METHODS IN ANALYSING DATA FROM A PHYSICIAN ATTITUDE SURVEY

T. To, M. Agha, SP. Pinfold, H.A. Llewellyn-Thomas, C. Sawka, and A.M. O'Connor

Teresa To, Institute for Clinical Evaluative Sciences in Ontario, G106, 2075 Bayview Avenue, North York, Ontario, M4N 3M5, Canada. Tel: (416)480-4055 Ext. 3783. Fax: (416) 480-6048. E-mail: teresa@ices.on.ca

Key words: Logistic Regression, Ordinal Variable, Physician Attitude

ABSTRACT

Study Objective and Design: Using data from a study on physicians' preferences for the treatment of node negative breast cancer, the agreement among physicians on the appropriateness of adjuvant systemic therapy was analyzed. The objectives of the current analysis were twofold: to evaluate the appropriateness of using the original 9-level response variable (versus 2-level scale); and, to evaluate the effect of using a 9-level response variable on the size and direction of relative risk estimates (β) when the data are sparse. Using a resampling procedure and a 10% sample of the survey data, further evaluation was conducted on the effect of using all response levels on the β estimates when the data are limited. Two forms of treatment were studied: Tamoxifen and chemotherapy, and it was found that the distributions of the estimated odds ratios for both treatments resemble a normal distribution.

Main Results: Findings of this study indicate that while reducing the number of response categories in the outcome variable resulted in changing the absolute value of the estimated odds ratios, it does not necessarily increase the value of estimates. The effect of the reduction in the number of response categories in the outcome variable was most notable with the more significant independent variables. The mean of the corresponding distributions of the odds ratio resulting from a resampling procedure is close to the odds ratios using the whole data, however, the distributions of the odds ratios for other variables are more skewed.

Conclusions: It was concluded that using all response categories, even when the sample size is small, still result in reliable estimates.

INTRODUCTION

The management of localized breast cancer has evolved considerably over the last two decades. This evolution resulted in placing an increased emphasis on the use of adjuvant systemic therapy in an attempt to eradicate micro metastatic disease. While little debate exists about the value of systemic adjuvant approach for axillary lymph node positive breast cancer, its effectiveness in the treatment of axillary node negative breast cancer remains controversial.

Sawka et al[1] conducted a survey to examine the variations in physicians' preferences for systemic adjuvant therapy in the treatment of node negative breast cancer patients. The detailed methods used in the Sawka et al[1] survey (hereafter referred to as the "main study") have been described elsewhere. One hundred and twenty-five of the eligible 144 physicians received a survey questionnaire and were asked to rate the appropriateness of systemic therapy by Tamoxifen, chemotherapy, or both, for the treatment of hypothetical patients (described in varying clinical profiles) with node negative breast cancer. The data from the Sawka et al[1] survey are used for this current study. In the survey questionnaire, the outcome variable, i.e., (appropriateness of each therapy) was a 9-level response variable. Responses were grouped into 2-levels for the purpose of increasing the cell counts. The grouping of response categories is a common strategy, especially when the sample size is small and data are sparse. This practice, could result in a loss of the details of collected information, possibly changing the size and direction(s) of risk estimates, and attaining a statistically significant test result in a multivariate analysis of the data.

This study evaluates the effect on the estimated odds ratios, when the use of all response categories is compared with the use of grouped categories. The effect of a small sample size on the estimated odds ratios was evaluated when using all response categories. The specific objectives of the current study are two-fold: 1) to evaluate the effect of using a 9-level response variable on the estimates of the odds ratios (β) from an ordinal logistic regression, and 2) to evaluate the bias in the β estimates when the data are sparse.

With the increasing usage of multi-level response variables in health related studies, it is important to evaluate the effect of the most common strategy, i.e., combining adjacent categories, on the estimated odds ratios and consequently the interpretation of the results.

METHODS

Variables

The independent variables were the five patient characteristics, which were used to construct the varying clinical profiles. These include: age (AGE), tumour size (SIZE), estrogen receptor status (ER), histologic grade (GRADE), and lymphatic invasion (LYMPH). The main dependent outcome variable was the perceived appropriateness (APR) of systemic therapy, as assessed on a 9-level ordinal scale (with 1 = "most inappropriate"

and 9 = "most appropriate").

Statistical Method

1. Univariate Analysis

Using all response levels, the effects of independent variables on the agreement among the physicians about the appropriateness of each treatment have been evaluated and compared with the results of the main study.

2. Multivariate Analysis

2.1 Ordinal Logistic Regression

Statistical models of the dependence of an ordinal variable (i.e., the rating of the appropriateness of a treatment) on one or more explanatory variables are termed "ordinal regression models"; sometimes they are also referred to as the "proportional odds model" or the "cumulative odds model."

