The proportional means regression model for the analysis of recurrent event data

Running head: Proportional means regression for recurrent events

Luisa Canal, Rocco Micciolo

Department of Psychology and Cognitive Sciences, University of Trento, Italy

Corresponding author: Rocco Micciolo

Department of Psychology and Cognitive Sciences

University of Trento

Corso Bettini, 31

38068 Rovereto (TN), Italy

e-mail: rocco.micciolo@unitn.it

Summary

This paper deals with the proportional means regression model for the analysis of recurrent event data proposed by Lawless and Nadeau (LN) and based on the mean function. Appealing features of this model are that it is simple and robust, being based on moment estimates, it allows a pictorial representation of the rate of recurrence and can be considered a straightforward extension of the Cox model to recurrence data. Furthermore, it is shown that, under particular circumstances, the LN regression model gives the same estimates for regression coefficients as the standard Poisson regression based only on the counts, without knowledge of the recurrence times. We apply the LN method to two real-life datasets from the bio-medical field and provide a set of functions, written in the open source language **R**, which expands the available tools for the applied researcher; these functions estimate the cumulative mean function as well as the parameters of the LN proportional means regression model.

KEY WORDS: *counting processes, cumulative mean function, mean function, Poisson processes, proportional means, recurrent data.*

Introduction

Survival analysis is a powerful and flexible method for identifying associations between an outcome and a number of prior exposures to risk factors. In this setting, Cox's proportional hazards model (1) is the most commonly used technique; being semiparametric, it is more attractive than a fully parametric model which places constraints on distributional assumptions. There is growing interest in the analysis within longitudinal study designs where the event of interest can occur repeatedly in the same individual. For example, a patient diagnosed with a skin cancer can relapse over time, or a subject with a psychiatric diagnosis can have multiple contacts with psychiatric health services.

It is not possible to apply the standard Cox model to multiple failure times, since the assumption of independence between event times within individuals would be violated. As a result, the time to first event is commonly used for events that occur repeatedly.

Alternatively, the time between repeated occurrences can be ignored and multiple event data can be analysed, considering only the total number of events (occurring in a fixed period of time) and resorting to statistical models for counts, such as the Poisson model or the negative binomial regression model.

Several models and methods have been proposed in the literature to deal with recurrent event data; see Lawless (2), Kelly and Lim (3), Therneau and Hamilton (4), Therneau and Grambsch (5), and, for a review and a discussion, Cook and Lawless (6). Among these methods, one, proposed in 1995 by Lawless and Nadeau (7), seems, in our opinion, particularly appealing for the clinical researcher, since, as we will see, it allows a pictorial representation of the rate of recurrence and can be regarded as a simple and robust straightforward extension of the Cox model to recurrence data. This approach, which is based on the cumulative mean of the events, does not involve a full probabilistic specification of the processes, but requires knowledge of the individual times at which events occur.

Despite these appealing characteristics, little use has been made of this technique in clinical research studies. A search of Scholar Google (http://scholar.google.it), focusing on articles published during 2007 and the first months of 2008 and using the phrases "mean function", "cumulative mean function" and "proportional means" gave over 600 references. However, only one article was published in a medical journal.

To expand the available tools for the applied researcher, and to encourage use of the Lawless and Nadeau (LN) approach to the analysis of recurrent event data, we present a set of functions written in the native **R** language (8), which provide both graphical tools as well as computational procedures for fitting the LN regression model. **R** is a free programming language and software environment for statistical computing. Being an open source, its code is freely distributed, under the GNU General Public Licence, through *CRAN* (http://cran.r-project.org/), which is an acronym for the Comprehensive R Archive Network.

Data examples

To illustrate and motivate the development of the LN method for the analysis of recurrent events, we shall use two real-life datasets from the personal experience of the authors.

Recurrence of cutaneous epitheliomas

Cutaneous epitheliomas are the most common malignant neoplasms in the Caucasian population; the most frequent are basal cell carcinoma (BCC) and squamous cell carcinoma (SCC). Both BCC and SCC are characterised by a relatively high frequency of recurrences. In the Italian province of Trento, a *Skin Cancer Registry* was established in 1992 with the aim of recording all cutaneous tumours occurring in the province's residents (9,10). We compare rates of occurrence of cutaneous epitheliomas according to gender and histotype, examining data available from this registry for the period January 1992 to December 1997. For each patient, the time of each new occurrence of skin cancer was recorded. A total of 2557 individuals were included in this study. During the follow up, 311 recurrences were observed in 226 patients,

while over 91% of subjects had no recurrence. The maximum number of recurrences in a single patient was eight, recorded in one subject. The mean number of recurrences per patient was 0.12 with a standard deviation of 0.48 (therefore the ratio between the variance and the mean was about 2:1).

Patterns of psychiatric contacts in a psychiatric case register

Psychiatric data collected in a psychiatric case register document contacts between residents and the psychiatry services of a selected geographical area. These data typically show a large number of subjects with a small number of contacts and, at the same time, a low number of subjects recording a high number of contacts. The data presented refer to patients entered in the South Verona Psychiatric Case Register (SVPCR) in the period 1 January 1979 to 31 December 1991 (11, 12). All the subjects were followed up for 13 weeks after the day of their first contact. For each patient, the total number of contacts in the 91 days of follow up was known, as was the day on which each contact took place. The following covariates were available: gender, occupational status, diagnosis, referral source of the first contact, type of first contact. A total of 3454 subjects were included in this study, recording a total of 6913 contacts. The mean number of contacts per patient (in the 91 days) was 2.0 with a standard deviation of 3.7 (therefore the ratio between the variance and the mean was about 7:1); 1589 subjects (46.0%) had no further contact, after the first one, during the study period while the highest number of contacts recorded for a single patient was 48; a total of 28 patients each had more than 20 contacts during the follow up.

Table 1 shows how these data were recorded in the first 14 patients. The total follow-up time and the total number of recurrences observed are given in the first two columns; the day on which each of the contacts took place is reported in the following columns. For example, the first subject had two contacts: the first after 41 days of follow up and the second 59 days after the start of the follow up. The fifth subject had no contact at all during the 91 days of follow up, so all the cells referring to the recurrence times are empty.

