Original article

Scand J Work Environ Health 2012;38(5):447-455

doi:10.5271/sjweh.3276

Hazard functions to describe patterns of new and recurrent sick leave episodes for different diagnoses

by Navarro A, Moriña D, Reis R, Nedel FB, Martín M, Alvarado S

Affiliation: Unitat de Bioestadística, Facultat de Medicina, Campus UAB, 08193 Cerdanyola del Vallès, Spain. albert.navarro@uab.cat

Refers to the following text of the Journal: 2007;33(3):233-239

The following article refers to this text: 2012;38(6):485-488

Key terms: hazard function; occupational health; occupational health; recurrence; risk; sick leave; statistical model; survival analysis

This article in PubMed: www.ncbi.nlm.nih.gov/pubmed/22286954

Additional material

Please note that there is additional material available belonging to this article on the *Scandinavian Journal of Work, Environment & Health* -website.

Original article

Scand J Work Environ Health. 2012;38(5);447–455. doi:10.5271/sjweh.3314

Hazard functions to describe patterns of new and recurrent sick leave episodes for different diagnoses

by Albert Navarro, PhD,[1] David Moriña, MSc,[1, 2] Ricardo Reis, MD, PhD,[1, 3] Fúlvio B Nedel, MD, PhD,[1, 4] Miguel Martín, PhD,[1] Sergio Alvarado, MSc [1, 5]

Navarro A, Moriña D, Reis R, Nedel FB, Martín M, Alvaredo S. Hazard functions to describe patterns of new and recurrent sick leave episodes for different diagnoses. *Scand J Work Environ Health.* 2012;38(5):447–455. doi:10.5271/sjweh.3276

Objectives This study aims to identify the hazard functions that describe the occurrence patterns of new and recurrent sick leave (SL) episodes for mental, respiratory, and musculoskeletal diagnoses.

Methods The data come from a cohort of workers in the Hospital das Clínicas da Universidade Federal de Minas Gerais, Brazil, including all employees working ≥20 hours per week, whose first employment relation with the hospital started between 1 January 2000 and 31 December 2007 (N=1579). We created 15 samples corresponding to combinations of diagnoses causing SL and the number of previous episodes already suffered. We fitted Weibull, log-normal, and log-logistic models by resampling and selected the model having the lowest Akaike information criterion in the greatest number of resamples.

Results Differences were observed in the probability distributions associated with the process generating a SL. Diagnosis showed important differences in terms of risk intensity: mental episodes were the least frequent. There were differences in risk intensity and shape of the function over time depending on the episode number, particularly between the first episode and recurrences. In addition, these differences varied by diagnosis.

Conclusions In most of the samples analyzed, we identified a mixture of distributions, implying a need to revise the statistical methods of analysis for SL occurrence with the aim of obtaining consistent estimates of the risk and the associated factors.

Key terms occupational health; recurrence; risk; statistical model; survival analysis.

Sick leave (SL) is a commonly used outcome in occupational epidemiological studies, and the statistical analysis of its occurrence involves certain difficulties that must be taken into account. First, as for any other probabilistic analysis of a phenomenon, its variability depends on the individual's characteristics and is tackled through statistical modeling. In order to explain the variability of some phenomenon reasonably well, the method used should be appropriate to the behavior and the number of factors considered sufficient to capture a relevant amount of the variability. Moreover, for events that may appear more than once in the same individual, recurrence must be taken into account in order to obtain accurate estimates and efficient inferences. Failing to take account of recurrence in the statistical analysis leads to falsely narrow confidence intervals of the estimates (1), which may in turn lead to factors being regarded as statistically significant when they are not. Also, in certain cases, it may lead to a bias in the estimates (2).

1 Grups de Recerca d'Amèrica i Àfrica Llatines (GRAAL), Unitat de Bioestadística, Facultat de Medicina, Universitat Autònoma de Barcelona, Barcelona, Spain.
2 Centre Tecnològic de Nutrició i Salut (CTNS), TECNIO, CEICS, Reus, Spain.
3 Serviço de Atenção à Saúde do Trabalhador, Universidade Federal de Minas Gerais, Belo Horizonte, Brasil.
4 Grups de Recerca d'Amèrica i Àfrica Llatines (GRAAL), Mestrado em Promoção da Saúde. Departamento de Biologia e Farmácia, Universidade de Santa Cruz do Sul, Santa Cruz do Sul, Brasil.
5 Escuela de Salud Pública, Facultad de Medicina, Universidad de Chile, Santiago de Chile, Chile.

Correspondence to: Albert Navarro, Unitat de Bioestadística, Facultat de Medicina, Campus UAB, 08193 Cerdanyola del Vallès, Spain. [E-mail: albert.navarro@uab.cat]

Given the recurrent nature of the phenomenon, the first problem that may arise is the possibility that the risk of SL varies depending on the number of previous episodes that the worker has suffered, casting doubt on the applicability of classical analysis methods (3). This phenomenon, known as "occurrence dependence", was quantified in detail in a recent article (4) where it was observed (in this same cohort but during a shorter period) that the risk of SL increases as the number of previous episodes increases.

