Clinical research of kidney diseases V: extended analytic models

Pietro Ravani1,2, Patrick Parfrey2, Veeresh Gadag3, Fabio Malberti1 and Brendan Barrett2

1Divisione di Nefrologia e Dialisi, Azienda Infermieristica F. Paganini, Cremona, Italy, 2Clinical Epidemiology Unit and 3Division of Community Health and Humanities, Faculty of Medicine, Memorial University of Newfoundland, Canada

Introduction

In some study designs the same epidaemiological unit is observed more than once. For example, in a cross-sectional study of the radial artery flow rate, several outcome values can be recorded on the same subject under different experimental conditions (e.g. exposure to different vasoactive substances). Some longitudinal studies typically monitor participants over time and both predictors (e.g. blood pressure) and outcomes (e.g. left ventricular mass index) are measured on different occasions in the same subject. In other designs, observations can fall into groups (clustered data), such as single measurements taken on a paired organ (e.g. the eye of the kidney) or single observations on different members of the same hospital/region or family. More complex designs may lead to a combination of clustering and repeated/longitudinal measurements on the same subject. In all these cases, measurement errors are no longer due to chance alone. In these situations traditional regression methods are inappropriate as they assume independent errors [1,2].

Multiple measurements on the same subject tend to be closer to each other than those obtained from different individuals. Similarly, single assessments of paired organs or members of the same hospital/region or family are correlated because different organs of the same subject and different individuals of the same community share biologic experiences, environmental exposures and genetic background. For some outcomes such as disease recurrence, previous experience may induce negative correlation. In all cases, once a measurement value has been obtained further values within the same individual/cluster can be more accurately ‘guessed’ and the corresponding measurement errors are no longer due to chance alone. In these situations traditional regression methods are inappropriate as they assume independent errors [1,2].

Two major analytical approaches exist for the analysis of correlated generalized linear (continuous, binary and count outcomes) and time-to-event data: random effect modelling and variance-corrected methods. The main assumption of these approaches is that the responses are correlated within cluster/subject, but independent between clusters/subjects.

Extended generalized linear models

Schrier et al. studied the effect of rigorous (120/80 mmHg) versus standard (135–140/85–90 mmHg) blood pressure (BP) control on left ventricular mass index (LVMI), in hypertensive subjects with the autosomal-dominant polycystic kidney disease (ADPKD) and left ventricular hypertrophy [3]. Echocardiograms were performed at baseline, at 1 and 7 years. Two analytical approaches are appropriate for this type of data: a standard linear ANCOVA model of the difference in LVMI at the study end by BP target, taking into account baseline levels (plus other inputs) and disregarding intermediate measurements; or a model for correlated outcome data, including longitudinal data. The two models answer the following questions: ‘What is the difference in LVMI after 7 years of rigorous treatment as compared to standard treatment?’ and ‘What is the (yearly) change in LVMI of rigorous and standard treatments as compared to baseline values?’. A ‘mixed model’ for longitudinal data analysis was chosen, revealing that rigorous BP control was significantly more effective in reducing LVMI. Unfortunately the paper fails to report how important biases were prevented/controlled (sampling, measurement, information, performance) and issues such as sample size estimation, model checking and coefficient estimates. For example, the reader would be interested in the effect of the assignment group (i.e. difference in the intercept at baseline); of standard treatment (slope); of rigorous intervention (difference in slope); of other covariates and interaction terms (possibly impacting the intercept and/or slope). In addition to these effects, generalized mixed models (GMM) estimate other parameters called ‘random effects’. These effects account for the variation in the response that the predictors of interest fail to explain (Figure 2). GMM include models not only for normally distributed responses as in this example (identity link function), but also for correlated responses with binomial (logit link) or Poisson (log link) distribution [1,2].