Follow-up times (days)	Number of recurrences	Time of 1st recurrence	2nd time	3rd time	4th time	5th time
91	2	41	59			
91	1	91				
91	3	17	42	80		
91	1	81				
91	0					
91	2	4	67			
91	3	8	39	88		
91	1	84				
91	2	26	68			
91	0					
91	0					
91	0					
91	5	40	45	54	68	88
91	3	1	4	10		

Table 1. Recurrence data recorded for the first 14 patients in the South Verona Psychiatric Case Register.

In both datasets presented, the high ratio between the variance and the mean makes the Poisson assumption untenable; in fact, some patients were more prone to recurrent events than others, which suggests that a non-parametric or a semiparametric procedure is more appropriate, particularly to test for treatment differences.

The cumulative mean function

Let us consider *k* individuals, each observed for the follow-up time τ_i $(i = 1,...,k)$. Let r_i denote the observed number of events (recurrences) for subject *i* over the interval $[0, \tau_i]$ and $t_{i1} \le ... \le t_{ir_i}$ the times of events.

For example, looking at the data reported in Table 1, the follow-up time for patient number 13 was $\tau_{13} = 91$ days, during which a total number of recurrences $r_{13} = 5$ were observed; the corresponding times $t_{13,1} \le t_{13,2} \le t_{13,3} \le t_{13,4} \le t_{13,5}$ were, respectively, 40, 45, 54, 68, and 88.

The cumulative mean function (CMF) of the number of recurrences $N_i(t)$ occurring for the *i*–th individual over the interval $[0, t]$ is defined as $M(t) = E[N_i(t)]$, where $M(t)$, which is supposed to be the same for all the subjects, is the sum (or the integral, depending on the time

scale) of the mean function $m(t)$, i.e. the expected value of the number of events experienced at time *t*. In what follows, for the sake of simplicity, we refer to the discrete-time case.

Under the assumption that the *k* individuals are mutually independent and that the $n_i(t)$'s (i.e. the number of events observed at time *t* for the *i*–th individual) are independent Poisson random variables with mean $m(t)$, the maximum likelihood estimate of $m(t)$ is $\hat{m}(t) = n.(t)/\delta.(t)$ (i.e. the mean number of events observed at time *t* over all the *k* individuals) and the estimate of $M(t)$ is $\hat{M}(t) = \sum_{s=0}^{t} n.(s)/\delta.(s)$, i.e. the sum of the mean number of events observed up to time *t*. In fact, $n.(s) = \sum_{i=1}^{k} n_i(s)$ is the total number of recurrences observed at time *s* and $\delta.(s) = \sum_{i=1}^{k} \delta_i(s)$ is the overall number of individuals under observation at time *s*, in which $\delta_i(s)$ is an indicator variable equal to one if individual *i* is under observation and "at risk" at time *s*, and zero otherwise.

For example, in the SVPCR data, all the subjects were followed up for 91 days after being entered in the register, so that $\delta.(s)$ is 3454 for all the times *s*, from *s*=1 to *s*=91 (in particular, $\delta.(1) = 3454$ and $\delta.(2) = 3454$). On the other hand, at day one from the start of the follow up, 181 contacts were observed, so that $n.(1) = 181$; at day two from the start of the follow up, 161 contacts were observed, so that $n.(2) = 161$. The mean number of events observed at time 1 was therefore $\hat{m}(1) = n.(1)/\delta.(1) = 181/3454$, while that observed at time 2 was $\hat{m}(2) = n.(2)/\delta.(2) = 161/3454$. The cumulative mean number of events observed at time 2 was therefore $\hat{M}(2) = \hat{m}(1) + \hat{m}(2) = 181/3454 + 161/3454$.

The plot of $\hat{M}(t)$ versus *t* yields information about the number of events expected by time *t* and whether two groups differ significantly in the expected number of events. The left panel of Figure 1 shows the estimated cumulative mean number of contacts recorded in the SVPCR data;

95% confidence intervals are also shown. The slope of $\hat{M}(t)$ can be considered a failure rate, thereby allowing the plot of $\hat{M}(t)$ to yield information on the event process. The occurrence of psychiatric contacts is higher in the first weeks and then declines, a pattern, well known to physicians (since patients need more attention, particularly in the initial phases of their illness), that is pictorially and quantitatively represented in Figure 1. After 4 weeks of follow up, the cumulative mean number of contacts is 1; instead, to obtain a cumulative mean of 2, thirteen weeks are needed. The differences in the cumulative means between subsequent weeks seem to show a steady decline. From a modelling point of view, the process of event occurrence considered here is a non-homogeneous Poisson process. Were the process homogeneous, the plot would show a straight line.

As far as the cutaneous epitheliomas are concerned (Figure 1, right panel), the cumulative means of the numbers of recurrences after 1, 3, and 5 years of follow up are, respectively, 0.054, 0.137 and 0.211. The differences in the cumulative means between subsequent years appear approximately constant (about 0.04 recurrence/year of follow up).

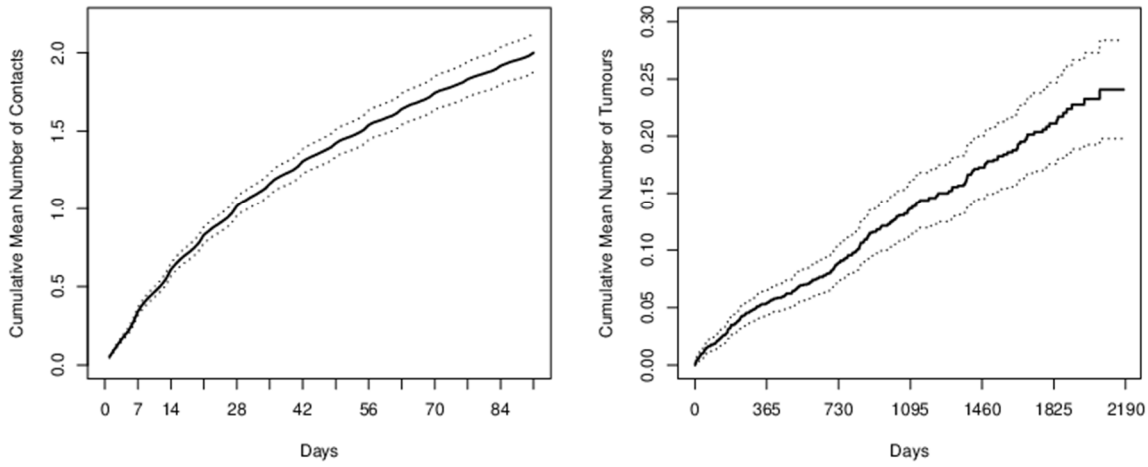

Figure 1. Estimated cumulative mean function, together with 95% confidence limits, for the psychiatric contacts in South Verona (left panel) and for the recurrences of cutaneous epitheliomas (right panel).

