Rev Esp Cardiol. 2011;**64(6)**:501–507

Focus on: Contemporary Methods in Biostatistics (I)

# Regression Modeling Strategies

Eduardo Núñez,[a,b,*] Ewout W. Steyerberg,[c] and Julio Núñez[a]

[a] *Servicio de Cardiología, Hospital Clínico Universitario, INCLIVA, Universitat de Valencia, Spain*
[b] *Cuore International, Reading, Pennsylvania, United States*
[c] *Department of Public Health, Erasmus MC, Rotterdam, The Netherlands*

*Article history:*
Available online 6 May 2011

*Keywords:*
Overfitting
Number of events per variable
Calibration
Discrimination

A B S T R A C T

Multivariable regression models are widely used in health science research, mainly for two purposes: prediction and effect estimation. Various strategies have been recommended when building a regression model: *a)* use the right statistical method that matches the structure of the data; *b)* ensure an appropriate sample size by limiting the number of variables according to the number of events; *c)* prevent or correct for model overfitting; *d)* be aware of the problems associated with automatic variable selection procedures (such as stepwise), and *e)* always assess the performance of the final model in regard to calibration and discrimination measures. If resources allow, validate the prediction model on external data.



## Estrategias para la elaboración de modelos estadísticos de regresión

*Palabras clave:*
Sobresaturación
Número de eventos por cada variable
Calibración
Discriminación

R E S U M E N

Actualmente los modelos multivariables de regresión son parte importante del arsenal de la investigación clínica, ya sea para la creación de puntuaciones con fines pronósticos o en investigación dedicada a generar nuevas hipótesis. En la creación de estos modelos, se debe tener en cuenta: *a)* el uso apropiado de la técnica estadística, que ha de ser acorde con el tipo de información disponible; *b)* mantener el número de variables por evento no mayor de 10:1 para evitar la sobresaturación del modelo, relación que se puede considerar una medida grosera de la potencia estadística; *c)* tener presentes los inconvenientes del uso de los procesos automáticos en la selección de las variables, y *d)* evaluar el modelo final con relación a las propiedades de calibración y discriminación. En la creación de modelos de predicción, en la medida de lo posible se debe evaluar estas mismas medidas en una población diferente.


## INTRODUCTION

Multivariable regression models are widely used in health science research. Data are frequently collected to investigate interrelationships among variables or to determine factors affecting an outcome of interest. It is here where multivariable regression models become a tool to find a simplified mathematical explanation between the candidate predictors and the outcome. The ultimate goal is to derive a parsimonious model that makes sense from the subject matter point of view, closely matches the observed data, and has valid predictions on independent data.

Due to advances in statistical software, which have made them friendlier to the user, more researchers with limited background in biostatistics are now engaged in data analysis. Thus, the goal of this review is to provide practical advice on how to build a parsimonious and more effective multivariable model. The overall steps in any regression model exercise are listed in Table 1. Due to limited space, only the most practical points are presented.

## DATA STRUCTURE AND TYPE OF REGRESSION ANALYSIS

Regression models share a general form that should be familiar to most, usually: response = weight$_1$ × predictor$_1$ + weight$_2$ × predictor$_2$ + ... weight$_k$ × predictor$_k$ | normal error term. The variable to be explained is called the dependent (or response) variable. When the dependent variable is binary, the medical literature refers to it as an outcome (or endpoint). The factors that explain the dependent variable are called independent variables, which encompass the variable of interest (or explanatory variable) and the remaining variables, generically called covariates. Not infrequently, the unique function of these covariates is to adjust for imbalances that may be present in the levels of the explanatory variable. Sometimes, however, the identification of the predictors for the response variable is the main study goal, and in this case, every independent variable becomes of interest. Models can be used for different tasks (Table 1) that can be summarized as

* Corresponding author: Epidemiology and Statistical Services, Cuore International, 2914 Leiszs Bridge Rd. Reading, PA 19605, United States.
*E-mail address:* enunezb@gmail.com (E. Núñez).

