STATISTICS IN MEDICINE
Statist. Med. **18**, 3355–3363 (1999)

RECURRENT INJURY EVENT-TIME ANALYSIS†

JAMES T. WASSELL,[1]* WILLIAM C. WOJCIECHOWSKI[1‡] AND DEBORAH D. LANDEN[2]

[1] *Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health, Division of Safety Research, 1095 Willowdale Road, Morgantown, WV 26505-2888, U.S.A.*
[2] *Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health, Pittsburgh Research Center, P.O. Box 18070, Pittsburgh, PA, 15236-0070, U.S.A.*

SUMMARY

Public health decision making based on data sources that are characterized by a lack of independence and other complicating factors requires the development of innovative statistical techniques. Studies of injuries in occupational cohorts require methods to account for recurrent injuries to workers over time and the temporary removal of workers from the 'risk set' while recuperating. In this study, the times until injury events are modelled in an occupational cohort of employees in a large power utility company where employees are susceptible to recurrent events. The injury history over a ten-year period is used to compare the hazards of specific jobs, adjusted for age when first hired, and race/ethnicity differences. Subject-specific random effects and multiple event-times are accommodated through the application of frailty models which characterize the dependence of recurrent events over time. The counting process formulation of the proportional hazards regression model is used to estimate the effects of covariates for subjects with discontinuous intervals of risk. In this application, subjects are not at risk of injury during recovery periods or other illness, changes in jobs, or other reasons. Previous applications of proportional hazards regression in frailty models have not needed to account for the changing composition of the risk set which is required to adequately model occupational injury data. Published in 1999 by John Wiley & Sons, Ltd. This article is a US Government work and is in the public domain in the United States.

INTRODUCTION

Data from occupational cohort studies may suffer from the same complications as other types of cohort or longitudinal studies. These include staggered accrual of subjects, loss to follow-up of subjects, changes in the composition of the risk-set due to changes in job status or recovery time from injury, changing values of important covariates over time and recurrent events. However, occupational injury data pose an additional problem requiring non-standard methods because individuals may have unobserved risk factors for injury. As a result, statistical models that account for unobserved effects provide substantial improvement compared to models that ignore individual random effects. The distribution of the unobservable heterogeneity is identifiable when there is a common subject-specific random effect for all the event-times recorded for one subject. Frailty models[1–7] can account for a complicated data structure and incorporate random effects to

* Correspondence to: James T. Wassell, Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health, Division of Safety Research, 1095 Willowdale Road, Morgantown, WV 26505–2888, U.S.A. E-mail: jtw2@cdc.gov

† This article is a US Government work and is in the public domain in the United States
‡ Currently at Rice University, Department of Statistics, Houston, TX 77251, U.S.A.

Published in 1999 by John Wiley & Sons, Ltd.

compare the injury hazards among different jobs using a method that is consistent with the natural history of occupational injuries.

Frailty models have been shown to be useful for analysis of correlated survival or event-time data. These methods propose that dependency between injury event-times for one individual is the result of an unobserved random variate that is common to all the injury event-times for one individual. This random variate has been termed 'frailty,' and acts as a multiplier which enhances or degrades the hazards for all times common to one individual. This approach has the advantage of modelling a possible 'mechanism' for dependent data as compared to other approaches[8–10] which adjust variance estimates to compensate for the effects of data dependency. Initially, frailty analyses were restricted to models that assume a parametric form for the 'baseline' (independent data models) survival time distributions. Parameters for these models may be estimated using traditional maximum likelihood methods, although the computing methods are complicated and rely on specialized computer programs.[11,12] Recent work has illustrated the use of the expectation maximization (EM) algorithm[13] and profile likelihood methods[14] to estimate parameters for frailty models not requiring parametric assumptions of the survival time distributions. In considering a counting process approach to frailty models, a recent paper by Nielsen *et al.*[15] describes frailty as 'accident proneness.' These models rely on iterative computing methods to obtain covariate estimates from Cox regression models assuming an arbitrary baseline hazard function in the presence of a frailty effect.

