W. P. Bergsma, E. M. D. Aris, and F. S. Tibaldi

Linear Categorical Marginal Modeling of solicited symptoms in vaccine clinical trials

Article (Accepted version) (Refereed)

Original citation:
Bergsma, W. P., Aris, E. M. D. and Tibaldi, F. S. (2013) *Linear Categorical Marginal Modeling of solicited symptoms in vaccine clinical trials.* Statistics in Biopharmaceutical Research, 5 (1). pp. 27-37. ISSN 1946-6315

DOI: 10.1080/10496491.2012.738111

Linear Categorical Marginal Modeling of Solicited Symptoms in Vaccine Clinical Trials

Bergsma, W.P., Aris, E.M.D, and Tibaldi, F.S.
Biometrics Department, GlaxoSmithKline Biologicals, Rixensart, Belgium
Department of Statistics, London School of Economics and Political Science, London, UK

Abstract

Analysis of the occurrence of adverse events, and in particular of solicited symptoms, following vaccination is often needed for the safety and benefit-risk evaluation of any candidate vaccine, and typically involves taking repeated measurements. In this article, it is shown that Linear Categorical Marginal Models are well-suited to take the dependencies in the data arising from the repeated measurements into account and provide detailed and useful information for comparing safety profiles of different products while remaining relatively easy to interpret. Linear Categorical Marginal Models are presented and applied to a Phase III clinical trial of a candidate meningoccocal pediatric vaccine.

KEYWORDS : Marginal models, repeated categorical data, vaccine development, safety.

1 Introduction

When developing new vaccines, it is necessary to show that new candidates have an acceptable safety profile. Typically, the clinical safety evaluation of the vaccine is performed regarding two specific aspects. First, the occurrence of a certain number of local or general symptoms is checked proactively via diary cards recording the occurrence or absence of the symptom during a certain number of days after the injection. These symptoms are usually called solicited symptoms. For ease of recording a standard intensity scale is often used and contains a certain number of possible intensity of the symptom, typically between 1 and 3 (see, for example Table 1). Subjects are then asked to fill in the maximum daily intensity of each reported solicited symptom during the entire solicited symptom follow-up period in the diary card. We will consider here a 4-day follow-up period, the day of vaccination being denoted as day 1.

In parallel to this solicited symptoms collection, the subject is asked to record any occurrence of adverse event experience that could also occur post vaccination. As there is no pre-specification of the type or medical classification of the symptoms for which information is requested, these symptoms are usually called unsolicited symptoms.

In order to avoid different types of biases, solicited and unsolicited symptoms occurring after the vaccination by the candidate vaccine (hereafter denoted as active group) are often compared to the ones obtained after injection of a licensed vaccine (control group) observed in the same experimental conditions.

This paper will focus on the analysis of solicited symptoms. As the outcomes of these symptoms are often collected as categorical variables, the analysis methods presented below will specifically take this aspect into account. Several ways for comparing the active and the control groups will be presented and compared. First, the data along with standard methods

Table 1: *Definition of solicited adverse events intensities*

Adverse Event	Intensity	Description
Pain*	0	Absent
	1	Minor reaction to touch
	2	Cries/protests on touch
	3	Cries when limb is moved/spontaneously painful
Redness*	0	Absent
	1	>0 to ≤ 10 mm
	2	>10 to ≤ 30 mm
	3	>30 mm
Irritability	0	Behavior as usual
	1	Crying more than usual/no effect on normal activities
	2	Crying more than usual/interferes with normal activities
	3	Crying that cannot be comforted/interferes with normal activities

*at injection site.

presenting results for each day or overall will be introduced in Section 2, where the difficulties caused by the fact that we have dependencies in the data due to the repeated measurements are outlined. To overcome these difficulties, Linear Categorical Marginal Models (LCMMs), which take these dependencies into account, will be proposed in Section 3. For the ease of presentation, we will first consider the analysis of the occurrence of any event regardless of the intensity. Analyses taking into account the several intensities will only be dealt with in Section 4. Results from simulated data evaluating the Linear Categorical Marginal Models will be presented in Section 5. Finally, the advantages and drawbacks of Linear Categorical Marginal Models obtained via a Maximum Likelihood estimation procedure will be discussed.

2 Case Study

The data analyzed in this manuscript is coming from a Phase III trial of a meningococcal vaccine in children. For confidentiality reasons only partial data of the trial are used to illustrate our methods. In this study, children at age 12 to 15 months are randomly assigned 3:1 to 2 groups to be either vaccinated by the candidate vaccine or by a control. The candidate vaccine should offer a broader protection to meningococcal infection, so the safety question of interest is whether the safety profile of the vaccine is or is not worse than the control. The information collected for the solicited local symptoms is summarized in the first six columns of Table 2. The vaccine is injected in the upper left thigh at day 1, and the parents of the subjects are asked to fill in diary cards indicating whether or not the vaccinee experienced either pain, redness, or irritability during the follow-up period of 4 days.