1885-5857/$ – see front matter © 2011 Sociedad Española de Cardiología. Published by Elsevier España, S.L. All rights reserved.
doi:10.1016/j.rec.2011.01.017

**Abbreviations**

ARD: absolute risk difference
EPV: events per variable
KM: Kaplan-Meier method
MAR: missing at random
MCAR: missing completely at random
MFP: multivariable fractional polynomial
NMAR: non-missing at random
NNT: number needed to treat

**Table 1**
Overall Steps in Multivariable Regression Modeling

| |
|---|
| **Determining the aim of the model** |
| *Prediction (prognostic models)* |
| *Effect size (or explanatory models)* |
| **Ascertainment of true outcome** |
| *Minimize endpoint misclassification error* |
| *Prefer hard outcomes for prognostic models* |
| *If combined endpoint, ensure the direction of the effect is the same for both components* |
| *Consider using new outcomes such as days alive and out of hospital in heart failure studies* |
| **Choosing appropriate statistical method dependent on outcome and prediction type** |
| *Continuous: linear regression* |
| *Binary: logistic regression* |
| *Binary with censored observations* |
| Cox proportional regression |
| Parametric survival regression |
| Competing risks |
| *Longitudinal & time-to-event endpoint: Joint modeling approach* |
| *Longitudinal data with interest in intermediate endpoints: multi-state Markov modeling* |
| **Proper model building, including internal validation** |
| *Parsimony versus complexity* |
| *Selection of the right variables. Caution on the inappropriate use of stepwise procedures. Use backward instead of forward. Tune up the stopping rule according to the sample size*[1] |
| *Avoidance of overffiting (EPV rule of thumb)* |
| *Do not assume linearity for continuous variables; transform them if necessary. Use FPF or RCS for complex nonlinear functions* |
| **Assessing model's performance** |
| *Internal validation (preferably bootstrapping). Parameters to be evaluated* |
| Overall performance measures ($R^2$, Brier score) |
| Discrimination ability (AUC, C-statistics, IDI, NRI) |
| Calibration (Hosmer-Lemeshow goodness of fit test, calibration plot, calibration slope, Gronnesby and Borgan test, calibration-in-the-large) |
| *External validation. The same parameters, but in external data* |
| **The need for regression coefficient's shrinkage** |
| *If calibration assessment shows overly optimistic coefficients, then* |
| Adjust shrinkage based on calibration slope, or |
| Use more complex penalization methods such as LASSO and MLE |
| **Presenting the results** |
| *Unadjusted versus adjusted* |
| *Relative metrics (OR, HR)* |
| *Absolute metrics (ARD, NNT)* |

ARD, absolute risk difference; AUC, area under the ROC curve; C-statistics, equivalent to AUC for censored data; EPV, number of events per variable; FPF, fractional polynomial function; HR, hazard ratio; IDI, integrated discrimination index; LASSO, least absolute shrinkage and selection operator; MLE, maximum likelihood estimation; NNT, number needed to treat; NRI, net reclassification index; OR, odds ratio; $R^2$, explained variation measure; RCS, restricted cubic splines.

prediction and/or effect estimation.[2] Table 2 summarizes the differences in strategies between prediction and effect estimation models.

## Models for Prediction

These models are created when the main goal is to predict the probability of the outcome in each subject, often beyond the data from which it originated. As an example, the clinical prediction rules derived from a model fitted to the Framingham data has been shown, after multiple external validations, to provide a quantitative estimation of the absolute risk of coronary heart disease in a general population.[3] For these types of models, the researcher needs to balance complexity (and accuracy) versus parsimony; in other words, how closely the model needs to fit the data at hand versus how generalizable the predictions will be in external populations. Complex models, such as those with multiple interactions, excessive number of predictors, or continuous predictors modeled through complex nonlinear relationship, tend to fit poorly in other populations.

