Tenets of Good Practice in Regression Analysis. A Brief Tutorial

Dana Pisică[1,2], Ruben Dammers[2], Eric Boersma[3], Victor Volovici[1,2]

BACKGROUND: **Regression analysis quantifies the relationships between one or more independent variables and a dependent variable and is one of the most frequently used types of analysis in medical research. The aim of this article is to provide a brief theoretical and practical tutorial for neurosurgeons wishing to conduct or interpret regression analyses.**

METHODS AND RESULTS: **Data preparation, univariable and multivariable analysis, choice of model, model requirements and assumptions are discussed, as essential prerequisites to any regression analysis. Four main types of regression techniques are presented: linear, logistic, multinomial logistic, and proportional odds logistic. To illustrate the applications of regression to real-world data and exemplify the concepts introduced, we used a previously reported data set of patients with intracranial aneurysms treated by microsurgical clip reconstruction at the Department of Neurosurgery of Erasmus MC University Medical Center Rotterdam, between January 2000 and January 2019.**

CONCLUSIONS: **Regression analysis is a powerful and versatile instrument in data analysis. This material is intended as a starter for those wishing to critically interpret or perform regression analysis and we recommend multidisciplinary collaborations with trained methodologists, statisticians, or epidemiologists.**

INTRODUCTION

Regression is a class of statistical tools almost ubiquitous in medical research data analysis. It quantifies the relationships between one or more independent (predictor, explanatory) variables (IVs) and a dependent (response, outcome) variable (DV). These relationships are mathematically determined by the parameters of a statistical model, which is fit to the observed data. Statistical software is used to obtain the best possible estimation of the model parameters, given the prespecified model and the data provided. Regression can be used to address various types of research questions, such as identification of new predictors, estimation of confounder adjusted causal effects, and prediction of future outcomes.

The mathematical groundwork of regression analysis was laid in the first decade of the nineteenth century by Adrien-Marie Legendre and Carl Friedrich Gauss, who separately described the method of least squares and applied it for the calculation of orbits of celestial bodies. Since then, both the mathematical complexity of regression techniques and their applications to almost all fields of scientific inquiry have greatly evolved. Medical research in

Key words
- Linear regression
- Logistic regression
- Multinomial regression
- Proportional odds logistic regression
- Regression analysis
- Study design

Abbreviations and Acronyms
CI: Confidence interval
DV: Dependent variable
GCS: Glasgow Coma Scale
HRQoL: Health-related quality of life
IV: Independent variable
LRT: Likelihood ratio test
MCA: Middle cerebral artery
mRS: modified Rankin Scale
OR: Odds ratio
PO: Proportional odds
SAH: Subarachnoid hemorrhage
SE: Standard error
VIF: Variance inflation factor
WFNS: World Federation of Neurosurgical Societies

From the [1]Center for Medical Decision Making, Department of Public Health and Departments of [2]Neurosurgery, [3]Cardiology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands

To whom correspondence should be addressed: Dana Pisică, M.D. [E-mail: d.pisica@erasmusmc.nl]

Supplementary digital content available online.

Citation: World Neurosurg. (2022) 161:230-239.
https://doi.org/10.1016/j.wneu.2022.02.112

*Journal homepage: **www.journals.elsevier.com/world-neurosurgery***

*Available online: **www.sciencedirect.com***

general and neurosurgical studies in particular frequently use regression as part of data analysis. However, improper conducting, reporting, and interpretation of regression analysis and results sometimes occur and can lead to inadequate conclusions and recommendations.

The aim of this article is to provide a brief theoretical and practical tutorial for neurosurgeons (and clinicians at large) wishing to conduct or interpret regression analyses, especially those with no formal statistical training and/or no statistical support available.

METHODS AND RESULTS

Scope

Many comprehensive textbooks, methodological articles, and online learning resources exist on regression and it is impossible to address all regression analysis intricacies in the span of a single article. Therefore, we focused on a few essential aspects, applicable to any regression analysis: data preparation, univariable and multivariable analysis, choice of model, model requirements and assumptions. Some key aspects, such as variable selection and imputation of missing data, were not addressed. Four main types of regression techniques are discussed and exemplified: linear, logistic, multinomial logistic, and proportional odds (PO) logistic. For each, the mathematical model, underlying assumptions, and particularities are presented.

Data

To illustrate the application of regression to real-world data, we used an existing, partially prospective and partially retrospective, data set of 609 patients with intracranial aneurysms treated by microsurgical clip reconstruction at the Department of Neurosurgery of Erasmus MC University Medical Center Rotterdam, between January 2000 and January 2019. Details about this cohort can be found elsewhere.[1]

The following variables were recorded and available for analysis: patient age, sex, history of hypertension, subarachnoid hemorrhage (SAH), Hunt and Hess grade, World Federation of Neurosurgical Societies (WFNS) score, Glasgow Coma Scale (GCS) score, aneurysm location, size and dome/neck ratio, in-hospital mortality, discharge destination, and functional outcome at 6-month follow-up, measured by the modified Rankin Scale (mRS).[2] Because no continuous outcome was recorded and strictly for illustrative purposes, we used baseline variables to simulate a DV ranging from 0 to 100, which could represent for example a health-related quality-of-life (HRQoL) metric.

Outcomes

To exemplify the 4 regression techniques, we considered the following variables as outcomes of interest: HRQoL, as measured by our simulated variable (continuous); in-hospital mortality (binary); discharge destination (unordered categorical); and mRS score at 6-month follow-up (ordered categorical). All research questions and results presented in this material are for illustrative purposes only, make no causal claims, and do not warrant clinical interpretation. Please consult our previous work for the clinically relevant information regarding this cohort.[1]

Data Preparation

A first step before conducting regression analysis is to obtain a sense of the data, using descriptive tables and figures. We can inspect the scales (continuous, discrete, or categorical) and distributions of variables (frequencies of categorical variables, normally/nonnormally distributed continuous variables) and present descriptive statistics for the entire sample (**Supplementary Table 1**) or, if desirable, for specific exposure or outcome groups (e.g., patients alive and dead at discharge; **Supplementary Table 2**). Apart from preliminary insight into the trends in the data, descriptive statistics also allow comparisons with other reported samples, with ramifications on the generalizability of results.

Another preliminary step is to properly define and, if necessary, transform the crude variables used in regression analyses. Continuous IVs (e.g., patient age) should be recorded and modeled as such. When included in a regression model, an average effect on the outcome (as expressed in the model) is computed for a unit increase along the scale of the IV (e.g., 1 year), over its entire range of possible values. Categorization of naturally continuous variables (e.g., age <65 vs. age ≥ 65 years), although frequently encountered, is not advised, because it causes loss of information and assumes implausible constant effects for all values above and below the arbitrary cutoff point(s), with effect leaps at the cutoff(s).[3-6]

Please note that in this material, we use the term "effect" as a substitute for statistical association, rather than causal effect.

Categorical IVs can be unordered (e.g., aneurysm location) or ordered (e.g., WFNS score). In a regression model, effect estimates are generated comparing each unordered IV category with a single category selected as reference (e.g., effects of each of the other possible locations relative to the reference middle cerebral artery [MCA] location). Variable categories should be defined before data collection and the reference category should be selected in accordance with the research question. Occasionally, because of sparse data in some categories, researchers might opt to combine multiple small categories into larger ones. However, this strategy can lead to loss of information and the dilution of significant category effects in the broader new category, especially if the combined categories are the most different from one another (e.g., posterior circulation and posterior communicating artery combined in a single category).[3]

For ordered categorical IVs, various options exist for handling the variable. Researchers may ignore the ordering and treat the variable as unordered, resulting in effect estimates relative to a reference category. In doing so, the number of degrees of freedom required for estimation increases, which is impractical for variables with many categories. Another option is to reduce the number of categories by collapsing adjacent groups (e.g., WFNS I–III vs. WFNS IV–V). However, this strategy might cause loss of information and special consideration should be given to selecting the cutoff(s). Alternatively, we may treat the ordered IV as continuous and obtain effect estimates for 1 level increase on the ordinal scale. In the latter case, the effect linearity must be investigated, because transitions from WFNS I to II might not be equivalent to transitions from WFNS IV to V. In our regression examples, we treated WFNS, Hunt and Hess, and GCS scores as continuous.

A potentially problematic situation that can be anticipated by examining the distribution of IVs across outcome levels is

separation. This situation occurs when the DV is perfectly predicted by 1 or more IVs for all patients (complete separation) or for a subset (quasi-complete separation).[7] For example, only 1 in 76 patients who died in hospital had an unruptured aneurysm (**Supplementary Table 2**), meaning that in-hospital mortality can be almost perfectly predicted by SAH. A similar situation occurs for posterior circulation location. Separation occurs either causally (the IV is a perfect predictor for the DV) or because of sparse data and can lead to extreme parameter estimates. When insufficient sampling is assumed, we may combine categories during data preparation (see earlier discussion) or use special techniques during the analysis phase (see section on "Logistic Regression").