Below, we first present some simple classical analyses involving Bonferroni-Holm corrections, and highlight the difficulties that arise due to the dependencies in the data as the measurements on the different days involve the same subjects.

In order to have an indication of the difference between the 2 groups, it is possible, for each day to test whether the difference in occurrence of the symptom is statistically significant. For example, Table 2 presents results from exact tests comparing the percentage of subjects

Table 2: *Differences between groups in percentage of subjects reporting a specified solicited local symptom during the 4-day post vaccination period.*

		Control (N=499)		Active (N=1381)		Control - Active 95%CI			p-value	
Symptom	Day	n	%	n	%	%	LL	UL	Raw	B-H
Pain	1	333	66.7	929	67.3	-0.54	-5.80	4.39	0.824	1.000
	2	252	50.5	613	44.4	6.11	1.00	11.41	0.021	0.084
	3	116	23.2	264	19.1	4.13	-0.31	9.05	0.051	0.154
	4	49	9.8	102	7.4	2.43	-0.80	6.35	0.102	0.204
	Any	366	73.3	1025	74.2	-0.87	-5.98	3.80	0.721	–
Redness	1	310	62.1	797	57.7	4.41	-0.70	9.59	0.090	0.090
	2	312	62.5	769	55.7	6.84	1.73	12.01	0.008	0.025
	3	214	42.9	504	36.5	6.39	1.30	11.69	0.013	0.027
	4	115	23.0	236	17.1	5.96	1.57	10.83	0.004	0.016
	Any	382	76.6	979	70.9	5.66	0.83	10.67	0.017	–
Irritability	1	221	44.3	570	41.3	3.01	-2.10	8.32	0.245	0.703
	2	218	43.7	587	42.5	1.18	-3.93	6.50	0.675	0.703
	3	164	32.9	413	29.9	2.96	-1.94	8.20	0.234	0.703
	4	115	23.0	246	17.8	5.23	0.83	10.12	0.012	0.047
	Any	279	55.9	763	55.2	0.66	-4.46	5.85	0.834	–

Note: Since the measurements on the four days and of the three symptoms are done on the same subjects, the differences between groups are correlated but this specific correlation is not taken into account here.

with the solicited symptom (p-values correspond to Fisher's exact test p-values). As several comparisons have been made, the p-values have been adjusted by symptom using the Bonferroni-Holm method, denoted the B-H p-value (see Holm, 1979). Note that, because within each subject the several p-values are likely to be correlated, this correction method may not be optimal. However, as it does not require any assumptions (model or distribution related) and its family-wise error rate does not exceed 5%, this test could be used here although it may be too conservative which may be problematic as it might mask a possible difference. For more complex settings with multiple doses or when several symptoms have to be considered simultaneously, other multiplicity correction methods such as the double false discovery rate (Mehrotra & Heyse, 2004) could be considered. Considering the unadjusted tests (regardless of intensity, i.e., the 'All' rows in Table 3), we would find statistically significant differences in the occurrence of pain (day 2), redness (days 2, 3 and 4) and irritability (day 4). However, when taking into account the multiplicity of the tests within a symptom via the B-H method, no statistically significant difference would be found for pain. Hence, as this example illustrates, it may not be uncommon to be in a situation in which several tests are significant when not adjusting for multiplicity, but no or fewer tests are significant if adjusted with an adjustment method that does not use all information in the data, leaving the user in doubt of which conclusion to draw especially as the B-H method is likely to overcorrect for multiplicity here. The problem magnifies if the intensities of the solicited symptoms experienced have to be taken into account such as in Table 3. There, as the B-H correction is applied per symptom, no B-H p-value is significant anymore for irritability, and only one out of the 6 significant p-values for redness remain significant when corrected for multiplicity.

An alternative analysis disregards intensity and only considers whether or not the symptom occurred on any of the four days. Here, a statistically significant difference would only be observed for redness (see Table 2). However, although this method is perfectly valid and circumvents the problem of repeated measures, substantial information may be lost. Indeed, it is possible that certain effects are visible during certain days but not during others, which could be the case for pain and irritability.

3 Linear Categorical Marginal Models

Table 2 shows the percentage differences of occurrence of various solicited symptoms between control and active groups are reported. In this section, we discuss some models for these data, using which we can answer various questions of interest, such as whether or not there are statistically significant differences between the responses on the four days. