Nephrol Dial Transplant (2007) 22: 3422–3430
doi: 10.1093/ndt/gfm777
Advance Access publication xx xx xxxx

Special Feature

NDT
Nephrology Dialysis Transplantation

Clinical research of kidney diseases III: Principles of regression and modelling

Pietro Ravani[1,2], Patrick Parfrey[2], Veeresh Gadag[3], Fabio Malberti[1] and Brendan Barrett[2]

[1]Divisione di Nefrologia e Dialisi, Azienda Instituti Ospitalieri di Cremona, Cremona, Italy, [2]Clinical Epidemiology Unit and [3]Division of Community Health and Humanities, Faculty of Medicine, Memorial University of Newfoundland, Canada

Keywords: confounding; interaction; interval estimate; point estimate; regression models

Introduction

Inappropriate data analysis is a source of measurement error in clinical studies [1]. *Descriptive* methods (graphs, summary statistics and relational plots) are used to assess variable distributions, identify possible outliers and reveal the form of the relationship of interest. For example, in a study of hyperparathyroidism in chronic kidney disease, researchers are interested in the sample mean and standard deviation (SD) of both parathyroid hormone and kidney function levels, and in the form of their possible relationship (i.e. whether it is present across all variable levels and whether it can be described by a line, a curve, etc.). The next step is to extend the conclusions beyond the immediate sample (*inference*) and estimate, for example, the amount of parathyroid hormone increase as kidney function declines. Statistical models are used to test whether an input–output relationship is supported by observed data and assess its direction and strength [1,2]. Most researchers and consumers of clinical research are familiar with the preliminary steps of data analysis. However, there is a growing interest in filling the gap between elementary notions and more advanced knowledge. The present paper provides introductory notes on general principles of statistical modelling, including how regression methods are chosen and used to address epidemiological phenomena such as confounding and interaction.

Regression analysis

Role of statistics

The task of statistics in the analysis of epidemiological data is to distinguish between chance findings and results that may be replicated upon repetition of the study [1]. For example, if a relationship between left ventricular mass (LVM) and systolic blood pressure (SBP) exists, LVM is expected to change by a certain amount as SBP changes. Data from a recent large-scale multicentre application of cardiac magnetic resonance in people without clinical cardiovascular disease were modelled to estimate the average change in LVM per unit change in SBP. LVM was 9.6 g greater (*point estimate*), 95% confidence intervals (CI) from 8.5 to 10.8 (*interval estimate*), per each 21 mmHg (SD) higher SBP [3]. The point estimate (fit) is the explained output variability, whereas the difference between recorded and predicted values (random error) is the variability unexplained by the model. This is used to calculate the 95% CI (measure of precision).

The residual error implies that the value of the response for an individual knowing his/her SBP (and other inputs in the multivariate model) can never be predicted with certainty. For example, considering the adjusted effects of SBP (9.6 g per SD) and body mass index (BMI, 11.7 g per 5 kg/m^2), the expected LVM of a subject with SBP of 147 mmHg and BMI of 30 kg/m^2 is $9.6 \times 7 + 11.7 \times 6 = 137.4$ g [3]. This LVM may not correspond to the observed value for a subject with SBP of 147 and BMI of 30 (if that subject exists). Further information can reduce this error. However, even including several inputs into the model the 'exact' response value can never be established. In other words, some amount of variation will remain unexplained after fitting a model to any data. Measures of the explained variability in the response, such as the overall R^2 statistic in linear models or equivalent (likelihood) measures in other models, inform on the clinical relevance of the effects as opposed to their statistical significance [2].

Finally, statistics only convey the effect of the chance element in the data but can neither identify nor reduce systematic errors [1,2]. The only bias that can be controlled during statistical analyses is 'measured' confounding.

Correspondence and offprint requests to: Pietro Ravani, Divisione di Nefrologia, Azienda Istituti Ospitalieri di Cremona, Italy, Largo priori 1, Cremona, 26100, Italy. E-mail: pietro.ravani@med.mun.ca

