World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:2, No:5, 2008

On Methodologies for Analysing Sickness Absence Data: An Insight into a New Method

Xiaoshu Lu, Päivi Leino-Arjas, Kustaa Piha, Akseli Aittomäki, Peppiina Saastamoinen, Ossi Rahkonen, and Eero Lahelma

***Abstract*—**Sickness absence represents a major economic and social issue. Analysis of sick leave data is a recurrent challenge to analysts because of the complexity of the data structure which is often time dependent, highly skewed and clumped at zero. Ignoring these features to make statistical inference is likely to be inefficient and misguided. Traditional approaches do not address these problems. In this study, we discuss model methodologies in terms of statistical techniques for addressing the difficulties with sick leave data. We also introduce and demonstrate a new method by performing a longitudinal assessment of long-term absenteeism using a large registration dataset as a working example available from the Helsinki Health Study for municipal employees from Finland during the period of 1990-1999. We present a comparative study on model selection and a critical analysis of the temporal trends, the occurrence and degree of long-term sickness absences among municipal employees. The strengths of this working example include the large sample size over a long follow-up period providing strong evidence in supporting of the new model. Our main goal is to propose a way to select an appropriate model and to introduce a new methodology for analysing sickness absence data as well as to demonstrate model applicability to complicated longitudinal data.

***Keywords*—**Sickness absence, longitudinal data, methodologies, mix-distribution model.

I. Introduction

THE increasing direct costs of work absences have challenged government's policymakers, public authorities, insurance companies and employers to find ways to reduce the heavy economic and social burdens. In Europe, sick leave policy is one of the top policy priorities [1]. The literature of sickness absence is increasing. Socio-economic, demographic, occupational status, and work-related and economic factors are important determinants of sickness absences [2]-[10].

Musculoskeletal and stress-related disorders cause a large proportion of sickness absences [11]-[17]. Lifestyle-related risk factors of sickness absences such as obesity, tobacco use, alcohol intake, and physical inactivity have also been identified [18]-[21].

A systematic review of earlier studies on the association between risk factors and sickness absence was presented in [22]. The review concluded that the knowledge of the causes and consequences of sick leave is still limited. In addition, it is difficult to generalize the overall results across studies from different countries because of the large differences in economic and social environments which affect both the studied predictors and the sick leave outcomes. Indeed, sickness absence is a dynamic temporal behavior and many risk factors vary over time. Much of the literature on this topic, however, lacks information on such aspects of the phenomenon of sickness absences and influential factors.

In this paper, we are going to particularly address some methodological difficulties in terms of statistical models with complicated sick leave data. We also demonstrate a new method by performing a longitudinal assessment of long-term absenteeism using a large registration dataset as a working example, available from the Helsinki Health Study for municipal employees from the City of Helsinki in Finland during the period of 1990-1999. The strengths of the material include a large sample size and a long follow-up period with ten waves of data collection.

Manuscript received April 29, 2008.

Xiaoshu Lu is with the Finnish Institute of Occupational Health, Topeliuksenkatu 41 a A, FIN-00250 Helsinki, Finland (corresponding author; phone: 358-30-4742505; fax: 358-30-4742008; e-mail: xiaoshu@cc.hut.fi).

Päivi Leino-Arjas is with the Finnish Institute of Occupational Health, Topeliuksenkatu 41 a A, FIN-00250 Helsinki, Finland (e-mail: päivi.leino-arjas@ttl.fi).

Kustaa Piha is with the Department of Public Health, University of Helsinki, Finland (e-mail: Kustaa.Piha@helsinki.fi).

Akseli Aittomäki is with the Department of Public Health, University of Helsinki, Finland (e-mail: Akseli.Aittomäki@helsinki.fi).

Peppiina Saastamoinen is with the Department of Public Health, University of Helsinki, Finland (e-mail: Peppiina.Saastamoinen @helsinki.fi).

Ossi Rahkonen is with the Department of Public Health, University of Helsinki, Finland (e-mail: Ossi.Rahkonen@helsinki.fi).

Eero Lahelma is with the Department of Public Health, University of Helsinki, Finland (e-mail: Eero.Lahelma@helsinki.fi).