The distributional form of the random effects or frailties determines the structure and characteristics of the dependency among event-times. Several different distributions for the frailties have appeared in the literature;[16] the first proposed frailty model assumed that the frailties were distributed as gamma random variates with an expected value equal to one and a variance parameter that reflects the degree of dependency in the data.[17] Because the gamma and other frailty distributions suffer from a lack of the proportional hazards property in the marginal distribution, Hougaard[18] proposed the use of the positive stable frailty distribution. This analysis of recurrent injury data is based on the positive stable frailty model to take advantage of the proportional hazards property which allows interpretation of the effects of covariates in a straightforward manner.

For subjects who have recurrent events, the counting process formulation[1,19] of the proportional hazards model provides some advantages over the standard Cox regression analysis. For each subject an interval of risk is defined from the start of an at-risk period, to the time of an injury event or censored time. This approach has the advantage that the risk set for any event only includes other individuals whose interval of risk includes the event-time of interest. The method described here is intended to model events in calendar time through the use of the counting process formulation and also includes unobserved random variates (subject-specific frailty) to analyze multiple event data.

Cohort description

Ten years of injury records of employees of an electrical utility corporation who worked from 1980 to 1989 were analyzed in this study. The cohort consisted of 608 employees aged 18 to 24 years on 1 January 1980. Employees' entry into the study was staggered, although 84 per cent of the cohort were actively employed on 1 January 1980. Job titles included in the cohort were lineman, cable splicer, and others (electrician, troubleman, foreman and serviceman). Injuries were included in the study that occurred on the job and required the attention of a physician.

More than 400 employees experienced at least one injury and more than 250 employees experienced two or more injuries in this cohort. Following injury, an employee might not return to work for a period of recuperation or might return to work at a temporary office assignment until fully ready to resume former job duties. Employee records also indicated withdrawal from the workforce due to other reasons (for example, job change). The counting process model formulation was used to effectively account for these 'not at risk' periods.

Covariates investigated in this study included indicator variables for race: Black, Hispanic and White (referent group), and age at start of the study as a continuous covariate. The focus of the analysis is to compare three different job categories (linemen, cable splicers and others) regarding injury hazard. Additional details regarding this study can be found in an analysis of this data based on a nested case-control methodology[20] and an analysis based on the Weibull distribution for event-times with covariates and gamma frailty.[12]

METHODS

The following description of frailty models is adapted from Klein[13] and Wang *et al.*[14] Frailty models are based on the assumption that, if the value of the frailty were known, all the event-times for injuries occurring to an individual would be independent. Event-times, T_{ij}, are indexed by $i = 1, \ldots, B$, to indicate the ith subject with one or more recorded event-times denoted by the index $j = 1, 2, \ldots, n_i$. For each event time, a censoring indicator is recorded, I_{ij}, with $I_{ij} = 1$ if T_{ij} is an injury time and $I_{ij} = 0$ otherwise. Additional covariates may be included that are subject characteristics or specific to a subject at the time of an injury, $\mathbf{Z}_{ij} = (Z_{ij1}, \ldots, Z_{ijk})$ with k recorded covariates at T_{ij}. The unobserved frailty random variable, which varies among subjects but is common to all event-times for a subject, is denoted as W_i. Assuming a proportional hazards model, the independence assumption is demonstrated by defining the hazard rate as $\lambda(t_{ij}|\mathbf{Z_{ij}}, W_i) = W_i\lambda_0(t_{ij})\exp(\beta\mathbf{Z}_{ij})$ with $\lambda_0(t_{ij})$ as the arbitrary baseline hazard function and β is a vector of coefficients to be estimated. The joint survival function for all event-times recorded for the ith subject, given the frailty, is

$$P[T_{ij} > t_{ij}, j = 1, \ldots, n_i | W_i, \mathbf{Z}_{ij}, j = 1, \ldots, n_i] = \exp\left\{ -\left[\sum_{j=1}^{n_i} \Lambda_0(t_{ij})\exp(\beta Z_{ij})\right] W_i \right\}.$$

The discontinuous intervals of risk for subjects are accommodated by estimating the hazard function

$$\lambda_0(t_{ij}) = \frac{d_{ij}}{\sum_{t_{ij}} Y_i(t_{ij}) W_i \exp(\beta \boldsymbol{Z}_{ij})}$$

using an indicator function, $Y_i(t_{ij})$, to specify whether or not subject i is 'at risk' at time t_{ij}, where d_{ij} is the number of events that occur at time t_{ij}. The cumulative hazard, $\Lambda_0(t_{ij}) = \sum_{t \leqslant t_{ij}} \lambda_0(t_{ij})$ is a sum of the hazard values associated with all times less than or equal to t_{ij}.