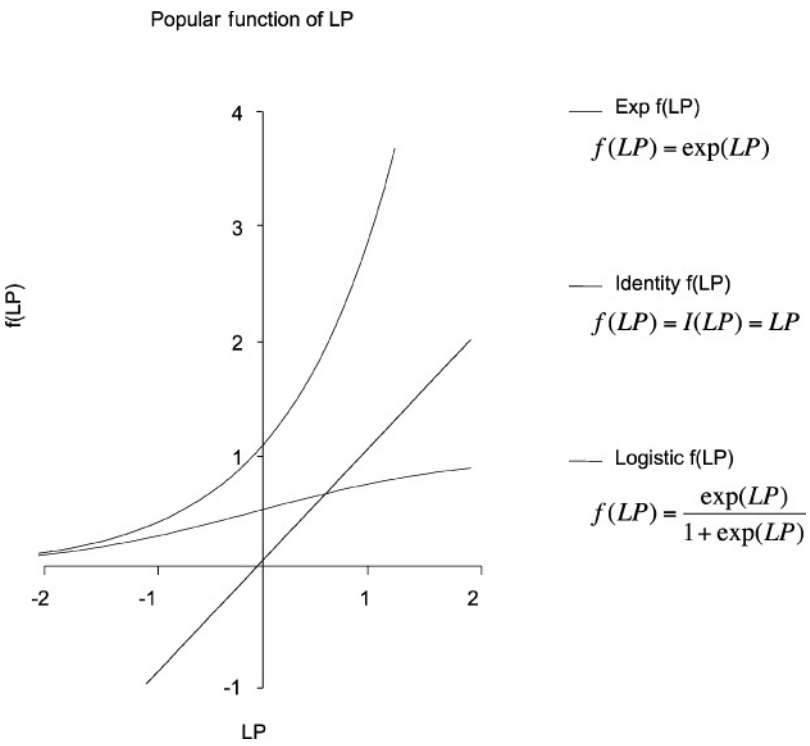

Fig. 1. Example of three common functions of the linear predictor (LP). The identity function does not change the LP (linear model) and yields graphs that are straight lines; the exponential function is the exponentiated LP (Poisson model); the logistic function is a sigmoid function of LP (logistic model). Note that different functions have not only different shapes but also different ranges.

However, the interpretation of the results would be wrong if the statistical tool is incorrect. This implies the choice of the proper function to model the data and regression technique.

Concept of function

Most clinical research can be simplified as an assessment of the relationship between exposure (*X*, independent variable) and disease (*Y*, dependent variable). For example, if the study hypothesis is that LVM depends on BMI, smoking habit, diabetes and SBP [3], then the observed values of LVM (y) are said to have a functional relationship with these four variables (x_1, x_2, x_3 and x_4). This implies a link between input(s) and output.

A function (equation) can be thought of as a 'machine' transforming some ingredients (inputs) into a final product (output). Technically the ingredients on which a function operates are the 'argument' of that function. Just as any machine produces a specific output and has a typical shape and characteristics, similarly any function has its specific response variable and a typical mathematical form and graphical shape.

A special 'ingredient' is the *linear predictor* (LP), which is the 'argument' of most statistical functions of interest to clinical epidemiology. LP contains one or more inputs (the 'Xs') combined in linear fashion (i.e. LP is a *linear function of the Xs*). Figure 1 shows three important functions of the LP: the identity function, which does not modify its argument and gives LP as output; the exponential function of LP and the logistic function of LP. The underlying mathematical structure is not important here. However, two aspects should be noted: first, different transformations change the 'shape' of the relationship between LP (input) and its function (output); second, although LP can range from $-\infty$ to $+\infty$ (allowing any type of input to be accommodated into it), its function can be constrained into a range between 0 and 1 (logistic function); it can have a lower limit of 0 (exponential function) or can just have the same range as the LP (identity function). These aspects are crucial as the model choice is based on the distribution of the response. Different responses require different ways of modelling LP.

Regression methods

Once a function has been chosen to describe the relationship of interest, its coefficients are estimated using regression strategies. Regression and correlation are often confused as they measure the degree of relationship between two or more variables in two related but different ways. Correlation (or more generally, *covariation*) measures the degree of association between two variables without distinction between input(s) and output. The variables can be two inputs or two outcome measures in the same subject. In *regression* analysis the output is modelled as a function of one or more inputs to predict its future values.

The term regression implies the tendency towards an average value. For example, if there is a linear relationship between age and 5-year mortality, the average change in mortality per unit change in age can be estimated using linear regression. This estimation task is accomplished by obtaining specific values (*estimates*) for the 'unknowns' (*parameters*) of the specific regression function. In the

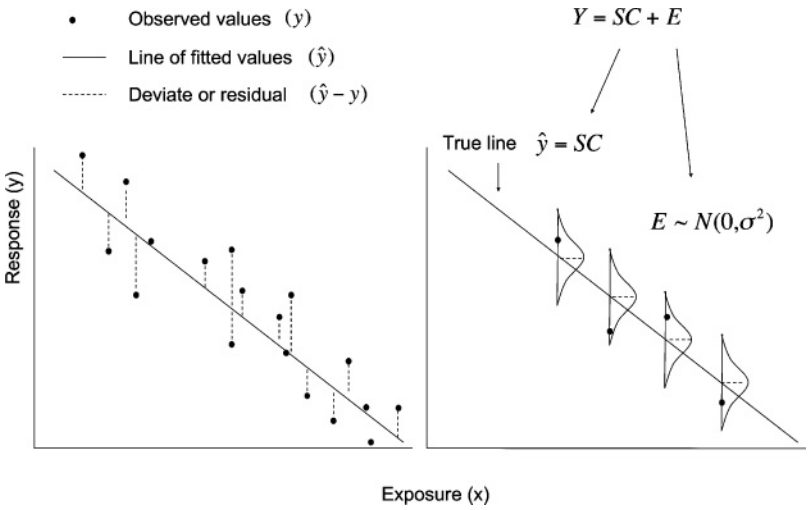

Fig. 2. Ordinary least-squares method and components of a statistical model. The regression line drawn through a scatter plot of two variables is 'the best fitting line' of the response (left). In fact, this line is as close to the points as possible providing the 'least sum of squares' deviates or residuals (vertical dashed lines). These discrepancies are the differences between each observation ('*y*') and the fitted value corresponding to a given exposure value ("$\hat{y}$"). The statistical model of the response (*Y*) includes a systematic component (SC) corresponding to the regression line (the linear predictor LP in linear regression or some transformation of the LP in other models) and an error term (*E*, right) characterized by some known distribution (for the linear model the distribution is normal, with mean = 0 and constant variance = σ^2).

above example, the linear function of mortality is mortality = LP + ε = $\beta_0 + \beta_{age} \times$ age + ε, where LP is the linear function $\beta_0 + \beta_x \times x$ (fit) and ε is the variability in the data unexplained by the model. The parameters of this univariable model are β_0, representing the intercept of the line describing the input–output relationship, and β_x, representing its slope (average change in mortality per year of age). For each 'β' the model provides a point estimate and 95% CI.