Missing data should also be investigated and reported. Because of relatively small percentages of missing data in our study (**Supplementary Table 1**), we performed complete case analysis in the examples that follow. This practice is generally not ideal, because part of the collected data is discarded and biased effects might occur when the missingness in the IVs is associated with the outcome (e.g., if the patients who are most severe at presentation and have the worst outcomes are also the ones with missing baseline variables). Multiple methods for handling missing data are available.[3,4,8-10]

Univariable versus Multivariable Regression

Univariable regression quantifies the relationship between a single IV and the outcome (e.g., hypertension and mRS score at 6 months). It is useful for effect estimation (direction and magnitude of the statistical association between variables, not necessarily causal in nature), hypothesis testing (compatibility of the observed data with a model in which no relationship exists between the 2 variables), and prediction (expected outcome given a certain IV value). Although informative, univariable regression ignores the influence of other IVs on the relationship being studied. For example, the univariable association between hypertension and mRS score might be partially attributed to the older age of hypertensive patients, also associated with lower functional status. It becomes important to disentangle effects on the outcome when multiple IVs are considered simultaneously. This goal is achieved by multivariable regression, which yields effect estimates for each IV, adjusted (corrected) for the effects of other IVs in the model.

In some studies, researchers are interested in the causal effect of a particular IV (exposure), adjusted for the effects of other variables potentially influencing the outcome. Confounding variables causally influence both the exposure and the outcome. Therefore, including them in the multivariable model is essential in obtaining valid effect estimates for the IV of interest.[11] When confounders are not adjusted for, spurious, distorted associations may occur between the exposure and the outcome. When causal effects are of interest, special consideration must be paid to the construction and interpretation of results from multivariable regression models. Causal diagrams should be used before modeling to study the underlying relationships between variables.[12-14] Alternatively, the scientific interest behind a multivariable analysis might lie in estimating the independent effects of multiple IVs of interest, either for identifying variables associated with the outcome (exploration) or for incorporating relevant variables in a model to predict future outcomes (prediction). Prediction modeling is not discussed in this material, because it requires extensive subject-matter knowledge and modeling considerations.[3,15]

When performing complete case analysis, the available sample sizes for univariable and multivariable analyses might differ according to missing values in the IVs. For example, in our data set, the multivariable model for HRQoL was fitted on a subset of 428 patients with all variables of interest recorded. However, the univariable analysis between age and HRQoL could have been performed on the entire sample. In this type of situation, we cannot interpret a change in effects of an adjusted analysis performed on a subset of patients (609 vs. 428). To avoid this situation, both univariable and multivariable analyses should be performed on the same sample.[3]

As a side note, the terms "univariable" and "multivariable" concern the number of IVs being studied and should not be confused with "univariate" and "multivariate," which concern the number of DVs included simultaneously in the model (e.g., a patient's mRS score at a single time point [univariate] vs. multiple correlated longitudinal recordings of a patient's mRS score [multivariate]).[16] Multivariate regression is not covered in this material and we focus on univariate analyses, both univariable and multivariable.

Choosing the Appropriate Regression Model

The 4 regression techniques presented in this material belong to a larger family called generalized linear models. All models in this family postulate that at the population level, the DV (Y) is mathematically related to a linear combination of the IVs (X_1, X_2, …, X_p). In this linear combination, each IV is multiplied by a coefficient or parameter $\beta(\beta_{1,2,\ldots,p})$, which represents the independent effect of the variable on the outcome Y, and which we estimate by fitting the model to our sample. The sum of the IVs multiplied by their respective coefficients $(\beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \cdots + \beta_pX_p)$ is called the linear predictor, and represents the mean value Y is expected to have, conditional on specific values of the predictor variables $(E(Y|X_{1,2,\ldots,p}) = \mu)$. An intercept β_0 is included in the linear predictor and is interpreted as the expected outcome when all continuous IVs have a value of zero and all categorical IVs are equal to their reference categories.

Because the DV can be of many types (continuous, discrete, or categorical), a link function (such as the logit function) is sometimes applied to it to transform the expected value of the outcome from its original scale to the continuous scale of the linear predictor. **Table 1** offers an overview of the 4 regression models selected for this material. The choice among them should be made according to the scale of the DV.

Common Requirements

Before presenting the particularities of each model, some important requirements applicable to all 4 techniques are discussed (**Table 2**). All models in this material assume independence of observations. This requirement means that observations in the data set are in no way correlated or clustered, either within patients (the same outcome is measured repeatedly, longitudinally for the same patient) or higher levels (e.g., patients are clustered into hospitals or countries). When this

Table 1. Overview of the Linear, Logistic, Multinomial Logistic, and Proportional Odds Logistic Regression Models

Regression Model	Scale of the Outcome Variable	Mathematical Model and Interpretation of Coefficients*
Linear	Continuous (health-related quality of life, measured in points)	$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_p X_{pi} + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$ where Y_i is the outcome of patient i, $X_{1i, 2i, .., pi}$ are the IVs of patient i, $\beta_{0, 1, 2, ..., p}$ are the regression coefficients and ε_i is the error term (residual) for patient i
		Continuous X_j: β_j = mean difference in outcome (in units on the outcome scale) for a unit increase on the scale of X_j Binary X_j, with 2 categories coded as 0 (the reference category) and 1: β_j = mean difference in outcome (in units on the outcome scale) comparing the nonreference category (1) with the reference category (0) of X_j
Logistic	Binary (in-hospital mortality yes/no)	$\log\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_p X_{pi}$ where $P(Y_i = 1)$ is the probability of the outcome event occurring for patient i, $X_{1i, 2i, .., pi}$ are the IVs of patient i and $\beta_{0, 1, 2, ..., p}$ are the regression coefficients
		Continuous X_j: $\exp(\beta_j)$ = the relative change in odds (odds ratio) of experiencing the outcome event for a unit increase on the scale of X_j Binary X_j, with 2 categories coded as 0 (the reference category) and 1: $\exp(\beta_j)$ = the relative change in odds (odds ratio) of experiencing the outcome event comparing the nonreference category (1) with the reference category (0) of X_j
Multinomial logistic	Categorical unordered (discharge destination: home/rehabilitation/ nursing home)	$\log\left(\frac{P(Y_i = 1)}{P(Y_i = K)}\right) = \beta_1 X_i$ $\log\left(\frac{P(Y_i = 2)}{P(Y_i = K)}\right) = \beta_2 X_i$ … $\log\left(\frac{P(Y_i = K-1)}{P(Y_i = K)}\right) = \beta_{k-1} X_i$ where K is the number of unordered categories of the outcome variable Y, Y_i is the outcome of patient i, X_i is the vector of IVs of patient $i (X_i = X_{1i, 2i, .., pi})$ and $\beta_1, \beta_2, \ldots, \beta_{k-1}$ are the vectors of regression coefficients for the $K-1$ models (e.g., $\beta_1 = (\beta_{10}, \beta_{11}, \beta_{12}, \ldots, \beta_{1p})$)
		For each of the $K-1$ regression models: Continuous X_j: $\exp(\beta_{ij})$ = the relative change in odds (odds ratio) of experiencing the outcome category i versus the reference category K for a unit increase on the scale of X_j Binary X_j, with 2 categories coded as 0 (the reference category) and 1: $\exp(\beta_{ij})$ = the relative change in odds (odds ratio) of experiencing the outcome category i versus the reference category K comparing the nonreference category (1) with the reference category (0) of X_j
Proportional odds logistic	Categorical ordered (modified Rankin Scale score at 6-month follow-up, 0<1<2<⋯<6)	$\log\left(\frac{P(Y_i \geq 2)}{P(Y_i < 2)}\right) = \alpha_1 + \beta X_i$ $\log\left(\frac{P(Y_i \geq 3)}{P(Y_i < 3)}\right) = \alpha_2 + \beta X_i$ … $\log\left(\frac{P(Y_i \geq K)}{P(Y_i < K)}\right) = \alpha_{k-1} + \beta X_i$ where K is the number of ordered levels of the outcome variable Y, Y_i is the outcome of patient i, X_i is the vector of IVs of patient $i (X_i = X_{1i, 2i, .., pi})$, β is the common vector of coefficients across the $K-1$ models ($\beta = (\beta_1, \beta_2, \beta_3, \ldots, \beta_p)$) and $\alpha_1, \alpha_2, \ldots, \alpha_{k-1}$ are model intercepts
		Continuous X_j: $\exp(\beta_j)$ = the relative change in odds (odds ratio) of experiencing higher, rather than lower categories on the outcome scale for a unit increase on the scale of X_j Binary X_j, with 2 categories coded as 0 (the reference category) and 1: $\exp(\beta_j)$ = the relative change in odds (odds ratio) of experiencing higher, rather than lower categories on the outcome scale comparing the nonreference category (1) with the reference category (0) of X_j

IV, independent variable.
*Coefficient(s) for categorical IVs with >2 categories have different interpretations, depending on the coding of the variable: with dummy coding/as continuous, with/without collapsing categories.