Positive Stable Frailty Model Estimation

If a random variable has a positive stable distribution, $W \sim \boldsymbol{PS}(\rho)$, $W \geqslant 1$, indexed by the parameter $\rho \in (0, 1]$, then we can take advantage of the Laplace transform: $\boldsymbol{E}[\exp(-sW)] = \exp(-s^{\rho})$. When applied to the joint survival function above, substituting

$s = -[\sum_{j=1}^{n_i} \Lambda_0(t_{ij}) \exp(\beta Z_{ij})]$ yields a joint unconditional survival probability suitable for constructing a log-likelihood function (shown in the Appendix) that can be used in parameter estimation. Note that independence of the event-times occurs if $\rho = 1$ and smaller values of ρ result in greater group dependency. Under independence, (that is, $\rho = 1$, so that all $w_i = 1$) the log-likelihood equation reduces to

$$\mathrm{LL}_{\text{independence}} = \sum_{j=1}^{n_i} I_{ij}[\beta Z_{ij} + \log[\lambda_0(t_{ij})]] - H_i$$

which provides the same estimates of regression coefficients as obtained from a Cox regression ignoring any grouping of the data.

To estimate parameters for a semi-parametric model with frailty an iterative method is required. Klein[13] proposed the EM algorithm approach to estimation for the gamma frailty proportional hazards regression analysis of family-grouped survival times. These analyses did not require the counting process approach for estimating covariate effects because family members (unlike members of an occupational cohort) do not leave the risk set for short periods of time.

In this study, we utilized a profile likelihood approach based on an EM algorithm for parameter estimation.[14] The profile likelihood requires determining the expected value of the frailties, given the current estimates of the regression parameters, the frailty parameter, and the baseline hazard rate. The profile likelihood estimation requires 'imputing' values for the individual frailties and using those estimates as an 'offset' term in a Cox regression estimation program. The estimation algorithm proceeds as follows:

1. Estimate the regression parameters under independence using the counting process formulation of the proportional hazards model.
2. Expectation step: calculate the expected value of the frailties given the regression estimates and the non-parametric estimate of the hazard (this hazard estimate depends on the current estimates of regression parameters).
3. Maximization step: use the group-specific estimates of the frailties as an 'offset' term in the proportional hazards model to update estimates of the regression parameters.
4. Iterate between steps 2 and 3 until parameter estimates converge, each time obtaining 'better' imputed values for the frailties.
5. Calculate the value of the log-likelihood that corresponds to the set of parameter estimates.
6. Find the set of parameter estimates that corresponds to the maximum of log-likelihood either graphically or with the use of a maximization routine.

Custom functions were written using S-plus statistical software[21] to perform the estimation for the positive stable frailty model. An optimization function (nlminb) was used to select the set of parameter estimates to maximize the profile likelihood. These functions (named 'frailty.stable' are available through StatLib (http://lib.stat.cmu.edu/S/).

Estimates of the standard error of the regression coefficients were obtained using the group jack-knife technique.[22,23] The group jack-knife estimates are obtained by multiple reanalysis of the data, each time omitting the group of injury records for one employee. The standard error estimates are equivalent to 'robust' estimates described elsewhere.[8] In addition, the group jack-knife estimates were compared to grouped bootstrap estimates based on a resampling of employee records, defining a group as all injury records for an employee. Hazard ratio estimates are not a linear combination of the coefficient estimates, therefore 95 per cent confidence intervals

for the hazard ratio estimates were obtained using the bias corrected accelerated percentile of the bootstrap replicates, rather than the estimated standard errors of the coefficients. A confidence interval for the frailty parameter was also obtained from the profile likelihood for comparison.