Following the notation of Armstrong and Sloan[2], the integers $1,\ldots,k$ are the labels for the k ordered response categories (here $1,\ldots,9$ from "most inappropriate" to "most appropriate"); and π_j, $j=1,\ldots,k$ are the multinominal probability of a response appearing in each category. In an ordinal regression model, the π_j depends on the values of a vector of explanatory variables x through regression parameters.

In a natural extension of the logistic model for binary response data, the proportional odds model can be described as follows:

$$\text{logit}(\gamma_j) = \ln(\gamma_j/(1-\gamma_j)) = \theta_j - \beta^T x, \quad j=1,\ldots,k$$

where $\gamma_j = \pi_1 + \ldots + \pi_j$ are the cumulative probabilities of being in one of the first j categories. The parameters θ_j represents the baseline logits of cumulative response probabilities in a person for whom $x=0$, and β represents the "regression" parameters through which the effects of the explanatory variables are mediated.

The above model was used to study the appropriateness ratings of treatment by Tamoxifen and chemotherapy (the dependent variable), and also the varying patients' characteristics (explanatory variables).

2.2 *Re-Sampling Procedure*

Grouping of the response categories is a common strategy when the sample size is small or when the data are sparse. Using all response categories under this situation will result in empty cells. Therefore, the question remains about the effect (absolute magnitude or bias) in using all response categories versus grouped categories when there are limited data. It has been suggested that, using the proportional odds model when the number of response categories is large and the data are sparse, the estimates of β_j may be too large in magnitude and result in biased estimates.[3]

In this study, one thousand samples with 600 observations each (1/10th of the number of observations for each treatment in the main study) were selected. The similar proportional odds model procedures were performed for each sample, (as described in the previous section) with the same explanatory variables included in the model.

Univariate analyses of this study were conducted on each of the clinical scenarios defined, while ordinal logistic regressions were used in the multivariate analyses. All analyses were performed using the Statistical Analysis System software.[4]

RESULTS

Forty-eight clinical scenarios (2x3x2x2x2) were defined as follows: AGE (<40 years/40+years), SIZE (small, medium, and large), ER (positive/negative), GRADE (well differentiated/poorly differentiated), and LYMPH (present/absent). For each profile, the distributions of ratings of the perceived appropriateness of Tamoxifen and chemotherapy were calculated in order to assess raters' agreement.

Results from both univariate and multivariate analyses, using all response levels, were in agreement with that of the main study. Where the response variable was grouped, AGE, ER and LYMPH were the deciding variables for evaluating the appropriateness of treatment.

Ordinal logistic regression analysis with a 9-level response variable

Table 1 shows the reported odds ratios in the main study[1] compared to the current analysis. The main study was based on a grouped binary outcome (appropriate vs inappropriate) whereas, the current study is based on a 9-level response variable.

McCullagh and Nelder[3] suggest that reducing the number of response categories could affect the parameter estimates and consequently the estimated odds ratios. Comparing the estimated odds ratios using a 2-level response with the original 9-level indicate that the changes in the estimates could be in either direction. Using all 9 levels the estimates remain unchanged for the majority of the variables (or changes in the same direction). This, however, is not always the case, specifically for more significant variables such as ER or tumour SIZE for the Tamoxifen treatment. Although these changes do not affect the general conclusion of the study, it may affect the intensity of the interpretation of the results.

Re-Sampling Procedure

Figures 1-1 to 1-6 show a normal distribution of the sample odds ratios for most of the five independent variables for the Tamoxifen treatment. The odds ratios calculated based on the parent data (N=6000), for most of

the variables, are close to the mean of the odds ratios estimated from the samples.

Table 2 shows the estimates of the odds ratios of each variable based on the parent data and those of the samples. It is clear from Table 2 that for the variables, SIZE, GRADE, and LYMPH, the difference is negligible. For more significant variables (AGE and ER), the distributions of the odds ratios are more skewed (Figures 1-1 to 1-6). In other words, the effect of reducing sample size seems to be larger on more significant variables such as AGE and ER. Since the distribution of ratings is highly skewed with respect to variables AGE and ER, it is not expected that a reduction in sample size results in smaller or even empty cells at the extreme levels.

Identical procedures were performed on the variables for chemotherapy treatment. Figures 2-1 to 2-6 show the distributions of the odds ratios for 1,000 samples of size 600 for chemotherapy. Again, almost all of the distributions resemble a normal distribution and the mean of the distributions is very close to the observed odds ratio based on the parent data (Table 3). Although there is a slight increase in the odds ratios for ER, GRADE and LYMPH, the difference seems to be negligible. The result is similar to that observed for Tamoxifen.