Table 2. Common and Specific Requirements for Linear, Logistic, Multinomial Logistic, and Proportional Odds Logistic Regression Models

Regression Model	Specific Requirements	Common Requirements
Linear	Normally distributed residuals, with mean 0 Homoscedasticity of residuals	Independence of observations Linearity of effects of continuous independent variables Additivity of effects No perfect multicollinearity Sufficient sample size Inclusion of important variables No highly influential observations
Logistic	No separation	
Multinomial logistic	Independence of irrelevant alternatives	
Proportional odds logistic	Proportional odds assumption	

assumption is not met, other statistical methods are necessary to account for the correlated nature of the data, such as generalized linear mixed models or generalized estimating equations.

Another common assumption states that relationships between continuous IVs and the outcome (as expressed in the model) are linear, meaning that incremental increases have the same effect at any point along the range of the continuous IV. For example, the effect of a 5-year difference in age on HRQoL is assumed to be the same for patients aged 30 versus 35 years and 60 versus 65 years (*red line* in **Supplementary Figure 1**). Any nonlinearity can be evaluated by scatter plots of the continuous IVs versus the DV, and by studying model residuals. Nonlinear relationships can then be modeled using various functions, such as splines or fractional polynomials (*green line* in **Supplementary Figure 1**).

Multivariable models further imply additivity of effects, meaning that IV effects on the outcome add to each other and are not influenced by the effect of one another. For example, the effect of aneurysm size on mortality, represented by β_1, in the logistic model

$$\log(\text{odds of mortality}) = \beta_0 + \beta_1 \times \text{aneurysm size} + \beta_2 \times \text{female sex}$$

is additive to and not influenced by the effect of (female) sex, represented by β_2 (**Supplementary Figure 2**, *left*). If evidence exists that the effect of aneurysm size is different in men and women, and thus an interaction between the 2 variables exists, then a multiplication term can be added to the linear predictor, as follows:

$$\log(\text{odds of mortality}) = \beta_0 + \beta_1 \times \text{aneurysm size} + \beta_2 \times \text{female sex} + \beta_3 \times \text{aneurysm size} \times \text{female sex}$$

The effect of aneurysm size on mortality in men (female sex = 0) is then represented by β_1, whereas the effect in women is represented by $\beta_1 + \beta_3$ (**Supplementary Figure 2**, *right*).

Another common requirement concerns the absence of multicollinearity between IVs. This situation occurs when 2 or more IVs are highly correlated ($\rho > 0.80$) and can thus be predicted from one another. Because highly correlated variables tend to change together, their independent contributions to the outcome are difficult to disentangle mathematically and consequently the precision of the estimated coefficients is reduced, with inflated standard errors (SEs), wide confidence intervals (CIs) and unreliable significance tests. For example, the WFNS score is to some extent a categorization of the GCS and the 2 are highly correlated ($\rho = -0.90$ in our sample). Dealing with multicollinearity depends on the aim of the analysis. If the aim is to estimate independent effects of highly correlated variables, then, to obtain accurate SEs and CIs, these should not be added simultaneously in a model. If the research question concerns prediction or if the highly correlated variables are introduced in the model as confounders, then, all may be included simultaneously, because their joint effect on the outcome is not affected. However, in these cases, no inferences should be made on their individual coefficients.[4] A formal way to assess multicollinearity is the variance inflation factor (VIF), which quantifies how much the variance of the coefficient of a variable is influenced (inflated) by its correlations with other variables in the model. VIF values >10 suggest high correlations.[3]

An adequate sample size is an essential requirement for a reliable estimation of IV effects. A simple and widely accepted rule of thumb states that a minimum of 10 subjects per variable for linear regression and 10 events per variable for logistic regression is required.[17] The per variable denominator factually concerns the number of degrees of freedom required to estimate all regression coefficients rather than the actual number of variables. When the number of patients with events exceeds the number of event-free patients, the latter is the limiting factor. Although general events per variable rules of thumb work in many situations, the method has shortcomings,[18-21] and alternative methods (e.g., penalized regression) should be considered in cases of sparse data.

Estimation of Model Parameters

Once a model is chosen, and assuming that its requirements are met, statistical software is used to estimate the population parameters β of the model, based on the sample data. This estimation is achieved through an iterative process, called maximum likelihood estimation. This method also generates SEs, reflecting the precision of the estimations and allowing for the construction of CIs and hypothesis testing. Detailed mathematical formulas for these calculations can be found elsewhere,[22] but an important aspect is that the precision of estimations is directly influenced by the variance in the data and sample size. In the following sections, the 4 types of regression models selected for this material are discussed.

Linear Regression

Linear regression models are useful to describe the relationships between 1 or more IVs and a continuous DV. As an illustration, we modeled HRQoL as a function of patient and aneurysm characteristics. HRQoL was not normally distributed (**Supplementary Figure 3**) in our sample. The normality assumption of the linear regression

Table 3. Univariable and Multivariable Linear Regression for Health-Related Quality of Life

Outcome: HRQoL (N = 428)	Interpretation of β Coefficient	Univariable Analysis: β Coefficient (95% CI, Wald Test *P* Value)	Multivariable Analysis: β Coefficient (95% CI, Wald Test *P* Value)
RCS (age, 3 knots)*	Per year increase	Not interpretable	Not interpretable
Sex	Male	Reference	Reference
	Female	−0.6 (−3.1 to 1.9, 0.615)	1.3 (0.6 to 2.0, <0.001)
Hypertension	No	Reference	Reference
	Yes	−3.7 (−6.2 to −1.3, 0.003)	−0.5 (−1.2 to 0.2, 0.136)
Subarachnoid hemorrhage	No	Reference	Reference
	Yes	−3.1 (−5.9 to −0.2, 0.034)	−2.1 (−3.6 to −0.6, 0.005)
Hunt and Hess score†	Per point increase	−1.6 (−2.4 to −0.8, <0.001)	−1.5 (−1.8 to −1.2, <0.001)
WFNS score†		−1.5 (−2.2 to −0.8, <0.001)	−1.3 (−1.5 to −1.1, <0.001)
GCS score†		0.6 (0.3 to 0.9, <0.001)	0.4 (0.3 to 0.6, <0.001)
Aneurysm location‡	Middle cerebral artery	Reference	Reference
	Anterior cerebral artery	−0.7 (−5.4 to 4.0, 0.761)	−0.1 (−1.4 to 1.2, 0.873)
	Posterior circulation	4.5 (−1.4 to 10.3, 0.135)	0.6 (−0.9 to 2.2, 0.426)
	Posterior communicating artery	−0.4 (−4.7 to 4.0, 0.866)	−0.6 (−1.8 to 0.5, 0.291)
	Internal carotid artery	2.9 (−1.2 to 7.1, 0.169)	−1.0 (−2.1 to 0.2, 0.102)
	Anterior communicating artery	−1.9 (−4.6 to 0.8, 0.176)	−0.4 (−1.2 to 0.4, 0.306)
Size of aneurysm	Per mm increase	−0.8 (−1.1 to −0.6, p<0.001)	−0.7 (−0.8 to −0.6, <0.001)
Dome-to-neck ratio	Per point in ratio increase	−0.7 (−1.5 to 0.2, 0.148)	0.0 (−0.2 to 0.3, 0.901)

HRQoL, health-related quality of life; CI, confidence interval; RCS, restricted cubic spline; WFNS, World Federation of Neurosurgical Societies; GCS, Glasgow Coma Scale.

*Age was the only continuous independent variable that showed a significant nonlinear relationship with the outcome, as reflected in **Supplementary Figure 1**. An RCS function with 3 knots was used to model this relationship flexibly. The flexible model was compared with a model in which the relationship was assumed to be linear, using an *F* test (thus formally testing the linearity assumption), and offered a better fit to the data. The RCS does not allow a straightforward interpretation of the effect of age. We can observe graphically that older age is associated with lower mean HRQoL and the effect of increased age is stronger for patients older than 50 compared with younger patients. The overall test of association between age modeled flexibly and HRQoL was significant ($P < 0.001$) in both univariable and multivariable analysis.