Several recommendations have been proposed for building these types of models,[2,4,5] the following being the most important: *a)* incorporate as much accurate data as possible, with wide distribution for predictor values; *b)* impute data if necessary as sample size is important; *c)* specify in advance the complexity or degree of nonlinearity that should be allowed for each predictor; *d)* limit the number of interactions, and include only those prespecified and based on biological plausibility; *e)* for binary endpoints, follow the 10-15 events per variable (EPV) rule to prevent overfitting. If not possible, then proceed to data reduction; *f)* be aware of the problems with stepwise selection strategies. If used, proceed with a backward elimination instead, and set the criterion for stopping rule equivalent to AIC ($P = .157$). With small samples, relax even more the stopping rule ($P = .25$ to .5). Use prior knowledge whenever possible; *g)* check the degree of collinearity between important predictors and use subject matter expertise to decide which of the collinear predictors should be included in the final model; *h)* validate the final model for calibration and discrimination, preferably using bootstrapping, and *i)* use shrinkage methods if validation shows over-optimistic predictions.

## Models for Effect Estimation

These models are created either as tools for effect estimation, or as a basis for hypothesis testing. Most articles published in biomedical literature are based on this type of models. Because there is a little concern for parsimony, the balance would be in favor of developing a more accurate and complex model that reflects the data at hand. However, always use principles that prevent overfitted estimates, and if necessary precede to data reduction methods. It is always a good principle to validate the final model based on calibration and discrimination measures.

## Types of Models Driven by the Structure of the Data

Another consideration when building a regression model is to choose the appropriate statistical model that matches the type of dependent variable. There is a multitude of variation in how the data is collected, and not infrequently the same data can be

**Table 2**
Differences in Strategies According to the Task Assigned to the Model

| Aims | Considerations | Validation |
|---|---|---|
| *Prediction* | Parsimony over complexity; the integrated information from all predictors is what matters | Calibration and discrimination using bootstrapping (internal validation) |
| Prediction of an outcome of interest (prognostic score) | Avoid overuse of cutpoints | Shrinkage of main effects (linear shrinkage or penalization) |
| Identification of important predictors | Multiple imputation if missingness > 5% | Calibration and discrimination on independent data (external validation) |
| Stratification by risk | Don't excessively rely on stepwise procedures. If too many predictors or insufficient subject matter knowledge, then use MFP algorithm | Update coefficients in the new data if necessary |
| | Use EPV rule of thumb to limit the number of predictors. Group predictors if needed (i.e., propensity score) | Convert regression coefficients into scores using appropriate algorithms |
| | Spend additional degrees of freedom on interactions and modeling nonlinear relationship of continuous predictors if prior knowledge suggested. Limit testing of individual terms; instead consider overall significance | Evaluate clinical usefulness of the derived score (nomogram, prognostic score chart, decision curve analysis) |
| | | If clinical decision making is in mind, also report absolute metrics (ARD or NNT) |
| *Effect estimation* | The weight between parsimony and complexity varies according to the research question | At least calibration and discrimination ability of the model should be reported; internal validation is a plus |
| Explanatory: understanding the effect of predictors | Always minimize continuous predictors categorization | If clinical decision making in mind, also report absolute metrics (ARD or NNT) |
| Adjustment for predictors in experimental design to increase statistical precision | Multiple imputation if missingness > 5% | |
| Focus on the independent information provided by one predictor (or explanatory variable) | Use MFP algorithm as the preferred selection variable approach | |
| | Always keep present the EPV rule of thumb. Group predictors in case of small sample size (i.e. propensity score) | |
| | If sample size allows, model nonlinear relationship of continuous predictors with no more than 4-5 degree of freedom (FPF or RCS). Final decision should be based on the overall significance (omnibus *P*-value) | |
| | As hypothesis generating study, test for interactions. Keep interaction terms only when the omnibus *P*-value is significant | |

ARD, absolute risk difference; EPV, number of events per variable; FPF, fractional polynomial function; MFP, multivariable fractional polynomial; NNT, number needed to treat; RCS, restricted cubic splines.

analyzed with more than one regression method. Table 1 shows the type of regression methods that match the most frequent types of data collected (based on a normal error assumption). A detailed explanation of each of these methods is beyond the scope of this paper, and therefore, only the most important aspects will be provided.

### *Linear Regression Analysis*

The main assumption is the linearity of the relationship between a continuous dependent variable and predictors. When untenable, linearize the relationship using variable transformation or apply nonparametric methods.

