Is it a risk factor, a predictor, or even both? The multiple faces of multivariable regression analysis

Groenwold, R.H.H.; Dekkers, O.M.

European Journal of Endocrinology, 2023, **188**, E1–E4
https://doi.org/10.1093/ejendo/lvac012
Advance access publication 18 January 2023
Methodology Editorial

Is it a risk factor, a predictor, or even both? The multiple faces of multivariable regression analysis

Rolf H.H. Groenwold[1,2,*] and Olaf M. Dekkers[1,3]

[1]Department of Clinical Epidemiology, Leiden University Medical Center, Albinusdreef 2, 2333 ZA, Leiden, the Netherlands
[2]Department of Biomedical Data Sciences, Leiden University Medical Center, Albinusdreef 2, 2333 ZA, Leiden, the Netherlands
[3]Department of Endocrinology, Leiden University Medical Center, Albinusdreef 2, 2333 ZA, Leiden, the Netherlands

***Corresponding author:** Department of Clinical Epidemiology, Leiden University Medical Center, Albinusdreef 2, 2333 ZA, Leiden, the Netherlands.
Email: r.h.h.groenwold@lumc.nl

Abstract

The medical research literature is abundant with regression analyses that include multiple covariates, so-called multivariable regression models. Despite their common application, the interpretation of their results is not always clear or claimed interpretations are not justified. To outline the distinctions between different interpretations, we describe several possible research objectives for which a multivariable regression analysis might be an appropriate way of analyzing the data. In addition, we describe caveats in the interpretation of results of multivariable regression analysis.

Keywords: regression analysis, etiology, prediction, epidemiology, multivariable regression

Significance

Multivariable regression analysis is widely used in medical research, with different objectives and different interpretations of results. A continuum of research objectives is described to indicate differences in key features and considerations when applying multivariable regression analysis.

Introduction

Regression analysis is applied in many biomedical data analyses and is named according to the type of outcome (or dependent) variable that is modeled and the presumed link with the covariates (or independent variables). Commonly used regression analyses include linear regression, binary logistic regression, and survival analysis, which model a continuous, a binary, and a time-to-event outcome, respectively. In case multiple covariates are included in a regression analysis, this is referred to as multivariable regression (which is not to be confused with multivariate analysis where multiple dependent variables are analyzed).

The application of regression analysis can serve different purposes in biomedical research. For example, in a study investigating whether thyroid autoimmunity causes metabolic syndrome, regression analysis was used to adjust for the potential confounders' age, sex, income, education, smoking, alcohol consumption, walking activity, thyroid-stimulating hormone, and free thyroxine.[1] Regression analysis can also be used to develop a multivariable risk prediction model, such as in a study that developed a model to predict the risk of post-surgical recurrence of pheochromocytoma.[2]

In these examples, the application of regression analysis was similar, yet the objective of the study and the interpretation of the results differ considerably. Still, in publications, the distinction between an exploratory study of etiologic factors and a study that aims to identify prognostic factors is often not clear.[3] In this paper, we describe a continuum of research objectives, for which multivariable regression analysis is applicable. We describe their main features and points to consider when applying multivariable regression analysis.

A continuum of research objectives

In this section, we distinguish between five possible research objectives. A summary is provided in Table 1.

Confirmatory research of an etiologic factor

Consider a study of the cumulative effect of radiation on the risk of thyroid cancer. Exposure to radiation is the variable of interest. Because those with, eg, high radiation exposure may differ from those with low radiation levels (differences other than the radiation exposure, eg, different mean age or comorbidities), researchers may want to control for such differences to prevent confounding.[4] Correction for measured confounding could be achieved by including potential confounding variables (confounders) as covariates in a

Received: October 15, 2022. **Revised:** November 30, 2022. **Editorial Decision:** December 2, 2022. **Accepted:** December 2, 2022

Table 1. Different applications of multivariable regression analysis in epidemiologic research and main considerations

Research objective	Explanation	Important considerations
Confirmatory etiologic research	The multivariable regression model includes the etiologic factor of interest and also potential confounders.	The relation between the exposure of interest and the outcome should be modelled correctly. Also, no confounding variables are omitted from the model, while their relations with the outcome are modelled correctly too. Regression coefficients of the confounding variables should not be given a causal interpretation.
Exploratory etiologic research	The multivariable regression model includes multiple etiologic factors as well as a general set of potential confounders.	In addition to the considerations for confirmatory etiologic research, the exploratory nature of the study should be reflected in the inferences made. With an increasing number of risk factors, the risk of a false-positive finding increases. Correction for multiple testing could be considered. P-values are not a measure to find the strongest risk factor.
Predictor finding study	The multivariable regression model includes multiple variables each of which might be predictive of the outcome.	The relations between each of the predictors and the outcome are modelled correctly, otherwise the predictive value of a variable may be missed. With an increasing number of possible predictors, the risk of a false-positive finding increases. P-value of the predictors do not capture the predictive performance of the entire model.
Multivariable risk prediction modelling	The multivariable regression model includes multiple possible predictors.	The relations between each of the predictors and the outcome are modelled correctly, otherwise the predictive value of the model is suboptimal. Assessment of the quality of the model is based on the predictive performance of the entire model, i.e., the combination of predictors, not on individual predictors. Variable selection could affect model performance at external validation (overfitting).
Research of added predictive value	The added value of one or more predictors is assessed in addition to a (existing) multivariable risk prediction model.	Assessment of the added value of the predictor(s) is based on the predictive performance of the extended model compared to the original model and not on, e.g., the p-value of the regression coefficient(s) of the additional predictor(s).