To understand the philosophy of this approach, it is useful to look at the study outcome variability as a mixture of different components [4]. The regression coefficients of the model covariates estimate the “explained variability”
Table 1. Examples of correlated (panel) data sets. Each study participant can be assessed once or in several occasions, in random sequence (repeated measures) or over time (longitudinal data). Each row in the panels represents an observation, with single measurements per subject (left) or updated values (right) of both predictors (e.g. presence of diabetes—X) and outcomes (e.g. mean arterial pressure values—Y). Predictors can be time invariant (such as gender, age at baseline or presence of diabetes—Xti) or time varying, i.e. assume different values (e.g. glucose levels—Xtv). In either case, observations can belong to clusters, such as families, schools, geographic areas or hospitals (lower row). In addition, there can be multilevel data, when clusters are nested in super-clusters, such as patients (Pt) in physicians (MD) in hospitals (H). In these situations, level 1 is the most detailed level (the single observation), level 2 the epidemiological unit (the patient) and level 3 or higher the next level of hierarchy (membership level).

<table>
<thead>
<tr>
<th>Clustering absent</th>
<th>Single measurements</th>
<th>Repeated/longitudinal data</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>MD</td>
<td>Pt</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>1</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>2</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>3</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>4</td>
</tr>
<tr>
<td>Clustering present</td>
<td>H</td>
<td>MD</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>7</td>
</tr>
</tbody>
</table>

Fig. 1. Type of correlation and epidemiological consequences. The term correlation implies the existence of some degree of association between two variables. Depending on the type of correlated variables different epidemiological phenomena may ensue [1]: multi-collinearity results from the association between different predictors (x1, x2) tested in the same outcome (y) model; confounding occurs when an independent variable (c) is associated with both the outcome and the exposure (x); outcome correlation occurs when the observations are not independent (y1, y2).
Fig. 2. Random coefficient models. Mixed models estimate both fixed and random effects. For example, fixed effects are differences in left ventricular mass index (LVMI) by blood pressure (BP) targets (continuous curves and lines). Random factors are responsible for the deviations from the average fixed effects (two subjects are indicated for simplicity with dashed curves in only one treatment group). Random effects (dashed lines) can affect either or both the intercept and the slope of the curve defining the input–output relationship. In other words, independent of the fixed effects, different subjects (in both treatment groups) may have different values of the response either or both at baseline (random intercept) and by level of exposure (random slope).

Rejected. The proportion of the total variance due to subject estimates the correlation in the data and the accuracy of the measurement tool (Figure 3).

Variance correction

Freedman et al. tested the association between log-coronary calcification score and log-albumin:creatinine ratio in 588 white participants with type 2 diabetes from 325 families [5].

‘Generalized estimating equations with exchangeable correlation and the sandwich estimator of the variance were used’ to model the association of interest controlling for the effects of extraneous variables and the correlation in the data due to the presence of familial clusters. They found that the adjusted log-calcification score was 0.1716 higher (standard error 0.0592, \( P = 0.0037 \)) per unit increase of log-albumin:creatinine ratio (going back to the original scales, this means that the expected calcification score was equal to the intercept times albumin:creatinine ratio to the power of 0.1716).

Generalized estimating equations (GEE) represent another possible approach to the analysis of correlated data. This method corrects the ‘model variance’ (i.e. the random part of the model) for the dependences in the data. To put it simply, the way the correlation in the data has ensued is taken into account for the estimation of the parameters (effects) and their standard errors. GEE have the same structure as standard regression models, i.e. a systematic component and a random component without specification of any additional random layer (Table 2). As generalized linear models [2] and GMM, GEE require the specification of one of the link functions from the generalized linear family (identity, logit, log). However, their estimation method requires the additional specification of a working correlation for the observed responses (Table 3). For example, Freedman et al. used an exchangeable correlation in their model [5]. This means that the standard errors of the GEE coefficients were corrected assuming that one single correlation coefficient (parameter) would describe the association of pairs of different responses (subjects) from the same cluster (family).

Which structure best describes the real data correlation is not always obvious, although the research design may help decide. However, GEE analysis requires only a rough estimate of this structure to get started. The final parameter estimates (fixed effects, their standard errors and the correlation \( \rho \) coefficients) are not usually dependent on the accuracy of the initial assumptions about the correlation matrix. In fact they are consistent (i.e. converge to the true value) as the number of clusters/subjects increases even if the initial structure is incorrectly specified, unless the
Table 2. Comparison of generalized linear models (GLM), generalized mixed models (GMM) and generalized estimating equations (GEE)

