

Chapter 33

Using Latent Class Models to Analyze Response Patterns in Epidemiologic Mail Surveys

Thomas Kohlmann and Anton K. Formann

Institute for Social Medicine, Medical University of Lübeck, Germany
Department of Psychology, University of Vienna, Austria

1. Introduction

Latent class analysis (LCA) does not belong to the „standard“ statistical methods in medical and epidemiologic research. During the past decade, however, LCA has been repeatedly applied to various research problems in medicine. Applications of the latent class model range from development and evaluation of diagnostic criteria in psychiatry and other medical disciplines to assessment of rater agreement, errors in categorical measurement and misclassification (for a review, see Formann & Kohlmann, 1996). Most of these studies were carried out in clinical medicine or clinical epidemiology. In contrast, there are only a few applications of LCA in population-based epidemiologic studies, although the model provides a framework within which many problems of categorical data analysis could be addressed in this field of research.

The purpose of this paper is to assess the utility of latent class models (cf. Goodman, 1974; Haberman, 1979; Formann, 1985; Clogg, 1988) in the analysis of epidemiologic data on self-reported musculo-skeletal symptoms. Data came from a large population-based study which utilized postal questionnaires. In particular, we examine if and how LCA can be applied to identify distinct patterns of symptom reporting (section 3), and we describe the potential contribution of LCA results in the selection of subjects for further diagnostic studies (section 4).

Our analysis is motivated by a major obstacle which researchers of the epidemiology of rheumatic diseases often face: The presence of musculo-skeletal complaints such as pain, swelling, or stiffness in joints and combinations thereof cannot, in general, be considered clear-cut indicators of specific rheumatic diseases. Many people experience rheumatic symptoms without being affected by manifest musculo-skeletal disorders. On the other hand, manifest musculo-skeletal disorders are often associated with multiple complaints none of which represents an unequivocal symptom of that particular disease. When self-reports of such complaints in a questionnaire, for example, are used for purposes of screening („case finding“), it is therefore difficult for the researcher to decide which pattern of symptoms he or she should accept as a positive „test result“.

Yet, if symptoms are used as criteria for inclusion in a further diagnostic evaluation, appropriate selection of these criteria is of crucial importance in order to minimize the false-positive rate (i.e., the proportion of persons included in the medical examination without having the disease of interest), and hence to reduce cost and time (Morrison, 1985).

2. Materials and Methods

The Hannover Rheumatoid Arthritis (RA) Study (Wasmus et al., 1989) was designed to determine the prevalence of RA (a chronic and often disabling inflammatory disease which primarily affects the connective tissue surrounding joints) in the adult population. In a first step, a random sample of the 25 to 74 year old German residents of Hannover (Germany) was screened for musculo-skeletal symptoms by means of a mailed questionnaire. Among others, this questionnaire contained five questions about the presence of five symptoms „today“ (back pain, neck pain, pain in one or several joints, joint swelling, morning stiffness) and about the occurrence of three symptoms during the past 12 months (pain in one or several joints, joint swelling, morning stiffness). Each item was to be answered in a simple yes/no response format. In a second step, participants with positive responses in two or more of the joint-related items were invited to a rheumatological examination at the Hannover Medical School (see section 4, for details). Based on this examination, participants were classified as RA cases or non-cases.

Of the 9,558 questionnaires mailed out between June 1986 and November 1988, 8,084 (86%) were returned. For the present study, we excluded 476 cases (6% of the survey sample) due to incomplete data in any of the study variables and another 446 cases due to inconsistent responses to the questionnaire item referring to joint pain „today“. This item actually consisted of two parts which asked about pain in one joint and pain in several joints, respectively. Cases with positive responses to both parts of the item were excluded. However, further analyses without exclusion of inconsistent observations showed essentially the same results as those presented below.