### *Logistic Regression Analysis*

The logistic regression model is appropriate for modeling a binary outcome disregarding time dimension. All we need to know about the outcome is whether it is present or absent for each subject at the end of the study. The resulting estimate of effect for treatment is the odds ratio (OR) adjusted for other factors included as covariates. Sometimes, logistic regression has been used inappropriately to analyze time-to-event data. Annesi et al.[6] demonstrated that, compared to Cox, the two methods yielded similar estimates, and an asymptotic relative efficiency of both models close to 1 only in studies with short follow-up and low event rate. Therefore, logistic regression should be considered as an alternative to Cox regression only when the duration of the cohort follow-up can be disregarded for being too short, or when the proportion of censoring is minimal and similar between the two levels of the explanatory variable.

### *Time-to-Event Design*

Survival regression methods have been designed to account for the presence of censored observations, and therefore are the right choice to analyze time-to-event data. Cox proportional hazard is the most common method used. The effect size is expressed in relative metrics as a hazard ratio (HR). A constant instantaneous hazard risk difference during follow-up is the main assumption. Several nonparametric alternatives to Cox regression have been proposed in the presence of nonproportionality.[7]

Parametric survival methods are recommended as follows: *a)* when the baseline hazard or survival function is of primary interest; *b)* to get more accurate estimates in situations where the shape of the baseline hazard function is known by the researcher; *c)* as a way to estimate the adjusted absolute risk difference (ARD) and number needed to treat (NNT) at prespecified time points; *d)* when the proportionality assumption for the explanatory variable is not tenable,[8] and *e)* when there is a need to extrapolate the results beyond the observed data.

In survival analysis, each subject can experience one of the several different types of events at follow-up. If the occurrence of one type of event either influences or prevents the event of interest, a competing risks situation arises. For example, in a study of patients with acute heart failure, hospital readmission, the event of interest, is prevented if the patient dies at follow-up. Death here is a competing risk, preventing that this patient could be readmitted. In a competing risks scenario, several studies have demonstrated that the traditional survival methods, such as Cox regression and Kaplan-Meier method (KM) are inappropriate.[9] Alternative methods have been proposed, such as the cumulative incidence functions and the proportional subdistribution hazards model by Fine and Gray.[10]

In summary, consider these methods: *a)* when the event of interest is an intermediate endpoint and the patient's death prevents its occurrence; *b)* when one specific form of death (such as cardiovascular death) needs to be adjusted by other causes of death, and *c)* to adjust for events that are in the causal pathway between the exposure and the outcome of interest (usually a terminal event). For instance, revascularization procedures may be considered in this category by modifying the patient's natural history of the disease, and therefore, influencing the occurrence of mortality.

### *Repeated Measures With Time-to-Event Endpoint*

Many studies in biomedical research have designs that involve repeated measurements over time of a continuous variable across a group of subjects. In cardiovascular registries, for instance, information is obtained from patients at each hospitalization and collected along their entire follow-up; this information may include continuous markers, such as BNP, left ventricle ejection fraction, etc. In this setting, the researcher may be interested in modeling the marker across time, by determining its trajectory and the factors responsible for it; or perhaps the effect of the marker on mortality becomes the main interest; or the marker may be simply used as a time-varying adjuster for a treatment indicated at baseline. A frequent and serious problem in such studies is the occurrence of missing data, which in many cases is due to a patient's death, leading to a premature termination of the series of repeated measurements. This mechanism of missingness has been called informative censoring (or informative drop-out), and requires special statistical methodology for analysis.[11–14] This type of approach is called Joint modeling regression, and started being implemented in standard statistical softwares.[15,16]

In a different setting, when the aim is to describe a process in which a subject moves through a series of states in continuous time, Multistate Markov modeling becomes the right analytical tool.[17,18] The natural history of many chronic diseases can be represented by a series of successive stages, and with an "absorbing state" at the end of the follow-up (usually death). Within this model, patients may advance into or recover from adjacent disease stages, or die, allowing the researcher to determine transition probabilities between stages, the factors influencing such transitions, and the predictive role of each intermediate stage on death.