<table>
<thead>
<tr>
<th></th>
<th>GLM</th>
<th>GMM</th>
<th>GEE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>$g(Y) = LP + \varepsilon$</td>
<td>$g(Y) = LP + \zeta + \varepsilon$</td>
<td>$G(Y) = LP + \varepsilon$</td>
</tr>
<tr>
<td>Meaning of $\beta$</td>
<td>Population-average (marginal) effect</td>
<td>Subject-specific (conditional) effect</td>
<td>Population-average (marginal) effect</td>
</tr>
<tr>
<td>Link function</td>
<td>Specified</td>
<td>Specified</td>
<td>Specified</td>
</tr>
<tr>
<td>Estimation method</td>
<td>Maximum-likelihood</td>
<td>Maximum-likelihood</td>
<td>Quasi-likelihood</td>
</tr>
<tr>
<td>Random effects $\zeta$</td>
<td>Not implied</td>
<td>Specified</td>
<td>Not implied</td>
</tr>
<tr>
<td>Residual errors $\varepsilon$</td>
<td>Independent</td>
<td>Independent</td>
<td>Correlated</td>
</tr>
<tr>
<td>Variance of $\zeta$</td>
<td>Not estimated</td>
<td>Estimated specifying a covariance structure</td>
<td>Not estimated</td>
</tr>
<tr>
<td>Variance of $\varepsilon$</td>
<td>Estimated without constrains</td>
<td>Estimated specifying a covariance structure</td>
<td>Estimated specifying a correlation structure</td>
</tr>
<tr>
<td>Robust method for the variance of $\varepsilon$</td>
<td>Can be used</td>
<td>Not used</td>
<td>Recommended</td>
</tr>
</tbody>
</table>

*See [2] for the meaning of the link function and error distributions from the exponential family.*

Fig. 3. Between (B) versus within (W) subject (j) correlation and variance components. Blood flow rate values recorded on two occasions in the same subject are correlated as they (1) tend to lie on the same side of the overall mean (linear predictor, LP), and (2) be closer to each other than those taken on different individuals (within subject variability < between subject variability). The response value ($y_{ij}$) of subject ($j$) in occasion ($i$) equals LP plus an error term ($\varepsilon_{ij}$). This includes two components: the variability due to subject ‘$j$’ (random effect $\zeta_j$ equal to the difference between the subject mean ($\mu_j$) and LP) and the variability due to measurement on occasion ‘$i$’ (effect of occasion $\varepsilon_j$ nested in subject equal to the difference between $\mu_j$ and each response measured on $j$, $y_{ij}$). Usually it is assumed that both these components are normally distributed ($\sim N$) with mean zero and some non-zero variance ($\psi$ and $\theta$). The intra-class correlation coefficient $\rho$ estimates the outcome correlation as the proportion of the total variance explained by the random effect ($\zeta_j$).

fraction of missing data is large or they are not missing at random. Although the correlation structure is not necessarily the same for all clusters/subjects, GEE assume one set of $\rho$ parameters common to all clusters/subjects to avoid estimating too many parameters.

Given the importance of the chosen correlation structure and the possibility of misspecification in real-life situations, a special method called 'sandwich estimator' is used to estimate the standard errors of the coefficients in these models. This method corrects the variance incorporating the dependences in the process of computations by removing one cluster at a time, and providing an honest estimate for correlated data whenever the observations left out at any step are independent of the observations left in. The standard errors of the coefficients are usually (but not always) larger, depending on the sign of the correlation in the
Table 3. Examples of correlation structures used in GEE. Each panel represents a correlation matrix between any two of four possible observations in the same subject taken at times 1, 2, 3, 4 (or in subjects 1, 2, 3, 4 of the same cluster). Each symmetric matrix has a value of 1 along the main diagonal (as each measure correlates perfectly with itself) and some non-1 value off the diagonal. In the absence of correlation (independent errors) the correlation structure is independent (identity matrix); it is exchangeable if there is only one parameter $\rho$ for any pair of measurements (e.g. in a cross-sectional study the order of measurement is arbitrary and it may be assumed that any two responses within a cluster have the same correlation); unstructured if there are $n \times (n-1)/2$ different parameters $\rho$ (e.g. if it is assumed that there are as many $\rho$ parameters as there are paired combinations of $n$ measurements); autoregressive if there is only one $\rho$ parameter raised to the power of the absolute difference between the response times (e.g. in longitudinal designs the order of measurement is not arbitrary and it is reasonable to assumed that the correlation is greater for observations taken closer in time than farther apart).