multivariable regression model. Here, the interest is specifically in the etiologic factor "radiation exposure", while the other covariates are included to improve the validity of the estimated regression coefficient that quantifies the effect of radiation on thyroid cancer risk. The study's conclusion may be that radiation increases the risk for thyroid cancer, independent of the other variables in the model; the effect estimates of the confounding variables are of no interest (see below).

Exploratory research of multiple etiologic factors

Particularly with recently discovered diseases or diseases with a renewed research scope, there may be a general search for (modifiable) risk factors. In such exploratory activities, there are generally multiple potential risk factors of interest, without a clear hierarchy. Although each of these risk factors may have its own set of potential confounders, this may be unclear due to the exploratory state of the research field. Different variables could be included in one multivariable regression model, with a general set of potential confounders (eg, age, sex, etc.). Such studies are not uncommon, but the interpretation and credibility of their results differ considerably from those of the confirmatory studies mentioned above and independent confirmation is needed.[5] Another challenge of exploratory research of multiple etiologic factors is that the risk of a false-positive finding increases with the number of etiologic factors studied ('multiple testing').[6]

Predictor-finding study

In research aiming to discover possible predictors of a certain outcome, the ultimate goal is not to find causal risk factors but to identify predictor variables (or prognostic factors) that are associated with the outcome. Subsequently, such predictor variables could be incorporated when developing a multivariable risk prediction model (see below). In predictor-finding studies, multiple potential predictor variables are included in a multivariable model, without a hierarchy of those variables. Again, given the exploratory nature of this kind of study, (external) validation is needed, for example through the development of a multivariable risk prediction model using a new (independent) data set.

Development of a multivariable risk prediction model

Suppose researchers want to develop a tool that can support a physician in quantifying the lifetime risk of developing thyroid cancer. Possible predictors include age, sex, cumulative radiation exposure, and a history of thyroid disease. By means of multivariable regression analysis, the combined information of those predictor variables is linked with the outcome (lifetime risk). Each variable contributes to the model and the quality of the model is judged on the predictive performance of the full model, not on its components. A central aim of prediction modeling is finding the appropriate set of predictor variables that—together—accurately predict the outcome.[7] Because no causal inferences are made, considerations regarding confounding are irrelevant in the context of risk prediction modeling. We note that the development of a multivariable prediction model does not end with running a regression analysis. For example, external validation and possibly updating of the model may be needed to ensure good performance in future use.[8]

Research into the added predictive value of a variable

In the case of an existing risk prediction model, it might be of interest to investigate improvements of that model. Consider the above-mentioned model of lifetime thyroid cancer risk. A researcher may think the predictive performance of that model could be improved by including "low iodine diet" as another predictor in the model. In this case, the multivariable regression model could be extended to include "low iodine diet" in addition to the four predictors already included in the model. The assessment of the added value of "low iodine diet" should be based on the improvement of the predictive performance of the extended model over the existing model, not on the regression coefficient of "low iodine diet" (or its *P*-value). Different measures are available to quantify the added value of an additional predictor.[9]

Warnings on the interpretation of multivariable regression analysis

There are also issues that require special attention, depending on the particular context and application. Here we describe some that we consider relevant.

Table 2 fallacy

As outlined above, when the aim is to confirm or discover the causal effects of exposures, the purpose of inclusion of covariates in a multivariable regression model is confounding control. Ideally, the set of covariates includes a sufficient number of variables such that no unmeasured confounding remains. The output of the multivariable regression analysis not only provides the estimated regression coefficient of interest, but also the regression coefficients of the confounders. However, interpreting the latter coefficients as estimates of causal effects is a fallacy, because each of those confounding variables may have its own set of confounders, which may not be included in the model.[10] It underlines that there is no common—or universal—set of confounders that suffices for each analysis; in fact, which variables should be considered as confounders depends on the research question and study design.