In view of this, another problem in applying classical methods of analysis to the study of SL occurrence would arise if the process generating the SL were also to change according to the number of previous episodes suffered by the individual. In this case, it is logical to think that the probability distribution associated with the process that generates the episodes could be different. If this was the case, we would most likely be facing a phenomenon that was the result of a mixture of distributions. The analysis of such a phenomenon would not be trivial. On the one hand, the analysis of SL occurrence associated with parametric methods would be highly suspicious. Indeed, if the relative performance of statistical methods differs across the generating processes, then studies based upon one process may be misleading (5).

On the other hand, we must take into account that the problems mentioned above may differ depending on the diagnosis associated to the SL being studied. The factors explaining the heterogeneity may not be the same or, even if they are, may have different effects depending on the diagnosis. The intensity of occurrence dependence will be a function of the diagnosis (4). Also, the process generating the SL may differ by diagnosis and be different from previous SL.

Indeed, in order to determine which of the available statistical alternatives is most appropriate for studying the risk of a SL, it would first be necessary to know the mixture of distributions involved in the process that generates the SL since this would allow simulation of realistic data (6) and thus lead to a reliable assessment of the suitability of their application. The development of methodological research seeking more appropriate strategies in this field should permit others working on this topic to conduct more precise analyses of the risk of SL and its associated factors.

The analysis presented here has been based on the assumption that the process generating a SL differs depending on the number of previous episodes suffered and the diagnosis involved. Thus, the objective of the study is to identify the processes resulting in a SL in a range of situations as determined by these two characteristics as a preliminary step to establishing reliable methods for epidemiological analysis in this field.

Methods

Design and participants

The data for this study come from a cohort of workers in the Hospital das Clínicas da Universidade Federal de Minas Gerais, Brasil (7). The general sample used in this paper comprises all employees working ≥20 hours per week, who first began working in the hospital between 1 January 2000 and 31 December 2007 (N=1579). Employees were followed up from their first day working in the hospital until either the conclusion of their contract or the end of the study period (ie, 31 December 2009), whichever came first.

The median follow-up time per worker was 39.0 months (the 10th and 90th percentiles were 6.6 and 85.2, respectively). The workers were mostly women (71.7%), young [75.0% <35 years old, mean=30.6 (SD=7.6) years], and with a high educational level (only 7.6% below secondary grade). Table 1 shows the occupational characteristics of the studied employees.

All SL were medically certified. A "new episode" was defined as the first SL that a worker experienced (ie, since they began work at the hospital) for a given International Classification of Diseases (ICD) diagnosis; "recurrence" was defined as any episode, other than the first, for a given diagnosis. SL absences approved within three days following the finalization of a preceding absence for the same diagnosis code were considered prolongations of the earlier SL and thus not considered a new episode. When we refer to SL due to "all causes", the above criteria are still applicable although the specific diagnosis producing each episode is not taken into consideration. Other aspects related to diagnosis classification are explained in detail in an earlier paper (4), in which a sample from the same cohort was analyzed although for a shorter period.

Analysis strategy

Sixteen different samples were generated, resulting from the combinations formed between the number of previous episodes suffered (0, 1, 2, or 3) and four different categories of reasons for the SL, based on ICD categories (8): (i) diseases of the respiratory system (ICD codes: J00–J99); musculoskeletal diseases (ICD codes: M00–M9); (iii) mental and behavioral disorders (F00–F99); and (iv) all causes (any ICD code, except for pregnancy, childbirth, and the puerperium, as these were excluded from the study).

For the first SL (ie, when there was no record of a previous episode), each sample consisted of all workers meeting the inclusion criteria. From the first episode onwards, the samples were smaller since they corresponded only to workers affected by the number of previous episodes associated with that diagnosis (see figure 1).

Table 1. Sample description

Characteristics	N	%	Worker-months
Occupation [a]			
Science or art professionals	268	17.0	9478.6
Middle level technicians	814	51.6	40435.8
Administrative workers	402	25.5	14929.3
Other workers	95	5.9	4170.1
Employee's working hours			
20–39 hours	592	37.5	21303.3
≥40 hours	987	62.5	47710.5
Employment relation			
Civil servant	473	30.0	28547.5
Outsourced	1106	70.0	40466.3
Type of work			
Medical assistance	846	53.6	39946.4
Supporting	602	38.1	23602.7
Infrastructure	131	8.3	5464.7

[a] According to the Brazilian Classification of Occupations (23)

The sample corresponding to workers who had suffered three previous episodes due to mental and behavioral disorders was excluded from the analysis due to the small number of workers in this situation (N=41). In order to ensure consistency of the estimations, follow-up was cut off at the point when there were <30 workers at risk.