II. Commonly Used Models and Model Applicability

To date, in the literature on sickness absences, cross-sectional studies prevail. Such a data design is not always appropriate for analyzing sickness absenteeism. Moreover, in both cross-sectional and longitudinal studies, sick leave data are often highly skewed and clustered at zero. In rare cases in which the dependent variable, for example the duration of absences, has a normal distribution, linear regression model is often a first choice.

In this section, we shall discuss some of the widely used statistical models employed for sick leave analysis. We provide a brief outline of the model equations without going into details. Since it is out of the scope of this paper to give a thoroughly review of these methods, we have chosen to focus on introducing representative models to highlight some important methodological issues we may face with sick leave data. However, we should, whenever possible, bear in mind that the model chosen is always dependent upon the problem specification and data at hand.

A. Linear Regression

Let the observation y_i be a realization of dependent variable Y_i which has a normal distribution with mean μ_i and variance σ^2 as

$$Y_i \sim \mathrm{N}(\mu_i, \sigma^2) \tag{1}$$

Suppose we have data on predictors $x_1,...,x_p$ which take values $x_{i1},...,x_{ip}$ for the ith unit. Then a linear model refers to a simple mapping between μ_i and the predictors as

$$\mu_i = \boldsymbol{x}_i'\theta \tag{2}$$

where $\boldsymbol{x}_i$ is a vector with the values of the p predictors for the ith unit and θ a vector containing the p regression coefficients. The maximum likelihood (ML) parameter estimates θ can be obtained with well-known least squares method.

To examine the trends in socio-economic differences in long sickness absence spells, a linear regression model was adopted for analyzing longitudinal sick leave data [23]. Serial correlation due to the repeated observations over the years was not taken into account in their studies.

B. Logistic Regression

We only illustrate the case where the observation y_i is the binary coded as zero or one for convenience. Then the dependent variable Y_i takes the values zero and one with probabilities π_i and 1-π_i. The distribution of Y_i is a *Bernoulli* as

$$Y_i \sim \mathrm{B}(\pi_i) \tag{3}$$

The logistic model supposes further that the *logit* of the probability π_i is a linear function of the predictors expressed as

$$logit(\pi_i) = \boldsymbol{x}_i'\theta \tag{4}$$

where vectors of predictors and regression coefficients are $\boldsymbol{x}_i$ and θ.

Note that the logistic model is a generalized linear model with link function *logit*. Hence the regression coefficients can be interpreted along the same lines as in linear model. However, as the left-hand-side of (4) is a *logit* but not a mean, θ_j presents the change in the *logit* of the probability associated with a unit change in the jth predictor if all other predictors are held constant. Logistic regression is a useful way of describing the relationship between risk factors and occurrences or incidences of sick leave [12], [16].

In examining the long and short-term economic incentives inherent in the sickness and unemployment insurances, sickness duration was modelled with a linear regression and the outcome of healthy and non-healthy was modelled with a logistic regression. These two models were not linked statistically [9].

This type of model is often employed for studies on cross-sectional or pooled sick leave data in most of the published research reports. Obviously, some interesting and, possibly, important time series issues involved in sick leave data can not be properly studied with cross-sectional or pooled data which can be an important risk factor. For example, in the study of absenteeism costs for 1284 hourly workers from a manufacturing company, the best single predictor of future absenteeism was claimed to be the past absenteeism [24]. Furthermore, there are strong heterogeneties among individuals for some risk factors as described in Introduction and found in literature review. These suggest that a longitudinal design is more appropriate. Longitudinal analysis becomes even more difficult when analyzing sickness absence data. Very often, the data have a large number of values centered at zero and skewing of the rest of the values. Furthermore, the observations are likely correlated as the data are collected over time for the same samples.

With regard to the above-described concerns, Poisson regression and zero-inflated Poisson models are often applied when the outcome of sick leave counts is the focus.

C. Poisson Regression and Zero-inflated Poisson Regression

Suppose a sample of observation y_i is a realization of the dependent Poisson variable Y_i which takes the integer values as 0, 1, 2…and has Poisson distribution with both mean and variance μ_i as

$$Y_i \sim \mathrm{P}(\mu_i) \tag{5}$$

The Poisson regression models the mean or variance as

$$log(\mu_i) = \boldsymbol{x}_i'\theta \tag{6}$$

where vectors of predictors and regression coefficients are $\boldsymbol{x}_i$ and θ.