Variable selection

As the name indicates, multivariable regression models include multiple variables. Which variables to include can be decided by the researcher, but also data-driven selection procedures can be applied such that out of a set of variables, particular variables are selected, or instead omitted from the model (ie, forward and backward selection). Reasons and consequences differ per research objective. A general recommendation is not to select variables based on univariable ('unadjusted') analysis.[11,12]

In etiologic factor research, a set of confounders is included in a regression model to control for confounding. Often, the set of confounders is decided prior to data analysis. Backward elimination of covariates from the model could then be considered to increase the precision of the regression coefficient of the etiologic factor of interest. Particularly omitting variables that have no (or a very weak) relation with the outcome may improve precision. At the same time, omitting potential confounders could introduce a bias due to unmeasured confounding. The benefit and potential risk of variable selection should be balanced.[13]

An important reason for variable selection in risk prediction modeling studies is to select a model that is easy to use in clinical practice: as few variables as possible and preferably those variables that are easy to measure. Selecting variables that predict the outcome, while omitting variables that do not, will obviously improve the predictive performance of the model. Nevertheless, variable selection in prediction modeling may result in a model that is perfectly tuned for the data used to develop the model, but it may not perform well in new data (or future users), a phenomenon called overfitting.[7] Methods to limit the impact of overfitting include shrinkage and penalized regression.[7]

Concluding remarks

We provided an overview of possible applications of multivariable regression analysis in medical research and discussed several points of attention. By no means do we claim that our overview is exhaustive. There may be situations in which the broad classification of research aims does not apply, yet multivariable regression analysis might still be a preferred option. Conversely, there may be research situations that correspond well with one of the situations we described, yet an alternative data analytical method, eg, classification trees, is used. Nevertheless, we hope that for the many situations in which one of the research options does apply, the differentiation between the categories is helpful to researchers, reviewers, and readers.

Funding

None declared.

Conflicts of interest: R.H.H.G. reports no conflicts of interest. O.M.D. is a Deputy Editor for European Journal of Endocrinology. He was not involved in the review or editorial process for this paper, on which he is listed as an author.

References

1. Kim HJ, Park SJ, Park HK, Byun DW, Suh K, Yoo MH. Thyroid autoimmunity and metabolic syndrome: a nationwide population-based study. *Eur J Endocrinol*. 2021;185(5):707-715. https://doi.org/10.1530/EJE-21-0634
2. Parasiliti-Caprino M, Bioletto F, Lopez C, *et al*. Development and internal validation of a predictive model for the estimation of pheochromocytoma recurrence risk after radical surgery. *Eur J Endocrinol*. 2022;186(3):399-406. https://doi.org/10.1530/EJE-21-0370
3. Ramspek CL, Steyerberg EW, Riley RD, *et al*. Prediction or causality? A scoping review of their conflation within current observational research. *Eur J Epidemiol*. 2021;36(9):889-898. https://doi.org/10.1007/s10654-021-00794-w
4. Groenwold RHH, Dekkers OM. METHODOLOGY FOR THE ENDOCRINOLOGIST: basic aspects of confounding adjustment. *Eur J Endocrinol*. 2020;182(5):E5-E7. https://doi.org/10.1530/EJE-20-0075
5. Luijken K, Dekkers OM, Rosendaal FR, Groenwold RHH. Exploratory analyses in aetiologic research and considerations for assessment of credibility: mini-review of literature. *BMJ*. 2022;377:e070113. https://doi.org/10.1136/bmj-2021-070113
6. Groenwold RHH, Goeman JJ, Cessie SL, Dekkers OM. Multiple testing: when is many too much? *Eur J Endocrinol*. 2021;184(3):E11-E14. https://doi.org/10.1530/EJE-20-1375
7. Steyerberg EW. *Clinical Prediction Models*. Springer; 2009.
8. Steyerberg EW, Moons KG, van der Windt DA, *et al*. Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med*. 2013;10(2):e1001381. https://doi.org/10.1371/journal.pmed.1001381
9. Cook NR. Quantifying the added value of new biomarkers: how and how not. *Diagn Progn Res*. 2018;2(1):14. https://doi.org/10.1186/s41512-018-0037-2

10. Groenwold RH, Klungel OH, Grobbee DE, Hoes AW. Selection of confounding variables should not be based on observed associations with exposure. *Eur J Epidemiol*. 2011;26(8):589-593. https://doi.org/10.1007/s10654-011-9606-1
11. Heinze G, Wallisch C, Dunkler D. Variable selection—a review and recommendations for the practicing statistician. *Biom J*. 2018;60(3):431-449. https://doi.org/10.1002/bimj.201700067
12. Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *Am J Epidemiol*. 2013;177(4):292-298. https://doi.org/10.1093/aje/kws412
13. Greenland S, Daniel R, Pearce N. Outcome modelling strategies in epidemiology: traditional methods and basic alternatives. *Int J Epidemiol*. 2016;45(2):565-575. https://doi.org/10.1093/ije/dyw040