The proportional means regression model

In randomised clinical trials as well as in epidemiological settings, it is frequently deemed interesting to make comparisons among groups, possibly accounting for the effects of a number of covariates. For example, in the SVPCR data, we are interested in evaluating whether the type of the first contact (unplanned vs planned) is associated with the recurrence rate of subsequent contacts. The plot of the CMFs for these two groups is shown in the left panel of Figure 2. A higher recurrence rate is clearly evident in patients who entered the SVPCR with an unplanned contact. On the other hand, it appears that the recurrence rates in males and females are similar (see the right panel of Figure 2). However, what we need is a formal test of significance together with a quantitative measure of the "difference" between recurrence rates in the two groups. A regression model can provide an appropriate answer to both these needs.

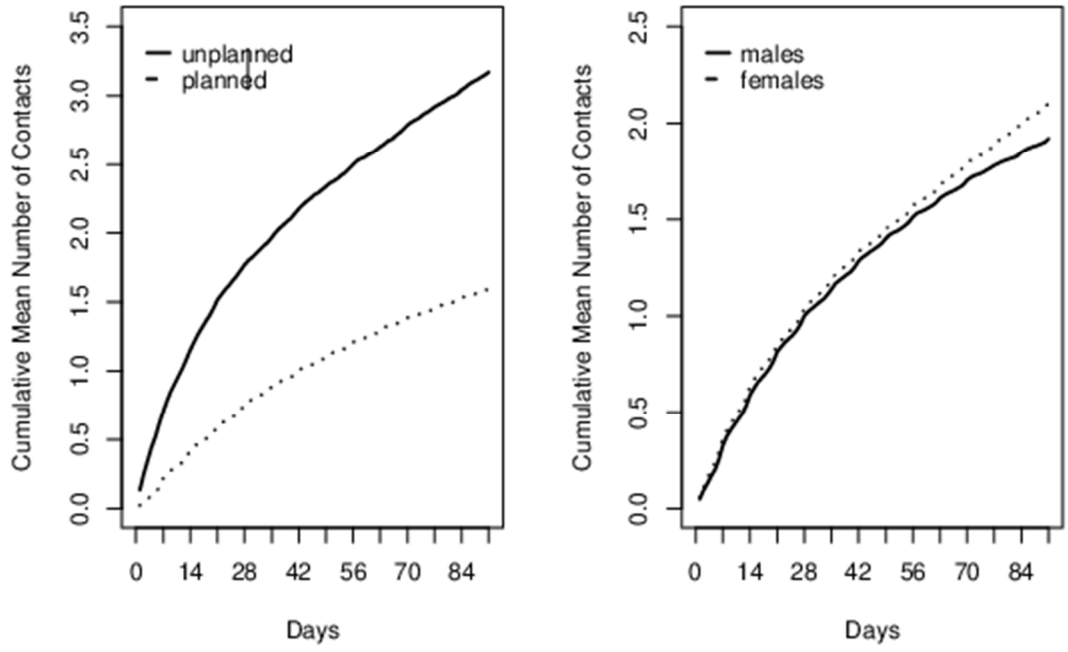

Figure 2. Estimated cumulative mean functions according to gender (right panel) and to the type of the first contact with the South Verona psychiatry services (left panel).

The LN model is a semiparametric proportional means regression model based on the mean function $m(t)$, analogous to the proportional hazards model for lifetime data. The regression

model was set up including a multiplicative effect of a $p \times 1$ vector $\mathbf{x}_i$ of fixed covariates (without the constant for the intercept term) on the mean function: $m_i(t) = m_0(t)\exp(\mathbf{b}'\mathbf{x}_i)$, where $m_0(t) \geq 0$ is a baseline mean function and **b** is a $p \times 1$ vector of regression coefficients. In their paper, Lawless and Nadeau (7) considered the more general case $m_i(t) = m_0(t)g(\mathbf{x}_i(t);\mathbf{b})$ where *g* is a positive-valued function and the covariates can be time-dependent.

Under the Poisson assumption, Lawless and Nadeau obtained the following estimating equations for the $m_0(t)'s$ and **b**:

$$\sum_{i=1}^{k} \delta_i(t)\{n_i(t) - m_0(t)\exp(\mathbf{b}'\mathbf{x}_i)\} = 0 \qquad t = 0,1,...,\tau_i \qquad [1]$$

and

$$\sum_{i=1}^{k}\sum_{s=0}^{\tau_i} \delta_i(s)\left\{\frac{n_i(s) - m_0(s)\exp(\mathbf{b}'\mathbf{x}_i)}{\exp(\mathbf{b}'\mathbf{x}_i)}\right\}\frac{\partial \exp(\mathbf{b}'\mathbf{x}_i)}{\partial \mathbf{b}} = \mathbf{0} \qquad [2]$$

where, as previously indicated, $n_i(t)$ represents the number of events that occur at time *t* for subject *i* and $\delta_i(t) = 1$ if $t \leq \tau_i$ and $\delta_i(t) = 0$ if $t > \tau_i$.

The authors noted that equation [1] gives

$$m_0(t) = \frac{n.(t)}{\sum_{i=1}^{k} \delta_i(t)\exp(\mathbf{b}'\mathbf{x}_i)} \qquad [3]$$

and inserting equation [3] in equation [2], they obtained the $p \times 1$ system of equations in **b**

$$\sum_{i=1}^{k}\sum_{s=0}^{\tau} \delta_i(s)\left\{n_i(s) - \frac{n.(s)\exp(\mathbf{b}'\mathbf{x}_i)}{\sum_{i=1}^{k} \delta_i(s)\exp(\mathbf{b}'\mathbf{x}_i)}\right\}\mathbf{x}_i = \mathbf{0}. \qquad [4]$$

This set of equations are Cox partial likelihood equations, meaning that packages that implement partial likelihood analysis of repeated events can be used to fit the proportional means model; for example, in the `survival` package of **R** (13), the function `coxph` can be used

to estimate the regression coefficients employing the so-called "start/stop" format for the recurrence data and the `cluster` option for the individuals in order to obtain robust standard errors. Otherwise, equation [4] can be solved iteratively using Newton's method.

