Multivariable Analysis

Marlies Wakkee[1], Loes M. Hollestein[1] and Tamar Nijsten[1]

Journal of Investigative Dermatology (2014) **134**, e20. doi:10.1038/jid.2014.132

In all observational research, one will sooner or later be confronted with the question of whether a certain exposure is related to an outcome. For example, is the risk of cutaneous melanoma affected by the use of nonsteroidal anti-inflammatory drugs (NSAIDs) or is psoriasis an independent predictor for the occurrence of cardiovascular diseases? Questions like these can be answered using multivariable regression analysis. This technique can be used in observational research to adjust for confounders, to assess the effect size of risk factors, or to develop prediction models.

Researchers with limited epidemiological background may feel uncertain how to appraise studies using multivariable regression analysis, let alone use multivariable regression analysis themselves. The objective of this article is therefore to provide a practical overview of the basic principles of multivariable analysis, illustrated with various examples.

WHAT IS A MULTIVARIABLE ANALYSIS?

Multivariable analysis is a statistical technique that can be used to simultaneously explore whether multiple risk factors (referred to as independent variables) are related to a certain outcome (referred to as dependent variable). The type of regression model that is selected depends mainly on the outcome variable and the role of time in the available data (Table 1). In this article we will describe the three most frequently used types of regression analysis: linear regression, logistic regression, and Cox proportional hazards regression analysis, which are generally sufficient to answer most research questions.

Multivariable linear and logistic regression

In cross-sectional and case–control studies, you can use either linear or logistic regression to analyze the data. If the outcome is continuous (e.g., weight), linear regression can be applied and relationships will be represented by β-coefficients. For dichotomous outcomes, such as the presence or absence of a disease, logistic regression is used to calculate odds ratios (ORs). Continuous variables may also be transformed into a dichotomous variable, such as weight into the absence or presence of obesity. Transforming data by grouping results can lead to a loss of information and precision, but the resulting risk estimates may be easier to interpret. If cases and controls are matched for certain risk factors (e.g., age), a conditional logistic regression model should be used. Based on their matched criteria, the cases and controls are linked to form a set to which the multivariable analysis used to adjust for other confounders can be applied. Discarding the matched design by adjusting for matched factors in an unconditional analysis leads to a bias toward the null. Interactions with the matching variables can still be taken into account.

In all models it is assumed that the independent variables have a linear relationship with the continuous outcome (linear regression) or logarithm of odds of the dichotomous outcome (logistic regression). In the case of nonlinear relationships, the independent variables can be forced into a normal distribution by taking the natural logarithm or by creating multiple dichotomous variables.

Multivariable Cox proportional hazards analysis

In cohort studies, where the exposure precedes the outcome, data can be analyzed using multivariable Cox proportional

WHAT MULTIVARIABLE REGRESSION ANALYSIS DOES

- Aims to explore how multiple risk factors are independently related to an outcome.
- Applies to various models depending on the distribution and temporal relationship of the outcome.
- Allows adjustment for known and available confounders.
- Enables accounting for statistical interaction between independent variables.

LIMITATIONS

- Multivariable regression analysis provides information on potential associations, but a significant association does not automatically imply causality.
- It only adjusts for measured confounding.

[1]*Department of Dermatology, Erasmus University Medical Center, Rotterdam, The Netherlands*

Correspondence: Marlies Wakkee, Postbus 2040, 3000 CA; Rotterdam, The Netherlands. E-mail: m.wakkee@erasmusmc.nl