				S 1 ¹⁾	N ²⁾	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y
				S 2	N	N	Y	Y	N	N	Y	Y	N	N	Y	Y	N	N	Y	Y
				S 3	N	N	N	N	Y	Y	Y	Y	N	N	N	N	Y	Y	Y	Y
				S 4	N	N	N	N	N	Y	N	N	Y	Y	Y	Y	Y	Y	Y	Y
S 5	S 6	S 7	S 8																	
N	N	N	N		3145	404	222	224	245	94	50	45	23	6	4	6	23	10	5	1
Y	N	N	N		24	9	6	16	11	11	4	9	1	2	1	1	2	2	1	4
N	Y	N	N		242	36	40	27	105	82	42	78	9	1	1	0	5	6	6	8
Y	Y	N	N		4	2	2	1	11	6	3	12	0	0	0	0	1	1	3	7
N	N	Y	N		46	11	6	3	13	4	3	4	26	10	6	7	21	12	2	9
Y	N	Y	N		1	1	0	0	1	0	0	1	1	0	0	0	3	0	0	2
N	Y	Y	N		49	6	2	8	16	10	9	12	18	1	1	2	25	22	10	8
Y	Y	Y	N		0	0	0	0	1	1	0	3	0	0	1	1	10	3	3	9
N	N	N	Y		37	11	8	16	14	10	9	8	1	0	0	0	1	1	0	0
Y	N	N	Y		24	14	8	13	16	12	7	16	1	0	0	3	3	2	1	4
N	Y	N	Y		64	15	12	20	30	37	23	63	2	0	1	0	4	5	2	7
Y	Y	N	Y		13	10	7	16	31	67	19	93	1	2	0	1	8	10	6	21
N	N	Y	Y		3	0	2	0	3	2	1	2	1	0	0	1	6	3	1	3
Y	N	Y	Y		1	0	1	0	2	1	0	2	1	3	1	1	12	2	2	3
N	Y	Y	Y		48	6	3	8	14	16	0	17	7	5	2	0	21	12	9	8
Y	Y	Y	Y		6	1	1	2	16	18	9	26	5	0	2	4	36	45	23	126

1) Symptoms „today“: S 1: back pain, S 2: neck pain, S 3: joint pain, S 4: swelling, S 5: stiffness; symptoms during past 12 months: S 6: joint pain, S 7: swelling, S 8: stiffness. S 3 is a combination of two item parts: pain in one single joint and/or pain in two or more joints; cases with inconsistent responses (pain in one and pain in two or more joints) were excluded.

2) N: no, Y: yes.

Table 1: Eight-item data set from the Hannover Rheumatoid Arthritis Study

Exclusion of cases with incomplete or inconsistent data resulted in an effective sample size of 7,162. Of these, 1,348 persons fulfilled the inclusion criteria and were invited to the medical examination; 956 (71%) actually participated. Thirty-two RA cases were identified.

Table 1 shows the multidimensional contingency table of manifest responses to the eight items of the postal questionnaire. To obtain parameter estimates for the latent class models described below, we used the program for linear logistic LCA written by Formann (1992) and the MLLSA module (Clogg, 1977) of the Categorical Data Analysis System (CDAS, Eliason 1988).

3. Identifying Patterns of Symptom Reporting

From a substantive point of view, manifest responses to the questionnaire items could be generated by various underlying patterns of symptom reporting. Of those, patterns suggesting the existence of specific „syndromes“ or indicating an ordering of latent classes on a one-dimensional scale would be of special interest for the epidemiologist.

Looking for such patterns in an exploratory way, we submitted all eight variables of the postal survey sample to unrestricted LCA (cf. Rost & Langeheine in this volume, eq. 34 in section 2.1). Models with two to eight latent classes were fitted. According to the suggestions given by Clogg (1977), we refitted these models at least twice, each time specifying a different set of start values to circumvent the problem of local maxima.

Pearson and Likelihood Ratio X^2 statistics indicated that none of the models was acceptable. Moreover, we found multiple (usually more than two) solutions as well as estimates at the boundary of the parameter space. It may be suspected that either more than eight latent classes are necessary to explain the observed association between symptoms or that the structure of association essentially cannot be analyzed within the LCA framework if all eight questionnaire items are included simultaneously. From a conceptual point of view, inclusion of items referring to different time windows („today“, „past 12 months“) could have been responsible for the above results. For example, symptoms experienced during the past 12 months could have occurred at different points in time, being only arbitrarily related to each other.

Number of Latent Classes	Pearson X^2	Likelihood Ratio X^2	df	Local Maxima	Boundary Solutions
2	666.1	554.2	20	no	no
3	170.9	162.4	14	no	no
4	8.5	8.4	8	no	no
5	0.3	0.3	2	yes ¹⁾	no

¹⁾ If local maxima were detected, entries refer to the solution with the smallest X^2 value (Pearson statistic).

Table 2: Goodness-of-Fit Statistics and Supplementary Information for Five-Item Latent Class Models

Therefore, we excluded the three items referring to the past 12 months and restricted the analysis to the complaints respondents experienced „today“. Goodness-of-fit tests for LCA applied to the reduced set of variables are shown in Table 2. While the models with two and three latent classes have to be clearly rejected, the four-class model fits the data well. Note that apart from the model with five latent classes, no local maxima or boundary solutions have been observed.