Data. Put simply, the statistical testing is more conservative (the confidence intervals larger) as compared to the corresponding generalized linear model applied to the same data as though each observation was independent (independent correlation structure). This empirical method is called robust because the variance estimation is consistent, even if the chosen correlation structure is incorrect (robustness to misspecifications) [6–8].

Model choice

The choice of the analytical tool for correlated generalized linear data can be guided by different considerations. As opposed to GMM, GEE are based on only one level of clustering, are not designed for inferences about the covariance structure (the working correlation structure is formulated with no distributional assumptions) and do not give predicted response values for each cluster. Using GMM involves making extra assumptions, but gives more efficient estimates, and allows estimating contributions to variability from different sources, including multilevel correlations.

Finally, GEE are marginal models as they assume a model holding over all clusters (population average). Therefore, the coefficients represent the average change in the response over the entire population for a unit change in the predictor. GMM are conditional models in that they assume a model specific to each cluster/subject. Therefore, the coefficients represent the average change in the response for each cluster/individual, given a unit change in the predictor. Although population effects can be derived averaging cluster effects, conditional models are most useful when the objective is to make inferences about clusters/individuals rather than the population.

Extended survival models

Correlated survival times

Correlation in the occurrence and timing of repeated events may occur when individuals experiencing a single event belong to groups or clusters, or where the subject experiences some event more than once due to a recurrent event process [9]. The correlation in the survival times may result from differences in the general tendency to fail across individuals and varying tendency to fail further once the recurrence process has started (Figure 4). Heterogeneity across subjects (unshared frailty) may be due to unknown, unmeasured or unmeasurable effects (different lifestyles, genetic traits, environmental factors and experiences), which influence the likelihood to succumb to disease. As a result, some individuals are more (and others less) prone to disease, experiencing their first, second, third, etc., recurrent episode more (less) quickly than others. Event dependence within a subject emerges when the threshold for further events changes once previous events have occurred (e.g. the baseline risk of thrombosis of the second and third bypass graft is progressively higher or lower than that of the first). Further events become more or less likely according to whether the process induces a biological weakening or strengthening of the organism and whether the subject is more or less frail (shared frailty). In either case the risk for an event is a function of previous occurrences. Medical
Clinical research of kidney diseases V

Fig. 4. Sources of correlation within multiple failure-time data. Unknown (or unmeasured) factors can be responsible for heterogeneity across individuals (with consequent different $\lambda(t)$, i.e. baseline risk across subjects) and within subject dependence of the failure events ($\lambda(t)$ varying within the subject during the recurrent process).

Research and clinical experience suggest that both individual unshared tendencies and varying shared susceptibility to fail during the recurrent process are likely to be the rule, rather than the exception, in the study of multiple events, and that each may enhance the effect of the other [9].

This correlation among events violates the assumption that the timing of events is independent and has two important consequences: the estimates of the coefficients and their standard errors are both biased (wrong) and inefficient (imprecise) in typical repeated events contexts. Variations of the Cox model (and other survival models), namely frailty (or random effect) models and variance-corrected methods, have been proposed to account for the correlation among event times.

Risk sets for survival analysis

Data layouts for survival analysis are complex as they define the risk set based on the three components of the response variable (time start, time stop and censor status), and possible distinction of different basal risk categories [2]. For an appropriate definition of the risk set different aspects should be considered [9]: classification of type and order of the failure events (whether the events are of different or the same type, whether they occur with or without natural order); definition of the time at risk (when the risk starts and ends); consideration of the mechanisms through which the predictor is involved in the process (whether/how the same predictor affects more outcomes) and definition of what is being modelled (the time to each event, the total course of a recurrent process or the time segments to each recurring event).

For unordered events of the same type (such as lesions of the eye [10]) and of different type (such as uraemia and mortality in a follow-up study of chronic kidney disease patients [11]) a risk set called marginal has been suggested (stratified for event of different type to allow different basal risks). For ordered events (such as catheter infections or dysfunctions, repeated peritonitis or transplant rejection episodes) four options have been proposed. Table 4 reports key characteristics of these risk sets.