If, as is the case with the SVPCR data, all the subjects are observed for the same follow-up time $(\tau_1 = ... = \tau_k = \tau)$ so that $\delta_i(t) = 1$ for all the times $t \le \tau$, regardless of the subject, equations [4] can be simplified and it is possible to arrive at a meaningful interpretation. In fact, summing over all the times and indicating $\sum_{s=0}^{\tau} n_i(s) = n_i$ (i.e. the total number of contacts of subject *i*) and $\sum_{i=1}^{k} n_i = n$ (i.e. the total number of contacts observed), equations [4] can be rewritten in the much more compact way $\sum_{i=1}^{k} n_i \mathbf{x}_i = \sum_{i=1}^{k} \hat{n}_i \mathbf{x}_i$, where $\hat{n}_i = w_i n$ and the w_i's are weights (which sum to one) defined as $w_i = \frac{\exp(\mathbf{b}'\mathbf{x}_i)}{\sum_{i=1}^{k} \exp(\mathbf{b}'\mathbf{x}_i)}$. In other words, the $\hat{n}_i$'s are the recurrences expected in subject *i* on the basis of his "risk score", given by $\exp(\mathbf{b}'\mathbf{x}_i)$ as a fraction of the "total risk score" $\sum_{i=1}^{k} \exp(\mathbf{b}'\mathbf{x}_i)$. In this case, equations [4] are analogous to the equations to be solved for the "standard" Poisson regression model, i.e. $\sum_{i=1}^{k} n_i \mathbf{x}_i = \sum_{i=1}^{k} \exp(\mathbf{b}'\mathbf{x}_i)\mathbf{x}_i$ where, however, $\mathbf{b}$ contains an intercept term so that the number of equations to be solved is $p+1$. In actual fact, apart from the intercept, the estimates of the *p* regression coefficients are the same for the LN model and the Poisson regression model. As a corollary, in the particular case considered, the estimates of the regression coefficients of the LN model are unaffected by knowledge of the individual recurrence times.

Robust variance estimates for the regression parameters $\hat{\mathbf{b}}$, accounting for the dependence structure of the recurrence times, can be computed as outlined in the appendix of the paper by Lawless and Nadeau (7).

As far as significance tests and confidence intervals are concerned, the LN model showed that under mild conditions $\sqrt{k}\left(\hat{\mathbf{b}}-\mathbf{b}\right)$ is asymptotically normal. The accuracy of the approximation depends on the number of subjects (*k*), on the average counts per individual, and on the degree of overdispersion. According to the LN model, the approximation can be considered satisfactory when *k* is equal to 30 or more and the average count per individual is greater than 4 except when overdispersion is very large (variance at least five times that of the Poisson model). In this case the approximation is satisfactory if *k* is 90 or more. Under these prescriptions the asymptotic approximations are sufficiently accurate for practical use.

The main assumption of the LN model is that, conditional on the covariate values, the end-of-observation times τ_i are determined independently of the event process. If this is not the case, then $\hat{M}(t)$ may be seriously biased.

Lin et al. (14) provided a rigorous justification of the LN procedure through a modern empirical process theory. Furthermore, they developed both graphical and numerical methods based on Gaussian processes for checking the adequacy of the fitted model.

Proportional means regression results

The LN proportional means regression model was fitted to the two datasets considered. Estimates of the regression coefficients as well as the associated standard errors were obtained employing the **R** `mfreg` function described in the appendix. This function has the peculiarity of not requiring data in the "start/stop" format, which can be an advantage as many datasets are not organised in this "long" format.

The results relative to cutaneous epitheliomas are set out in Table 2. Males showed a higher number of recurrences than females (about 1.7 times that recorded in women), while a non-significant effect of histotype (as well as of the interaction between sex and histotype) was found. The finding of a similar biological behaviour between BCC and SCC is somewhat unexpected, since it is well known to dermatologists that BCC is a cancer with a higher probability of recurrence than SCC. However, in this study, multiple synchronous tumours were considered a single multifocal lesion (i.e. two or more tumours of the same histotype diagnosed in the same subject on the same day were considered a single recurrence); in this case (i.e. when only metachronous tumours were considered), the recurrence pattern of the two histotypes was quite similar. On the other hand, a higher recurrence rate in males is well known; however, employing the regression model we were able to quantify the gender effect (in univariate as well as in multivariate analyses), both as a point estimate and as a 95% confidence interval (1.2, 2.3).

Covariate	Parameter estimate	SE	z
Univariate Analysis			
Histotype (BCC vs SCC)	0.086	0.195	0.441
Sex (F vs M)	-0.503	0.165	3.052
Multivariate Analysis			
Histotype (BCC vs SCC)	0.100	0.197	0.511
Sex (F vs M)	-0.505	0.166	3.045
Histotype (BCC vs SCC)	0.285	0.226	1.262
Sex (F vs M)	-0.129	0.373	0.347
interaction	-0.494	0.413	1.197

Abbreviations: BCC = basal cell carcinoma; SCC = squamous cell carcinoma

Table 2. Parameter estimates for the proportional means regression on the recurrences of cutaneous epitheliomas

Table 3 shows the results of the analysis of the pattern of contacts with psychiatry services in South Verona. As far as the univariate analysis is concerned, since all the subjects were followed up for 91 days (so that the same number of subjects was at risk at each time), an exact solution can be obtained for the estimating equations of the LN model regression coefficients; for a categorical variable with k levels, coded with $k-1$ dummies, the estimate of the j-th regression coefficient is $\ln\left[\left(n_0 N_j\right)/\left(n_j N_0\right)\right]$, where n_j is the number of subjects in the category j+1 and N_j is the number of contacts had by the subjects in the category j+1 (the deponent 0 indicates the reference category). With the exception of gender, all the other variables considered were associated with the number of contacts.