Table 1
Estimated Odds Ratios in Main Study and Current Study

		Tamoxifen		Chemotherapy	
Variable	**Reference Category**	**Main Study (2 - Levels)**	**Current Study (9 - Levels)**	**Main Study (2 - Levels)**	**Current Study (9 - Levels)**
Age	Premenopause	3.68	3.65	0.18	0.26
Tumour Size - Large	Small Size	1.48	2.33	2.89	6.66
ER[1] Status	Negative	16.09	7.69	0.52	0.30
Grade[2]	Well-differentiated	0.77	0.86	1.47	2.58
Lymphatic[3]	Without invasion	0.79	0.80	1.18	1.99

[1] ER: Estrogen Receptor, [2] Grade: Histologic Grade, [3] Lymphatic: Lymphatic Invasion

Figure 1-1

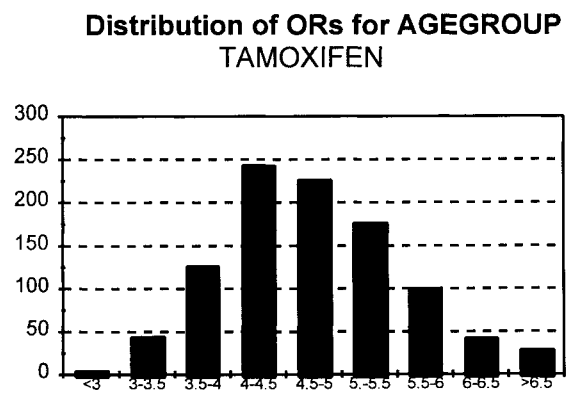

Figure 1-2

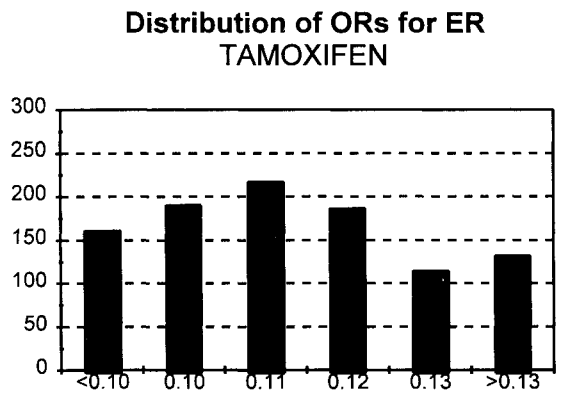

Figure 1-3

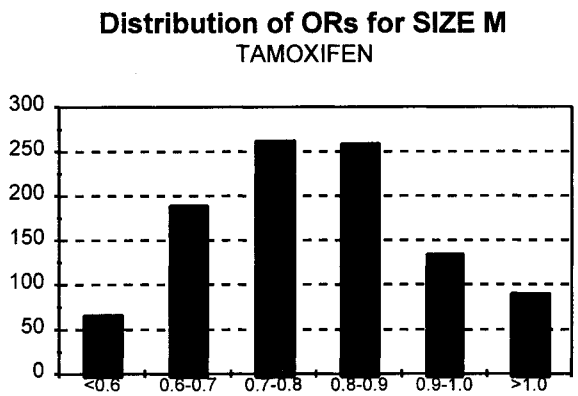

Figure 1-4

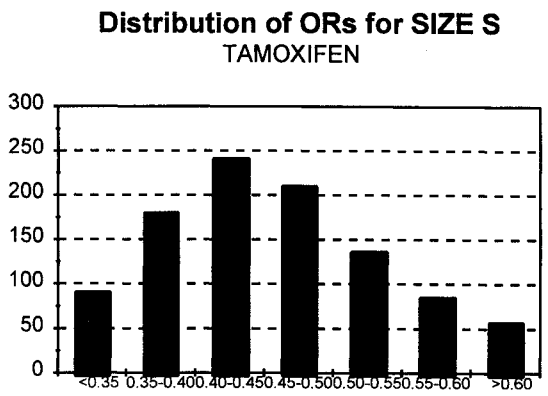

Figure 1-5

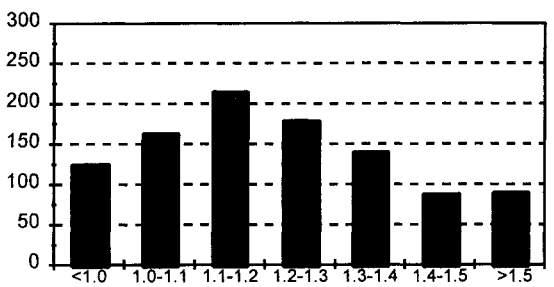

Figure 1-6

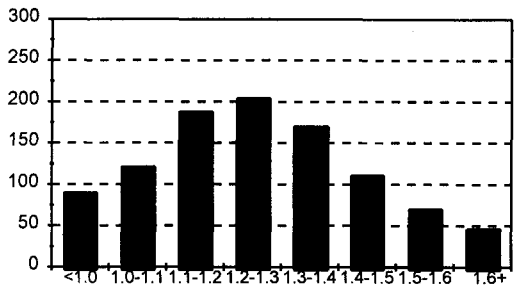

Table 2
Estimated Odds Ratios (OR) based on the whole sample and mean of the distribution

		Tamoxifen	
Variable	**Reference Category**	**OR (Whole Sample) (N= 6000)**	**OR (Mean of the distribution) (1000 samples of 600 observations)**
Age	Premenopause	3.65	4.66
Tumour Size - Small	Large size	0.43	0.44
Tumour Size - Medium	Large size	0.74	0.78
ER[1] Status	Positive	0.13	0.11
Grade[2]	Poorly-differentiated	1.16	1.19
Lymphatic[3]	With invasion	1.24	1.24

[1] ER: Estrogen Receptor [2] Grade: Histologic Grade [3] Lymphatic: Lymphatic Invasion

