–1000. <https://doi.org/10.1038/nmeth.3627> PMID: 26824102
35. Altman N, Krzywinski M. Association, correlation and causation. *Nat Methods*. 2015; 12(10):899–900. <https://doi.org/10.1038/nmeth.3587> PMID: 26688882
36. Altman N, Krzywinski M. Analyzing outliers: influential or nuisance? *Nat Methods*. 2016; 13(4):281–2. <https://doi.org/10.1038/nmeth.3812> PMID: 27482566
37. Altman N, Krzywinski M. Regression diagnostics. *Nature Methods*. 2016; 13(5):385–6. <https://doi.org/10.1038/nmeth.3854>
38. Krzywinski M, Altman N. Multiple linear regression. *Nat Methods*. 2015; 12(12):1103–4. <https://doi.org/10.1038/nmeth.3665> PMID: 26962577
39. Krzywinski M, Altman N. Classification and regression trees. *Nature Methods*. 2017; 14(8):755–6.
40. Lever J, Krzywinski M, Altman N. Regularization. *Nature Methods*. 2016; 13(10):803–4. <https://doi.org/10.1038/nmeth.4014>
41. Lever J, Krzywinski M, Altman N. Model selection and overfitting. *Nature Methods*. 2016; 13(9):703–4. <https://doi.org/10.1038/nmeth.3968>
42. Lever J, Krzywinski M, Altman N. Logistic regression. *Nature Methods*. 2016; 13(7):541–2. <https://doi.org/10.1038/nmeth.3904>
43. Bertani A, Di Paola G, Russo E, Tuzzolino F. How to describe bivariate data. *J Thorac Dis*. 2018; 10(2):1133–7. <https://doi.org/10.21037/jtd.2018.01.134> PMID: 29607192
44. Brembilla A, Olland A, Puyraveau M, Massard G, Mauny F, Falcoz PE. Use of the Cox regression analysis in thoracic surgical research. *J Thorac Dis*. 2018; 10(6):3891–6. <https://doi.org/10.21037/jtd.2018.06.15> PMID: 30069391
45. Liu RZ, Zhao ZR, Ng CSH. Statistical modelling for thoracic surgery using a nomogram based on logistic regression. *J Thorac Dis*. 2016; 8(8):E731–E6. <https://doi.org/10.21037/jtd.2016.07.91> PMID: 27621910
46. Bertolaccini L, Pardolesi A, Davoli F, Solli P. Nanos gigantium humeris insidentes: the awarded Cox proportional hazards model. *J Thorac Dis*. 2016; 8(11):3464–5. <https://doi.org/10.21037/jtd.2016.11.84> PMID: 28066628
47. Mengual-Macencle N, Marcos PJ, Golpe R, Gonzalez-Rivas D. Multivariate analysis in thoracic research. *J Thorac Dis*. 2015; 7(3):E2–E6. <https://doi.org/10.3978/j.issn.2072-1439.2015.01.43> PMID: 25922743

48. Bewick V, Cheek L, Ball J. Statistics review 14: Logistic regression. *Crit Care*. 2005; 9(1):112–8. <https://doi.org/10.1186/cc3045> PMID: 15693993
49. Bewick V, Cheek L, Ball J. Statistics review 12: Survival analysis. *Crit Care*. 2004; 8(5):389–94. <https://doi.org/10.1186/cc2955> PMID: 15469602
50. Bewick V, Cheek L, Ball J. Statistics review 7: Correlation and regression. *Crit Care*. 2003; 7(6):451–9. <https://doi.org/10.1186/cc2401> PMID: 14624685
51. Zou KH, Tuncali K, Silverman SG. Correlation and simple linear regression. *Radiology*. 2003; 227(3):617–22. <https://doi.org/10.1148/radiol.2273011499> PMID: 12773666
52. Gareen IF, Gatsonis C. Primer on multiple regression models for diagnostic imaging research. *Radiology*. 2003; 229(2):305–10. <https://doi.org/10.1148/radiol.2292030324> PMID: 14595133
53. Streiner DL. Statistics Commentary Series: Commentary No. 32: Multiple Regression: What Can Possibly Go Wrong? *J Clin Psychopharmacol*. 2019; 39(3):200–2. <https://doi.org/10.1097/JCP.0000000000001040> PMID: 30921100
54. Streiner DL. Statistics Commentary Series: Commentary #16-Regression Toward the Mean. *J Clin Psychopharmacol*. 2016; 36(5):416–8. <https://doi.org/10.1097/JCP.0000000000000551> PMID: 27496345
55. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Linear and logistic regression analysis. *Kidney Int*. 2008; 73(7):806–10. <https://doi.org/10.1038/sj.ki.5002787> PMID: 18200004
56. van Dijk PC, Jager KJ, Zwinderman AH, Zoccali C, Dekker FW. The analysis of survival data in nephrology: basic concepts and methods of Cox regression. *Kidney Int*. 2008; 74(6):705–9. <https://doi.org/10.1038/ki.2008.294> PMID: 18596734
57. de Mutsert R, Jager KJ, Zoccali C, Dekker FW. The effect of joint exposures: examining the presence of interaction. *Kidney Int*. 2009; 75(7):677–81. <https://doi.org/10.1038/ki.2008.645> PMID: 19190674
58. Curran-Everett D. Explorations in statistics: regression. *Adv Physiol Educ*. 2011; 35(4):347–52. <https://doi.org/10.1152/advan.00051.2011> PMID: 22139769
59. Tolles J, Meurer WJ. Logistic regression relating patient characteristics to outcomes. *Jama-J Am Med Assoc*. 2016; 316(5):533–4. <https://doi.org/10.1001/jama.2016.7653> PMID: 27483067