Parameter estimates of the four-class model as shown in Figure 1 suggest a „syndromes“ model for the postal survey data: one class represents a „symptom-free group“ (class I) whose members have low probability of reporting any symptom; class II is mainly characterized by pain, swelling, and stiffness in the joints; class III comprises subjects who tend to report pain in the back, neck, and in the joints but no swelling and stiffness. Class IV is complementary to the „symptom-free group“; members of this class are characterized by high probabilities of reporting each symptom.

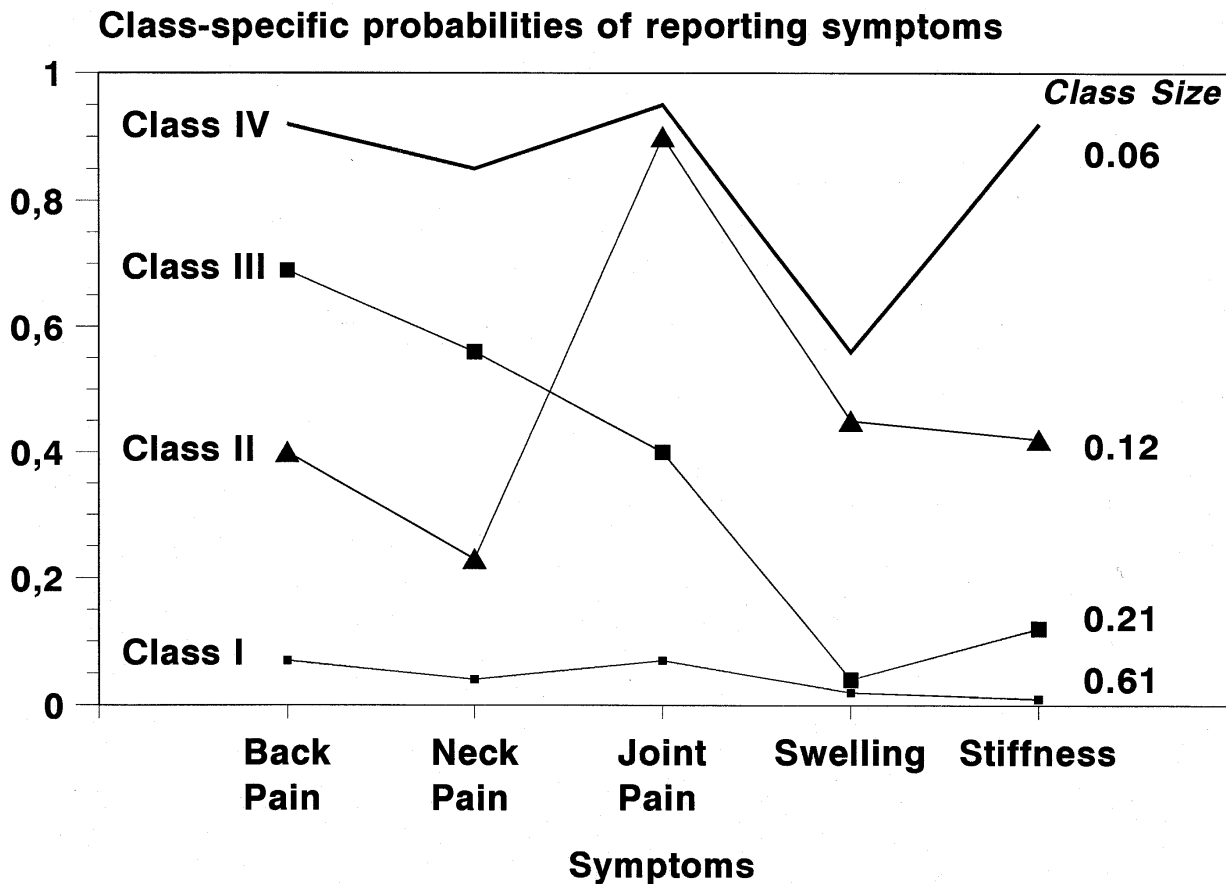


Figure 1: Results of latent class analysis of five musculo-skeletal symptoms „today“ (N = 7,162).