For each of the 15 samples used, 1000 resamples the same size as the original were generated using the Bootstrap re-sampling technique. In each of these, the regression models of interest were fitted, recording the estimated parameters and Akaike criterion value (AIC) (9) for each case. The AIC is used to choose between non-nested models, which are considered better as the value decreases. Thus we compared the AIC for Weibull, log-normal, and log-logistic models in each resample, noting which one yielded the lowest value and selecting the one that was the best-fitting model in the highest percentage of subsamples. The parameters of the distribution were estimated using the median of the coefficients estimated in the resamples.

Parametric models utilized

Parametric survival models were fitted, specifically those for Weibull, log-normal, and log-logistic regression, parameterized as accelerated failure-time (AFT) models (10), the general expression for which is:

$$ln(t_j) = \beta_o + X_j\beta_x + ln(\tau_j) \quad \text{Equation 1}$$

where t_j denotes the random variable for time until occurrence of a SL for worker j, β_0 is the constant of the model, X_j and β_x are vectors of covariates and regression coefficients, respectively, and $ln(\tau_j)$ is the error term associated with the particular regression model. Thus, τ_j follows a Gumbel distribution with form parameter p for the Weibull regression, a standard normal distribution with mean 0 and deviation σ for the log-normal distribution, and a log-logistic distribution with mean 0 and deviation $\pi\gamma/\sqrt{3}$ for the log-logistic distribution.

Since the models in this paper do not include covariates, equation 1 reduces to:

$$ln(t) = \beta_o + ln(\tau) \quad \text{Equation 2}$$

Table 2 presents the functions associated with these distributions. The Weibull model assumes that the hazard increases (p>1) or decreases (0<p<1) monotonically. If p=1, the Weibull model is equivalent to the exponential model, which assumes that the hazard is constant throughout follow-up. The log-normal distribution fits hazard rates that increase over time up to a maximum and then decrease. The log-logistic model with γ<1 produces a hazard function similar to that for the log-normal, whereas with $\gamma \geq 1$ it fits a monotonically decreasing function.

In order to check the fit of the chosen distributions, we graphically compared the cumulative hazard function estimated empirically using the Nelson-Aalen estimator and Cox-Snell residuals obtained from fitting with the chosen distributions and their median parameters. If the model fits the data, the plot should be a straight line with a slope of 1 (10).

The resampling process and the analyses were performed using the Stata statistical package, version 11 (StataCorp, College Station, TX, USA) and the graphics facilities of R 2.12.1 (R Foundation for Statistical Computing, Vienna, Austria).

Results

In total there were 7872 SL episodes and 69 013.8 months of follow-up. The crude incidence rate (IR) was 11.41 SL per 100 worker-months. Figure 1 presents the rates by diagnosis and previous episodes, where it may be seen that for a given diagnosis the rate increases as the number of previous episodes increases. Thus, for example, the IR for the first SL for mental and behavioral disorders is 0.31 SL per 100 worker-months; this increases more than ten-fold when preceded by one or two SL (IR=3.14 and IR=3.47 for the second and third SL, respectively). There was also a pattern whereby as the number of previous episodes increases, the median survival time decreases. For example, for diseases of the respiratory system, the medians were 44.2, 22.6, 17.2 and 14.7 months until the first, second, third, and fourth SL, respectively.

The result of the resampling process is presented in table 3 along with the specification of the chosen dis-

Table 2. Functions involved in the regression models.

Distribution	Survival function	Density function	Parameterization [a]
Weibull [b]	$exp(-\lambda t^p)$	$\lambda p t^{p-1} exp(-\lambda t^p)$	$\lambda = exp(-p\beta_0)$
Log-normal [c, d]	$1 - \Phi\left(\frac{log(t) - \mu}{\sigma}\right)$	$\frac{1}{t\sigma\sqrt{2\pi}} exp\left[\frac{-1}{2\sigma^2}\{\log(t) - \mu\}^2\right]$	$\mu = \beta_0$
Log-logistic [e]	$\frac{1}{1 + (\lambda t)^{1/\gamma}}$	$\frac{\lambda^{1/\gamma} t^{1/\gamma - 1}}{\gamma\{1 + (\lambda t)^{1/\gamma}\}^2}$	$\lambda = exp(-\beta_0)$