Note that the Poisson model is again another type of generalized linear model with link function *log*. In the model, the regression coefficient θ_j presents the expected change in the *log* of the mean per unit change in the predictor x_j. Increasing x_j by one unit is associated with an increase of θ_j in the *log* of the mean.

Using a register-based cohort of all live-born in Norway between 1967 and 1976, the extent to which musculoskeletal sickness absence was influenced by a range of circumstances concerning family background and health in early life was investigated with Poisson regression model [11]. In accounting for non-normal distribution of the dependent variable of absence spells in most of the cases, Poisson regression has been found to be superior to linear regression in sick leave absence studies [25]-[26].

Indeed, the often encountered non-normal distribution of the outcome is also a threat to the validity of the commonly adopted statistical analysis as we mentioned before. Treating the data as they were normally distributed is inappropriate which may lead to the wrong conclusions. One obvious and simple way to avoid such difficulty is to use distribution-free or nonparametric approaches. Therefore, many studies performed logistic regression models as cited before (e.g. [3], [17], [27]). Note that this model discards the important 'duration' information of sick leave counts which is obviously important for us to understand sick leave behaviors.

As a result, zero-inflated Poisson (ZIP), zero-inflated binomial (ZIB) and zero-inflated negative binomial (ZINB) models are often the best way to model zero-clustered sick leave data. A literature review shows that only ZINB model has been employed to study sick leave data. Using cross-sectional data, the associations between self-reported health problems and sickness absence from work were analyzed [2]. ZINB model was used in their statistical analysis [2].

The link functions of the models of ZIP, ZIB, ZINB are the same as that of Poisson regression model. The only difference among these models is the count probability distributions in which ZIP, ZIB and ZINB allow for over-dispersion. Such feature can be modified further to best suit particular data structures and study aims. We previously tried ZIP and ZINB models for our data (see below for description) with the WinBUGS software. Unfortunately, WinBUGS could not handle such an oversized dataset.

III. A New Model: Two-Part Mixed-Distribution Model

Here we propose a more flexible model: the two-part mixed-distribution model originated in econometrics studies [28]-[29]. The original model presented one part for the probability of occurrence of nonzero observations (a probit or logit model) and one part for the probability distribution of the nonzero observations. The two parts were assumed to be not connected. Recently, the model has been extended as a mixed-distribution model for longitudinal data. Random effects have been included in the two parts which are allowed to be linked [30]-[32]. Tooze et al. also implemented the model in a SAS Macro called MIXCORR (available from the first author) [32].

The dependent variable Y_{it} has continuous distribution and takes the observed value y_{it} for subject i at time t. Let R_{it} denote the occurrence variable defined as

$$R_{it} = \begin{cases} 0, \text{ if } Y_{it} = 0 \\ 1, \text{ if } Y_{it} > 0 \end{cases} \tag{7}$$

with conditional probabilities

$$\Pr(R_{it} = r_{it} \mid \theta_1) = \begin{cases} 1 - p_{it}(\theta_1), \text{ if } r_{it} = 0 \\ p_{it}(\theta_1), \text{ if } r_{tj} = 1 \end{cases} \tag{8}$$

where $\theta_1 = (\beta'_1,\ u_{1i})'$ is a vector of fixed (β_1) and random occurrence (u_{1i}) effects.

One part of the two-part mixed-distribution model, logistic model, for occurrence takes the form

$$logit(p_{it}(\theta_1)) = X'_{1it}\beta_1 + u_{1i} \tag{9}$$

where X_{1it} is a vector of covariates for occurrence.

Another part of the two-part mixed-distribution model, log-normal model, for duration variable $Y^{+}_{it} \equiv [Y_{it} | R_{it} = 1]$ takes the following form

$$log(Y^{+}_{it}|\theta_2) \sim \mathrm{N}(X'_{2it}\beta_2 + u_{2i},\ \sigma^2_e) \tag{10}$$

where X_{2jt} is a vector of covariates for duration and $\theta_2 = (\beta'_2, u_{2i})'$ is a vector of fixed (β_2) and random intensity (u_{2i}) effects.