	Proportional Means Univariate Analysis			Proportional Means Multivariate Analysis			Poisson Multivariate Analysis			Negative Binomial Multivariate Analysis		
	b	se	z	b	se	z	b	se	z	b	se	z
Gender												
Females vs Males	-0.090	0.063	-1.427	-0.057	0.072	-0.783	-0.057	0.026	-2.138	-0.127	0.058	-2.183
Occupational status												
Unempl. Vs Empl.	0.758	0.110	6.919	0.478	0.109	4.363	0.478	0.040	12.090	0.564	0.104	5.416
Other vs Empl.	0.106	0.065	1.630	0.138	0.072	1.927	0.138	0.028	4.918	0.152	0.060	2.526
Diagnosis												
Affective dis. vs Schiz.	-0.892	0.101	-8.849	-0.812	0.100	-8.156	-0.812	0.038	-21.276	-0.788	0.109	-7.226
Organic psych. vs Schiz.	-0.699	0.206	-3.402	-0.648	0.210	-3.084	-0.648	0.071	-9.066	-0.532	0.181	-2.940
Alc. / pers. dis. vs Schiz.	-0.745	0.124	-6.013	-0.667	0.118	-5.640	-0.667	0.042	-15.886	-0.687	0.118	-5.808
Neurotic dis. vs Schiz.	-1.210	0.105	-11.524	-1.061	0.102	-10.416	-1.061	0.041	-26.174	-1.012	0.110	-9.236
Other dis. vs Schiz.	-1.348	0.111	-12.174	-1.227	0.108	-11.398	-1.227	0.044	-28.054	-1.201	0.113	-10.669
Referral Source												
GPs vs Self-referral	-0.265	0.085	-3.108	-0.010	0.085	-0.115	-0.010	0.040	-0.245	-0.008	0.089	-0.087
Others vs Self-referral	-0.510	0.068	-7.460	-0.350	0.078	-4.508	-0.350	0.029	-12.180	-0.377	0.063	-6.009
First contact												
Unplanned vs Planned	0.683	0.066	10.355	0.410	0.073	5.626	0.410	0.028	14.844	0.385	0.065	5.904

Abbreviations: dis. = disorder; Schiz. = Schizophrenia; psych. = psychosis; Alc./pers.dis. = Alcoholism/personality disorder.

Table 3. Parameter estimates for the proportional means regression on the recurrences of contacts with psychiatric services.

This result was confirmed when the joint effect of all the considered variables was evaluated in a multivariate analysis (Table 3). A significantly higher number of contacts was found for unemployed subjects, for patients with an unplanned first contact, and for those who were self-referred (or referred by relatives). As far as diagnosis is concerned, a higher number of contacts was found for schizophrenic patients.

For comparison, Table 3 also gives the estimates of the regression coefficients from the Poisson model and from the negative binomial regression model estimated taking into account only the

total number of contacts and discarding the recurrence times. It is worth noting the striking difference in the estimated standard errors between the LN and the Poisson models – 2- to 3-fold greater with the former than the latter – leading to more significant statistical tests (more significant than they actually are) and overly narrow confidence intervals when employing the naïve Poisson model. On the other hand, if we calculate standard errors for the Poisson model employing a robust sandwich estimator, the values obtained are the same as those obtained with the LN model. Furthermore, the estimates of the regression coefficients of the Poisson and of the LN models, too, were the same, as expected. We recall that this happens because all the follow-up times are the same, which is not generally the case. However, from a computational point of view, a Poisson regression model can be fitted to the recurrence data to obtain initial estimates for the regression coefficients in order to speed up the convergence of the Newton algorithm. The standard errors from the negative binomial regression are comparable with those obtained from the LN model.

Informal graphical techniques developed for checking the adequacy of the Cox model in survival analysis can be employed for the LN model, too. A simple graphical evaluation of the proportionality assumption can be obtained by plotting the predicted CMF against the observed one for different groups. The left panel of Figure 3 shows this comparison for the SVPCR data relative to the type of the first psychiatric contact. Although there is no doubt about the prognostic role of this variable, it nevertheless appears that the proportionality assumption does not hold, particularly for the first weeks of follow up, where the model underestimates the cumulative mean number of contacts for patients with an unplanned first contact. Once again we can interpret this finding from a clinical point of view; in fact, the variable considered is a proxy for the severity of the psychiatric illness and patients with a more severe disease need more attention in the initial phases of their illness. However, as far as the Cox model is concerned, this violation does not invalidate the comparison between the two groups considered, since the observed lines are always well separated and do not cross. Therefore,

although a violation of the proportionality assumption is present, this violation does not appear crucial, hence we can employ the LN estimate of the "average" effect of the type of the first psychiatric contact as a quantitative tool for comparison with other studies. In the right panel of Figure 3, a similar comparison is shown for the effect of gender on the recurrences of cutaneous epitheliomas. In this case there is no evident departure from the proportionality assumption, as was also suggested by plots of the natural logarithm of $\hat{M}(t)$ against time in males and females, which were roughly vertical translations of one another (data not shown).