## DATA MANIPULATION

Not infrequently, the data require clean-up before fitting the model. Three important areas need to be considered here:

1. Missing data. This is a ubiquitous problem in health science research. Three types of missingness mechanisms have been distinguished[19]: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR) (Table 3). Multiple imputation was developed for dealing with missing data under MAR and MCAR assumptions by replacing missing values with a set of plausible values based on auxiliary information available on the data. Despite the fact that in most cases the missing mechanism is untestable and seldom MCAR, most contemporary statisticians are in favor of imputing missing values with complex multiple imputation algorithms, particularly when the missingness is equal to or greater than 5%.
2. Variables coding. Variables must be modeled under appropriate coding. Try to collapse categories for an ordered variable if data reduction is needed. Keep variables continuous, as much as possible, since their categorization (or even worse, their dichotomization) would lead to an important loss of prediction information, to say nothing about the arbitrariness of the chosen cutpoint. Therefore, in the case of variable dichotomization, provide arguments on how the threshold was chosen, or if it was based on an acceptable cutpoint in the medical field.
3. Check for overly influential observations. When evaluating the adequacy of the fit model, it is important to determine if any observation has a disproportionate influence on the estimated parameters, through influence or leverage analysis. Unfortunately, there is no firm guidance regarding how to treat influential observations. A careful examination of the corresponding data sources may be needed to identify the origin of the influence.

**Table 3**
Missing Mechanisms

| Mechanism | Description | Example | Effects |
|---|---|---|---|
| MCAR | Probability of missing not related either to observed or unobserved data | Accidental loss of patient records by a fire. Patient lost to follow-up because of new job | Loss of statistical power. No bias on estimated parameters |
| MAR | Given the observed data, the probability of missingness does not depend on unobserved data | Missingness related to known patient characteristics, time, place, or outcome | Loss of statistical power and bias with CC analysis on data with > 25% missingness.[20] Bias can be minimized with the use of multiple imputation (with missingness between 10-50%) |
| NMAR | The probability of missingness depends on unobserved variables and/or missing values in the data | Missingness related to the value of the predictor, or characteristics not available in the analysis | Loss of statistical power. Bias cannot be reduced in this case, and sensitivity analyses have to be conducted under various NMAR assumptions |

CC, completed case; MAR, missing at random; MCAR, missing completely at random; NMAR, not missing at random.

## MODEL BUILDING STRATEGIES

Variable selection is a crucial step in the process of model creation (Table 1). Including in the model the right variables is a process heavily influenced by the prespecified balance between complexity and parsimony (Table 3). Predictive models should include those variables that reflect the pattern in the population from which our sample was drawn. Here, what matters is the information that the model as a whole represents. For effect estimation, however, a fitted model that reflects the idiosyncrasy of the data is acceptable as long as the estimated parameters are corrected for overfitting.

Overfitting is a term used to describe a model fitted with too many degrees of freedom with respect to the number of observations (or events for binary models). It usually occurs when the model includes too many predictors and/or complicated relations between the predictors and the response (such as interactions, complex nonlinear effects) that may indeed exist in the sample, but not in the population. As a consequence, predictions from the overfitted model will not likely replicate in a new sample, some selected predictors may be spuriously associated to the response variable, and regression coefficients will be biased against the null (over-optimism). In other words, if you put too many predictors in a model, you are very likely to get something that looks important regardless of whether there is anything important going on in the population. There are a variety of rules of thumb to approximate the sample size according to the number of predictors. In linear multiple regression, a minimum of 10 to 15 observations per predictor has been recommended.[21] For survival models, the number of events is the limiting factor (10 to 15).[22] For logistic regression, if the number of non-events is smaller than the number of events, then it will become the number to be used. In simulation studies, 10 to 15 events per variable were the optimal ratio.[23,24]

Additional measures have been proposed to correct for overfitting: *a)* use subject matter expertise to eliminate unimportant variables; *b)* eliminate variables whose distributions are too narrow; *c)* eliminate predictors with a high number of missing values; *d)* apply shrinkage and penalization techniques on the regression coefficients, and *e)* try to group, by measures of similarity, several variables into one, either by using multivariate statistical techniques, an already validated score, or an estimated propensity score.