Different regression methods exist. The method commonly used in linear regression, for example, is the *ordinary least-squares method* (OLS). In lay words this method chooses the values of the function parameters (β_0, β_{age}) that minimize the distance between the observed values of the response *y* and their mean per unit of *x* (thus minimizing 'ε'). Graphically this corresponds to finding a line on the Cartesian axes passing through the observed points and minimizing their distance from the line of the average values towards which the observed measures are 'regressed' (Figure 2, left). Other estimation methods exist for other types of data, the most important of which is *maximum likelihood estimation* (MLE). As opposed to OLS, MLE works well for both normally (Gaussian) and non-normally distributed responses (for example Binomial or Poisson). However, all estimation procedures choose the most likely values of the parameters given the data, those that minimize the amount of error or difference between what is observed and what is expected.

Statistical models

Definition

Models are representations of essential structures of objects or real processes. For example, the earth may be approximated to a sphere in geographic calculations although it is flattened at the poles. Given a reasonably linear relationship between kidney function and haemoglobin concentration, a linear model may be used to study anaemia in chronic kidney diseases. Even in the presence of mild deviations from ideal circumstances, the representation of a process by means of a simple model, such as the linear model, helps grasp the intimate nature and mechanisms of that process. Obviously, critical violations of model assumptions would make the model inappropriate. The linear model would be wrong if the relationship was exponential. Similarly, the sphere would not be an acceptable model if the earth were a cone. A useful model is a good compromise between appropriateness of the chosen function and interpretability of the effects (β), while leaving as little unexplained variability in the data as possible (ε).

Indeed biologic phenomena, as opposed to deterministic phenomena of physics or chemistry, are characterized by considerable variability, yielding different results when repeated in the same experimental conditions. Probabilistic rather than deterministic models are applied to biomedical sciences as they include indexes of uncertainty around the population parameters estimated using samples (e.g. 95% CI). These characteristics of statistical models are reflected in their fit and random components. For example, the fit portion of a linear model is a line, and the errors are distributed normally with mean equal to zero (Figure 2, right). In other models the fit portion has different shapes and the residuals have different distribution.

Model choice

The most appropriate statistical model to fit the data depends on the type of response variable because this determines the shape of the systematic portion and the typical error distribution. Previous literature often provides useful information to guide on the model choice before data

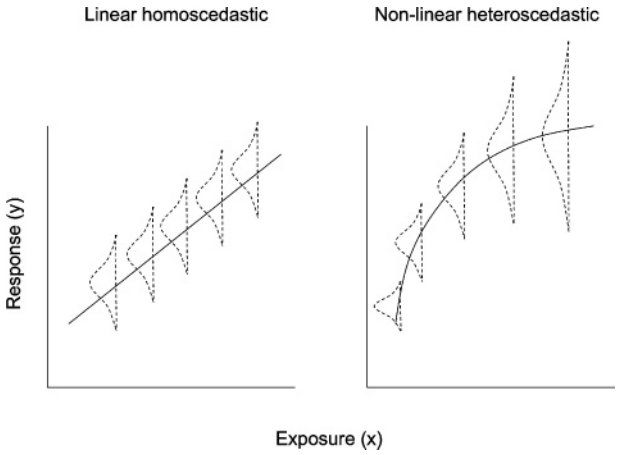

Fig. 3. Linearity and equal variance. In the left panel, the response is linearly related to the exposure and has constant variance (homoscedasticity). In the right plot, two possible important violations are depicted: non-linearity and unequal variance (heteroscedasticity).

are collected. Once the model has been built, its systematic and random components are verified graphically, using formal tests based on residuals in order to ensure that the chosen model fits the data well. These procedures are called model checks (including influential observations and outliers) and assumption verification. They will not be discussed in this paper. However, three major principles can be summarized using the linear model as an example.

First, the relationship between input and output must reflect the mathematical form of that model. For example, to use the linear model the relationship must be linear. In other models, the functional form of the relationship is describable by other curve shapes (Figure 1) and the meaning of the parameters is different. This 'shape' assumption pertains to the systematic component (Figure 3). The other two conditions to use a statistical model pertain to its random component. First, the residuals must follow a distribution compatible with the specific model: *normal* in linear regression, *binomial* in logistic regression and *Poisson* in Poisson regression. For example, in a study of asymmetric dimethylarginine (ADMA) and glomerular filtration rate (GFR), the observed GFR was approximately symmetrically distributed above and below the fitted line of GFR over ADMA (error mean = zero), with equal (constant) variance along the whole line [4]. Second, the residuals must be *independent*. This is possible only if the observations are independent. This condition is violated if repeated measures are taken on the same subjects or if there are clusters in the data, i.e. some individuals sharing some experience/conditions that make them not fully independent. Consequently, once some measurements have been made, it becomes possible to more accurately 'guess' the values of further measurements within the same individual/cluster, and the corresponding errors are no longer due to chance alone. This final assumption must be satisfied in the study design. In the presence of correlation, appropriate statistical techniques are required.