The two-part models are assumed to be correlated with the following correlation matrix as

$$\begin{bmatrix} u_{1i} \\ u_{2i} \end{bmatrix} \sim \mathrm{BVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^{\ 2} & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^{\ 2} \end{bmatrix} \right) \tag{11}$$

through the coefficient ρ. If $\rho = 0$, the two parts are not correlated.

It is easy to verify that the p.d.f., $f(y_{it}|\theta)$, has mixed distributions. The likelihood is maximized to get the estimated β_1 , β_2, σ_1, σ_2, σ_e, ρ. Akaike's Information Criterion (AIC) can be used to compare the model's goodness of fit which is constructed by penalising the log-likelihood for the number of parameters [33].

IV. Working Examples

A. Data

Since our main goal is to propose a way to select an appropriate model and to introduce a new methodology for analysing sickness absence data as well as to demonstrate model applicability to such complicated longitudinal data, we choose a large registration sick leave dataset with long-term outcomes as a working example to demonstrate the effectiveness of our proposed new model. The follow-up consists of ten waves of data collection. From a methodological point of view, our focus is to show superior performance of the introduced new model. Based on this argument, the data, data variables and analysis results will be reported briefly. More details about the data background can be found in [23].

Data are from the Helsinki Health Study on health and well-being for municipal employees from the City of Helsinki in Finland which cover all employees' information on sickness absence and other individual level information. Spells of sickness absences are grouped as short-term sickness absences with 0–3 days and medically confirmed long-term sickness absences with over 3 days. We select long-term sickness absence as an outcome which is further processed as a continuous variable in order to meet the requirements of two-part mixed-distribution model in the following way: The annual sickness absence rate is expressed as a percentage of total working days less holidays, including public holidays, i.e. the percentage of the number of absence days (short or long-term) due to work divided by total possible working days (maximum value is 100). Socio-economic, demographic and occupational characteristics of the employees are selected as independent variables. In particular we investigate weather differences in socio-economic and occupational characteristics are important explanatory factors.

Table I briefly describes the selected variables studied in this analysis. Fig. 1 shows average temporal changes for long-term sickness absences, indicating that the changes are nonlinear. Therefore we add $\text{YEAR}^2 = \text{YEAR*YEAR}$ as an extra predictor.

TABLE I
BRIEF DESCRIPTION OF THE DEPENDENT AND INDEPENDENT VARIABLES

Variables	Total sample number = 50256
ABSENCE	annual rate of long-term absence (0-100), continuous
YEAR	year (1990-1999), continuous
AGE	age (18-64), continuous
GENDER	gender, categorical M: male F: female
SES	socio-economic class (1-4),categorical 1: managers and professionals 2: semi-professionals 3: routine non-manuals 4: manual worker
EDU	educational background (1-5),categorical
INCOME	logarithm of annual income, continuous
CONTRACT	employment contract type (1-4), categorical

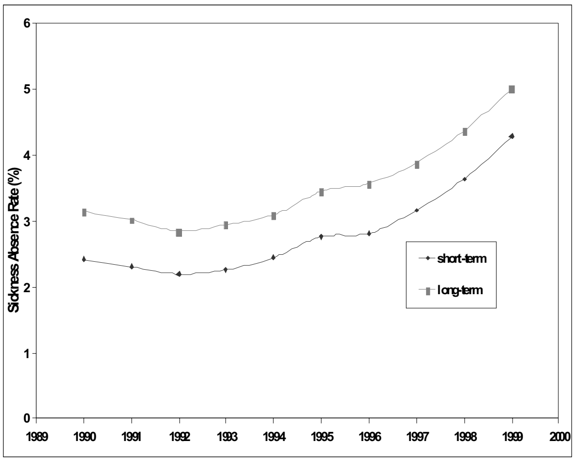

Fig. 1 Average rates of short and long-term sickness absences from 1990 to 1999

Because many employees did not have any absenteeism, the dependent variables have many zero observations. Take the year 1999 as an example, the percentages of zeroes are 55% and 66% for male and female, respectively.

B. Model Comparison Results

For comparison purpose, a general linear model is also selected based on the following reasons: firstly it is one of the commonly used models for analyzing sickness absence data as reviewed previously. Secondly, logistic models are not considered due to the unavailability of 'duration' information when they are applied. Such information is very important for understanding sickness absence performance. Thirdly, as mentioned before zero-inflated Poisson models are difficult to implement through WinBUGS owing to the oversize of our working data.