### Automatic Selection of Variables

Most statistical software offers an option to automatically select the "best model" by sequentially entering into and/or removing predictor variables. In forward selection, the initial model comprises only a constant, and at each subsequent step the variable that leads to the greatest (and significant) improvement in the fit is added to the model. In backward deletion, the initial model is the full model including all variables, and at each step a variable is excluded when its exclusion leads to the smallest (nonsignificant) decrease in the model fit. A "combination" approach is also possible, which begins with forward selection but after the inclusion of the second variable it tests at each step whether a variable already included can be dropped from the model without a significant decrease in the model fit. The final model of each of these stepwise procedures should include a set of the predictor variables that best explains the response.

The use of stepwise procedures has been criticized on multiple grounds.[2,24–27] Stepwise methods frequently fail by not including all variables that actually have influence on the response, or by selecting those with no influence at all. Relaxing the $P = .05$ value used as the stopping rule improves the selection of important variables in small datasets.[1] Stepwise procedures have also been associated with icreased probability of finding at least one of the variables significant by chance (type I error), due to multiple testing and no error-level adjustment. A forward selection with 10 predictor variables performs 10 significance tests in the first step, 9 significance tests in the second step, and so on, and each time includes a variable in the model when it reaches the specified criterion. In addition, stepwise procedures tend to be unstable, meaning that only slight changes in the data can lead to different results as to which variables are included in the final model and the sequence in which they are entered. Therefore, these procedures are not appropriate for ranking the relative importance of a predictor within the model.

To overcome the drawbacks associated with stepwise procedures, Royston[28] has developed a procedure called multivariable fractional polynomials (MFP), which includes two algorithms for fractional polynomials model selection, both of which combine backward elimination with the selection of a fractional polynomials function for each continuous predictor. It starts by transforming the variable from the most complex permitted fractional polynomials,and then, in an attempt to simplify the model, reduces the degree of freedom toward 1 (or linear). During this linearization process, the optimal transformation of the variable is set by statistical testing. It has been claimed that the default algorithm resembles a closed-test procedure by maintaining the overall type I error rate at the prespecified nominal level (usually 5%). The alternative algorithm available in MFP, the sequential algorithm, is associated with an overall type I error rate about twice that of the closed-test procedure, although it is believed that this inflated type I error rate confers increased power to detect nonlinear relationships.

Other sophisticated alternatives have been proposed to improve the process of variable selection, such as best subset regression,[29] stepwise techniques on multiple imputed data,[30,31] using automated selection algorithms that make appropriate corrections, such as the LASSO (least absolute shrinkage and selection operator) method,[32] and maximum likelihood penalization.[33]

Recently, the use of bootstrap resampling techniques has been advocated as a means to evaluate the degree of stability of the models resulting from stepwise procedures.[34–37] Such bootstrap samples mimic the structure of the data at hand. The frequency of the variables selected in each sample, called bootstrap inclusion fractions (BIF), could be interpreted as criterion for the importance of a variable. A variable that is weakly correlated with others and significant in the full model should be selected in about half of bootstrap samples (BIF≥50%). With lower *P*-values, the BIF increases toward 100%.

## FINAL MODEL EVALUATION

Central to the idea of building a regression model is the question of model performance assessment. A number of model performance metrics have been proposed, although they can be grouped in two main categories: calibration and discrimination measures (Tables 1 and 3). Independent of the aim for which the model was created, these two performance measures need to be derived from the data which gave origin to the model, or even better, through bootstrap resampling (known as internal validity). With bootstrapping, it is possible to quantify the degree of over-optimism and the amount of shrinkage necessary to correct the model's coefficients. However, if the goal is to evaluate the model's generalizability, a crucial aspect in prediction models, then these performance measures need to be estimated on external data.

However, this is not always possible due to lack of resources. For a comprehensive review of this topic, refer to "Clinical Prediction Models" by Steyerberg.[38]

Calibration refers to the agreement between observed outcomes and model predictions. In other words, it is the ability of the model to produce unbiased estimates of the probability of the outcome. The most common calibration measures are calibration-in-the-large, calibration slope (both derived from calibration plots), and the Hosmer-Lemeshow test (or its equivalent for Cox regression, the Gronnesby and Borgan test[39]).