When the necessary conditions to use a certain model are clearly violated, they can be carefully diagnosed and treated. For instance, often non-linearity and unstable variance of a continuous response can be at least partially corrected by some mathematical transformations of the output and/or the inputs in order to permit use of the linear model. Urinary protein excretion, for example, is often log-transformed both when it is treated as output [5] and input [6]. However, any data transformation changes the meaning of the model parameters and their interpretation may become obscure. Reports often fail to explain clearly the meaning of the parameters of some complex models [5–8].

These three conditions have the following meaning. Once the model has been fitted to the data (1) it must be possible to quantify the amount of change of the output per unit change of the input(s), i.e. the parameter estimates are *constant* and apply over the whole range of the predictors; (2) what remains to be explained around the fit is unknown independent of the input(s) values and (3) the measurement process. For more detailed discussion on applied regression the interested reader is referred to specific texts [9].

Multivariable versus univariable analysis

Multiple regression models contain more than one input. Therefore, they estimate more effects simultaneously. A graphical approach using the linear model may help understand this concept.

When only the response variable is considered, e.g. the overall mean and standard deviation of LVM [3], the largest possible variability is observed in the data (*unconditional* response). The output variability becomes smaller if the response is studied as a function of one input at a time or, better, two inputs at the same time (*conditional* distribution of the response). As the systematic component of a multivariable model contains more information on the variability of the response, the amount of unexplained variability gets smaller (Figure 4, left). The intercept and standard error of the model without input variables ('null model', i.e. $y = \beta_0 + \varepsilon$) are the parameters of the unconditional distribution of the response (mean and SD).

Figure 4 (right) shows the multidimensional consequences of introducing more inputs. With two quantitative predictors such as SBP and BMI, the fitted values of LVM lie on a plane in the three-dimensional space, the plane that minimizes the residuals. The addition of a third quantitative variable would create a hyper-plane in the multidimensional space and so on. Of note, qualitative inputs, such as diabetes, separate the fitted values on more planes, one per each level of the independent variable. This plane would have some sophisticated shape in other models, but the multidimensional meaning of multivariable analysis would be the same.

Modelling issues

Confounding

Definition. A *confounder* is an 'extraneous' variable associated with both exposure and response without lying in the pathway between them, (Figure 5). Conversely, a *marker* or proxy is only related to the exposure, an *intermediate variable* explains the outcome, and two inputs are *colinear* when they carry the same or at least similar information. For

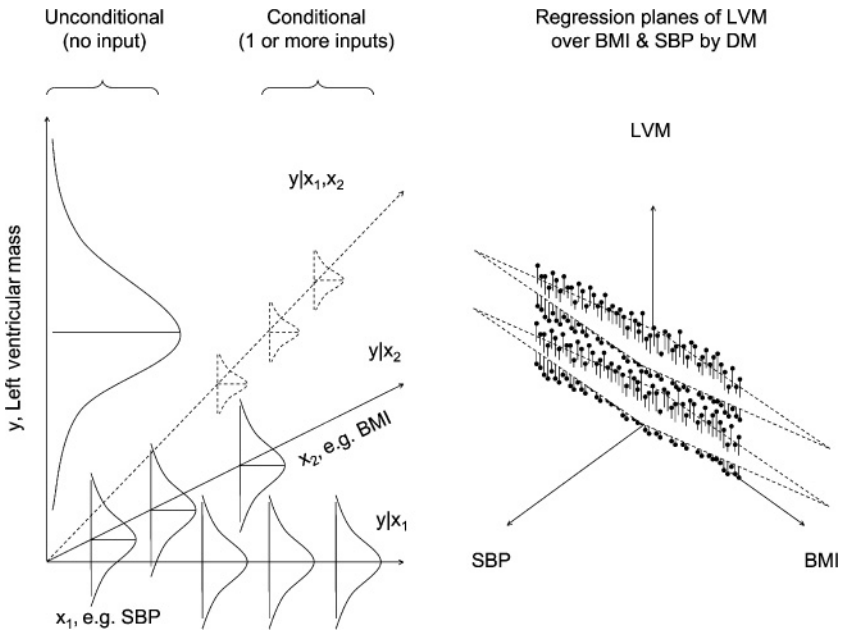

Fig. 4. Information gain and residual variance and three-dimensional representation of the linear model. The residual variance of left ventricular mass (LVM) gets progressively smaller in comparison to the unconditional response (distribution of the response without any knowledge about exposure) as more informative inputs are introduced into the model. The inputs are systolic blood pressure, SBP (x_1, in mmHg) and body mass index, BMI (x_2, in kg/m^2). These quantitative predictors generate a plane in the three-dimensional space. The number of fitted planes increases with the number of levels of a qualitative input (e.g. diabetes, DM).