RESULTS

Parameter estimates and standard errors for the positive stable and for the independence model are shown in Table I. The estimated frailty parameter is 0·443 for the positive stable model. A comparison of the values of the log-likelihoods at its maximum for the frailty parameter (− 3445·97 at rho = 0·443) and for the null model of no correlation (− 5558·90 at rho = 1) indicates that the model which accounts for the dependency in the data is significantly better. The profile log-likelihood 95 per cent confidence interval for the frailty parameter of the positive stable model is (0·420, 0·465) as indicated in Figure 1. The profile log-likelihood confidence interval is 23 per cent shorter than a 95 per cent confidence interval based on the group jack-knife standard error estimate of 0·015 (0·414, 0·472) and 26 per cent shorter than the BCa bootstrap confidence interval using percentiles of 250 bootstrap samples (0·416, 0·477).

With significant evidence of dependency in the data as indicated by the likelihood ratio tests of the frailty parameter, the interpretation of the covariate effects in the marginal distribution is affected. For example (under independence), a comparison of the hazard for the job title 'splicer' and the referent group (electrician, others) is 1·448 [exp(0·370) = 1·448] with a 95 per cent BCa bootstrap confidence interval of (1·092, 1·820), suggesting this job has a significantly higher hazard of injury. Given subject-specific random effects, the hazard estimate based on the positive stable frailty model is attenuated to 1·050 [exp(0·443 × 0·110) = 1·050] with a 95 per cent BCa bootstrap confidence interval of (0·950, 1·150), indicating that there is really no significantly higher hazard of injury for splicers. Table I illustrates other differences between the stable frailty model and hazard ratio estimates that ignore dependency; for example, the frailty model indicates that Hispanic workers have significantly higher injury hazard that is undetected by ignoring random effects.

DISCUSSION

Occupational cohort data on injury incidence and recurrence require analytical methods that account for the unique features of this data. In addition to some characteristics common to other longitudinal cohort studies, injury cohort data demonstrate the usefulness of random effect event-time models. The random effects account for the notions of unequal liability that have been long recognized by injury data investigators.[24] Frailty models account for an individual random effect that degrades or enhances the hazard function for injury for an individual. In contrast to modelling approaches that account for recurrent events through robust variance estimates, frailty models propose a mechanism for dependency in the data. In a cohort study, it is not possible to identify all risk factors for occupational injury, but additional unobserved factors influencing individual risk for injury may be accounted for by the use of frailty models.

An iterative method for estimation of parameters, using principles of the EM algorithm and based on a profile likelihood approach, has been developed. The algorithm iterates between expected values of the frailties, given regression estimates, and expected values of regression parameters, given frailty estimates, for a fixed value of the frailty distribution parameter. In addition to the counting process approach to analysis of multiple-event-time data, frailty models

Table I. Parameter estimates, standard errors, hazard ratios and 95 per cent confidence intervals obtained from the positive stable frailty model for correlated event-times and the Cox proportional hazards regression model for independent event-times

Effect	Positive stable frailty model				Cox regression model			
	Estimate	Std.Err.*	HR† = exp($\rho\beta$)	BCa 95% CI‡	Estimate	Std.Err.§	HR† = exp(β)	BCa 95% CI‡
Frailty (ρ)	0·443	0·015	—	(0·416, 0·477)	1·000	—	—	—
Black	0·076	0·092	1·034	(0·957, 1·119)	0·018	0·095	1·018	(0·845, 1·230)
Hispanic	0·288	0·135	1·136	(1·033, 1·265)	0·185	0·108	1·203	(0·963, 1·494)
Lineman	0·287	0·099	1·135	(1·031, 1·239)	0·408	0·112	1·503	(1·196, 1·804)
Splicer	0·110	0·110	1·050	(0·950, 1·150)	0·370	0·128	1·448	(1·092, 1·820)
Age	− 0·051	0·017	0·978	(0·964, 0·991)	− 0·039	0·017	0·961	(0·923, 0·994)

* Standard error estimates obtained from a group jack-knife procedure
† Hazard ratio point estimates
‡ Bias corrected accelerated percentile confidence intervals based on 250 bootstrap replicates
§ Standard error estimated from the information matrix

Published in 1999 by John Wiley & Sons, Ltd.