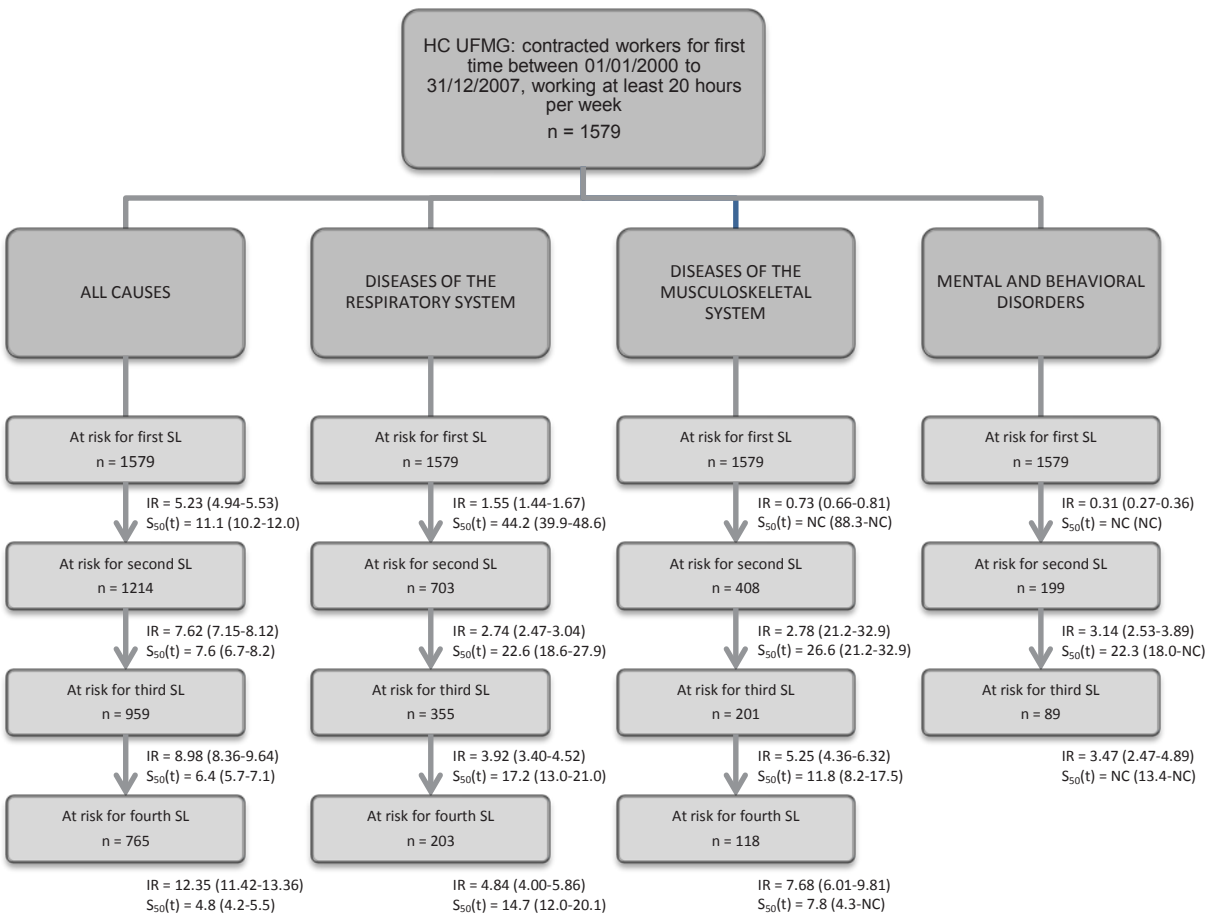
[a] In models with covariates: λ is λ_j, μ is μ_j and τ is τ_j

[b] Weibull: $\lambda_j = exp\{-p(\beta_0 + X_j\beta_x)\}$

[c] Log-normal: $\mu_j = \beta_0 + X_j\beta_x$

[d] Φ is the standard normal cumulative distribution.

[e] Log-logistic: $\lambda_j = exp\{-(\beta_0 + X_j\beta_x)\}$



NC: Not calculable.

Figure 1. Resampling process: workers at risk, observed incidence rate (IR) in worker-months, observed median survival time ($S_{50}(t)$) in months and 95% confidence intervals, in each sample.

Table 3. Summary of results obtained from the Bootstrap resampling process. [AIC= Akaike criterion value.]