†WFNS, Hunt and Hess, and GCS scores were highly correlated (variance inflation factors of 11, 8, and 8, respectively) and were included together in the full multivariable model to adjust the effects of all other variables. Individual effects for them are derived from reduced multivariable models, excluding the other 2 correlated variables, to avoid unstable effects caused by multicollinearity. To show the impact of multicollinearity: in the full model, Hunt and Hess score had an estimated effect of 0.0 (95% CI, −0.6 to 0.6; $P = 0.98$), which became −1.5 (−1.8 to −1.2; $P < 0.001$) when excluding WFNS and GCS scores. The reduced multivariable model with the lowest Akaike Information Criterion was the one containing only WFNS score, suggesting that its addition to a model offered the best fit, compared with the other 2 variables.

‡The overall test of association of aneurysm location with HRQoL was not significant (*F* test with 5 degrees of freedom: *P* value 0.2 in univariable analysis, 0.4 in multivariable analysis).

model does not apply to the DV but to the outcome residuals. The following variables were selected as IVs: age; aneurysm size; dome/neck ratio (continuous); Hunt and Hess, WFNS, and GCS scores (categorical ordered, treated as continuous); sex; hypertension; SAH (categorical binary); and location of aneurysm (categorical unordered, 6 categories). We considered possible interactions between age and all other predictors. The available sample size was 428 patients, meaning that we could reliably estimate a maximum of 42 degrees of freedom.

Mathematical Model.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_p X_{pi} + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2)$$

where Y_i is the outcome of patient i, $X_{1i, 2i, \ldots, pi}$ are the IVs of patient i, $\beta_{0, 1, 2, \ldots, p}$ are the regression coefficients and ε_i is the error term (residual) for patient i, representing the difference between the observed outcome and the outcome predicted by the corresponding linear predictor.

The linear model states that for a given set of values for the IVs, the outcome is expected to follow a normal distribution with mean equal to the linear combination $\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_p X_{pi}$ and variance σ^2. The formula reflects how the model includes a predicted, explained component (the sum score) and a noise, random component (the error term).

Interpretation of Coefficients. For a continuous IV X_k, the estimated β_k coefficient is interpreted as the mean difference in outcome (in units on the DV scale) per unit difference on the scale of the IV. For example, in our sample, a 1 mm larger aneurysm is associated with 0.8 points lower mean HRQoL (95% CI, −1.1 to −0.6). When adjusted for (i.e., for subjects with the same) age, sex, hypertension, SAH, Hunt and Hess, WFNS and GCS scores, dome/ratio

and aneurysm location, the mean difference is −0.7 points (95% CI, −0.8 to −0.6) (Table 3).

For a binary IV X_k, the estimated β_k coefficient is interpreted as the difference in mean outcome between the 2 categories. For example, the unadjusted difference in mean HRQoL between patients with and without SAH (the reference) is −3.1 points (95% CI, −5.9 to −0.2), whereas the adjusted mean difference is −2.1 points (95% CI, −3.6 to −0.6).

For categorical IVs with >2 categories, such as aneurysm location, multiple β coefficients are estimated and are interpreted as the differences in mean outcome between the various categories and the reference category (MCA). For example, patients with posterior circulation aneurysms have 4.5 points (95% CI, −1.4 to 10.3) higher mean HRQoL compared to patients with MCA aneurysms.

Hypothesis Testing. Besides the β coefficients, the population parameter σ^2 is also estimated when fitting the linear model, which is needed to calculate SEs and CIs for the estimates. For each estimated β, a Wald test is usually conducted to test the null hypothesis that the estimated coefficient is equal to 0 (no effect of the variable on the outcome). The Wald test and the CI are mathematically related because all effect sizes that would produce a test P value >0.05 and thus fail to reject the null hypothesis correspond to the 95% CI. A 95% CI containing 0 is always accompanied by a nonsignificant Wald test P value at a 0.05 significance level. It is not recommended to report P values alone and make strict inferences accordingly. Point and interval estimates for the regression coefficients should always be reported to offer insight into the magnitude of effects and their precision.[23,24]

The F test can be used to test multiple coefficients simultaneously (compare nested models). For example, to test whether interaction terms between age and the other IVs were associated with HRQoL, a full model including the IVs and the interactions with age was compared with a reduced model containing only the main effects. In our case, the P value of the F test was 0.5, and thus the null hypothesis that the coefficients for all interaction terms were 0 was not rejected.

For categorical IVs with >2 categories (location) and continuous IVs modeled flexibly (e.g., with spline functions [age]), the overall association of the variable with the outcome should be tested using an F test, not a Wald test, even in univariable analysis. This requirement is because both types of variables need >1 parameter to be estimated (5 for location and 2 for age modeled with restricted cubic splines with 3 knots). The association of the variable with the outcome requires testing the null hypothesis that all estimated parameters are zero. Although age modeled flexibly has a significant effect on HRQoL ($P < 0.001$) in our example, location does not ($P = 0.4$). Given the lack of effect of the entire location variable, no inferences should be made on the relative effects comparing different aneurysm locations with the MCA reference, the P values of which are derived from Wald tests.

A global measure of goodness of fit of the entire linear model is the R^2, defined as the proportion of the total variability of the outcome that is explained by the model. For multivariable models, an adjusted R^2, obtained by penalizing the R^2 for the number of variables in the model, can also be calculated. Our full model has an adjusted R^2 of 0.92, meaning that 92% of the observed variability in HRQoL is explained by the model. This finding is to be expected, given that we have simulated our outcome variable based on the available predictors.

Specific Requirements. Two assumptions specific to linear regression are normality and homoscedasticity of residuals, captured in the $\varepsilon_i \sim N(0, \sigma^2)$ expression. They mean that the residuals of all data points come from a normal distribution (normality), with mean 0 and a constant variance σ^2, independent of the values of the IVs (homoscedasticity). These assumptions are required to calculate SEs and perform hypothesis tests. We can assess them by plotting the residuals (residuals Q-Q plot, residuals vs. predicted outcome; Supplementary Figure 4). Some degree of deviation from the normality assumption is acceptable when linearity and homoscedasticity hold, especially in large samples.[25] When the homoscedasticity assumption is violated, but the linearity assumption holds, point estimates of regression coefficients are accurate, but more advanced methods such as bootstrapping or robust methods are required to compute the SEs.[4]

Logistic Regression

Logistic regression can be used to model the relationships between 1 or more IVs and a binary DV. To exemplify logistic regression, we studied the effect of patient sex on in-hospital mortality, for patients with SAH treated with microsurgical clip reconstruction, adjusted for known predictors of mortality (age, size of aneurysm, and WFNS score).[26] Our sample included 470 patients, of whom 70 died in hospital (15%), meaning that we could reliably estimate a maximum of 7 coefficients. Interactions between sex and the other variables were considered.

Mathematical Model. As with linear regression, the outcome (dead at discharge, $Y = 1$ or alive, $Y = 0$) is linked to a linear combination of IVs. The binary DV is expressed in the model by its probability (probability of mortality, $p = P(Y = 1)$), which, like all probabilities, has a value in the [0−1] interval. To bring the probability to the scale of the linear predictor, which may take any value between $(-\infty, +\infty)$, the logit link function is applied:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right), \quad \text{where } p \text{ is a probability}$$

The ratio between the probability of an event occurring (probability of mortality, p) and the probability of an event not occurring (probability of being alive, 1−p) is called odds and is a measure of the likelihood of an event. Thus, the logit function is also called log-odds, because it represents the natural logarithm of the odds. The mathematical equation of the model becomes:

$$\log\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \log(\text{odds}_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_p X_{pi}$$

where $P(Y_i = 1)$ is the probability of the outcome event occurring for patient i, $X_{1i, 2i, \ldots, pi}$ are the IVs of patient i and $\beta_{0, 1, 2, \ldots, p}$ are the regression coefficients.

Interpretation of Coefficients. Effect estimates from logistic regression are usually reported as odds ratios (ORs). The β_k coefficient of a continuous IV X_k represents the difference in log-odds of experiencing the outcome for a unit increase on the scale of the IV, corrected

or not for other IVs. By exponentiating the estimated β_k coefficient, we obtain estimated ORs.

$$OR = e^{\beta_k}$$

Positive β coefficients result in ORs >1, representing increased odds of experiencing the outcome, and negative β coefficients result in ORs <1, representing decreased odds of experiencing the outcome. Null effects ($\beta = 0$) have ORs of 1. For example, in our sample, a year difference in age is estimated to increase the odds of in-hospital mortality by 1.02 (95% CI, 1.00–1.04), adjusted for sex, aneurysm size, and WFNS score (**Supplementary Table 3**). The point estimate can also be restated as an increase of 2% in odds of mortality for every year increase in age for patients with the same sex, WFNS score, and aneurysm size.