Discrimination is the model's ability to assign the right outcome to a randomly selected pair of subjects; in other words, it allows the model to classify subjects in a binary prediction outcome setting. The area under the ROC curve (AUC) is the most common performance measure used in the evaluation of the discriminative ability for normal-error models with a binary outcome. The equivalent for censored data is the C-statistic.[40]

In the context of translational research (omics & biomarkers era), the evaluation of the added predictive value for a predictor is perhaps as important as the validation of the prediction accuracy of the model as a whole. Several approaches have been proposed, the most important being net reclassification improvement, integrated discrimination improvement,[41,42] and decision curve analysis.[43]

In summary, a good model calibration and discrimination, evaluated through bootstrapping, is currently considered an important prerequisite for the application of any prediction model, assuming that independent data testing is not feasible.

## RESULTS PRESENTATION

The final consideration in the process of model creation is how the estimated parameters will be presented. Commonly, the statistical packages, when comparing two groups on a binary outcome, express the effect size of the explanatory variable in relative metrics. For logistic and Cox regressions, the OR and HR are the traditional metrics to indicate the degree of association found between a factor and the outcome. Because these are a ratio of proportions, the information conveyed about their effect size is relative to each other. Similarly, in randomized controlled trials, the relative risk, which is no more than a ratio of proportions, is often used to summarize the association between the intervention and the outcome. The main limitation inherent to these relative measures is that they are not affected by variations in baseline event rate. For instance, in Cox regression the HR does not take into account the baseline hazard function, and thus must be interpreted as a constant effect of the exposure (or intervention) during the follow-up. As such, there is a concern that relative measures provide limited clinical information, while absolute measures, by incorporating the information about the baseline risk of the event, will be more relevant for clinical decision making.

The most common absolute measures are: ARD and NNT. It must follow that NNT is just a reciprocal of the ARD. Because of randomization, in randomized controlled trials it is relatively straightforward to estimate the ARD as simply the difference in the outcome proportions, measured at the end of the trial, between treated and untreated subjects. Even when the outcome is time-to-event in nature, the differences in survival probabilities can be estimated at different durations of follow-up with the KM survival curves.[44] However, with observational studies, subjects in the 2 groups of the exposure variable often differ systematically in prognostically important baseline covariates, which in turn leads to the application of statistical methods that allow the calculation of adjusted ARD (and NNT).[45]

In summary, in observational cohort studies, ARD and NNT can be derived from adjusted logistic and Cox regression models, and as much as possible such measures should supplement the reporting of the traditional regression estimates.

## CONFLICTS OF INTEREST

None declared.