For the epidemiologist, these results contain valuable information as to the internal structure of the data. Especially the distinction between classes II and III, the „joint-type“ and the „back pain-type“ groups, and the identification of a group with „widespread complaints“ (class IV) appear to be useful in descriptive epidemiology of rheumatic complaints. Without use of multivariate methods such as LCA, prevalence figures would have to describe the frequency of complaints on a symptom-per-symptom basis. With the application of suitable statistical models, these figures can be based on empirically derived syndromes. In this context, comparison of findings from LCA with results of other multivariate methods aiming at the identification of patterns in categorical data - specifically Configural Frequency Analysis (CFA) and its recently developed variants (von Eye, 1990) - could be very informative.

4. Defining Inclusion Criteria for Further Diagnostic Evaluation

During the first year of the Hannover study, a respondent was invited to a medical examination if he or she had reported two or more of the joint-related symptoms (pain, swelling, stiffness). This resulted in a group of 386 examinees, eight of which were diagnosed as RA cases. As the efficiency of this procedure appeared to be rather low (nearly 50 persons had to be examined in order to identify one RA case), the study group narrowed the inclusion criterion. For the rest of the study, not only were two positive answers with respect to joint complaints required for inclusion in the examination survey, but additionally, one of the symptoms had to be joint swelling (either „today“ or in the past 12 months). With this new inclusion rule, efficiency could be considerably increased. Of the 570 persons examined after the inclusion rule had been modified, 24 were found to have RA. Thus, to identify one case of RA only 24 persons had to be examined.

What would have happened had the researchers based their new inclusion criterion on LCA results of mail survey data collected during the first study year? To answer this question, we replicated the analyses of the five symptoms „today“ using only data collected before modification of the inclusion criterion (Sample A, $N = 2,219$). In this sample, the four-class model fits well ($LR X^2 = 2.6$, $df = 8$), while models with two and three latent classes were not acceptable ($LR X^2 = 167.9$, $df = 20$ for the two-class model; $LR X^2 = 64.5$, $df = 14$ for the three-class model). On the whole, parameter estimates of the four-class model were well within the range of those obtained with the full data set. We also found a „symptom-free“ class and classes characterized by „joint-type“, „back pain-type“, and „widespread“ symptoms. Respondents were assigned to one of these classes employing the prediction rule based on the modal probabilities. The same prediction rule was used in those surveyed after the inclusion criterion had been modified (Sample B, $N = 4,943$).

RA is a disorder primarily affecting the joints and their surrounding tissue. Therefore, a reasonable decision of the researchers could have been to invite those assigned to the „joint-type“ group or to the group with „widespread“ symptoms to a further diagnostic evaluation. This would exclude those with „back pain-type“ complaints and those of the „symptom-free“ group. In fact, all RA cases in Sample A had been assigned to the „joint-type“ or to the „widespread symptoms“ class. Had the researchers implemented this criterion in Sample B, only 373 respondents would have been included in the examination survey instead of 570 according to the criterion actually used. However, since one RA case was not assigned to either the „joint-type“ or the „widespread symptoms“ class, sensitivity of this procedure would have decreased. The cost of examining 197 persons less than the number based on the original criterion applied in Sample B would have amounted to a loss of one RA case in the examination survey.

We further evaluated the loss introduced by using the LCA-based inclusion rule against the loss implied by other ways of narrowing the original inclusion criterion in Sample B. For several reasonable alternatives, Table 3 shows a) the number of respondents who would have been examined, and b) the number of RA cases lost due to modification of the inclusion rule.

The reduction of sample size in the examination survey varies between 84 (inclusion rule C-2) and 305 subjects (C-3), the largest reduction in sample size being associated with the highest loss of RA cases. Applying inclusion rule C-1 (B plus pain in joints „today“) would have resulted in a sample size reduction of 25 % without losing any RA case. Hence, this alternative might be considered best among those evaluated. We do not know whether the

researchers would have accepted losing one RA case; if so, the LCA-based inclusion rule would have resulted in a reduction of sample size of 35 %. As compared to these two, the other alternatives appear to be rather unattractive because either the reduction of sample size is too small (C-2), too many RA cases are lost (C-3), or the ratio of lost RA cases to sample size reduction is relatively unfavorable (C-4).

In comparing these results, it is noteworthy that in the LCA-based inclusion rule only data on complaints „today“ were used (including symptoms such as back and neck pain, which were considered irrelevant by the study group). In contrast, the other criteria also utilized information about symptoms in the past 12 months.