For categorical IVs, the estimated β coefficients relate differences between categories similarly to linear regression: a category is selected as reference (e.g., male sex) and relative effects for the remaining categories are computed (e.g., female vs. male sex). The resulting OR represents the relative change in odds of experiencing the outcome between the categories being compared. For example, the adjusted OR for mortality of female sex versus male sex is 1.1 (95% CI, 0.6–2.1), meaning that the odds of mortality increase by 1.1 (10% increase) for females compared with males, adjusted for age, size of aneurysm, and WFNS score.

Hypothesis Testing. Two main statistical tests are frequently used in logistic regression and its extensions (multinomial and PO logistic regression): the Wald test and the likelihood ratio test (LRT). Both can be used to test a single or multiple β coefficients at a time, although the LRT is preferred for testing multiple coefficients. When testing a single parameter, the Wald test assesses the null hypothesis that the estimated regression coefficient is equal to 0, similar to linear regression. The LRT is similar in application to the F test from linear regression, and can be used to compare the goodness of fit of nested models (compare subgroups of coefficients simultaneously), using the ratio of the likelihoods of the observed data under the models being compared. We used the LRT to test for possible interactions between sex and the other IVs. Comparing a full model (containing main effects and interactions) with a reduced model containing only the main effects is equivalent to testing the null hypothesis that the regression coefficients for all interaction terms are 0, the alternative being that at least one is different from 0. The *P* value of the LRT was 0.4 and thus all interaction terms were discarded.

Specific Requirements. A requirement specific to logistic regression is lack of separation, as explained in the section on "Data Preparation." In our sample, only 1 patient with an unruptured aneurysm died in hospital. Had we fitted the logistic model on the entire sample, with SAH as one of the IVs, the OR for SAH would have been 14.5 (3.1–258.2) in univariable analysis and 5.6 (1.1–103.5) with adjustment. These extreme estimates are the effect of quasi-separation. This finding is to be expected, given that patients with ruptured and unruptured aneurysms represent different populations, with different prognoses. Consequently, we performed the analysis only in the SAH subgroup. In analyses in which a clear causal relationship is not warranted, separation becomes a sparse data problem and can be dealt with using special methods such as exact logistic regression, Firth penalization, or regularization methods.[7]

Multinomial Logistic Regression

Multinomial or polytomous logistic regression can be used to model the relationships between 1 or more IVs and an unordered categorical DV (with $K > 2$ categories), such as discharge destination ($K = 3$: home, nursing home, rehabilitation center). Only patients who survived hospitalization were analyzed. Our hypothetical aim was to identify patient and aneurysm characteristics associated with discharge destination, using multivariable multinomial regression (exploratory analysis). Based on clinical experience and the available sample size, the following IVs were selected: age, sex, hypertension, SAH, aneurysm size, and WFNS score. Our sample included 480 patients, of whom 272 were discharged home, 111 to a rehabilitation center, and 97 to a nursing home.

Mathematical Model. The multinomial logistic model is a collection of $K-1$ logistic models simultaneously applied to the data. One of the K outcome categories is selected as reference category (in our example, home discharge) against which all other $K-1$ categories are regressed against (rehabilitation center vs. home; nursing home vs. home). The model thus generates $K-1$ separate sets of β coefficients for the IVs.

$$\log\left(\frac{P(Y_i = 1)}{P(Y_i = K)}\right) = \beta_1 X_i$$
$$\log\left(\frac{P(Y_i = 2)}{P(Y_i = K)}\right) = \beta_2 X_i$$
$$\ldots$$
$$\log\left(\frac{P(Y_i = K-1)}{P(Y_i = K)}\right) = \beta_{k-1} X_i$$

where K is the number of unordered categories of the DV Y, Y_i is the outcome of patient i, X_i is the vector of IVs of patient i($X_i = X_{1i, 2i, .., pi}$) and $\beta_1, \beta_2, \ldots, \beta_{k-1}$ are the vectors of regression coefficients for the K−1 models (e.g., $\beta_1 = (\beta_{10}, \beta_{11}, \beta_{12}, \ldots, \beta_{1p})$).

Interpretation of Coefficients. For each of the $K-1$ logistic models, effect estimates are computed for each IV and usually reported as ORs, with a similar interpretation as in logistic regression. For example, for every point increase in WFNS score, adjusted for the other variables, the odds of being discharged to a rehabilitation center versus being discharged home increase by 1.5 (95% CI, 1.3–1.8) and the odds of being discharged to a nursing home versus being discharged home increase by 1.8 (95% CI, 1.5–2.2). Females, compared with males, have 0.8 (95% CI, 0.5–1.3) times lower odds of being discharged to a rehabilitation center versus home and 0.9 (95% CI, 0.5–1.5) times lower odds of being discharged to a nursing home versus home, adjusted for the other variables (**Supplementary Table 4**).

Specific Assumptions. An assumption specific to the multinomial logistic model is the independence of irrelevant alternatives, which states that the relative likelihood of choosing between any 2 options (e.g., rehabilitation vs. nursing home) is independent of any additional alternatives in the set of choices (the choice for

home discharge) and is thus influenced only by the characteristics of the initial 2 options. This assumption can be tested with the Hausman-McFadden test, which compares estimated coefficients from a model with all the available outcome alternatives with estimated coefficients from constrained models applied on subsets of the alternatives. When the assumption is violated, alternative models should be considered, such as the multinomial probit model or nested logit model.[27,28]

Proportional Odds Logistic Regression

Many neurosurgical studies contain ordinal categorical DVs, especially for the assessment of functional outcome (mRS, Glasgow Outcome Scale, and Glasgow Outcome Scale Extended). The PO logistic regression model is an extension of logistic regression that allows modeling of ordinal outcomes, without having to dichotomize them (e.g., into favorable and unfavorable). Although frequently encountered in the literature, dichotomization of the outcome scale at various cutoffs is not recommended, because it discards valuable information contained in the natural ordering of the scale and decreases the power to detect significant effects.[29-32]

To exemplify this model, we studied the effect of patient sex on 6-month functional outcome assessed with the ordinal mRS, for patients with SAH treated with microsurgical clip reconstruction, adjusted for known predictors of outcome (age, hypertension, WFNS score, aneurysm size, and location).[33] Our sample included 337 patients. The distribution of 6-month mRS scores was mRS 0 in 49 patients, mRS 1 in 77, mRS 2 in 80, mRS 3 in 36, mRS 4 in 23, mRS 5 in 2, and mRS 6 in 70. Because the mRS 5 category was small, we combined adjacent categories 4 (moderate severe disability) and 5 (severe disability), thus creating a 6-point ordinal scale.

Mathematical Model. Similar to the multinomial logistic model, the PO logistic model is a collection of $K-1$ logistic regression models estimated simultaneously for an outcome variable Y with $K > 2$ ordered categories. For each possible dichotomization of the ordinal scale, the log-odds of Y being equal to or higher than the cutoff level are modeled as a constant linear combination of the IVs, plus a different intercept for each model. Thus, the PO model generates a single common set of regression coefficients for the IVs, across all possible dichotomizations of the DV scale. For each cutoff point on the outcome scale, a different intercept is estimated.

$$\log\left(\frac{P(Y_i \geq 2)}{P(Y_i < 2)}\right) = \alpha_1 + \beta X_i$$
$$\log\left(\frac{P(Y_i \geq 3)}{P(Y_i < 3)}\right) = \alpha_2 + \beta X_i$$
$$\cdots$$
$$\log\left(\frac{P(Y_i \geq K)}{P(Y_i < K)}\right) = \alpha_{k-1} + \beta X_i$$

where K is the number of ordered levels of the DV Y, Y_i is the outcome of patient i, X_i is the vector of IVs of patient i($X_i = X_{1i,\ 2i,\ \ldots,\ pi}$), β is the common vector of coefficients across the $K-1$ models ($\beta = (\beta_1, \beta_2, \beta_3, \ldots, \beta_p)$) and $\alpha_1, \alpha_2, \ldots, \alpha_{k-1}$ are model intercepts.

Interpretation of Coefficients. Effect estimates are computed for each IV across all $K-1$ cutoffs of the outcome scale and are reported as (common) ORs. These ORs are interpreted as pooled effects over the entire ordinal outcome scale. For example, patients with hypertension have 2.0 (95% CI, 1.2–3.2) times higher odds of higher mRS scores at 6 months (worse functional outcome) compared to patients without hypertension, adjusted for the other variables. For each point increase in WFNS score, adjusted for the other variables, the odds of having higher mRS scores increase by 1.7 (95% CI, 1.5–2.0) (**Supplementary Table 5**). Common ORs apply to all possible dichotomizations across the mRS (e.g., a point increase in WFNS score is associated with 1.7 times higher odds of being mRS 6 vs. mRS 0–5, and 1.7 times higher odds of being mRS 3–6 vs. mRS 0–2).