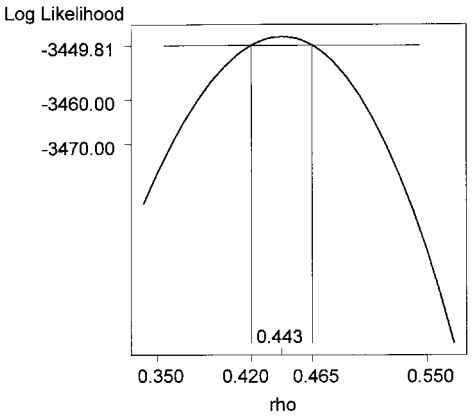

Figure 1. Log-likelihood profile confidence interval for the positive stable frailty distribution parameter (rho) based on a likelihood ratio test with 1 degree of freedom.. The log-likelihood function is maximum (-3445.97) at rho $= 0.443$. See the discussion for a comparison of this confidence interval with confidence intervals based on the group jack-knife variance estimate and bootstrap BCa confidence intervals

demonstrate additional improvement in estimating covariate effects. The additional complications of possibly censored times, variable 'at risk' periods and staggered accrual of subjects are accounted for in this analysis through the use of the counting process formulation of the proportional hazards model.

Additional research is needed regarding the appropriate choice of frailty distribution for a particular data set. Ideally, methods should be developed that permit the data to dictate the most suitable frailty distribution. This would provide valuable insight into the process generating data dependency in a given situation, similar to work which has focused on the gamma frailty models.[25] Further studies are needed to assess the sensitivity of estimation methods to the degree of correlation that may be present in realistic data. With significant frailty, the interpretation of the estimated coefficients depends on the choice of frailty distribution and the estimate of the frailty parameter so that the proportional hazards interpretation is possible only with the positive stable frailty model. Additional research is needed to develop methods to assess the validity of the proportional hazards assumption for the marginal distributions of multivariate failure time data. Although there is considerable computing effort required in fitting these models, these methods are becoming more widely available for data analysis.

APPENDIX

The log-likelihood for event-time data with positive stable frailty can be written as

$$\mathrm{LL} = \sum_{i=1}^{B} [D_i[\log(\rho) + (\rho - 1)\log(H_i)] - H_i^{\rho} + \log[\mathrm{J}(D_i, H_i)] + \sum_{j=1}^{n_i} I_{ij}[\boldsymbol{\beta Z_{ij}} + \log[\lambda_0(t_{ij})]]$$

where

$$J(D_i, H_i) = \sum_{m=0}^{D_i - 1} C_{D_i, m} H_i^{-m\rho}$$

based on a recursive function

$$C_{k,0} = 1$$

$$C_{k,m} = C_{k-1,m} + C_{k-1,m-1}[(k-1)\phi - (k-m)] \quad (m = 1, \ldots, k-2)$$

$$C_{k,k-1} = (\phi - 1)(2\phi - 1) \ldots [(k-1)\phi - 1] = \phi^{k-1}\Gamma(k-\rho)/\Gamma(1-\rho)$$

$$\phi = 1/\rho.$$

Also, $\lambda_0(t_{ij})$ is the hazard function at time t_{ij} described in the text. $H_i = \sum_{j=1}^{n_i} \Lambda_0(t_{ij}) \exp(\boldsymbol{\beta Z_{ij}})$, for the cumulative hazard function $\Lambda_0(t_{ij})$ described in the text, and $D_i = \sum_j I_{ij}$ is the number of events that occurred to the ith subject, where there are a total of B subjects with n_i event-times for each subject. The expected value of the frailty, given current estimates of other parameters and the data is

$$E[W_i|\boldsymbol{\beta}, \lambda_0(t_{ij})] = \frac{E[W_i^{D_i+1}\exp(-H_iW_i)]}{E[W_i^{D_i}\exp(-H_iW_i)]}$$

$$E[W^q \exp(-H_iW)] = \rho H_i^{\rho-1} \exp(-H_i^{\rho})J(q, H_i), \quad q = 0, 1, \ldots; \quad H_i > 0$$

using the J function described above.