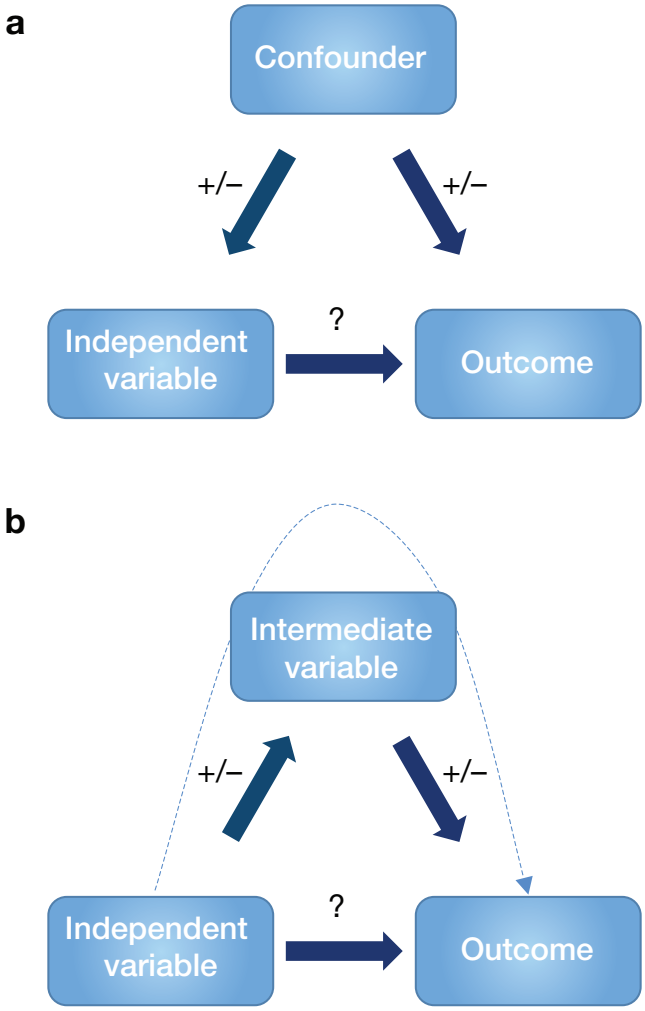

Figure 1. The role of a confounder versus an intermediate variable in the relationship between an independent variable and an outcome. (**a**) The association between an independent variable and an outcome may be confounded. That is, the confounder predicts the independent variable, predicts the outcome, and is not part of the causal pathway, leading to a triangular relationship. Thereby, the association between the independent variable and the outcome is (partly) explained by the confounder. (**b**) The role of an intermediate variable in the relation between an independent variable and an outcome. The intermediate variable is part of the causal pathway in the relationship between the independent variable and the outcome.

hazards regression analysis, also known as survival analysis. In Cox proportional hazards models the effect of independent variables on survival time is assessed and represented by hazard ratios (HRs). The advantage of Cox proportional hazards analysis is that it includes all observation-years available for each participant until the studied outcome (e.g., a cardiovascular event) or death, or otherwise to the end of the follow-up time, whichever comes first. It also takes into account the exposure time to a risk factor (e.g., days of sun exposure), which may be shorter than the total included follow-up time. The prospective means of data collection in cohort studies result in more precise data with a temporal component. This is in contrast to linear or logistic regression, where subjects are compared at one point in time or over a comparable timeframe (e.g., a two-year period), which implies retrospective data collection or excluding subjects without sufficient follow-up time and therefore losing valuable information. Finally, with Cox proportional hazards analysis it is assumed that all independent variables change linearly with the logarithm of the hazard.

Risk estimates

The β-coefficient obtained from linear regression is directly interpretable as the slope, which denotes the change in dependent variable per unit change in independent variable (Table 1). If the 95% confidence interval (CI) includes the value 0, this represents a nonsignificant association, because a slope of 0 means there is no association (or a non-linear relationship). In the case of logistic or Cox proportional hazards analysis, the ORs or HRs are an exponentiation of this β-coefficient, which results in an outcome that cannot extend below 0 but ranges from 0 to infinity. For ORs or HRs, the 95% CI must exclude the value 1 to demonstrate a significant association (Table 1) because a ratio of 1 means that the odds or hazards are the same for the two groups you are comparing. Reporting 95% CIs is preferred over reporting *P* values because reporting 95% CIs has the advantage of directly including both an effect size (point estimate) as well as the range of values in which the true value lies (width of 95% CI), rather than stating only whether a statistically significant difference is observed.

Examples of multivariable analysis

Multivariable conditional logistic regression was applied in an age- and gender-matched case–control study investigating the association between cutaneous melanoma (CM) as a dichotomous dependent variable and exposure to NSAIDs as an independent variable (Curiel-Lewandrowski *et al.*, 2011). Matching for age and gender was done by including at least one community-based control subject from the same 5-year age group and gender for every subject with CM in this study. The odds of CM were significantly lower among those using aspirin, with a crude OR of 0.75 and a 95% CI of 0.57–0.97, which actually represents an adjusted OR because it already includes adjustment for age and gender by matching the case and control subjects. This effect of aspirin was even stronger after adjusting for the number of sunburns during childhood in the multivariable model, resulting in an adjusted OR of 0.72 (95% CI 0.55–0.94) (Curiel-Lewandrowski *et al.*, 2011).