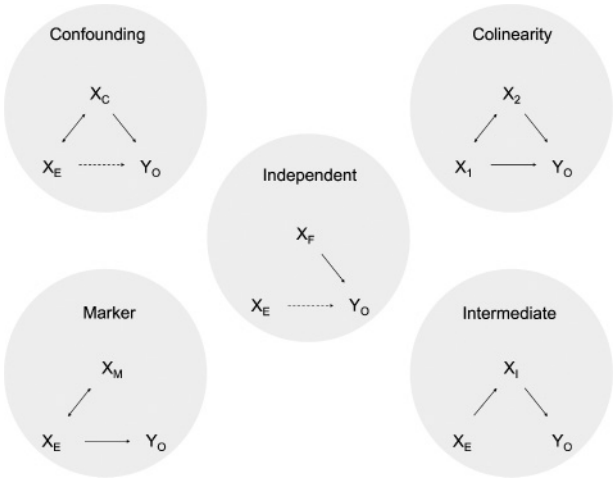

Fig. 5. Possible relationships between input and outcome. A confounding factor (X_C) is independently associated with the response (Y_O) even in the absence of the exposure of interest (X_E), and with X_E without being the consequence of X_E (e.g. in a study of coffee drinking, X_E, and coronary artery disease, CAD, smoking, X_C, is an independent risk factor for CAD, and related to coffee drinking without being in the causal chain between them). A marker or proxy (X_M) is associated with the exposure only and has no direct relationship with the outcome (e.g. yellow fingers are related to smoking but not to CAD). Two inputs (X_E and X_F) may also have an independent association with the response; this is the ideal situation as it maximizes the information in the data (e.g. both male gender and age are independently related to CAD). Colinearity is a phenomenon whereby two inputs (X_1 and X_2) carry (at least partially) the same information on the response (e.g. age and kidney function estimates including age). An intermediate variable (X_I) lies in the pathological path leading to the outcome (e.g. cholesterol levels in the causal chain between diet and CAD). Inclusion of X_I in a multivariable model can be useful to assess the amount of change in the estimated effect of X_E ($\hat{\beta}_E$) on Y_O mediated by X_I. However, the adjusted $\hat{\beta}_E$ is biased as (at least) part of the effect of X_E is due to an effect of X_E on X_I rather than confounding.

example, in a recent ultrasound study of renal resistance indices (RI) in chronic kidney disease, carotid intima–media thickness was significantly associated with RI in baseline models that did not include age [10]. However, older patients had thicker carotid artery walls and once age was entered into the model intima–media thickness lost its predictive power (confounding by age). In the final model of RI the introduction of phosphate 'lowered' the coefficient of GFR (another input). However, phosphate increase may be one mechanism through which kidney function reduction contributes to higher RI, i.e. be in the causal chain between exposure and response (intermediate variable). Adjustment of the effect of GFR for phosphate (which is affected by GFR) may be biased, as it does not merely reflect confounding. Although modelling strategies help identify multiple relationships, their direction and temporal sequence should be made explicit in the design and ideally tested in experimental studies [1,2]. A further challenge of longitudinal data is that some covariates may play the dual role of confounders and intermediates over time [18,19]. For example, in a study of the effect of obesity on mortality, the development of clinical cardiac or respiratory disease is an independent predictor of both mortality and subsequent weight loss and is influenced by prior weight gain. In a study of anti-proteinuric agents and mortality, the time-dependent covariate proteinuria is both an independent predictor of survival and initiation of therapy and is itself influenced by prior treatment.

Control. Confounding can be prevented using randomization, restriction or matching in the design phase [1,2] and controlled through stratification or modelling during analysis (Table 1). Stratification refers to cross-tabulation

Table 1. Confounding and interaction in multiple regression models

	Confounding	Interaction
Definition	Spurious exposure-response association due to a lurking variable (belongs to study)	Reciprocal effect modification of two (or more[a]) inputs (belongs to nature)
Mechanism	Differential distribution of the confounder by level of the exposure and its independent association with the response	Reciprocal effect strengthening (synergism) or weakening (antagonism) of two inputs (biologic mechanism)
Consequences	Biased estimate of the exposure effect (in either direction) unless the confounding variable is accounted for	Estimates varying by combination of interacting variable levels; ignoring interaction is wrong even if the estimate of the main effects is correct
Identification	Study of (1) the distribution of the confounding variable by level of the exposure, (2) its association with the response and (3) the effect change of the exposure considering the confounder (stratification)	The formal test of interaction is based on the creation of a product term obtained by multiplying the two main terms [7,11]; the interaction effect is tested in the presence of the main effects (even if not statistically significant)
Prevention at the time of study design	Restriction of subjects to one level of known confounding (inefficient); randomization to evenly distribute known/unknown confounders [1,2]; matching or equal representation of subjects with known confounders in study groups (to be coupled with matched analysis)	Hypothesized at the time of study design based on biological plausibility
Treatment during data analysis	Stratification: information aggregated with pooling or standardization [12]; classification of patients on levels of propensity score [13]. Modelling: confounding variable kept in the model, even if its effect is not significant, provided that the confounding phenomenon is of clinical relevance[b]	Interaction terms kept in the model if statistically significant, considering a more generous P value of 0.1

[a]When the interacting variables are more than two, there are higher order product terms in addition to the first-order interaction terms, e.g. AB, AC, BC, ABC when the main terms are A, B, C, etc.
[b]The amount of the acceptable change in the exposure effect in the presence or absence of the confounder in the model can be a matter of debate and the adopted policy should be explained in the reporting.