Variance-corrected models

Variance-corrected models represent one way to deal with the problems produced by heterogeneity across individuals and failure-time dependences. Variations within the family of variance-corrected models are based on different definitions of the risk sets including whether they allow for event-specific baseline hazards using stratification. In these models (marginal, counting process and conditional risk sets) the robust (cluster) variance estimator is used as in GEE analysis. Variance-corrected models do not incorporate any random effect into the estimates themselves.

Frailty models

In contrast to the variance-corrected models, frailty models do incorporate the heterogeneity between clusters/subjects
Table 4. Risk sets for survival analysis

<table>
<thead>
<tr>
<th>Risk set</th>
<th>Order/type</th>
<th>Example</th>
<th>Strata</th>
<th>Time zero</th>
<th>Modeling</th>
<th>Assumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Un-stratified marginal risk sets</td>
<td>Unordered/same</td>
<td>The same lesion in paired organs such as the eye [10]</td>
<td>No</td>
<td>Time measured from subject enrolment</td>
<td>Total t to each event</td>
<td>The same process for all events</td>
</tr>
<tr>
<td>Competing risk sets</td>
<td>Unordered/different</td>
<td>Uraemia and mortality in chronic kidney disease [11]</td>
<td>Yes</td>
<td>Time measured from subject enrolment</td>
<td>Total t to each event</td>
<td>Predictor involved in competing processes</td>
</tr>
<tr>
<td>Counting process</td>
<td>Ordered/same</td>
<td>Catheter infections or dysfunctions; fistula thrombosis; repeated peritonitis or transplant rejection episodes: they are ordered events in that they may be seen in a study that records the time to first, second, third event, and so on, and the subject is not at risk for further events until a prior one has occurred. Four layout options are available for ordered recurrences</td>
<td>No</td>
<td>Each observation time is event defined</td>
<td>Total t course of the recurrent process</td>
<td>Order not important; no tied times; same baseline hazard</td>
</tr>
<tr>
<td>Stratified marginal risk sets [9,12]</td>
<td>Ordered/same</td>
<td></td>
<td>Yes</td>
<td>Time measured from subject enrolment</td>
<td>Total t course of the recurrent process</td>
<td>Events as independent processes; order ignored</td>
</tr>
<tr>
<td>Conditional risk sets from entry</td>
<td>Ordered/same</td>
<td></td>
<td>Yes</td>
<td>Enrolment (time measured continuously from entry)</td>
<td>Total t course of the recurrent event process</td>
<td>No risk for further event until a prior has occurred</td>
</tr>
<tr>
<td>Conditional risk sets from event (gap)</td>
<td>Ordered/same</td>
<td></td>
<td>Yes</td>
<td>Clock set to zero after each event</td>
<td>T segments between events</td>
<td>No risk for further event until a prior has occurred</td>
</tr>
</tbody>
</table>

\[*In the absence of correlation and dependent censoring each observation continues until the first event that occurs giving the same results as the time to combined event analysis (competing risk model of Lunn–McNeil [13]). In the presence of correlation the observations continue beyond the first event that occurs (marginal model of Wei–Lin–Weissfeld [12]), each event can occur only once per subject, and all subject are at risk for all events.\]
into the estimated portion of the model by making assumptions about its distribution [16–19]. This latent ‘random effect’ varies across individuals but is assumed to be constant over time and shared by a single individual (or all members of a cluster). As a result, under frailty models the event times are assumed to be independent conditional on the patient’s underlying frailty and inference can be made in the standard fashion. Frailty models estimate the variance of this random effect. When this variance is significantly different from zero, the model supports the hypothesis of a significant heterogeneity in the data.

The risk set of the standard (unconditional) frailty model is the same as the conditional risk set from entry (Table 4), but without stratification. This has been viewed as a limitation in the presence of event dependence, which is controlled instead by stratified variance-corrected methods and therefore these may be preferred in the presence of event dependence without heterogeneity. Since repeated events processes are usually characterized by both event dependence and heterogeneity (or it is often unclear which feature of the data mostly underlies the correlation), a stratified (conditional) frailty model has recently been proposed with the same risk set as the gap time risk set [17].

Model choice

The choice of the analytical tool for correlated survival data is dictated by the type and order of the failure events and the clinical question to be answered (Table 4).