In a population-based cohort study assessing the association between psoriasis and cardiovascular events, multivariable Cox proportional hazards analysis was applied to adjust for other cardiovascular risk factors, including the significantly younger age of the psoriasis cohort (Dowlatshahi *et al.*, 2013). The crude HR showed a borderline significantly decreased cardiovascular risk with a HR of 0.69 and a 95% CI 0.48–1.00 for psoriasis patients compared to reference subjects. In this case the HR is called "borderline significant" because the upper limit of the CI includes 1; this is also reflected in the *P* value of 0.05. After adjusting for the total cardiovascular risk profile, the HR was 0.73 with a 95% CI between 0.50 and 1.06, which is not significant because the value 1 lies

Table 1. Multivariable regression techniques

Multivariable technique	Dependent variable (outcome)	Example of dependent variable	Reported measure of association	Statistically significant CI	No. of independent variables (predictors) allowed for consideration	Example of interpretation of associations
Linear regression	Continuous	Intima media thickness	β = change in the dependent variable	If the CI does not include 0	No. of subjects/10	Psoriasis: β = change in IMT of psoriasis patients compared with controls of the same age and sex
Logistic regression	Dichotomous	Occurrence of a cutaneous melanoma	Odds ratio (OR)	If the CI does not include 1.0	No. of events/10	NSAID use: OR of melanoma for NSAID users compared to no NSAID users adjusted for age, gender, and sunburn
Cox proportional hazards regression	Time until event	Time to occurrence of a CVD	Hazard ratio (HR)	If the CI does not include 1.0	No. of events/10	Psoriasis: HR of a CVD for psoriasis patients compared to patients without psoriasis adjusted for other CV risk factors

CI, confidence interval; CVD, cardiovascular disease; IMT, intima media thickness; NSAID, nonsteroidal anti-inflammatory drug.

within the 95% CI; this is also reflected in the *P* value of 0.10. Further examples of multivariable linear regression, such as the association between psoriasis and measures of subclinical atherosclerosis (for example, carotid intima media thickness), can also be found in this study (Dowlatshahi *et al.*, 2013).

CONFOUNDING AND INTERACTION

Confounding

A confounder is a variable that explains (part of) the association between exposure and outcome (Rothman and Greenland, 1998), but a confounder cannot be part of the causal pathway (Figure 1a). As depicted in Figure 1a, the confounder has a triangular relationship with both the exposure and the outcome. Confounders should not be confused with intermediate variables, which are part of the causal pathway (Figure 1b), because adjusting for an intermediate variable would attenuate the measure of association. To test whether a variable is a confounder, the data can be stratified for a suspected confounder to examine stratum-specific risk estimates. If data on confounders are available, they can be corrected for using multivariable regression models. Usually in clinical research not all confounders are available or known; this is referred to as unmeasured or residual confounding.

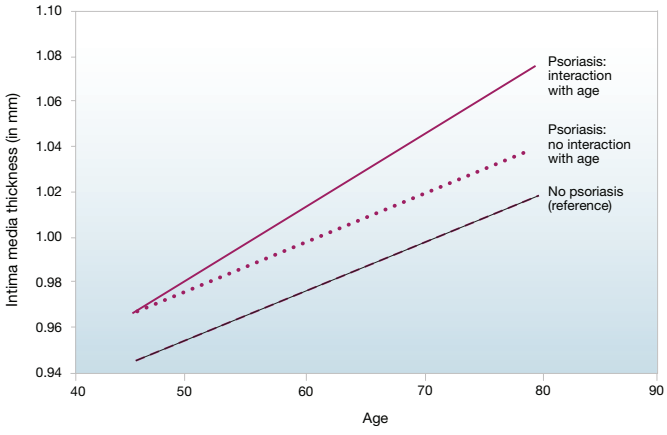

Figure 2. An example of interaction in linear regression between psoriasis and age for the difference in carotid intima media thickness. If no interaction is present, the difference in intima media thickness will be equal for all ages. In the case of statistical interaction, the difference in intima media thickness depends on age.

Interaction

Statistical interaction means that the effect of an independent variable is affected by a second independent variable in the multivariable model. This implies that the effect of two independent variables on the outcome is different than would be expected based on the separate effect of the independent variables. Statistical interaction may indicate the presence of a biological interaction, which is the effect of two or more factors in a causal mechanism to develop a disease.

There are two types of interaction. Additive interaction in linear regression means that the effect of two interacting variables is stronger or weaker than would be expected by adding the effect of two variables. If there is statistical interaction in logistic or Cox proportional hazards regression models, this is multiplicative interaction, meaning there is a multiplicative effect between two independent variables on the OR or HR.

An example of interaction in linear regression is shown in Figure 2, which is based on the study by Dowlatshahi *et al.* (2013), where no significant difference in carotid intima media thickness was found in psoriasis patients compared to healthy controls. In this hypothetical figure the potential effect of interaction between psoriasis and age on the carotid intima media thickness is demonstrated. If no interaction is present, the difference in intima media thickness will be equal for all ages. In the case of statistical interaction, the difference in intima media thickness is dependent on age.