Specific Assumptions. The inherent assumption behind the PO model is that the regression coefficients are the same for all $K-1$ dichotomizations of the ordinal scale. This is called the PO assumption or parallel lines assumption and can be tested in various ways.[4,34] However, when agreement exists regarding the ordering of the categories of the outcome metric (strictly increasing/decreasing degree of impairment/severity), the common OR is robust and can be interpreted as a summary measure of the different effects across all dichotomizations of the outcome, even when the assumption is violated. Reporting the binary ORs resulting from logistic regression models for every cutoff is also recommended.[30,31,35]

CONCLUSIONS

Regression analysis is a powerful and versatile tool in any researcher's armamentarium. We have discussed a few of its essential aspects and offered examples for 4 types of regression frequently used: linear, logistic, multinomial logistic, and PO logistic regression. Advances in statistical software speed and availability have made data analysis more accessible than ever before. We hope that this material will serve as a starter for those wishing to venture into regression analysis. We recommend consulting our listed references for further information, and more importantly, pursuing multidisciplinary collaborations with trained methodologists, statisticians, or epidemiologists.

CRediT AUTHORSHIP CONTRIBUTION STATEMENT

Dana Pisică: Conceptualization, Methodology, Formal analysis, Writing – original draft. **Ruben Dammers:** Investigation, Resources, Writing – review & editing. **Eric Boersma:** Conceptualization, Methodology, Writing – review & editing, Supervision. **Victor Volovici:** Conceptualization, Methodology, Investigation, Writing – review & editing.

ACKNOWLEDGMENTS

We thank Hester F. Lingsma for her valuable comments on the manuscript.

REFERENCES

1. Volovici V, Verploegh ISC, Vos JC, et al. Can young vascular neurosurgeons become proficient in microsurgical clip reconstruction in the endovascular era? A Rotterdam cohort spanning 2 decades with propensity score matching for complexity. *World Neurosurg*. 2020;144:e780-e788.

2. van Swieten JC, Koudstaal PJ, Visser MC, Schouten HJ, van Gijn J. Interobserver agreement for the assessment of handicap in stroke patients. *Stroke*. 1988;19:604-607.

3. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Switzerland: Springer International Publishing; 2019.

4. Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Switzerland: Springer International Publishing; 2015.

5. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med*. 2006;25:127-141.

6. Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ*. 2006;332:1080.

7. Mansournia MA, Geroldinger A, Greenland S, Heinze G. Separation in logistic regression: causes, consequences, and control. *Am J Epidemiol*. 2018;187:864-870.

8. Van Buuren S. *Flexible Imputation of Missing Data*. Second Edition. New York: Chapman and Hall/CRC; 2018. Taylor & Francis Group.

9. Rubin DB. Inference and missing data. *Biometrika*. 1976;63:581-592.

10. Marshall A, Altman DG, Royston P, Holder RL. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC Med Res Methodol*. 2010;10:7.

11. VanderWeele TJ. Principles of confounder selection. *Eur J Epidemiol*. 2019;34:211-219.

12. Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *Am J Epidemiol*. 2013;177:292-298.

13. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology*. 1999;10:37-48.

14. Gaskell AL, Sleigh JW. An introduction to causal diagrams for anesthesiology research. *Anesthesiology*. 2020;132:951-967.

15. Mijderwijk HJ, Steyerberg EW, Steiger HJ, Fischer I, Kamp MA. Fundamentals of clinical prediction modeling for the neurosurgeon. *Neurosurgery*. 2019;85:302-311.

16. Tsai AC. Achieving consensus on terminology describing multivariable analyses. *Am J Public Health*. 2013;103:e1.

17. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996;49:1373-1379.

18. Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A, Perneger TV. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. *J Clin Epidemiol*. 2011;64:993-1000.

19. van Smeden M, de Groot JA, Moons KG, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Med Res Methodol*. 2016;16:163.

20. van Smeden M, Moons KG, de Groot JA, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. *Stat Methods Med Res*. 2019;28:2455-2474.

21. de Jong VMT, Eijkemans MJC, van Calster B, et al. Sample size considerations and predictive performance of multinomial logistic prediction models. *Stat Med*. 2019;38:1601-1619.

22. McCullagh P, Nelder JA. *Generalized Linear Models*. Second Edition. New York: Taylor & Francis; 1989.

23. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*. 2016;31:337-350.

24. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. *Am Stat*. 2016. https://doi.org/10.1080/00031305.2016.1154108.

25. Schmidt AF, Finan C. Linear regression and the normality assumption. *J Clin Epidemiol*. 2018;98:146-151.

26. Risselada R, Lingsma HF, Bauer-Mehren A, et al. Prediction of 60 day case-fatality after aneurysmal subarachnoid haemorrhage: results from the International Subarachnoid Aneurysm Trial (ISAT). *Eur J Epidemiol*. 2010;25:261-266.

27. Benson AR, Kumar R, Tomkins A. *On the relevance of irrelevant alternatives. Proceedings of the 25th International Conference on World Wide Web, April 11—15, 2016, Montréal, Québec, Canada*. 2016.

28. Kwak C, Clayton-Matthews A. Multinomial logistic regression. *Nurs Res*. 2002;51:404-410.

29. Roozenbeek B, Lingsma HF, Maas AI. New considerations in the design of clinical trials for traumatic brain injury. *Clin Investig (Lond)*. 2012;2:153-162.

30. Roozenbeek B, Lingsma HF, Perel P, et al. The added value of ordinal analysis in clinical trials: an example in traumatic brain injury. *Crit Care*. 2011;15:R127.

31. McHugh GS, Butcher I, Steyerberg EW, et al. A simulation study evaluating approaches to the analysis of ordinal outcome data in randomized controlled trials in traumatic brain injury: results from the IMPACT Project. *Clin Trials*. 2010;7:44-57.

32. Ceyisakar IE, van Leeuwen N, Dippel DWJ, Steyerberg EW, Lingsma HF. Ordinal outcome analysis improves the detection of between-hospital differences in outcome. *BMC Med Res Methodol*. 2021;21:4.

33. Jaja BNR, Saposnik G, Lingsma HF, et al. Development and validation of outcome prediction models for aneurysmal subarachnoid haemorrhage: the SAHIT multinational cohort study. *BMJ*. 2018;360:j5745.

34. Brant R. Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*. 1990;46:1171-1178.

35. Senn S, Julious S. Measurement in clinical trials: a neglected issue for statisticians? *Stat Med*. 2009;28:3189-3209.

Conflict of interest statement: The authors declare that the article content was composed in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received 30 November 2021; accepted 21 February 2022

Citation: World Neurosurg. (2022) 161:230-239. https://doi.org/10.1016/j.wneu.2022.02.112

Journal homepage: ***www.journals.elsevier.com/world-neurosurgery***

Available online: ***www.sciencedirect.com***

SUPPLEMENTARY DATA

Supplementary Table 1. Patient and Aneurysm Baseline Characteristics and Outcomes

Variable	Overall Sample (N = 609)	Percentage Missing
Female sex, n (%)	427 (70.2)	0.2
Age (years), median (IQR)	53.2 (45.4—61.8)	0.0
Hypertension, n (%)	432 (73.0)	2.8
Subarachnoid hemorrhage, n (%)	520 (85.4)	0.0
Hunt and Hess score, median (IQR)	2 (2—4)	13.6
World Federation of Neurosurgical Societies score, median (IQR)	1 (1—3)	2.8
Glasgow Coma Scale score, median (IQR)	15 (13—15)	0.3
Aneurysm location, n (%)		0.0
Middle cerebral artery	261 (42.9)	
Anterior communicating artery	171 (28.1)	
Posterior communicating artery	60 (9.9)	
Internal carotid artery	57 (9.4)	
Anterior cerebral artery	39 (6.4)	
Posterior circulation	21 (3.4)	
Size of aneurysm (mm), median (IQR)	6.7 (4.5—9.4)	3.4
Dome/neck ratio, median (IQR)	2.0 (1.3—2.7)	3.9
In-hospital mortality, n (%)	76 (12.9)	3.6
Discharge destination, n (%)		3.6
Home	288 (49.1)	
Rehabilitation	119 (20.3)	
Nursing home	104 (17.7)	
Dead on discharge	76 (12.9)	
Modified Rankin Scale score, 6 months, n (%)		31.4
0	58 (13.9)	
1	102 (24.4)	
2	102 (24.4)	
3	42 (10.0)	
4	28 (6.7)	
5	2 (0.5)	
6	84 (20.1)	
Health-related quality of life, simulated, median (IQR)	70.3 (60.2—78.0)	5.6

IQR, interquartile range.