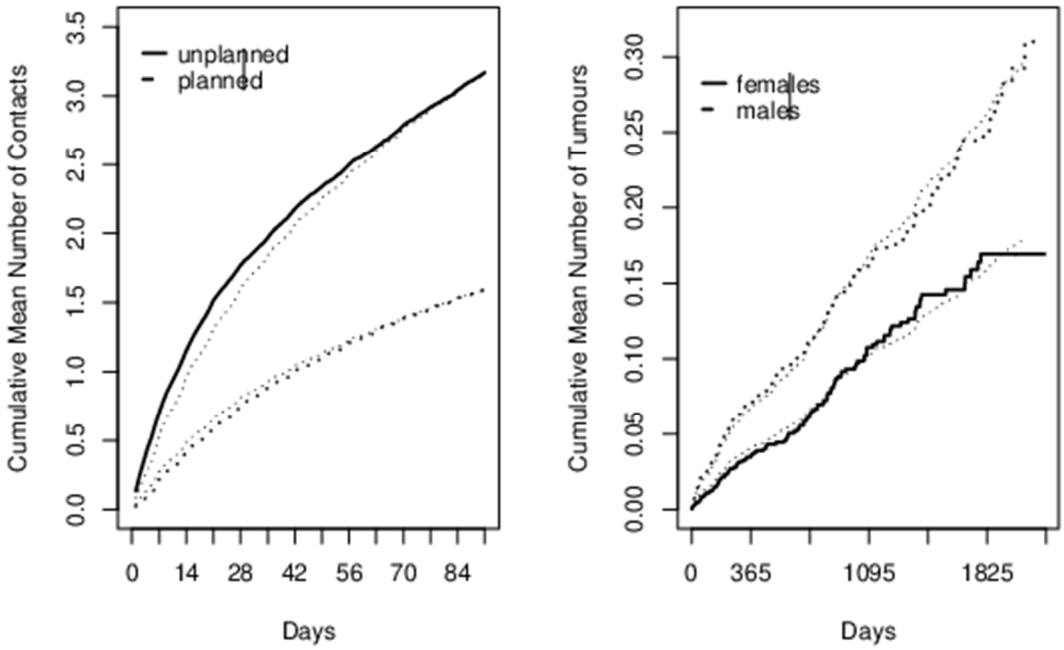

Figure 3. Left panel: estimated cumulative mean number of contacts with the South Verona psychiatry services according to the type of the first contact. Dotted lines represent predicted cumulative mean number of contacts according to the proportional means regression (see results reported in Table 3). Right panel: estimated cumulative mean number of recurrences of cutaneous epitheliomas according to gender. Dotted lines represent predicted cumulative mean number of contacts according to the proportional means regression (see results reported in Table 2).

Since the mean number of recurrences was less than 4 in the two datasets analysed and the overdispersion in the SVPCR data was very large, we wonder whether the asymptotic approximation can be considered sufficiently accurate, given that the number of subjects was quite large. A bootstrap estimate of the sampling distribution of the regression coefficients performed with both datasets revealed quite good agreement with the normal distribution and a

bootstrap variance quite similar to the robust one. A second check was performed running four simulations employing the SVPCR dataset as a population from which random samples of different size *k* (60, 80, 100, 120) were repeatedly extracted (using the type of the first psychiatric contact as covariate). Figure 4 shows the normal quantile-quantile plots of the simulated sampling distribution of the regression coefficients for the different values of sample sizes considered. As can be seen, the simulated sampling distribution was different from the Gaussian for samples of size 60 and 80; however, when the number of subjects considered was 120, the normal approximation appeared satisfactory.

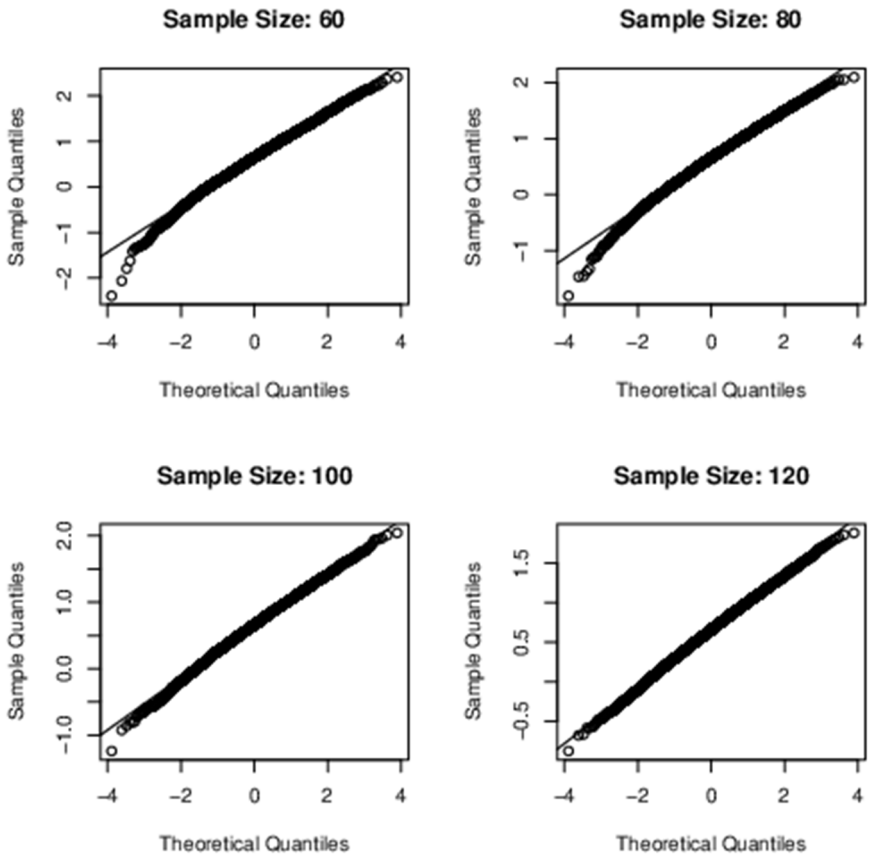

Figure 4. Normal quantile-quantile plots of the simulated sampling distribution of the regression coefficients for the type of the first psychiatric contact for 4 different values of sample sizes.

Discussion

Recurrent events arise frequently in medical settings and a number of different approaches have been developed to deal with multiple event survival data. However, despite the growing interest in the analytical techniques, these methods have not been commonly applied in the analysis of

data arising from clinical trials and/or observational studies published in medical journals, perhaps as a result of their complex structure. There is a general tendency to use simplistic methods employing the time to first event as the end point of the analysis. However, discarding information on subsequent events implies a loss of efficiency in the analysis and can provide a rather narrow perspective on the event process and covariate effects.

In this paper, we have reviewed a simple and robust method that can be considered a conceptually straightforward counterpart of the Kaplan-Meier estimate and of the Cox model for the analysis of datasets with multiple failures per subject. This method relies on the cumulative mean function $M(t)$ and on a multiplicative effect of the covariates. Although one could specify a parametric form for $M(t)$, the non-parametric estimates of $M(t)$ and of its variance proposed by Lawless and Nadeau are robust since they are moment estimates. Also the regression coefficients are based on Poisson maximum likelihood estimates which are valid quite generally because they are generalised least squares, or quasi-likelihood, estimates, provided that, conditional on the covariate values, the τ_i's are determined independently of the event processes.