REFERENCES

1. Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. *Statistical Models Based on Counting Processes*, Springer-Verlag, New York, 1993.
2. Costigan, T. M. and Klein, J. P. 'Multivariate survival analysis based on frailty models', *in* Basu, A. P. (ed.), *Advances in Reliability*, Elsevier Science Publishers B.V., Oxford, 1993, pp. 43–58.
3. Klein, J. P. and Moeschberger, M. L. *Survival Analysis*: *Techniques for Censored and Truncated Data*, Springer-Verlag, New York, 1997.
4. Oakes, D. 'Frailty models for multiple event times' *in* Klein, J. P. and Goel, P. K. (eds), *Survival Analysis*: *State of the Art*, Kluwer Academic Publishers, Dordrecht, 1992, pp. 371–379.
5. Pickles, A. and Crouchley, R. 'Generalizations and applications of frailty models for survival and event data', *Statistical Methods in Medical Research*, **3**, 263–278 (1994).
6. Wassell, J. T., Kulczycki, G. W. and Moyer, E. S. 'Modeling frailty in manufacturing processes' *in* Jewell, N. P., Kimber, A. C., Lee, M-L. T. and Whitmore, G. A. (eds), *Lifetime Data*: *Models in Reliability and Survival Analysis*, Kluwer Academic Publishers, Dordrecht, 1996, pp. 353–361.
7. Wassell, J. T., Kulczycki, G. W. and Moyer, E. S. 'Frailty models of manufacturing effects', *Lifetime Data Analysis*, **1**, 161–170 (1995).
8. Therneau, T. M. and Hamilton, S. A. 'rhDNase as an example of recurent event analysis', *Statistics in Medicine*, **16**, 2029–2047 (1997).
9. Lin, D. Y. 'Cox regression analysis of multivariate failure time data, the marginal approach', *Statistics in Medicine*, **13**, 2233–2247 (1994).
10. Li, Q. H. and Lagakos, S. W. 'Use of the Wei-Lin-Weissfeld method for the analysis of a recurring and a terminating event', *Statistics in Medicine*, **16**, 925–940 (1997).
11. Wassell, J. T. and Moeschberger, M. L. 'A bivariate survival model with modified gamma frailty for assessing the impact of interventions', *Statistics in Medicine*, **12**, 241–248 (1993).
12. Wassell, J. T. and Kulczycki, G. W. 'Frailty analysis of repeated injuries', *Proceedings of the Epidemiology Section of the American Statistical Association*, 130–134 (1995).
13. Klein, J. P. 'Semiparametric estimation of random effects using the cox model based on the EM algorithm', *Biometrics*, **48**, 795–806 (1992).
14. Wang, S. T., Klein, J. P. and Moeschberger, M. L. 'Semi-parametric estimation of covariate effects using the positive stable frailty model', *Applied Stochastic Models and Data Analysis*, **11**, 121–133 (1995).

15. Nielsen, G. G., Gill, R. D., Andersen, P. K. and Sørensen, T. I. A. 'A counting process approach to maximum likelihood estimation in frailty models', *Scandinavian Journal of Statistics*, **19**, 25–43 (1992).
16. Pickles, A. and Crouchley, R. 'A comparison of frailty models for multivariate survival data', *Statistics in Medicine*, **14**, 1447–1461 (1995).
17. Clayton, D. G. 'A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic heart disease', *Biometrika*, **65**, 141–151 (1978).
18. Hougaard, P. 'A class of multivariate failure time distributions', *Biometrika*, **73**, 671–678 (1986).
19. Fleming, T. R. and Harrington, D. P. *Counting Processes and Survival Analysis*, Wiley, New York, 1991.
20. Landen, D. D. 'A case-control study of electrical injury among a cohort of line mechanics', *American Journal of Industrial Medicine*, 1999, in press.
21. *S-PLUS* 4 *Guide to Statistics*. Data Analysis Products Division, Mathsoft, Seattle, WA, 1997.
22. Efron, B. and Tibshirani, R. J. *An Introduction to the Bootstrap*, Chapman and Hall, London, 1993.
23. Shao, J. and Tu, D. *The Jackknife and Bootstrap*, Springer-Verlag, New York, 1995.
24. Mackenzie, G. 'A proportional hazard model for accident data', *Journal of the Royal Statistical Society*, *Series A*, **149**, 366–375 (1986).
25. Shih, J. H. and Louis, T. A. 'Assessing gamma frailty models for clustered failure time data' *in* Jewell, N. P., Kimber, A. C., Lee, M-L. T. and Whitmore, G. A. (eds), *Lifetime Data*: *Models in Reliability and Survival Analysis*, Kluwer Academic Publishers, Dordrecht, 1996, pp. 371–379.