Figure 2-1

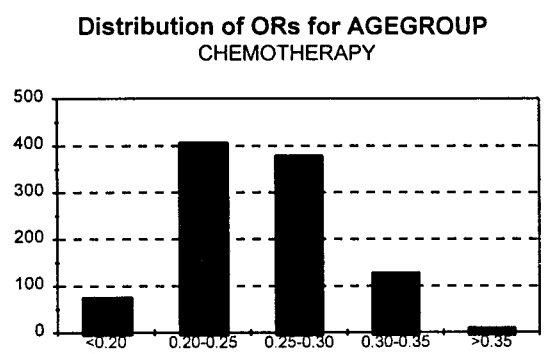

Figure 2-2

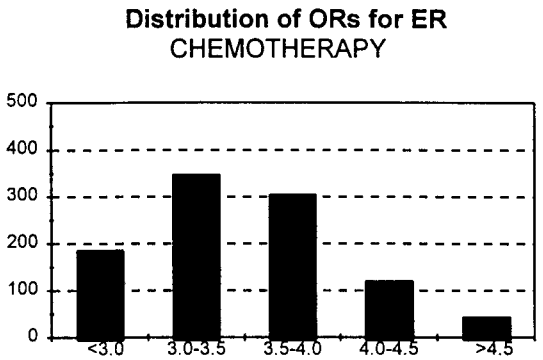

Figure 2-3

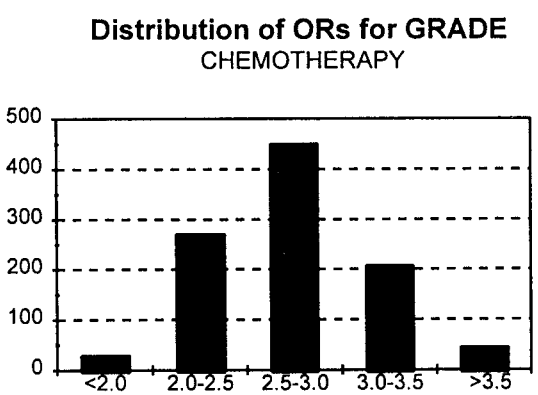

Figure 2-4

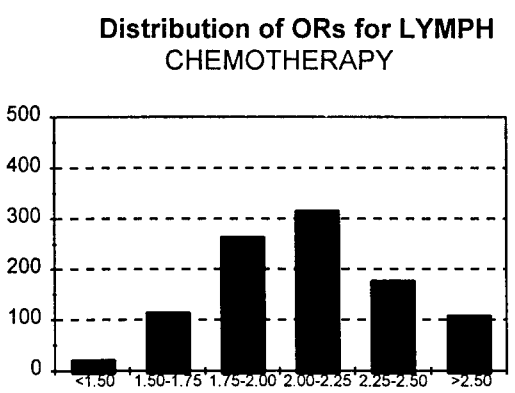

Figure 2-5

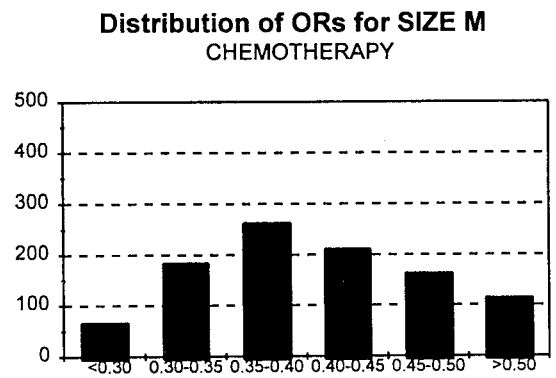

Figure 2-6

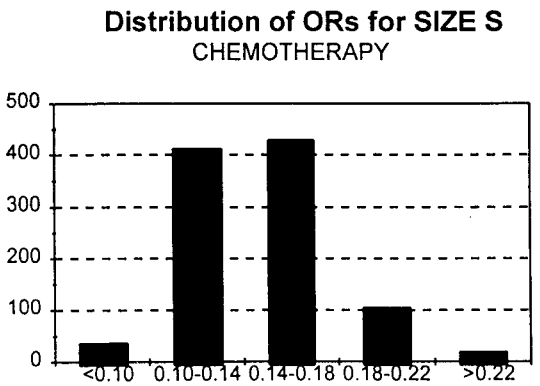

Table 3
Estimated Odds Ratios (OR) based on the whole sample and mean of the distribution
Chemotherapy

Variable	Reference Category	OR (Whole Sample) (N= 6000)	OR (Mean of the distribution) (1000 samples of 600
Age	Premenopause	0.26	0.25
Tumour Size - Small	Large size	0.15	0.14
Tumour Size - Medium	Large size	0.39	0.39
ER[1] Status	Positive	3.32	3.54
Grade[2]	Poorly-differentiated	0.39	0.37
Lymphatic[3]	With invasion	0.50	0.49

[1] ER: Estrogen Receptor [2] Grade: Histologic Grade [3] Lymphatic: Lymphatic Invasion

DISCUSSION

Ordinal scales are among the most frequently used measurement scales with applications in many areas, such as determination of physical or mental well-being, and classification of radiographs or rating applications. In many of these applications, a new category is usually formed to overcome the problem of sparse data or empty cells by combining adjacent categories of the original scale. The reduction of response categories will normally reduce the available information, change the estimate, and/or the attained significant level. It is essential that any conclusions on findings should not be affected by the number of response categories used.

This study is an exercise based on a 2- or a 9-level response category, using data from Sawka et al[1]. While results from the univariate analysis (using 9-levels) are in agreement with those of the reported main study (2-levels), the estimated odds ratios from multivariate analysis are not exactly identical. Reducing the number of response categories resulted in both an increase or a decrease in the estimated odds ratios. The effect of the reduction is more noticeable in more significant independent variables such as AGE and ER in this study. One possible explanation is that when the data are sparse, the parameter estimates may be too large in magnitude and unstable[3].