Samples	% of resamples obtaining the lowest AIC			Best	β_0 [a]			Ancillary [b]		
	Weibull	Log-normal	Log-logistic		Median	P2.5	P97.5	Median	P2.5	P97.5
All causes										
First episode	0	2.9	97.1	Log-logistic	5.843	5.782	5.912	0.700	0.668	0.736
Second episode	77.9	0	22.1	Weibull	5.944	5.864	6.032	0.797	0.761	0.840
Third episode	50.2	0	49.8	Weibull	5.782	5.693	5.877	0.822	0.779	0.870
Fourth episode	91.6	0	8.4	Weibull	5.469	5.378	5.561	0.858	0.808	0.909
Diseases of the respiratory system										
First episode	0	64.2	35.8	Log-normal	7.195	7.107	7.290	1.498	1.410	1.589
Second episode	11.5	0.4	88.1	Log-logistic	6.583	6.443	6.720	0.924	0.845	1.003
Third episode	83.1	0.1	16.8	Weibull	6.678	6.526	6.842	0.923	0.822	1.048
Fourth episode	97.2	0.7	2.1	Weibull	6.430	6.253	6.612	1.051	0.896	1.243
Musculoskeletal system										
First episode	0.4	44.8	54.8	Log-logistic	7.974	7.853	8.100	0.836	0.776	0.911
Second episode	84.0	9.4	6.6	Weibull	7.109	6.938	7.309	0.758	0.682	0.856
Third episode	5.5	93.3	1.2	Log-normal	5.853	5.540	6.182	1.989	1.739	2.224
Fourth episode	0	100	0	Log-normal	5.495	5.048	6.058	2.204	1.885	2.569
Mental and behavioral disorders										
First episode	0.7	66.7	32.6	Log-normal	8.924	8.702	9.215	1.545	1.377	1.746
Second episode	11.6	88.4	0	Log-normal	6.650	6.202	7.146	2.399	2.090	2.745
Third episode	6.4	93.3	0.3	Log-normal	6.696	6.098	7.480	2.246	1.750	2.778

[a] β_0 expressed in days.
[b] Ancillary: p for the Weibull, σ for Log-normal and γ for log-logistic model.

tribution and the estimated parameters. Figure 2 shows the estimated hazard rate functions based on the distributions identified in the resampling process, projected to a maximum of 120 months.

For SL due to diseases of the respiratory system, four hazard function curves are drawn corresponding to baseline hazards differing in intensity. The form of the function is rather different depending on whether we are dealing with the first SL (for which it increases up to month 12 then falls very gradually), the second SL (rises for months 1 and 2 then falls more steeply than the preceding SL), or the third and fourth SL where one increases and the other decreases over time, although so gradually that the form parameter *p* of the corresponding Weibull distributions is not statistically different from 1.0 (indicative of an exponential model with constant hazard).

SL due to diseases of the musculoskeletal system present certain differences, particularly between the first SL and the rest. The first SL, with much lower incidence than the recurrences, increases gradually with time, whereas the rest decline. The second falls more gently than the third and fourth with the consequence that, after about the 20th month, the hazard rate is greater after suffering one previous SL than after suffering two or three.

With regard to the SL caused by mental and behavioral disorders, there is a clear difference between the patterns over time for the first SL versus the rest. The first presents a sustained increase over time, although remaining within the lowest risk range, between 0.0005–0.003 SL per person-month. The risk functions for the second and third SL are almost identical, increasing up until the fourth month then decreasing monotonically.

The hazard rate for the first episode, independently of the reason, increases until the sixth month and subsequently declines gradually. The hazard for the second SL, higher than that for the first, declines until it practically coincides with the hazard for the first SL between months 9–20, and subsequently continues to decline, although more gradually than for the first. The functions for the third and fourth SL decrease, gradually tending to converge with time.

The supplementary material (http://www.sjweh.fi/data_repository.php) includes a graphical assessment of the goodness-of-fit of the distributions chosen for each case (see table 2) and the original samples. The curves obtained confirm the "good fit" of the various estimated distributions. In addition, the supplementary material also shows the results of the censoring of the distributions.

Discussion

The statistical analysis of the occurrence of sick leave absences is a complex task. One of the various difficulties is occurrence dependence, in other words that the risk of a SL episode depends on the previous episodes suffered by the worker, and moreover with intensities that differ