## REFERENCES


1. Steyerberg EW, Eijkemans MJ, Harrell Jr FE, Habbema JD. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Med Decis Making. 2001;21:45–56.
2. Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer-Verlag; 2001.
3. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998;97:1837–47.
4. Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. BMJ. 2009;338:b604.
5. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer; 2009.
6. Annesi I, Moreau T, Lellouch J. Efficiency of the logistic regression and Cox proportional hazards models in longitudinal studies. Stat Med. 1989;8: 1515–21.
7. Martinussen T, Scheike TH. Dynamic regression models for survival data. New York: Springer-Verlag; 2006.
8. Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. Stata J. 2009;9:265–90.
9. Pintilie M. Competing risks: a practical perspective. New York: John Wiley & Sons; 2007.
10. Fine JP, Gray RJ. A proportional hazard model for the subdistribution of a competing risk. J Am Stat Assoc. 1999;94:496–509.
11. Ibrahim JG, Chu H, Chen LM. Basic concepts and methods for joint models of longitudinal and survival data. J Clin Oncol. 2010;28:2796–801.
12. Rizopoulos D. Joint modelling of longitudinal and time-to-event data: challenges and future directions. In: 45th Scientific Meeting of the Italian Statistical Society. Padova: Università di Padova; 2010.
13. Touloumi G, Babiker AG, Kenward MG, Pocock SJ, Darbyshire JH. A comparison of two methods for the estimation of precision with incomplete longitudinal data, jointly modelled with a time-to-event outcome. Stat Med. 2003;22: 3161–75.
14. Touloumi G, Pocock SJ, Babiker AG, Darbyshire JH. Impact of missing data due to selective dropouts in cohort studies and clinical trials. Epidemiology. 2002;13:347–55.
15. Rizopoulos D. JM: An R package for the joint modelling of longitudinal and time-to-event data. J Stat Soft. 2010;35:1–33.
16. Pantazis N, Touloumi G. Analyzing longitudinal data in the presence of informative drop-out: The jmre1 command. Stata J. 2010;10:226–51.
17. Meira-Machado L, De Una-Álvarez J, Cadarso-Suárez C, Andersen PK. Multi-state models for the analysis of time-to-event data. Stat Methods Med Res. 2009;18:195–222.
18. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Stat Med. 2007;26:2389–430.
19. Little RJA, Rubin DB. Statistical Analysis with Missing Data, 2nd ed., Hoboken: J.W. Wiley & Sons; 2002.
20. Marshall FA, Altman FDG, Holder FRL. Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. BMC Med Res Methodol. 2010;10:112.
21. Green SB. How many subjects does it take to do a regression analysis? Multivar Behav Res. 1991;26:499–510.
22. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol. 1995;48:1503–10.
23. 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;49:1373–9.
24. Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol. 1999;52:935–42.
25. Altman DG, Andersen PK. Bootstrap investigation of the stability of a Cox regression model. Stat Med. 1989;8:771–83.
26. Steyerberg EW, Eijkemans MJ, Harrell Jr FE, Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med. 2000;19:1059–79.
27. Austin PC, Tu JV. Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. J Clin Epidemiol. 2004;57:1138–46.
28. Royston P, Sauerbrei W. MFP: multivariable model-building with fractional polynomials. In: Royston P, editor. Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Chichester: John Wiley & Sons; 2008. p. 79–96.

29. Roecker EB. Prediction error and its estimation for subset-selected models. Technometrics. 1991;33:459–68.
30. Vergouwe Y, Royston P, Moons KG, Altman DG. Development and validation of a prediction model with missing predictor data: a practical approach. J Clin Epidemiol. 2010;63:205–14.
31. Wood AM, White IR, Royston P. How should variable selection be performed with multiply imputed data? Stat Med. 2008;27:3227–46.
32. Tibshirani R. Regression shrinkage and selection via the LASSO. J Roy Stat Soc B Stat Meth. 2003;58:267–88.
33. Steyerberg EW. Modern estimation methods. In: Steyerberg EW, editor. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer; 2009. p. 231–40.
34. Beyene J, Atenafu EG, Hamid JS, To T, Sung L. Determining relative importance of variables in developing and validating predictive models. BMC Med Res Methodol. 2009;9:64.
35. Royston P, Sauerbrei W. Model stability. In: Royston P, editor. Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Chichester: John Wiley & Sons; 2008. p. 183–99.
36. Royston P, Sauerbrei W. Bootstrap assessment of the stability of multivariable models. Stata J. 2009;9:547–70.
37. Vergouw D, Heymans MW, Peat GM, Kuijpers T, Croft PR, De Vet HC, et al. The search for stable prognostic models in multiple imputed data sets. BMC Med Res Methodol. 2010;10:81.
38. Steyerberg EW. Evaluation of performance. In: Steyerberg EW, editor. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer; 2009. p. 255–79.
39. May S, Hosmer DW. Advances in survival analysis. In: Balakrishnana N, Rao CR, editors. Hosmer and Lemeshow type goodness-of-fit statistics for the Cox proportional hazards model.. Amsterdam: Elsevier; 2004. p. 383–94.
40. Harrell Jr FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361–87.
41. Pencina MJ, D'Agostino Sr RB, D'Agostino Jr RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. 2008;27:157–72.
42. Pencina MJ, D'Agostino Sr RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med. 2011;30:11–21.
43. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26:565–74.
44. Altman DG, Andersen PK. Calculating the number needed to treat for trials where the outcome is time to an event. BMJ. 1999;319:1492–5.
45. Laubender RP, Bender R. Estimating adjusted risk difference (RD) and number needed to treat (NNT) measures in the Cox regression model. Stat Med. 2010;29:851–9.