Supplementary Table 2. Patient and Aneurysm Baseline Characteristics, According to In-Hospital Mortality

Variable	Alive at Discharge (N = 511)	Dead at Discharge (N = 76)	Percentage Missing
Female sex, n (%)	356 (69.8)	53 (69.7)	0.2
Age, median (IQR)	52.8 (45.3—61.0)	56.8 (48.4—64.7)	0.0
Hypertension, n (%)	362 (71.7)	56 (81.2)	2.8
Subarachnoid hemorrhage, n (%)	425 (83.2)	75 (98.7)	0.0
Hunt and Hess score, median (IQR)	2 (2—3)	4 (2—5)	13.6
World Federation of Neurosurgical Societies score, median (IQR)	1 (1—2)	4 (2—5)	2.3
Glasgow Coma Scale score, median (IQR)	15 (13—15)	9.5 (3—14)	0.3
Aneurysm location, n (%)			0.0
Middle cerebral artery	226 (44.2)	31 (40.8)	
Anterior communicating artery	140 (27.4)	20 (26.3)	
Posterior communicating artery	47 (9.2)	8 (10.5)	
Internal carotid artery	44 (8.6)	12 (15.8)	
Anterior cerebral artery	34 (6.7)	4 (5.3)	
Posterior circulation	20 (3.9)	1 (1.3)	
Size of aneurysm (mm), median (IQR)	6.6 (4.3—9.0)	7.7 (4.9—12.0)	3.4
Dome/neck ratio, median (IQR)	1.9 (1.3—2.6)	2.1 (1.3—3.3)	3.9

IQR, interquartile range.

Supplementary Table 3. Univariable and Multivariable Logistic Regression for In-Hospital Mortality

Outcome: In-Hospital Mortality (n = 470)	Univariable Analysis: OR (95% CI, Wald Test *P* Value)	Multivariable Analysis: OR (95% CI, Wald Test *P* Value)
Age		
Per year increase	1.0 (1.0—1.0, 0.039)	1.0 (1.0—1.0, 0.066)
Size of aneurysm		
Per mm increase	1.1 (1.1—1.2, <0.001)	1.1 (1.0—1.2, 0.002)
World Federation of Neurosurgical Societies score		
Per point increase	1.6 (1.4—1.9, <0.001)	1.6 (1.4—1.9, <0.001)
Sex		
Male	Reference	Reference
Female	1.1 (0.6—1.9, 0.777)	1.1 (0.6—2.1, 0.736)

OR, odds ratio; CI, confidence interval.

Supplementary Table 4. Multivariable Multinomial Logistic Regression for Discharge Destination

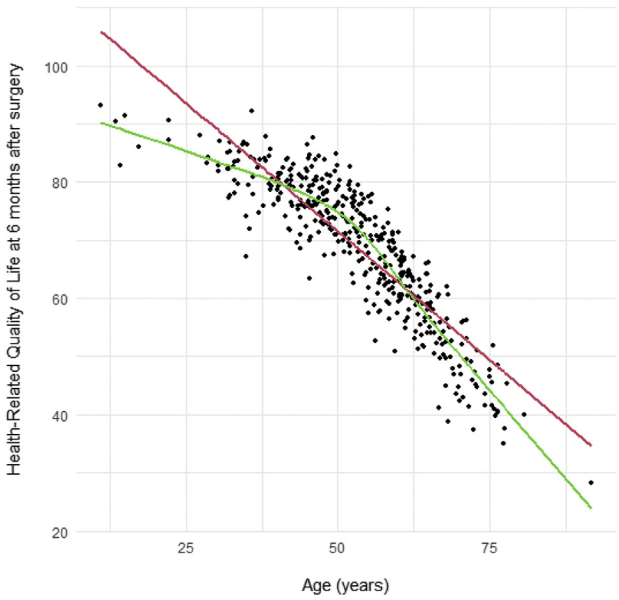
	Multivariable Analysis	
Outcome: Discharge Destination (N = 480)	**OR (95% CI, Wald Test *P* Value) for Being Discharged to a Rehabilitation Center versus Home**	**OR (95% CI, Wald Test *P* Value) for Being Discharged to a Nursing Home versus Home**
Sex		
Male	Reference	Reference
Female	0.8 (0.5—1.3, 0.399)	0.9 (0.5—1.5, 0.610)
Age		
Per year increase	1.0 (1.0—1.1, 0.006)	1.1 (1.0—1.1, <0.001)
Hypertension		
No	Reference	Reference
Yes	1.3 (0.8—2.2, 0.329)	1.6 (0.9—3.0, 0.142)
Subarachnoid hemorrhage		
No	Reference	Reference
Yes	0.9 (0.5—2.0, 0.871)	1.4 (0.5—3.9, 0.473)
Size of aneurysm		
Per mm increase	1.0 (1.0—1.1, 0.255)	1.0 (1.0—1.1, 0.347)
World Federation of Neurosurgical Societies score		
Per point increase	1.5 (1.3—1.8, <0.001)	1.8 (1.5—2.2, <0.001)

OR, odds ratio; CI, confidence interval.

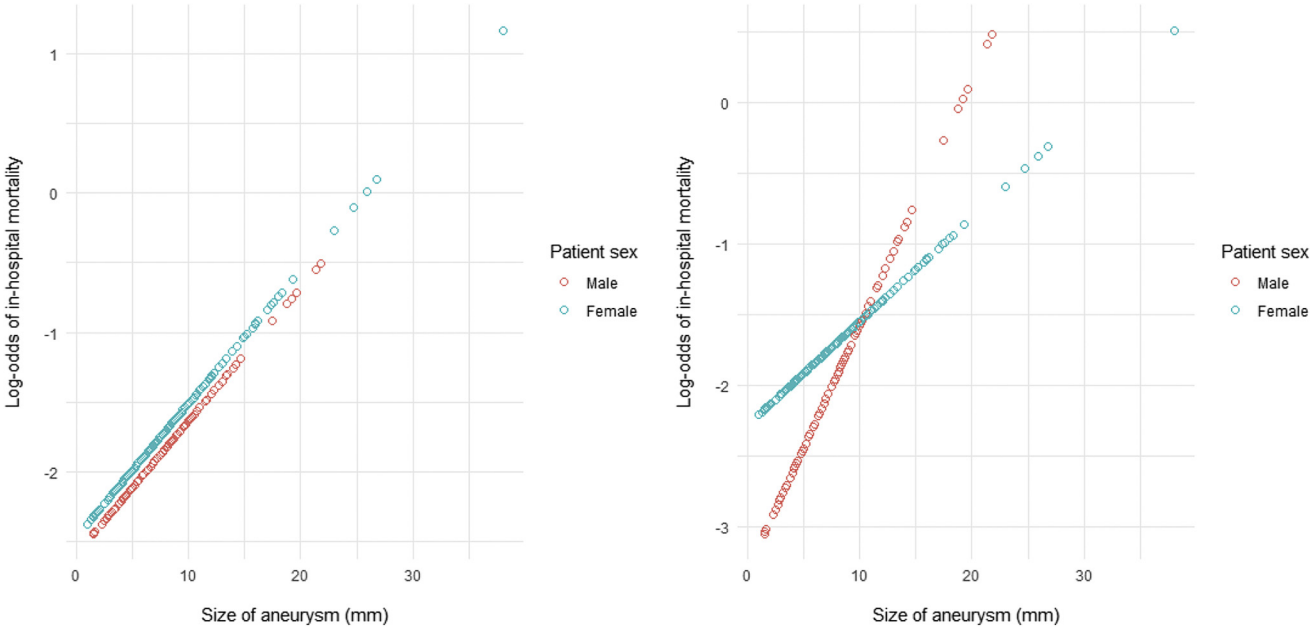
Supplementary Table 5. Univariable and Multivariable Proportional Odds Logistic Regression for Functional Outcome at 6 months after Surgery