Inclusion Rule		Number in Examination Survey	Reduction of Sample Size	Number of RA Cases Lost
LCA	Subject assigned to class II or IV in the four-class latent class model of five symptoms.	373	35 %	1
B ¹⁾	Two or more of the following: joint pain, swelling, stiffness - irrespective of time reference („today“ or during past 12 months), one symptom has to be swelling.	570	0 %	0
C-1	As B plus: one symptom has to be joint pain „today“.	426	25 %	0
C-2	As B plus: one symptom has to be joint pain during past 12 months.	486	15 %	0
C-3	As B plus: one symptom has to be joint stiffness „today“.	265	54 %	4
C-4	As B plus: one symptom has to be joint stiffness during past 12 months.	355	38 %	2

¹⁾ Inclusion rule B was actually used in Sample B of the Hannover study.

Table 3: Number of cases in examination survey and RA cases lost according to different inclusion rules

5. Conclusion

When analyzing „soft data“ from population-based surveys, epidemiologists generally face the problem of insufficient prior knowledge of basic characteristics of the measures they apply. Aggregation of information from several items and definition of relevant subgroups of the sample often rely on procedures whose adequacy is not empirically tested.

This paper examined the utility of LCA in a two-stage population survey of musculoskeletal symptoms in which part of the respondents were invited to a medical examination. A pervasive problem in choosing inclusion criteria for medical examination surveys is the trade-off between sensitivity and specificity of these criteria. In our example, we were able to demonstrate that results of LCA could have been successfully applied to this problem. The application of LCA to the survey data and subsequent definition of an inclusion rule based on its results, could have enabled the researchers to achieve both: closer insights into the internal

structure of symptom reporting behavior and considerable reduction of sample size based on an explicit model of that behavior.

However, implementing an inclusion rule based on LCA implies the subdivision of the whole postal survey sample into two parts (one part in which the latent class model is estimated, a second part in which the LCA-based inclusion rule is employed). This requires careful planning of data collection with appropriate definition of the number of cases in both parts of the postal survey sample.

It might be argued that it seems unreasonable to lay too much emphasis on attempts at reducing sample size by one or two hundred cases in a large medical examination survey. However, since seven cases per week on the average were medically examined in the Hannover study, sample size reduction according to the LCA-based inclusion rule would have saved more than half a year of data collection time.

LCA has several possible applications in epidemiologic surveys beyond the analyses presented in this paper. Cases with missing values, for example, which were discarded from our analysis could have been included using appropriately specified models. Constrained models representing meaningful substantive hypotheses and multi-group analyses could also have been implemented in a straightforward way. We presume that LCA, for which several computer programs are now available, is well suited for a considerable portion of data analysis problems in studies similar to our example, and we therefore recommend its use to epidemiologists and survey researchers.

Acknowledgements

The authors are indebted to Kim Bloomfield for her support in editing this paper, and to the principal investigators of the Hannover Study on Rheumatoid Arthritis (A. Wasinus, W. Mau, H. Raspe) for permission to use their data. We gratefully acknowledge the helpful and constructive comments of two anonymous reviewers.

References

- Clogg, C. C. (1977). *Unrestricted and restricted maximum likelihood latent structure analysis*. University Park, PA: Pennsylvania State University.
- Clogg, C. C. (1988). Latent class models for measuring. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 173-205). New York: Plenum.
- Eliason, S. R. (1988). *The categorical data analysis system*. Version 3.00A user's manual. University Park, PA: Pennsylvania State University.
- Formann, A. K. (1985). Constrained latent class models: Theory and applications. *British Journal of Mathematical and Statistical Psychology*, 38, 87-111.
- Formann, A. K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association*, 87, 476-486.
- Formann, A. K. and Kohlmann, T. (1996). Latent class analysis in medical research. *Statistical Methods in Medical Research*, 5, 179-211.
- Goodman, L. A. (1974). The analysis of qualitative variables when some of the variables are unobservable. Part I - a modified latent structure approach. *American Journal of Sociology*, 79, 1179-1259.
- Haberman, S. J. (1979). *Analysis of qualitative data. Vol. II: New developments*. New York: Academic Press.

Morrison, A. S. (1985). *Screening in chronic disease*. New York: Oxford University Press.

von Eye, A. (1990). *Introduction to configural frequency analysis. The search for types and antitypes in cross-classifications*. Cambridge: Cambridge University Press.

Wasmus, A., Kindel, P., Mattussek, S. and Raspe, H. H. (1989). Activity and severity of rheumatoid arthritis in Hannover/FRG and in one regional referral center. *Scandinavian Journal of Rheumatology, Suppl. 79*, 33-44.