Table 2. Hill's criteria of causation (Hill, 1965), which can be used as supporting evidence to demonstrate a causal relationship in observational research

	Criteria	Explanation
1	Strength of association	A larger association (e.g., larger odds ratio) increases the likelihood of a causal relationship.
2	Temporality	The exposure precedes the effect.
3	Consistency	The same association is replicated in different study settings and environments.
4	Biological plausibility	The presence of a rational and theoretical basis.
5	Dose–response relationship	Greater amount of exposure results in an increased risk.
6	Experimental evidence	The presence of experiments to make a causal interference more plausible.
7	Coherence	The association does not conflict with the existing theory and knowledge.
8	Specificity	The outcome has only one cause.
9	Analogy	Evidence from one research area can be applied to another area.

Some criteria may not be applicable to all research questions.

SELECTION OF INDEPENDENT VARIABLES FOR THE MULTIVARIABLE MODEL

Sufficient sample size

The number of confounders that you can adjust for depends on the sample size. This is reflected in large CIs in studies where the sample size is too small for the number of confounders. There are some rules of thumb to determine the required sample size (Table 1). For multiple logistic regression and Cox proportional hazards analysis, it is recommended that for every independent variable screened for association there are at least 10 events (Harrell, 2001). In multivariable linear regression it is recommended that for every independent variable approximately 10 subjects are included (Harrell, 2001).

Selection process of independent variables

In addition to having a sufficiently large sample size, you must also decide which potential confounders are eligible to enter into the model. Variable selection can be based on literature, clinical expertise, the influence on risk estimates, or statistical significance. Including all available variables that do not add much to the model can lead to unnecessarily large models. Using biologically plausible reasons (i.e., literature or clinical expertise) to select the variables for the model is a justifiable and often advised technique. However, including only biologically plausible variables rules out the option of finding unexpected confounders. Furthermore, a small sample size may require further reduction of the independent variables, which can be achieved through various techniques. One option is bivariable analysis, where confounders that change the studied association by 10% or more are included in the final model (Vandenbroucke *et al.*, 2007). Variables can also be selected based on a certain *P* value (e.g., <0.05) in univariable analysis, although a disadvantage of this technique is that variables that are not important in the univariable association, and are therefore excluded, can be important in the full model. Another option is to allow the statistical program to choose the variables by forward or backward selection. In this technique the role of each independent variable is evaluated stepwise based on statistical significance. The most significant variables are added one by one, starting with an empty model (forward selection), or nonsignificant variables are removed stepwise (backward selection), starting with a full model. This can lead to unpredictable effects because the significance may depend on the order of adding or removing covariates. In addition, this technique may also lead to the exclusion of variables you might have preferred to keep in your model based on clinical reasons.

Model performance

Once the final multivariable model has been built, there are various measures to assess the performance of the model. An easy-to-interpret measure of overall performance is the R^2. The R^2 represents the proportion of variance in the outcome that is explained by the independent variables in the model. The value of R^2 ranges from 0 to 1, with the value 1 representing the perfect model, where independent variables entirely account for the outcome. In linear regression this is represented by the (adjusted) R^2, and in logistic regression the R^2 of linear regression is best approximated by Nagelkerke's R^2. Unfortunately, an easily interpretable R^2 cannot be calculated for Cox proportional hazards regression models.

LIMITATIONS OF MULTIVARIABLE ANALYSIS

As discussed before, multivariable analysis only adjusts for measured confounding. This is a significant difference compared to randomized controlled trials, where the randomization process results in an equal distribution of all potential confounders, known and unknown, thereby removing any relation with the exposure group and thus the effect on the risk estimate.

Once a significant association is found by multivariable analysis, always consider whether statistical significance also implies clinical relevance. It can be helpful to determine the clinical relevance of a significant association by calculating the absolute risk difference or the number needed to treat or screen to prevent one event. In 1995 the British Committee on Safety of Medicines issued a warning that third-generation oral contraceptives were associated with a twofold higher risk of venous thromboembolisms, with second-generation oral contraceptives serving as the reference group. This news caused great anxiety among women who used oral con-

traceptives. There was, unfortunately, less attention for the absolute risk increase for venous thromboembolisms, which was, according the studies the warning was based on, about 1 in 7,000 for women who took the pill from the previous generation compared to 2 in 7,000 for women using a third-generation pill. The subsequent "pill scare" was estimated to result in an additional 13,000 abortions the following year in England and Wales (Furedi, 1999). This example illustrates that researchers should always try to place their results in the bigger perspective and should not always focus only on a statistically significant difference. Finally, significant associations found by multivariable regression analysis, for example, the association between psoriasis and cardiovascular death based on observational studies, do not automatically prove causality (Samarasekera *et al.*, 2013). Randomized controlled trials are the gold standard to prove causality. However, in observational studies causality can be made more plausible by collecting supportive evidence such as the Bradford–Hill criteria for causation, which are listed in Table 2 (Hill, 1965).