Outcome: Modified Rankin Scale Score at 6 Months (N = 337)	**Univariable Analysis: OR (95% CI, Wald Test *P* Value)**	**Multivariable Analysis: OR (95% CI, Wald Test *P* Value)**
Age		
Per year increase	1.0 (1.0—1.0, 0.001)	1.0 (1.0—1.0, 0.046)
Hypertension		
No	Reference	Reference
Yes	1.9 (1.2—3.0, 0.006)	2.0 (1.2—3.2, 0.005)
World Federation of Neurosurgical Societies score		
Per point increase	1.7 (1.5—1.9, <0.001)	1.7 (1.5—2.0, <0.001)
Aneurysm location*		
Middle cerebral artery	Reference	Reference
Anterior cerebral artery	1.0 (0.5—2.0, 0.905)	1.5 (0.7—3.2, 0.339)
Posterior circulation	0.7 (0.3—2.0, 0.565)	1.1 (0.4—3.1, 0.906)
Posterior communicating artery	1.5 (0.7—3.1, 0.273)	2.3 (1.1—4.9, 0.027)
Internal carotid artery	1.1 (0.5—2.3, 0.802)	1.7 (0.8—3.5, 0.179)
Anterior communicating artery	0.9 (0.6—1.4, 0.688)	1.5 (0.9—2.4, 0.137)
Size of aneurysm		
Per mm increase	1.1 (1.0—1.1, 0.007)	1.1 (1.0—1.1, 0.024)
Sex		
Male	Reference	Reference
Female	1.1 (0.7—1.7, 0.703)	1.1 (0.7—1.8, 0.605)

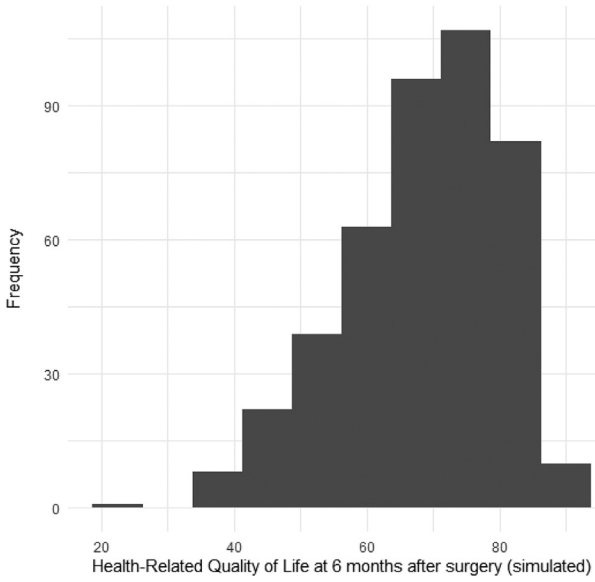
OR, odds ratio; CI, confidence interval.
*The overall test of association of aneurysm location with 6-month modified Rankin Scale score was not significant (likelihood ratio test with 5 degrees of freedom: *P* value 0.8 in univariable analysis, 0.3 in multivariable analysis).



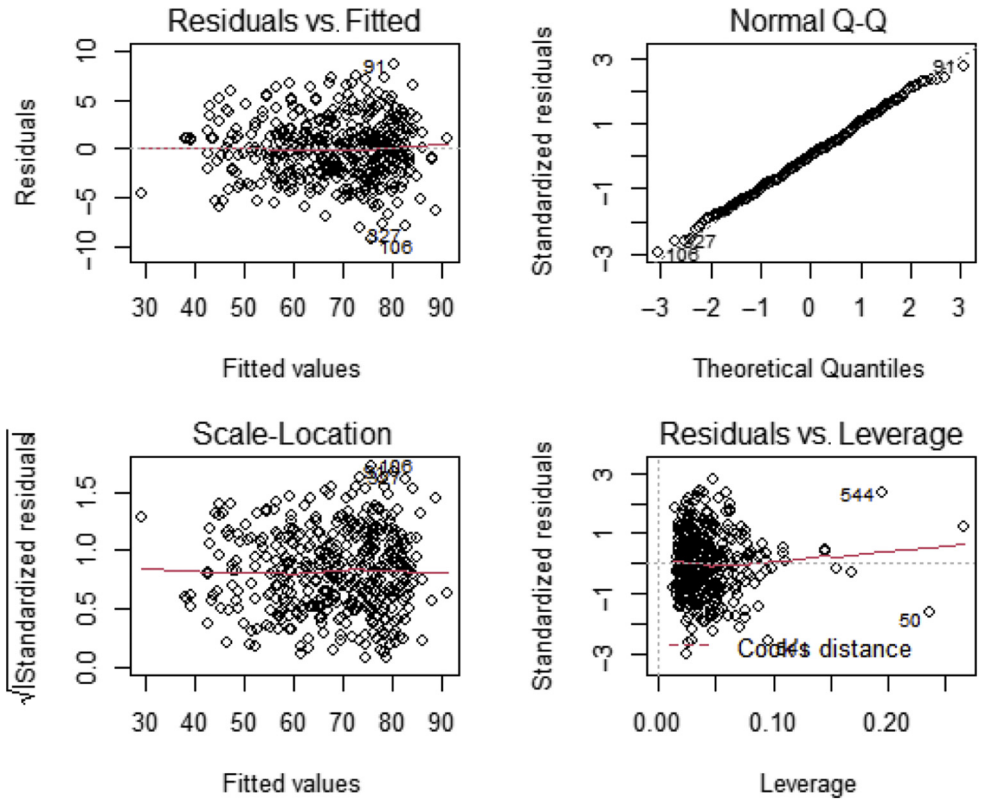
Supplementary Figure 1. Relationship between age and health-related quality of life (linearity assumption). Continuous independent variables (IVs) that show a nonlinear relationship with the outcome may be modeled flexibly using various functions, to allow for different curved relationships. The restricted cubic spline (RCS) function is recommended when no previous knowledge of a specific nonlinear relationship exists.[1] The initial linear model (*red regression line*) can then be compared with one that allows nonlinearity (*green regression line*), using a goodness of fit test for nested models, such as the *F* test or the likelihood ratio test. When significant nonlinearity is confirmed, the RCS function itself, another transformation of the IV suggested by the RCS function/outcome plot[1] or piecewise regression may be used in lieu of the crude IV to satisfy the linearity assumption.



Supplementary Figure 2. Relationship between size of aneurysm, patient sex and in-hospital mortality (additivity assumption). (*Left*) The first model assumes additivity of aneurysm size and patient sex effects. This is equivalent to 2 parallel regression lines between size of aneurysm and log-odds of mortality, both with slope β_1: one for males (with intercept $\beta_0 + \beta_2 \times 0 = \beta_0$) and the other for females (with intercept $\beta_0 + \beta_2 \times 1 = \beta_0 + \beta_2$). The coefficient β_2 is the vertical distance between the lines and represents the difference in log-odds of mortality between females and males, for any given aneurysm size. (*Right*) When an interaction term is added to the model, the 2 regression lines have different slopes ($\beta_1 + \beta_3 \times 0 = \beta_1$ for males and $\beta_1 + \beta_3 \times 1 = \beta_1 + \beta_3$ for females), translating to different effects of aneurysm size between the groups. An interaction between independent variables (IVs) can be tested by comparing a model containing only the main IV effects with a more complex model also containing the interaction term, using an *F* test or a likelihood ratio test. This is equivalent to testing the null hypothesis that the interaction term is not associated with the outcome ($\beta_3 = 0$). In models with >2 IVs, it is recommended that all clinically plausible interactions be considered simultaneously, followed by a global test for all interaction terms.[1]



Supplementary Figure 3. Distribution of a simulated health-related quality-of-life metric.



Supplementary Figure 4. Residual plots from multivariable linear regression for health-related quality of life (normality and homoscedasticity assumptions). Residual plots can help in assessing model fit and identifying violations of model assumptions. The residuals versus predicted outcome (fitted values) plot can highlight violations of the linearity assumption, which manifest as a curved pattern (*red line*) and suggest that nonlinear terms should be considered. Plotting residuals against each independent variable in the model could help identify the nonlinear relationship. A roughly horizontal line around zero, with no curved pattern, means the linearity assumption is met. Homoscedasticity can also be examined on this plot or alternatively on the scale-location plot. In both plots, when the assumption holds, the residuals are uniformly spread out along the fitted values, with no diverging patterns and a roughly horizontal line on the scale-location plot. When heteroscedasticity is observed, transformations of the outcome variable or robust methods for calculating standard errors might be considered. The normality assumption can be visually checked with a normal Q-Q plot of the residuals, in which the residuals should follow the reference diagonal line. Outliers (observations with large residuals) and leverage points (observations with extreme combinations of independent variables values) can be identified by examining the residuals versus leverage plot. Individual observations such as these have the potential to be highly influential on the estimation by the entire model of at least 1 regression coefficient/standard error, which can be checked by computing Cook's distance. Highly influential points merit careful inspection for data entry errors. When errors are ruled out, they may be removed from the data set, resulting in a model that would better fit the overall data. However, this strategy might also induce bias in the results, because the underlying population about which we are trying to make inferences likely includes extreme subjects.[1]

SUPPLEMENTARY REFERENCE

1. Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Switzerland: Springer International Publishing; 2015.