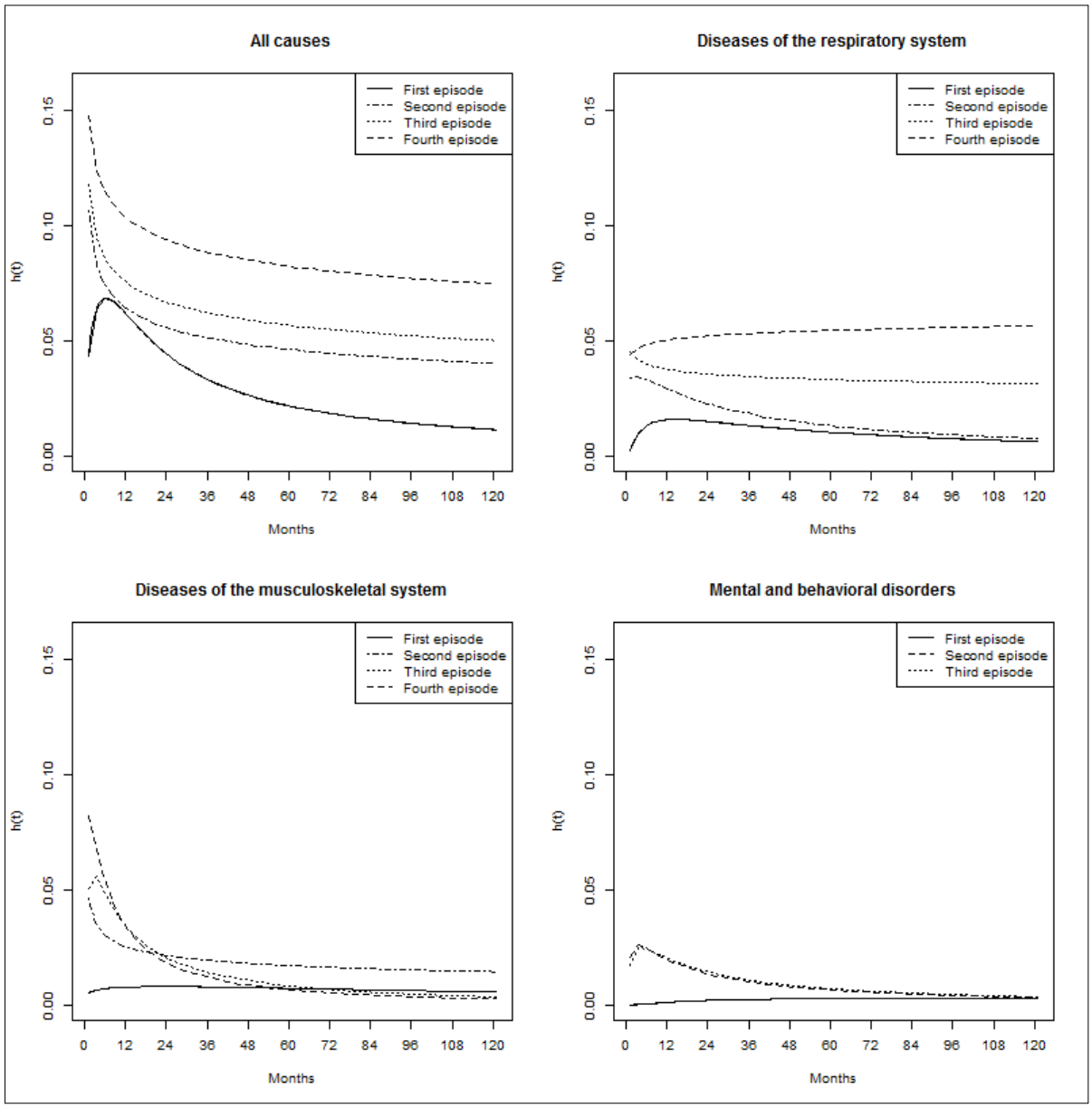


Figure 2. Estimated hazard rate functions, expressed in sick leave (SL) per worker-month.

depending on diagnosis (4). If, in addition, the process generating the SL differs by diagnosis and the previous SL, it implies even greater complexity in tackling the analysis. This paper provides evidence in this respect.

When the cause of the SL is not taken into account, the first episode is best fitted by a log-logistic distribution, while the recurrences are best fitted by a Weibull distribution. It may be observed that, for the first SL episode, the hazard rate rises during the first six months, and subsequently declines gradually, whereas the hazard functions for subsequent occurrences decline rapidly after the first month. This is probably a consequence of the fact that the most vulnerable workers suffer a SL relatively soon, so the curve corresponds to the healthier workers. After the sixth month, the form of the hazard functions for the various episodes do not appear to differ greatly, suggesting that the distribution of risk over time is similar for the different causes, although their intensities differ and the hazards all diminish over time.

It must be pointed out that the hazard functions presented in this article represent the group at risk and not a particular individual. That the hazard function decreases over time does not necessarily mean that the

risk for an individual does. This phenomenon is known as the "frailty effect" (11) and arises since the more frail individuals have a greater hazard and are more likely to suffer the SL earlier. Consequently, over time, the "at risk group" has an increasing proportion of less frail individuals, decreasing the population average hazard without this necessarily meaning a decline in the "individual" hazard.

With regard to SL due to respiratory diseases, the majority of which are acute upper-respiratory infections, we note that the curves of the hazard function represent more moderate intensities than is the case for the other diagnoses. Compared with other diagnoses, no important frailty effect may be observed. This seems logical given that the incidence of respiratory disease-related SL is generalized to the set of all workers and recurrence is common due to the fact that we are dealing with repeated exposures that do not entail subsequent immunity.

For SL due to diseases of the musculoskeletal system or to mental and behavioral disorders relatively similar patterns of risk may be observed: a relatively constant and low intensity hazard is estimated for the first episode, while the hazard declines fairly steeply over time for recurrences. For SL associated with the musculoskeletal system, the first recurrence has a more moderate decline than the second and third. Recurrences associated with mental and behavioral disorders present almost identical functions that decline steeply after the third month. Workers who manifest musculoskeletal problems or mental and behavioral disorders have a high propensity for recurrences, even though, in our data, there do not appear to be great differences in the patterns of risk for the various recurrences. For SL related to musculoskeletal or mental/behavioral disorders, we would be dealing with a "hurdle" phenomenon (12) when using terminology specific to count models, whereby occurrence of SL have two components. The first determines who "jumps over the hurdle", in other words, the component explains why some workers have an initial episode. The other component – with a completely different behavior – refers only to "sick" workers and implies a much higher risk with some individuals having high frailty.