The following notations are adopted in the model: i denotes the ith individual; t denotes the tth year. We shall assess the accuracy of the proposed model by comparing prediction performances from these three models. The models are fit separately to each gender.

• Model 1: General linear model:

$$log(\text{ABSENCE}_{it})= \theta_0+\theta_1\text{YEAR}_{it}+\theta_2\text{YEAR}^2_{it}+\theta_3\text{AGE}_{it}+\theta_4\text{SES}_{it}+\theta_5\text{EDU}_{it} +\theta_6\text{INCOME}_{it}+\theta_7\text{CONTRACT}_{it}+\varepsilon_{it}$$

$$\text{where } \varepsilon_{it}\sim \text{N}(0, \sigma^2) \tag{12}$$

• Model 2: Uncorrelated model:

Denote

$$Y_{it} = \text{ABSENCE}_{it} \tag{13}$$
$$X_{1it} = (\text{YEAR}_{it}, \text{YEAR}_{it}^2, \text{AGE}_{it}, \text{SES}_{it}, \text{EDU}_{it}, \text{INCOME}_{it}, \text{CONTRACT}_{it}) \tag{14}$$
$$X_{2it} = (\text{YEAR}_{it}, \text{YEAR}_{it}^2, \text{AGE}_{it}, \text{SES}_{it}, \text{EDU}_{it}, \text{INCOME}_{it}, \text{CONTRACT}_{it}) \tag{15}$$

The logistic equation is presented in (9) and log-norm equation in (10) with $\rho = 0$ in (11).

• Model 3: Correlated model:

Equations of both the logistic part and the log-normal part are the same as (13)-(15) but $\rho \neq 0$ in (11).

The comparison results are displayed in Table II for male staff. We only show some of the most relevant results here.

TABLE II
MODEL COMPARISON AND STATISTICS FOR LONG-TERM SICKNESS ABSENCES; MALE

Variables	Model 1 Parameter value (SE)	Model 2 Parameter value (SE)	Model 3 Parameter value (SE)
		Logistic Model	Logistic Model
σ^2		2.3389^{***} (0.0822)	2.3318^{***} (0.0809)
-2 Res Log AIC		110932.0 110974.0	
		Log-Normal	Log-Normal
σ^2	44.7076^{***} (0.3091)	0.4106***(0.0133)	0.3843^{***} (0.0120)
$\rho\sigma_1\sigma_2$			0.5655^{***} (0.0246)
-2 Res Log AIC	277824.9 277826.9	46849.02 46889.02	
-2 Res Log AIC		sum = 157781 sum = 157863	sum = 157114.1 sum = 157198.1 Diff in -2ll 666.91 in p<0.0001

$^{***}p<0.001$, $^{**}p<0.05$, $^{*}p<0.1$.

Table II illustrates that the AIC estimates of log-normal models for Model 2 is smaller than that of the general linear Model 1 indicating that Model 2 fits the data better. Note that

Model 1 is also a *log* linear regression. So the comparison is valid. The value '*Diff in -2ll 666.91 in p<0.0001*' shown in the right down column in Table II indicates that the two-part mixed-distribution Model 3 fits the data significantly better than the uncorrelated Model 2 does. This claim can be also concluded from the value $\rho\sigma_1\sigma_2$ (0.5655^{***}) also, reporting that the two-part logistic and log-norm models are significantly correlated. Overall, we demonstrate significantly better performance of the proposed new model. It is worth noting that similar model comparison results as Table II are also obtained for female employees.

C. Analysis Results

TABLE III
BRIEF FIT STATISTICS FOR LONG-TERM SICKNESS ABSENCES; MALE SAMPLE NUMBER = 12011; FEMALE SAMPLE NUMBER = 38245

Variables	Uncorrelated model Parameter value (SE) Male Model 2	Correlated model Parameter value (SE) Male Model 3
	Logistic Model	Logistic Model
Intercept	–	–
$YEAR^2$	$–^{***}$	$–^{***}$
	Log-Normal	Log-Normal
YEAR	$+^{***}$	$+^{***}$
$YEAR^2$	–	–
Model 1 Linear model	YEAR: $+^{*}$	
Variables	Uncorrelated model Parameter value (SE) Female Model 2	Correlated model Parameter value (SE) Female Model 3
	Logistic Model	Logistic Model
Intercept	$–^{***}$	$–^{***}$
$YEAR^2$	–	–
	Log-Normal	Log-Normal
YEAR	+	+
$YEAR^2$	$+^{***}$	$+^{***}$
Model 1 Linear model	$YEAR^2$ $+^{*}$	

$^{***}p<0.001$, $^{**}p<0.05$, $^{*}p<0.1$.