The most important assumption of this method is that the end of observations times τ_i's must be independent of the event processes. It is easy to think of situations in which this would not hold. For example, if we were studying system failures and systems with many failures had earlier been withdrawn from service. On the other hand, in the examples discussed, it is likely that the independence assumption is satisfied, since censoring occurred at a fixed time in all the subjects (in the skin tumour dataset, the end of the follow-up period was the same for all the subjects, while in the SVPCR dataset all the patients were observed for 91 days). However, it is possible to check, at least informally, this independence for the skin cancer dataset, grouping the subjects according to their end-of-observation times into two groups (up to two years of follow up, with a dummy covariate x_i equal to 0; between 3 and 5 years of follow up, with a

dummy covariate x_i equal to 1) and then testing that the $M(t)$'s in the two groups are equal. The estimate of the regression coefficient (–0.1415) together with its associated standard error (0.2134) gave no indication that the τ_i's are not independent of the patterns of recurrences. As Lawless and Nadeau pointed out, the covariate $x_i = \tau_i$ provides a more sensitive check of the independence of the τ_i's. Once again, in the case presented, the estimate of the regression coefficient (–0.0404) was comparable with the associated standard error (0.0514), meaning that the independence assumption cannot be rejected, as expected. If we specify a full probabilistic model for event processes, then the need for independent τ_i's can be removed, but variance estimates for parameters would be less robust than the ones given in (7) if the specification of the model is not correct.

With the hope of making the LN method more accessible to medical researchers, so that it can be a valuable addition to the set of statistical tools for the analysis of failure time data, we have provided a set of **R** functions which allows both a graphical display of the recurrence data and a more sensitive inference concerning the effect of covariates on the recurrence rate.

References

1. Cox DR. Regression models and life tables (with discussion). Journal of the Royal Statistical Society B 1982; 34: 187–220.
2. Lawless JF. Regression methods for Poisson process data. Journal of the American Statistical Association 1987; 82: 808-815.
3. Kelly PJ, Lim LL. Survival analysis for recurrent event data: an application to childhood infectious diseases. Stat Med 2000;19:13–33.
4. Therneau T, Hamilton SA. rhDNase as an Example of recurrent event analysis. Stat Med 1997; 16: 2029–2047.
5. Therneau TM, Grambsch PM. Modeling survival data: extending the Cox model. New York, NY: Springer, 2000.
6. Cook RJ, Lawless JF. Analysis of repeated events. Stat Methods Med Res 2002; 11: 141–166.
7. Lawless JF, Nadeau C. Some simple robust methods for the analysis of recurrent events. Technometrics 1995; 37: 158–168.
8. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria 2006. ISBN 3-900051-07-0, URL http://www.R-project.org.
9. Boi S, Cristofolini M, Micciolo R, Polla E, Dalla Palma P. Epidemiology of skin tumors: data from the cutaneous cancer registry in Trentino, Italy. J Cutan Med Surg 2003; 7: 300–305.
10. Canal L, Micciolo R, Boi S, Cristofolini M. Recurrence analysis of cutaneous epitheliomas. Statistica Applicata 2004; 16: 143–157.

11. Tansella M (ed). Community-based psychiatry: long-term patterns of care in South Verona. Psychological Medicine Monograph Supplement 19. Cambridge: Cambridge University Press, 1991.

12. Canal L, Micciolo R. Regression models for the analysis of psychiatric data. In: Biggeri A, Dreassi E, Lagazio C, Marchi M, eds Statistical Modelling. 19th International Workshop on Statistical Modelling. Firenze*:* Firenze University Press, 2004: 351–355.

13. Lumley T, Therneau T. The survival Package. The Comprehensive R Archive Network 2003, http://cran.r-project.org.

14. Lin DY, Wei LJ, Yang I, Ying Z. Semiparametric regression for the mean and rate functions of recurrent events. Journal of the Royal Statistical Society B 2000; 62: 711–730.

Appendix

R functions

To estimate the cumulative mean function as well as the regression coefficients of the LN proportional means model a set of functions are provided, all written using the native **R** language for maximum portability. The functions, which can be downloaded from the URL http://hostingwin.unitn.it/micciolo/pmr/index.html, are:

`cmfplot`	estimates the cumulative mean function (CMF) and produces a graphical plot. The user must supply three arguments, i.e. a vector `tau`, which contains the follow-up times, a matrix `tempi` with the recurrence times in the columns, a vector `nr` with the number of recurrences; optional arguments are the confidence level (set as default to 95%), the output (set to `FALSE`), the graphical plot of the CMF (set to `TRUE`) and of the confidence bands (set to `TRUE`). If the output argument `out` is set to `TRUE`, a matrix is given as output of the function to be employed for further analyses; the rows of the matrix are the times considered (with increments of 1 unit), while the columns have the following meanings: 1) the time *t* 2) the estimate of the CMF at the time *t* 3) the estimate of the standard error of (2) 4) the lower limit of the confidence interval 5) the upper limit of the confidence interval 6) the number of subjects at risk at time *t* (i.e. $\delta.(t)$) 7) the total number of recurrences observed at time *t* (i.e. $n.(t)$) Note that $\delta_i(t) = 1$ when $t \le \tau_i$ and $\delta_i(t) = 0$ otherwise.
`mfreg`	estimates the regression parameters of the LN proportional means model. The user must supply four arguments; the first three (`tau`, `tempi`, `nr`) are the same as those

	described above; the fourth argument is the matrix `xcov` containing the values of the covariates (for categorical variables, the corresponding dummies must be provided). A further argument (`betastart`) can be set to `TRUE` if the user wants the estimates of the regression coefficients from Poisson regression on counts (without considering the recurrence times) to be used as starting values for the Newton algorithm. The output of `mfreg` is a list containing the following elements: 1) `$estimates` with the estimates of the regression coefficients, their standard errors and the associated significance 2) `$asvar` with the elements of the covariance matrix of the regression coefficients calculated according to formula 3.10 in (7) 3) `$basemf` with the estimate of the baseline mean function $m_0(t)$ 4) `$times` with the corresponding times.
`nsolve`	**R** does not implement a function to solve a system of non-linear equations; on the other hand, the general-purpose optimisation function `optim` is available to find the minimum of a function. It is possible to use `optim` to find the solution of equations [4] by minimising the squared-norm of the set of functions (Ravi Varadhan has written a simple function `nlsolve` which performs this task by calling `optim` and using the quasi-Newton algorithm BFGS within `optim` and makes it available for **R** users). Here we propose a naïve algorithm which implements the standard Newton method in the **R** function `nsolve` which is called within `cmfreg`. The first two arguments of `nsolve` are two functions (see below), since in **R** function can be passed as arguments to functions. The first argument (`fun`) calculates and returns the values of the set of functions in equations [4] at a guessed value $\tilde{\mathbf{b}}$, while the second (`jac`) calculates the Jacobian matrix using the

	current guess $\tilde{\mathbf{b}}$ for solution; the third argument is the vector of guessed values and the fourth is a list containing ancillary data needed for performing the previous calculations. The last two arguments of `nsolve` control the convergence of the algorithm. Some information on the control flow is written on the terminal. On exit, `nsolve` returns the numerical values of the functions at the proposed final solution.
`fun`	is the first argument of `nsolve`; calculates the values of the functions in the equations [4] at a guessed value $\tilde{\mathbf{b}}$.
`jac`	is the second argument of `nsolve`; calculates the Jacobian matrix using the current guess $\tilde{\mathbf{b}}$ for solution.