It is possible that this pattern is repeated for SL associated with other reasons. In the process that governs recurrence, various factors may coincide. In some cases, the main cause may be the history of the disease itself that generates the SL, for example in the case of chronic diseases. At other times, it may be mainly attributable to the occupational risks to which the worker is exposed or due to the coping behavior of the worker, aiming to maintain his or her health and working capacity (13). And we must not forget that these factors act differently depending on social class (14, 15); moreover, they do not act in isolation at different times in the life of the worker, but rather interact in a complex manner. In the field of statistical analysis of SL occurrence, there are two articles (16, 17) which propose the use of certain non-parametric models – modifications of the classic Cox model – for the analysis of recurrent phenomena. However, neither article identifies the process generating the episodes. Christensen et al (16), without giving a reason for the choice, conduct simulations assuming that the time until the first episode follows an exponential distribution, whereas all subsequent episodes are generated depending on the duration of the immediately preceding episode. In contrast, Navarro et al (2) used empirical data combined with resampling techniques. On the other hand, Koopmans et al (17) studied the distribution associated with sickness absence in the particular context of long duration episodes (>6 consecutive weeks) and without differentiating between new and recurrent episodes; their conclusion was that the exponential distribution appears to be a good choice due to its simplicity.

In any case, the choice of modified Cox models for the study of recurrent phenomena would appear to represent a logical and conservative option since they provide for the handling of occurrence dependence in various ways and, being based on non-parametric models, do not require the form of baseline risk to be preestablished. Parametric survival analysis models, on the other hand, are stricter in this sense since each assigns a particular distribution to the baseline risk, the regression model being named according to this distribution. Thus the application of either of these models would require specifying the distribution that generates the process at the outset, but this is unknown for occurrences of SL. The nontrivial advantage of the parametric models, provided the baseline risk distribution correct, is that they permit obtaining more efficient estimators and hence more precise inferences can be made.

Thus it is fundamental to assess which is the best analytical strategy for studying the risk of a SL. Having more exact estimates of the associated risk factors would provide a better basis for the actions to be developed. It is important that simulation studies of statistical methods for recurrent events include simulated data sets based on a range of models for event generation (5). The present article presents the distributions associated with SL in various settings, as well as those corresponding to censored data (see Appendix), and hence may serve as a basis for future research in this area.

The results of this study have been obtained in a hospital setting and as a result caution must be exercised in their generalization to other occupational settings. Regulations governing SL vary between countries and the specific definitions of what is considered a new episode or a recurrence affect the quantification of the hazard functions. How a "new" episode is defined is important to consider. All workers in this study were first employed

in the hospital during the follow-up period. This means that a "new" episode, strictly speaking, refers to the first SL in this job, even though some workers may have had sick leave episodes in previous jobs. In any case, it should be noted that, for the most part, we are dealing with young workers with high levels of education, tending to imply they would have fairly short occupational histories. Moreover, this classification of "new episode" versus "recurrence" has already been demonstrated to have good discriminatory power in a previous article (4). In fact, from an operational point of view, it is more useful – and much more feasible – to know about the first episode in the present job than to know how many episodes each worker has had during their entire occupational life. Despite the limitations mentioned, our paper provides valuable, new information in relation to the process generating the SL and cautions about the need to revise the way the quantitative analysis is tackled.

In summary, the risk of occurrence of a SL differs depending on the diagnosis and the number of episodes previously suffered. In general, there are notable differences between the hazard function curves for the first episode and recurrences, both in terms of the intensity of risk and their shape, suggesting that they represent different phenomena. It is very likely that these differences are due to the configuration of the risk set: for the first SL, it includes all workers, whereas for recurrences it mostly consists of "sick" workers (those who had already recorded a SL). Thus in studying the risk of SL occurrence and its associated factors, the methods employed should enable the tackling of the presence of new episodes and recurrences in the data. There are currently some methods that incorporate this information and allow for an integrated analysis, for example using survival models with stratified baseline risk (18, 19) or the use of frailty terms (20–22).

Therefore, in the study of SL risk – depending on the diagnosis and whether the episode is new or a recurrence – it is advisable to take into account that the behavior of the phenomenon is different and consequently cannot be studied using the same methods of analysis. The present results expose these differences and offer a way to evaluate which specific methods would perform best in each of the different situations considered. This should serve as a reference for tackling the statistical modeling in subsequent studies investigating SL absences.

Acknowledgments

This study was partially funded by the Fondo de Investigación Sanitaria, Instituto de Salud Carlos III, Ministerio de Ciencia e Innovación, Gobierno de España (project PI080703).