Table 2. Hypothetical cardiac event data by the presence of diabetes (DM versus non-diabetics, ND) with age ($A < 65$ versus $A \geq 65$) as a potential confounder

	Age				Total	
	$A < 65$		$A \geq 65$		(without considering age)	
	DM	ND	DM	ND	DM	ND
Cases	8	4	24	10	32	14
Controls	98	116	74	75	172	191
OR		2.36		2.43		2.53

Age is a potential confounder of the relationship between diabetes (exposure) and cardiac events (response) as (1) there are more diabetics among the elderly; (2) it is independently associated with the response; and (3) it is not a consequence of diabetes. In fact the crude odds ratio (OR) associated with DM is 2.53 which overestimates the stratum-specific OR estimates and the adjusted (pooled Mantel–Haenszel) $\mathrm{OR_{MH}} = \left[\frac{8\times116}{226} + \frac{24\times75}{183}\right] \Big/ \left[\frac{4\times98}{226} + \frac{10\times74}{183}\right] = 2.41$, 95% CI from 1.22 to 4.77. Results of logistic regression are essentially the same (2.41, 95% CI 1.23 to 4.73). The number of strata increases (and their size decreases) with the number of potential confounders. Multivariable analysis is more efficient and avoids categorization of continuous variables. However, stratified analysis helps study data distribution by levels of key variables and guide modelling choice [12].

of data on exposure and response by categories of one or more potential confounders (Table 2). Adjusted estimates are obtained aggregating stratum specific information using pooling or standardization [12]. The way multivariable regression removes the association between confounder and outcome (the necessary condition for confounding) is straightforward. Consider the following linear model of the response (Y) including exposure (E) and a confounder (C): $Y = \beta_0 + \beta_E E + \beta_C C + \varepsilon$. The difference between Y and the effect of C left in the model gives the effect of the E: $y - \beta_C C = \beta_0 + \beta_E E + \varepsilon$. The right-hand part of the equation is a simple regression. The same applies to other models. This is the epidemiological concept of independence: 'independent' means purified from the effects of other inputs kept in the model.

Interaction

Definition. An interaction between two inputs is a *modification of the effect* of one input in the presence of the other (Table 1). For example, in the CARE trial, inflammation was associated with higher progression rate of kidney disease, whereas pravastatin treatment was associated with slower progression rate only in the presence of inflammation [11]. Inflammation modified the effect of pravastatin (and vice versa). The interacting variables (*main terms*) can be of the same type (qualitative or quantitative) or different type. The interaction effect can be qualitative (antagonism) or quantitative (synergism). For example, in one study [7] both BMI and HbA1C were directly related to the log albumin/creatinine ratio (output) when considered separately.

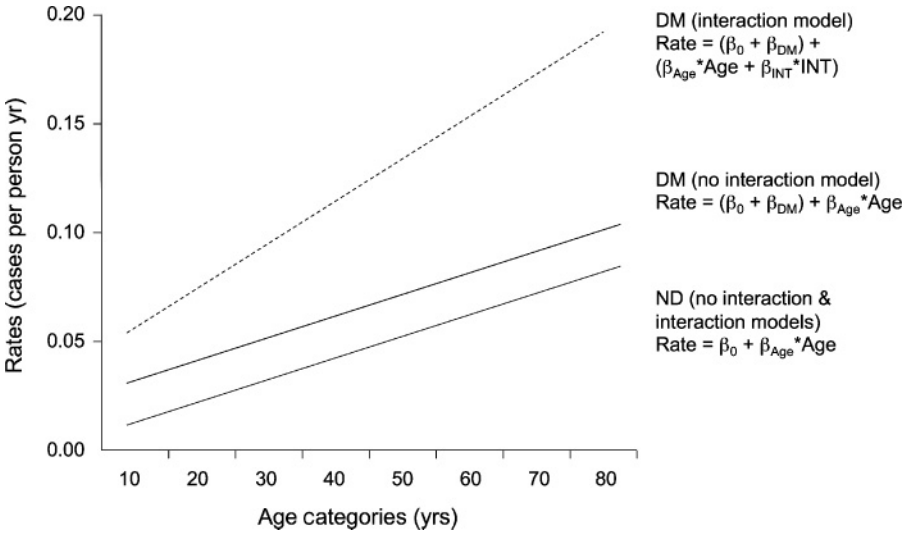

Fig. 6. Interaction parameter as a measure of the departure from the underlying form of a model. The plot shows two models of some event rate as a function of age and diabetes without interaction and with their interaction term. When diabetes is absent (ND, bottom line) the event rate is explained by age only in both models. When diabetes is present (DM) the fitted line of the event rate depends on age and diabetes according to the no interaction model (middle line) and on age, diabetes and their product (INT) in the interaction model (top line). In the no interaction model the effect of diabetes consists in shifting the event rate by a certain amount quantified by the coefficient of diabetes (change in the intercept of the line). In the interaction model, the (dashed) fitted line is not only shifted apart for the effect of diabetes but also diverging from the bottom line (absence of diabetes). The amount of change in the slope is the effect of the interaction between age and diabetes and is a measure of the departure from the underlying additive form of the model.

However, the interaction coefficient had a negative sign indicating that the total change in the response in the presence of one unit increase of both inputs was lower than the sum of the two main effects, i.e. $(0.1535 + 0.0386) - 0.0036$.

The term interaction is challenging as it describes both the biologic interdependence of two factors in exerting their effects and statistically the necessity for a new term in a model.

Statistical assessment versus epidemiological interpretation of interaction. The formal test for the presence of interaction tests whether there is a deviation from the underlying form of that model (Figure 6). For example, if the effect of age and diabetes on some event rate are respectively $\beta_{AGE} = 0.001$ (per year) and $\beta_{DM} = 0.02$ and there is no (significant) interaction, then the two fitted lines corresponding to the presence and absence of diabetes are constantly 0.02 rate units apart but have the same slope (*additive* model). Conversely, if there is an interaction effect $\beta_{INT} = 0.001$ the two lines of the interaction model are also diverging by a certain amount due to the further rate change per year of age in diabetics, graphically a difference in slope (*multiplicative* model). Statistically, the interaction coefficient estimates the amount of departure from the underlying form of the model. Epidemiologically, the coefficient of interaction is a difference between differences (in terms of LP). For example, there is a rate difference of 0.001 to consider if a subject is 1 year older and diabetic in addition to the differences of the main effects (Figure 6). In linear models, interactions between two continuous variables would change the slope of the fitted line without affecting the model intercept [7]. Interactions involving only qualitative inputs change the intercept of the line.