Table III illustrates some fit statistics using our three candidate models. For simplicity, only inconsistent results with respect to male and female employees are displayed with only signs. The consistent results are presented in the following paragraph. A positive sign (+) indicates a positive association and a negative sign (−) indicates a negative association of the predictors.

Table III illustrates in general that the incidence and duration of long-term sickness absences for male and female follow different temporal trends linearly or nonlinearly. It also demonstrates that inexact predictions are obtained in Model 1: no significant correlation of YEAR with long-term sickness duration is claimed from Model 1 for male staff for example, however Model 2 and Model 3 conclude differently. This shows that statistical inference can be inefficient and misguided if inappropriate model is adopted.

Finally, let's briefly summarize the consistent results we get for both male and female municipal employees. Considering the associations of the covariates with the probability of long-term sickness absences, younger staff had significantly lower absence probabilities. Higher-income subjects tended to have less probability of long-term sickness absences. Employees' socio-economic class had significantly effect on the long-term absence incidences. Manual worker had the highest absence incidence. The descending order of long-term absence incidence according to the category of socio-economic class is manual worker, routine non-manuals, semi-professionals, managers and professionals, which indicates that employees performing technical-manual work had higher incidence of long-term work absence than those performing mental work. There is an association between employment contract types and the incidence of long-term sickness absences. Employees who had a temporary working contract had a significantly lower incidence of work absence.

Turning to the associations of the covariates with the duration of long-term sickness absences (i.e. employees who did have long-term work absences or long-term absence rates were positive). All covariates are significant predictors. The accumulative absence days increased according to the following category of employees' socio-economic class: managers and professionals, semi-professionals, routine non-manuals, and manual workers.

V. CONCLUSION

Absenteeism is a major concern in our society. Even though extensive research has focused on relevant risk-based investigations, knowledge of the causes and consequences of sick leave is still limited. In addition, sickness absence is a dynamic temporal behavior and many influential factors vary over time. It is obvious that analysis of cross-sectional data, which is the most common technique in sick leave studies, is not enough for understanding the dynamic performances of sickness absences. There is a lack of studies on the dynamic process of sickness absence behaviors in literature. The hampering factors exist in both data surrounding and analysis methodologies. In this paper, we have focused on addressing these difficult issues.

Firstly, we have identified common data characteristics of sickness absence such as time dependent, highly skewed and clumped at zero, which challenges the traditional models. Ignoring these features to make statistical inference is likely to be inefficient and misguided. Take the example of long-term sickness absence rates referring to Table III, the time variable YEAR is predicted differently and errorly by general linear model.

Secondly, we have discussed commonly employed approaches used in sickness absence research to empirically address the methodology issues and proposed a way for selecting proper models. We have introduced the two-part mixed-distribution model for analysing longitudinal sickness absence data. This is one of the main purposes of this paper. An application of the model has been presented by using a large registration dataset from the Helsinki Health Study for municipal employees during the period of 1990-1999. Calculation results have demonstrated that the proposed model perform superior to other commonly adopted models in the literature.

Finally, to summarize and conclude the analysis results, the basic conclusion is that there is strong relationship between socio-economic and occupational background and long-term sickness absences. The revealed findings, through an application of the proposed two-part mixed-distribution model, are consistent with the literature.

REFERENCES

[1] D. Gimeno, F.G. benavides, J. Benach and B.C. Mick III, "Distribution of sickness ansence on the European Union countries". Occupational and Environmental Medicine, vol 61, pp867-869, 2004.