60. Meurer WJ, Tolles J. Logistic regression diagnostics understanding how well a model predicts outcomes. *Jama-J Am Med Assoc*. 2017; 317(10):1068–9. <https://doi.org/10.1001/jama.2016.20441> PMID: 28291878
61. Richardson AM, Joshy G, D'Este CA. Understanding statistical principles in linear and logistic regression. *Med J Australia*. 2018; 208(8):332. <https://doi.org/10.5694/mja17.00222> PMID: 29716508
62. Stel VS, Dekker FW, Tripepi G, Zoccali C, Jager KJ. Survival analysis II: Cox regression. *Nephron Clin Pract*. 2011; 119(3):C255–C60. <https://doi.org/10.1159/000328916> PMID: 21921637
63. Boscardin WJ. The use and interpretation of linear regression analysis in ophthalmology research. *Am J Ophthalmol*. 2010; 150(1):1–2. <https://doi.org/10.1016/j.ajo.2010.02.022> PMID: 20609702
64. Hosmer DW Jr., Lemeshow S. Survival analysis: applications to ophthalmic research. *Am J Ophthalmol*. 2009; 147(6):957–8. <https://doi.org/10.1016/j.ajo.2008.07.040> PMID: 19463538
65. Lemeshow S, Hosmer DW Jr. Logistic regression analysis: applications to ophthalmic research. *Am J Ophthalmol*. 2009; 147(5):766–7. <https://doi.org/10.1016/j.ajo.2008.07.042> PMID: 19376329
66. Bland JM, Altman DG. Statistics notes 1. Correlation, regression, and repeated data. *Brit Med J*. 1994; 308(6933):896. <https://doi.org/10.1136/bmj.308.6933.896> PMID: 8173371
67. Bland JM, Altman DG. Statistics notes 2. Regression towards the mean. *Brit Med J*. 1994; 308(6942):1499. <https://doi.org/10.1136/bmj.308.6942.1499> PMID: 8019287
68. Bland JM, Altman DG. Statistics notes 7. Some examples of regression towards the mean. *Brit Med J*. 1994; 309(6957):780. <https://doi.org/10.1136/bmj.309.6957.780> PMID: 7950567
69. Altman DG, Royston P. Statistics notes. The cost of dichotomising continuous variables. *Brit Med J*. 2006; 332(7549):1080–. <https://doi.org/10.1136/bmj.332.7549.1080> PMID: 16675816
70. Tripepi G, Heinze G, Jager KJ, Stel VS, Dekker FW, Zoccali C. Risk prediction models. *Nephrol Dial Transpl*. 2013; 28(8):1975–80. <https://doi.org/10.1093/ndt/gft095> PMID: 23658248
71. van Diepen M, Ramspek CL, Jager KJ, Zoccali C, Dekker FW. Prediction versus aetiology: common pitfalls and how to avoid them. *Nephrol Dial Transpl*. 2017; 32:1–5. <https://doi.org/10.1093/ndt/gfw459> PMID: 28339854
72. Anderson WN. Statistical techniques for validating logistic regression models. *Ann Thorac Surg*. 2005; 80(4):1169. <https://doi.org/10.1016/j.athoracsur.2005.06.049> PMID: 16181834
73. Blackwell E, Pagano M. Survival analysis. *Nutrition*. 1996; 12(6):459–60. [https://doi.org/10.1016/s0899-9007\(96\)00183-9](https://doi.org/10.1016/s0899-9007(96)00183-9) PMID: 8875548

74. Pagano M. Logistic regression. *Nutrition*. 1996; 12(2):135. [https://doi.org/10.1016/s0899-9007\(97\)85056-4](https://doi.org/10.1016/s0899-9007(97)85056-4) PMID: 8724390
75. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2021.
76. Nelder JA. Glim (Generalized Linear Interactive Modeling Program). *Roy Stat Soc C-App*. 1975; 24(2):259–61.
77. SAS Institute Inc. The SAS system for Windows. Release 9.4. Cary, NC: SAS Institute Inc.; 2021.
78. IBM Corporation. IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corporation; 2020.
79. Wynants L, Bouwmeester W, Moons KGM, Moerbeek M, Timmerman D, Van Huffel S, et al. A simulation study of sample size demonstrated the importance of the number of events per variable to develop prediction models in clustered data. *J Clin Epidemiol*. 2015; 68(12):1406–14. <https://doi.org/10.1016/j.jclinepi.2015.02.002> PMID: 25817942
80. van Smeden M, de Groot JAH, Moons KGM, Collins GS, Altman DG, Eijkemans MJC, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *Bmc Med Res Methodol*. 2016; 16. <https://doi.org/10.1186/s12874-016-0267-3> PMID: 27881078
81. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. *Brit Med J*. 2020; 368. <https://doi.org/10.1136/bmj.m441> PMID: 32188600
82. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Brit Med J*. 2015; 350. <https://doi.org/10.1136/bmj.g7594> PMID: 25569120
83. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *J Clin Epidemiol*. 1996; 49(8):907–16. [https://doi.org/10.1016/0895-4356\(96\)00025-x](https://doi.org/10.1016/0895-4356(96)00025-x) PMID: 8699212
84. Austin PC, Tu JV. Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *J Clin Epidemiol*. 2004; 57(11):1138–46. <https://doi.org/10.1016/j.jclinepi.2004.04.003> PMID: 15567629
85. Wiegand RE. Performance of using multiple stepwise algorithms for variable selection. *Statistics in Medicine*. 2010; 29(15):1647–59. <https://doi.org/10.1002/sim.3943> PMID: 20552568