References

1 Glynn RJ, Stukel TA, Sharp SM, Bubolz TA, Freeman JL, Fisher ES. Estimating the variance of standardized rates of recurrent events, with application to hospitalizations among the elderly in new england. Am J Epidemiol. 1993;137:776–86.

2. Navarro A, Reis RJ, Martin M. Some alternatives in the statistical analysis of sickness absence. Am J Ind Med. 2009;52:811–6. http://dx.doi.org/10.1002/ajim.20739.

3. Navarro A, Ancizu I. Analyzing the occurrence of falls and its risk factors: Some considerations. Prev Med. 2009;48:298–302. http://dx.doi.org/10.1016/j.ypmed.2008.12.019.

4. Reis RJ, Utzet M, La Rocca PF, Nedel FB, Martin M, Navarro A. Previous sick leaves as predictor of subsequent ones. Int Arch Occup Environ Health. 2011;84:491–9. http://dx.doi.org/10.1007/s00420-011-0620-0.

5. Metcalfe C, Thompson SG. The importance of varying the event generation process in simulation studies of statistical methods for recurrent events. Stat Med. 2006;25:165–79. http://dx.doi.org/10.1002/sim.2310.

6. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Stat Med. 2006;25:4279–92. http://dx.doi.org/10.1002/sim.2673.

7. Reis RJ, de Freitas La Rocca P, Basile L, Navarro A, Martin M. Cohort profile: The hospital das clinicas cohort study, Belo Horizonte, Minas Gerais, Brazil. Int J Epidemiol. 2008;37:710–5.

8. World Health Organisation. International statistical classification of diseases and related health problems: Tenth revision. Geneva: WHO; 2005.

9. Akaike H. A new look at the statistical model identification. IEEE Trans Automatic Control. 1974;19:716–23. http://dx.doi.org/10.1109/TAC.1974.1100705.

10. Cleves MA, Gould WW, Gutierrez RG. An introduction to survival analysis using stata. Rev ed. College Station, TX: Stata Press; 2004.

11. Kleinbaum DG, Klein M. Survival analysis: A self-learning text. 2nd ed. New York, NY: Springer Science+Business Media, Inc.; 2005.

12. Mullahy J. Specification and testing of some modified count data models. J Econom. 1986;33:341–65. http://dx.doi.org/10.1016/0304-4076(86)90002-3.

13. Kristensen TS. Sickness absence and work strain among danish slaughterhouse workers: An analysis of absence from work regarded as coping behaviour. Soc Sci Med. 1991;32:15–27. http://dx.doi.org/10.1016/0277-9536(91)90122-S.

14. Kristensen TR, Jensen SM, Kreiner S, Mikkelsen S. Socioeconomic status and duration and pattern of sickness absence. A 1-year follow-up study of 2331 hospital employees. BMC Public Health. 2010;10:643. http://dx.doi.org/10.1186/1471-2458-10-643.

15. Vahtera J, Virtanen P, Kivimaki M, Pentti J. Workplace as an origin of health inequalities. J Epidemiol Community Health. 1999;53:399–407. http://dx.doi.org/10.1136/jech.53.7.399.

16. Christensen KB, Andersen PK, Smith-Hansen L, Nielsen ML, Kristensen TS. Analyzing sickness absence with statistical models for survival data. Scand J Work Environ Health. 2007;33:233–9. http://dx.doi.org/10.5271/sjweh.1132.

17. Koopmans PC, Roelen CA, Groothoff JW. Parametric hazard rate models for long-term sickness absence. Int Arch Occup Environ Health. 2009;82:575–82. http://dx.doi.org/10.1007/s00420-008-0369-2.

18. Kelly PJ, Lim LL. Survival analysis for recurrent event data: An application to childhood infectious diseases. Stat Med. 2000;19:13–33. http://dx.doi.org/10.1002/(SICI)1097-0258(20000115)19:1<13::AID-SIM279>3.0.CO;2-5.

19. Therneau TM, Grambsch PM. Modeling survival data: Extending the cox model. New York: Springer; 2000.

20. Clayton D. Some approaches to the analysis of recurrent event data. Stat Methods Med Res. 1994;3:244–62. http://dx.doi.org/10.1177/096228029400300304.

21. Hougaard P. Frailty models for survival data. Lifetime Data Anal. 1995;1:255–73. http://dx.doi.org/10.1007/BF00985760.

22. Hougaard P. Analysis of multivariate survival data. New York: Springer; 2000. http://dx.doi.org/10.1007/978-1-4612-1304-8.

23. Ministério do Trabalho e do Emprego. Classificação Brasileira de Ocupações: CBO-2002 [Brazilian classification of occupations: CBO-2002]. 2nd ed. Brasília: Ministério do Trabalho e do Emprego; 2002.

Received for publication: 16 June 2011