For multiple events of different type (e.g. the same person may be observed to develop kidney failure, a stroke and then die) the variance-corrected marginal model is often a good choice. This is true when the model includes factors plausibly involved in the mechanism of more than one event type (e.g. hypertension). Frailty models can be used to specify and account for the sources of correlation in the data [16–19].

For ordered recurrent events of the same type (e.g. episodes of thrombosis in a vascular access) there are more choices, though most often the order condition and the difference in the baseline risks are important issues to be accounted for. The counting process is useful if there is no reason to believe that the baseline risk varies, as it is not stratified. The marginal risk model may be more appropriate to model repeated hospitalizations (where the reason for hospitalization has no natural order) than repeated bypass graft thrombosis or peritonitis episodes. Conversely, when the clinical course of repeated events supports the conditional assumption, one can either model the entire time course of the disease (from entry) or model the time segments between failures (from previous event). However, variance-corrected methods may still provide biased results in the presence of heterogeneity since they do not incorporate any random effect in the model.

Heterogeneity and event dependence can be considered components of a latent random effect inducing biased estimates if not taken into account. Both sources of correlation in the data may simultaneously underlie most of the recurrent event processes, although one may prevail over the other. In the presence of event dependence without heterogeneity the true variance of the frailty is zero. In these cases stratified variance-corrected methods perform well, whereas the traditional (un-stratified) frailty model may detect the presence of a random effect that is probably the consequence of event dependence rather than heterogeneity. In the presence of heterogeneity without event dependence, stratification may not be necessary since the baseline risk should not change by event number. In this case variance-corrected models may be inefficient and the unconditional frailty model would perform better. Yet, since repeated events data are very likely to exhibit both sources of correlation, a modelling strategy that is robust to heterogeneity and event dependence may be necessary [17].

Time-dependent effects and time-varying covariates

Another issue to consider when defining a risk set is related to the values and effects of the input variables. The term time dependent is more appropriately used to define the effect associated with an input and the term time varying is used for a covariate with updated values over time. For example, an input variable measured at baseline (e.g. recent myocardial infarction) can have different effects during different follow-up periods that can be modelled as a step-function of time (e.g. the relative risk for death from the infarct declines over time). In a follow-up study, baseline values of renal function were associated with increased risk for death only during the first year of observation and not thereafter [11]. These estimated time-dependent effects must satisfy the proportionality assumption when using the Cox’s model. Conversely a variable measured only once (at baseline) may interact with time and thus have an effect that changes with time, as was found for serum albumin in the HEMO study [20]. By definition, this effect will not satisfy the proportionality assumption. Another possibility is that the risk set contains updated values of a variable. For example, in a study of Urotensin II (a vasoactive substance) in chronic kidney disease patients, end-stage renal disease status (not yet on dialysis versus already on dialysis) had a different effect on cardiovascular events [21]. This input variable was treated as a time-varying covariate as subjects could change their status during follow-up. These input specific effects must also satisfy the proportionality assumption.

Special topics

Random effect modelling and variance-corrected methods are general approaches to model quantitative responses, categorical data, counts and survival times. The advantage of these methods is that they are natural and very flexible extensions of standard techniques, easily applicable to different circumstances. Special methods exist for specific analytical issues. Examples are repeated measures ANOVA for continuous responses and categorical exposures [22], and multivariate ANOVA (MANOVA) to study the simultaneous change of more quantitative outcomes in response to an exposure [23]. Several other models for longitudinal
designs are available, such as Markov Chains’ models to study the probability of a state change in a population [24], and time series to study observations at successive time intervals [25].

Conclusion

Clinical researchers are interested in describing how the study outcome varies in response to the effect of some exposure of interest (i.e. a therapy, a diagnostic test, a risk factor, etc.). In general, study design and sample size estimation, study conduct and statistical analysis should be consistent with the research hypothesis and objective. Examples of discrepancies among these domains are not uncommon in the medical literature. Often the correct approach to the assessment of the outcome variability requires multiple measurements of either or both predictors and outcome. This should be considered in the formulation of the study question and the design phase, including the choice of the correct analytical strategy to assess the correlation in the data and the role of its sources.

Acknowledgements. P.R. held a young investigator award from the Italian Society of Nephrology for the year 2005–2006 and received funding from data and the role of its sources.

Conflict of interest statement. None declared.

References


Received for publication: 10.1.08
Accepted in revised form: 20.2.08