ACKNOWLEDGMENTS

We thank E.A. Dowlatshahi for careful revision of the manuscript.

CONFLICT OF INTEREST

The authors state no conflict of interest.

CME ACCREDITATION

This CME activity has been planned and implemented in accordance with the Essential Areas and Policies of the Accreditation Council for Continuing Medical Education through the Joint Sponsorship of ScientiaCME and Educational Review Systems. ScientiaCME is accredited by the ACCME to provide continuing medical education for physicians. ScientiaCME designates this educational activity for a maximum of one (1) AMA PRA Category 1 Credit. Physicians should claim only credit commensurate with the extent of their participation in the activity.

To take the online quiz, follow the link below:
http://www.classmarker.com/online-test/start/?quiz=7gk52d99077c9990

SUPPLEMENTARY MATERIAL

A PowerPoint slide presentation appropriate for journal club or other teaching exercises is available at http://dx.doi.org/10.1038/jid.2014.132.

REFERENCES

Curiel-Lewandrowski C, Nijsten T, Gomez ML *et al.* (2011) Long-term use of nonsteroidal anti-inflammatory drugs decreases the risk of cutaneous melanoma: results of a United States case–control study. *J Invest Dermatol* 131:1460–8

Dowlatshahi EA, Kavousi M, Nijsten T *et al.* (2013) Psoriasis is not associated with atherosclerosis and incident cardiovascular events: the Rotterdam Study. *J Invest Dermatol* 133:2347–54

Furedi A (1999) The public health implications of the 1995 "pill scare." *Hum Reprod Update* 5:621–6

Harrell FE (2001) *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer: New York

Hill AB (1965) The environment and disease: association or causation? *Proc R Soc Med* 58:295–300

Rothman KJ, Greenland S (1998) *Modern Epidemiology*, 2nd edn. Lippincott–Raven: Philadelphia, PA

Samarasekera EJ, Neilson JM, Warren RB *et al.* (2013) Incidence of cardiovascular disease in individuals with psoriasis: a systematic review and meta-analysis. *J Invest Dermatol* 133:2340–6

Vandenbroucke JP, von Elm E, Altman DG *et al.* (2007) Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *Epidemiology* 18:805–35

SUGGESTIONS FOR FURTHER READING

Katz MW (2011) Multivariable analysis: a practical guide for clinicians. (easy)

Harrell FE (2001) Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. (advanced)

Institute for Digital Research and Education. Statistical computing: http://www.ats.ucla.edu/stat (output SPSS, STATA, SAS)

Steyerberg EW (2009) Clinical prediction models: a practical approach to development, validation, and updating. (for prediction modeling)

QUESTIONS

This article has been approved for 1 hour of Category 1 CME credit. To take the quiz, with or without CME credit, follow the link under the "CME ACCREDITATION" heading.

1. In a multivariable survival analysis:

A. Subjects without a complete follow-up time must be excluded.

B. The dependent variable is the time until event.

C. Statistical interaction between independent variables is additive.

D. The hazard ratio is significant if the confidence interval excludes 0.

2. What type of analysis was used to investigate whether subjects with psoriasis have an increased carotid intima media thickness (IMT) compared with subjects without psoriasis in the article by Dowlatshahi *et al.*? Start by determining the independent (risk factor) and dependent (outcome) variable.

A. Independent variable = psoriasis, dependent variable = IMT; linear regression.

B. Independent variable = IMT, dependent variable = psoriasis; proportional hazards analysis.

C. Independent variable = IMT, dependent variable = psoriasis; logistic regression.

3. Which statement on confounding is correct?

A. A confounder affects the relationship between two independent variables.

B. In addition to multivariable regression techniques, there are other options to correct for confounding.

C. It is always possible to adjust for confounding.

4. Approximately how many events are needed in logistic regression analysis to adjust for one independent variable?

A. This does not depend on the number of events, but on the number of study subjects.

B. For every independent variable approximately 10 events are required.

C. As long as there is a biologically plausible reason to adjust for a potential confounder, no assumptions on the sample size are required.