To illustrate the use of the proposed functions, we employed the cutaneous epithelioma dataset. This dataset is supplied as an **R** workspace file (with extension `.rdata`).

First, we (the user) had to load into the **R** workspace the set of functions: `source("mfreg.txt")`.

Next, we loaded the data set with the skin cancer data: `load("cutepi.rdata")`. In **R** workspace there are now the vectors `tau` (follow-up times) and `nr` (total number of recurrences), each with 2557 elements, the matrix `tempi` (recurrence times) and the matrix `Xcov`, with 3 columns: gender (0 = males; 1 = females), histotype (0 = SCC; 1 = BCC) and the interaction.

A look at the matrix `tempi` shows that the input of the data does not follows the "standard" counting process style of input. According to this style, a unit which has two recurrences and a censoring time has three observations; each observation has a start time, a stop time, and an indicator of whether the stop time is a recurrent event time or a censored time. As shown in Table 1, we preferred to store the follow-up times and the total number of recurrences observed

in each subject in two vectors, and the recurrence times in a matrix (with most of the cells possibly empty).

The estimate of the cumulative mean number of recurrences, together with 95% confidence limits, can be obtained employing the function `cmfplot`. The following script displays the plot shown in the right panel of Figure 1.

Script 1.

```
allpts <- cmfplot(tau,tempi,nr,out=TRUE,plot=FALSE)
plot(c(0,2180),c(0,0.3),type="n",xlab="Days",ylab="Cumulative Mean Number
of Contacts")
x <- allpts[,1]
y <- allpts[,2]; lines(x,y,lty=1,lwd=2)
y <- allpts[,4]; lines(x,y,lty=3,lwd=1.5)
y <- allpts[,5]; lines(x,y,lty=3,lwd=1.5)
```

The plot shown in the right panel of Figure 3, comparing the cumulative mean number of recurrences in males and females, can be reproduced using script 2. The first row selects male subjects (with the second dummy covariate equal to 0) and then the corresponding cumulative mean function is estimated (calling `cmfplot` and storing the result in `mal`). The third row selects females and then the corresponding cumulative mean function is estimated (calling `cmfplot` and storing the result in `fem`). The rows that follow plot the two "curves".

Script 2.

```
ok <- which(Xcov[,2]==0)
mal <- cmfplot(tau[ok],tempi[ok,],nr[ok],out=TRUE,plot=FALSE)
ok <- which(Xcov[,2]==1)
fem <- cmfplot(tau[ok],tempi[ok,],nr[ok],out=TRUE,plot=FALSE)
plot(c(0,2180),c(0,0.32),type="n",xlab="Days",ylab="Cumulative Mean Number
of Tumours",axes=FALSE)
axis(2)
axis(1,at=c(0,seq(365,365*6,by=365)))
box()
lines(fem[,1],fem[,2],type="l",lwd=2)
lines(mal[,1],mal[,2],type="l",lty=3,lwd=2)
legend(0,0.30,c("females","males"),lty=c(1,2),lwd=2)
```

To obtain the parameter estimates of the LN proportional means regression model shown in Table 2, script 3 was executed. Lines 1-3 evaluate the histotype effect; lines 4-6 evaluate the gender effect; lines 7-9 evaluate the joint effect of histotype and gender; lines 10-12 evaluate the effect of histotype, gender and of the interaction term. In all cases the result of the call to `mfreg` (which is a **R** "list") is stored in the object `fit`; lines 3, 6, 9, 12 of script 3 extract from `fit` the parameter estimates (together with the associated standard errors).

Script 3.

```
xcov <- matrix(Xcov[,1],ncol=1)
fit <- mfreg(tau,tempi,nr,xcov)
fit$estimates
xcov <- matrix(Xcov[,2],ncol=1)
fit <- mfreg(tau,tempi,nr,xcov)
fit$estimates
xcov <- Xcov[,1:2]
fit <- mfreg(tau,tempi,nr,xcov)
fit$estimates
xcov <- Xcov[,1:3]
fit <- mfreg(tau,tempi,nr,xcov)
fit$estimates
```

Finally, to obtain the predicted cumulative mean number of recurrences in males and females and to plot them together with observed ones, script 4 can be employed. The first 14 rows are taken from scripts 2 and 3. In line 15 the predicted cumulative mean number of recurrences for males (whose dummy was coded 0) is calculated (employing the **R** function `cumsum`) after having extracted from the result of the fit the estimated baseline mean function (`fit$basemf`). In line 17 the predicted cumulative mean number of recurrences for females (whose dummy was coded 1) is calculated from the estimated baseline mean function and the gender regression coefficient (`fit$basemf*exp(fit$est[1,1])`. Lines 16 and 18 plot predicted "curves" as dotted lines.

Script 4.

```
ok <- which(Xcov[,2]==0)
mal <- cmfplot(tau[ok],tempi[ok,],nr[ok],out=TRUE,plot=FALSE)
ok <- which(Xcov[,2]==1)
fem <- cmfplot(tau[ok],tempi[ok,],nr[ok],out=TRUE,plot=FALSE)
plot(c(0,2180),c(0,0.32),type="n",xlab="Days",ylab="Cumulative Mean Number
of Tumours",axes=FALSE)
axis(2)
axis(1,at=c(0,seq(365,365*6,by=365)))
box()
lines(fem[,1],fem[,2],type="l",lwd=2)
lines(mal[,1],mal[,2],type="l",lty=3,lwd=2)
legend(0,0.30,c("females","males"),lty=c(1,2),lwd=2)
xcov <- matrix(Xcov[,2],ncol=1)
fit <- mfreg(tau,tempi,nr,xcov)
x <- fit$times; y <- cumsum(fit$basemf)
lines(x,y,lty=3)
x <- fit$times; y <- cumsum(fit$basemf*exp(fit$est[1,1]))
lines(x,y,lty=3)
```