Table 3. Hypothetical cardiac event data expressed as an incidence rate ratio (IRR) by level of two risk factors: smoking and hypertension, where there is no interaction on a multiplicative scale but there is on an additive scale

		Hypertension	
		Absent	Present
Smoking	Absent	**1** Ref.	**10** IRR 10 (1.2, 77.6)
	Present	**5** IRR 5 (0.5, 42.5)	**50** IRR?

The 2 × 2 table shows the number of cardiac events per 1000 person-years by the presence of hypertension and smoking habit. As compared to subjects exposed to neither factor, the event rate in the presence of hypertension only is 10 times as high; in the presence of smoking is only 5 times as high and in the presence of both exposures is 50 times as high. On a multiplicative scale, there is no interaction as there is no departure from the multiplicative model (e.g. Poisson), i.e. 5 × 10 is exactly 50. Testing the parameter of the product term, the IRR is 1 (95% CI 0.5, 42.5).

However, risk ratios can be assessed also on an additive scale, where the IRR is 50 (6.9, 359). The two models have the same log-likelihood (−355.7). The only difference is the 'contrast' defining the null hypothesis. In the multiplicative model, the interaction term is a product term assuming the value of 1 for exposed to both and 0 otherwise. The null hypothesis is the absence of deviation from multiplicative risk (and of course it is not rejected). In the additive formulation, biologic interaction can be assessed using categories of covariate combination with exposed to none as reference (factored set of terms). The null hypothesis is the absence of difference on an *additive scale* (departure from additivity). The null hypothesis is rejected because the difference between exposed to both and exposed to neither prove to be larger than the sum of the other two differences, i.e. $(50 - 1) - [(10 - 1) + (5 - 1)] = 36$.

Measurement scale and biological implications. The definition of interaction as a measure of the departure from the underlying form of the model meets both statistical and

biological interpretation of the phenomenon as an amount of effect unexplained by the main terms. However, when this effect is measured, the interpretations differ, depending on the model scale [14]. In linear models, the statistical and biological perspectives coincide: input effects are (untransformed) differences. Interaction parameters are differences chosen to measure departure from an additive model: antagonistic interaction results in a change lower than expected (under-additive) [7]; synergistic interaction results in a change greater than expected (over-additive) (Figure 6). Statistical testing of this departure also measures the biologic phenomenon. Conversely, Cox's, logistic and Poisson regressions are multiplicative models because the joint effect of two or more factors is the product (rather than the sum) of their effects as LP is the argument of some non-identity function. For example, if the risk of death associated with diabetes is twice as high as in non-diabetics and is three times as high in men as in women, diabetic men have a risk six times higher than non-diabetic women. In these models effects are measured as ratios and interaction parameters are ratios chosen to measure departures from a multiplicative model: antagonistic interaction results in a change lower than expected (under-multiplicative), whereas a synergistic interaction results in a change greater than expected (over-multiplicative). Statistical assessment of this departure tests whether there is a departure from multiplicativity and not the existence of a biologic phenomenon [14]. Thus, from the statistical viewpoint, interaction depends on how the effects are measured. However, lack of evidence of deviation from the multiplicative scale supports the existence of biologic interaction, as the resulting change in the response is greater than the sum of the effects (over-additivity). This requires a biological explanation. For example, if diabetic men have a risk six times as high as non-diabetic women and the relative risks associated with the main effects are 3 and 2, there is no deviation from the multiplicative scale but there is over-additivity because $6 - 1 > (3 - 1 + 2 - 1)$. On the other hand, the choice of the model depends on the distribution of the response variable and cannot be dictated by the need to study interaction. There are ways to use multiplicative models and still assess the biological meaning of the phenomenon (Table 3 and 4).

Table 4. Hypothetical cardiac event data expressed as an incidence rate ratio (IRR) by level of two risk factors: smoking and hypertension, where there is antagonism on a multiplicative scale and synergism on an additive scale

		Hypertension	
		Absent	Present
Smoking	Absent	**1** Ref.	**7** IRR 7 (0.8, 56)
	Present	**3** IRR 3 (0.3, 28)	**14** IRR?

The 2 × 2 table shows the number of cardiac events per 1000 person-years by the presence of hypertension and smoking habit. As compared to subjects exposed to neither factor (absence of smoking and hypertension), the event rate in the presence of hypertension only is 7 times as high; in the presence of smoking is only 3 times as high and in the presence of both exposures is 14 times as high. If the risk in the interaction cell is less than multiplicative, the estimate of the risk ratio in a multiplicative model is less than 1, giving the misleading impression of a qualitative interaction (antagonism). The formal test for interaction gives an IRR of 0.6 (0.05, 7); using a factored set of terms IRR is 14 (1.8, 106). The additive model supports a quantitative interaction because the number of cases in the exposed to both group is larger than the sum of the two differences, i.e. $14 - 1 - [(3 - 1) + (7 - 1)] = 5$. The two Poisson models have the same log-likelihood of −166.7.