In the second part of this study, a resampling procedure was used to generate the distribution of the β estimates based on 1000 samples each with 600 observations. The results indicate that for most of the variables, the distribution of the β estimates obtained by the ordinal logistic regression, resemble a normal distribution with the mean being very close to the estimate based on the parent data.

Similar to McCullagh and Nelder's[3] suggestion, our observations imply that the use of the ordinal logistic regression with the original multi-level response categories (i.e., without collapsing categories), induces changes in the values of the estimated odds ratios, but does not change the overall conclusions. The estimates of the parameters of interest are still reliable.

Variables used in epidemiological studies or health surveys are often measured on an ordinal scale. Data of these types are sometimes analyzed as numerical scores, but such an approach is strictly valid only if intervals between consecutive points on the scale can be considered equivalent. To avoid this assumption, the scales are often dichotomized and analyzed, using standard techniques for binary data. Although valid, this approach loses information by collapsing some categories of the original scale. Statistical methods which respect the ordinal nature of this kind of response data have been developed and are available in standard statistical packages.

Our study suggests that using all response categories, even when the sample size is not large, will still result in reliable estimates.

REFERENCES

1. Sawka CA, O'Connor AM, Llewellyn-Thomas HA, To T, Pinfold SP. Harrison-Woermke D. Appropriateness of adjuvant systemic therapy for axillary node negative breast cancer: A physician opinion survey. J Clin Oncol 1995;13(6):1459-1469.

2. Armstrong B, Sloan M. Ordinal regression models for epidemiologic data. Am J Epidemiol 1989;129(1):192-203.

3. McCullagh P; Nelder JA. Generalized linear model. 2nd ed. London, England: Chapman & Hall; 1990. 149p.

4. Statistical Analysis System Institute Inc.(SAS), SAS Campus Drive, Cary, North Carolina, 27513, U.S.A. 1996.