[2] S. Taimela, E. Läärä, A. Malmivaara, J. Tiekso, H. Sintonen, S. Justén and T. Aro, "Self-reported health problems and sickness absence in different age groups predominantly engaged in physical work", Occupational and Environmental Medicine, vol 64, no 11, pp739-746, nov. 2007.

[3] K. Kondo, Y. Kobayashi, K. Hirokawa, A. Tsutsumi, F. Kobayashi, T. Haratani, S. Araki and N. Kawakami, "Job strain and sick leave among Japanese employees: a longitudinal study", Int Arch Occup Environ Health., vol. 79, no 3, pp213-219, mar. 2006.

[4] M.L. Nielsen, R. Rugulies, L. Smith-Hansen, K.B. Christensen and T.S. Kristensen, "Psychosocial work environment and registered absence from work: estimating the etiologic fraction", Am J Ind Med., vol 49, no 3, pp187-196, mar. 2006a.

[5] M.L. Nielsen, R. Rugulies, K.B. Christensen, L. Smith-Hansen and T.S. Kristensen, "Psychosocial work environment predictors of short and long spells of registered sickness absence during a 2-year follow up", J Occup Environ Med., vol 48, no 6, 591-598,jun. 2006b.

[6] J. Head, M. Kivimaki, P. Martikainen, J. Vahtera, J.E. Ferrie and M.G. Marmot, "Influence of change in psychosocial work characteristics on sickness absence: The Whitehall II Study", J Epidemiol Community Health, vol 60, no 1, pp55-61, jan. 2006.

[7] A. Notenbomer, C.A. Roelen and J.W. Groothoff, "Job satisfaction and short-term sickness absence among Dutch workers", Occup Med (Lond), vol. 56, no 4, pp279-81, jun. 2006.

[8] C.A. Schroer, M. Janssen, L.G. van Amelsvoort, H. Bosma, G.M. Swaen, F.J. Nijhuis and J. van Eijk, "Organizational characteristics as predictors of work disability: a prospective study among sick employees of for-profit and not-for-profit organizations", J Occup Rehabil, vol 15, no 3, pp435-445, sep. 2005.

[9] L-G. Engström and T. Eriksen, "Can differences in benefit levels explain duration and outcome of sickness absence?", Disability and rehabilitation, vol. 24, no 14, pp713-718, sep. 2002.

[10] M. Goldberg, J. F. Chastang, A. Leclerc, M. Zins, S. Bonenfant, I. Bugel, N. Kaniewski, A. Schmaus, I. Niedhammer, M. Piciotti, A. Chevalier, C. Godard, and E. Imbernon, "Socioeconomic, Demographic, Occupational, and Health Factors Associated with Participation in a Long-term Epidemiologic Survey: A Prospective Study of the French GAZEL Cohort and Its Target Population", American Journal of Epidemiology, vol 154, no. 4, pp373-384, 2002.

[11] P. Kristensen, T. Bjerkedal and L.M. Irgens, "Early life determinants of musculoskeletal sickness absence in a cohort of Norwegians born in 1967–1976", Social Science & Medicine, vol 64 no 3, pp646-655, feb. 2007.

[12] J. Sundquist, A. Al-Windi, S-E. Johansson and K. Sundquist, "Sickness absence poses a threat to the Swedish Welfare State: a cross-sectional study of sickness absence and self-reported illness", BMC public health, vol 7, pp45, jan. 2007

[13] M. Hansson, C. Boström and K. Harms-Ringdahl, "Sickness absence and sickness attendance—What people with neck or back pain think?", Social Science & Medicine, vol. 62, no 9, pp2183-2195, 2006.

[14] A. Burdorf and J.P. Jansen, "Predicting the long term course of low back pain and its consequences for sickness absence and associated work disability", Occup Environ Med., vol. 63, no 8, pp522-529. aug. 2006.

[15] A. Torres Lana, A. Cabrera de León, M.T. Marco García and A. Aguirre Jaime, "Smoking and sickness absence among public health workers", Public Health, vol 19, no 2, pp144-149, 2005.

[16] W. Jzelenberg and A. Burdorf, "Risk factors for musculoskeletal symptoms and ensuing health care use and sick leave", Spine, vol 30, no 13, pp1550-1556, 2005.