Analysis power

The study size should be much larger than the number of input variables in the model. Most authors recommend that there should be at least 10 to 20 times as many observations as there are coefficients in the model; otherwise the estimates are very unstable [15]. Models of binary outcomes require at least 10 events per parameter [16]. For example, age as continuous input will have one coefficient, three age categories will have two parameters (one reference category) and so on.

Reporting

The reporting of statistical methods and results in the medical literature is often suboptimal. A few tips are summarized in Table 5. More detailed checklists for reading

Table 5. Reporting statistical methods and regression results

	Methods	Results
Question	Define the relationship of interest	Report main result first
Study design[a]	Recruitment plan (subjects, centres), sample size estimation, multiple measurements and clusters	Overall description of how the study was conducted, subjects' participation and adherence
Response	How the outcome was defined and measured	Distribution of the response (unconditional)
Exposure	Main predictor of interest	Effect of the exposure
Other inputs	Other potential explanatory variables (independent, markers, intermediate), confounders, interacting variables	How and why they are or not in the model; if there is an interaction also the main terms must be in the model
Effects	Explain the epidemiological meaning of the parameters of the chosen function (difference or ratios per unit change of the input)	Summary of the relationship of interest (fit component); clinical relevance (point estimate); statistical significance (confidence intervals)
Model check	Which regression method was used, how the model was built, its assumptions verified and treated, outliers looked for and sensitivity analyses carried out	Report whether there are violations and result robustness; summarize the amount of variability of the response unexplained by the model (e.g. $1 - R^2$)

[a]Design issues have also to be addressed in the Methods and results sections [1,2].

and reporting statistical analyses are available in textbooks [17].

Acknowledgements. P.R. held a young investigator award from the Italian Society of Nephrology for the year 2005–2006 and received funding from the EU (Marie Curie Actions-OIF, proposal 021676) for the year 2006–2007.

Conflict of interest statement. None to declare.

References

1. Ravani P, Parfrey PS, Curtis B *et al*. Clinical research of kidney diseases I: researchable questions and valid answers. *Nephrol Dial Transplant* 2007; 22: 2459–68
2. Ravani P, Parfrey PS, Dicks E *et al*. Clinical research of kidney diseases II: problems of study design. *Nephrol Dial Transplant* 2007; 22: 2785–94
3. Heckbert SR, Post W, Pearson GD *et al*. Traditional cardiovascular risk factors in relation to left ventricular mass, volume, and systolic function by cardiac magnetic resonance imaging: the multiethnic study of atherosclerosis. *J Am Coll Cardiol* 2006; 48: 2285–2292
4. Ravani P, Tripepi G, Malberti F *et al*. Asymmetrical dimethylarginine predicts progression to dialysis and death in patients with chronic kidney disease: a competing risks modeling approach. *J Am Soc Nephrol* 2005; 16: 2449–2455
5. Palatini P, Mormino P, Dorigatti F *et al*. HARVEST Study Group. Glomerular hyperfiltration predicts the development of microalbuminuria in stage 1 hypertension: the HARVEST. *Kidney Int* 2006; 70: 578–584
6. Malik AR, Sultan S, Turner ST *et al*. Urinary albumin excretion is associated with impaired flow- and nitroglycerin-mediated brachial artery dilatation in hypertensive adults. *J Hum Hypertens* 2007; 21: 231–238
7. Kohler KA, McClellan WM, Ziemer DC *et al*. Risk factors for microalbuminuria in black Americans with newly diagnosed type 2 diabetes. *Am J Kidney Dis* 2000; 36: 903–913
8. Verhave JC, Hillege HL, Burgerhof JG *et al*. PREVEND Study Group: cardiovascular risk factors are differently associated with urinary albumin excretion in men and women. *J Am Soc Nephrol* 2003; 14: 1330–1335
9. Glantz SA, Slinker BK. *A Primer of Applied Regression and Analysis of Variance, 2nd edn*. New York: McGraw-Hill, 2001
10. Heine GH, Reichart B, Ulrich C *et al*. Do ultrasound renal resistance indices reflect systemic rather than renal vascular damage in chronic kidney disease? *Nephrol Dial Transplant* 2007; 22: 163–170
11. Tonelli M, Sacks F, Pfeffer M *et al*. Biomarkers of inflammation and progression of chronic kidney disease. *Kidney Int* 2005; 68: 237–245
12. Rothman KJ. Controlling confounding by stratifying data. In: *Epidemiology: An Introduction*. Oxford: Oxford University Press, 2002, 144–167
13. Winkelmayer WC, Kurth T. Propensity scores: help or hype? *Nephrol Dial Transplant* 2004; 19: 1671–1673
14. Rothman KJ. Measuring interaction. In: *Epidemiology: An Introduction*. Oxford: Oxford University Press, 2002, 168–180
15. http://www.statsoft.com/textbook/stmulreg.html
16. Hosmer DW, Lemeshow S. Special topics. In: *Applied Logistic Regression*. New York: Wiley, 2000, 339–351
17. Altman DG, Machin D, Bryant TN *et al*. (eds) *Statistics with Confidence, 2nd edn*. London: BMJ Books, 2000
18. Robins J. The control of confounding by intermediate variables. *Stat Med* 1989; 8: 679-7-1
19. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; 11: 550–560

Received for publication: 31.7.07
Accepted in revised form: 4.10.07