[17] E. L Horneij, I. B Jensen, E. B Holmström and C. Ekdahl, "Sick leave among home-care personnel: a longitudinal study of risk factors", BMC Musculoskelet Disord, vol. 5, no 38.. 8. doi: 10.1186/1471-2474-5-38, 2004.

[18] W.N. Burton, C.Y. Chen, A.B. Schultz and D.W. Edington, "The economic costs associated with body mass index in a workplace", Journal of Occupational and Environmental Medicine, vol 40, no 9, pp786-792, 1998.

[19] P. Lundborg, "Does smoking increase sick leave? Evidence using register data on Swedish workers", Tobacco Control, vol 16, pp114–118, 2007.

[20] J. Vahtera, K. Poikolainen, M. Kivimäki, L. Ala-Mursula and J. Pentti, "Alcohol Intake and Sickness Absence: A Curvilinear Relation", Am. J. Epidemiol., vol 156, pp969 – 976, nov. 2002.

[21] M. Labriola, T. Lund and H. Burr, "Prospective study of physical and psychosocial risk factors for sickness absence", Occup Med (Lond), vol 56, no 7, pp469-474, oct. 2006.

[22] P. Allebeck and A. Mastekaasa, "Risk factors for sick leave -general studies", Scand J Public Health, vol 32, Suppl 63, pp49-108, 2004.

[23] K. Piha, P. Martikainen, O. Rahkonen, E. Roos and E. Lahelma, "Trends in socioeconomic differences in sickness absence among Finnish municipal employees 1990-99", Scandinavian Journal of Public Health, vol. 35, pp348-355, 2007.

[24] L.T. Yen, D.W. Edington and P. Witting, "Prediction of prospective medical claims and absenteeism costs for 1284 hourly workers from a manufacturing company", J Occup Med, vol 34, no 4, pp428-435, 1992.

[25] P.G. Smulders and F.J.N. Nijhuis, "The job demands--job control model and absence behaviour: results of a 3-year longitudinal study", Work & Stress, vol 13, pp115-131, 1999.

[26] M. Kivimaki, J. Vahtera, L. Thomson, A. Griffith, T. Cox, J. Pentti, "Psychosocial factors predicting employee sickness absence during economic decline", J Appl Psychol., vol 82, pp858-872, 1997.

[27] V. Kujala, T. Tammelin, J. Remes, E.E. Vammavaara and J. Laitinen, "Work ability index of young employees and their sickness absence during the following year", Scand J Work Environ Health., vol 32, vol 1, 75-84, feb. 2006.

[28] N. Duan, W.G. Manning, C.N. Morris and J.P. Newhouse, "A comparison of alternative models for demand for medical care", Journal of Economic and Business Statistics, vol 1, pp115-126, 1983.

[29] W. Manning, N. Duan and W. Rogers, "Monte Carlo evidence on the choice between sample selection and two-part models", Journal of Econometrics, vol 35, pp59-82, 1987.

[30] G.K. Grunwald and R.H. Jones, "Markov models for time series with mixed distribution", Environmetrics, vol 11, pp327-329, 2000.

[31] P. Lachenbruch, "Utility if regression analysis in epidemiologic studies of the elderly", in The epidemiologic study of the elderly, R. Wallace and E. Woolson eds. Oxford University Press, 1992, pp. 371-381.

[32] J.A. Tooze, G.K. Grunwald and R.H. Jones, "Analysis of repeated measures data with clumping at zero", Statistical Methods in Medical Research, vol 11, pp341-355, 2002.

[33] H. Akaike, "Information theory as an extension of the maximum likelihood principle", in B.N. Petrov and F. Csaksi, ed. 2nd International Symposium on Information Theory. Akademiai Kiado, Budapest, Hungary.1973, pp267-281.

X. Lu was born in Changchun of China. She graduated from Jilin University in China with BSc and MSc degrees in mathematics and PhD from Helsinki University of Technology in Finland in civil and environmental engineering.

She is currently a Specialized Researcher with Finnish Institute of Occupational Health in Finland. She has published numerous articles in various international journals, most of them in the fields of mathematics, physics and engineering. She has recently received the Napier Shaw Bronze Medal for the best scientific paper, awarded by the Chartered Institution of Building Services Engineers (CIBSE) in UK. She is also a member of the Editorial Boards of Journal of Intelligent Buildings International (publisher: Earthscan in London).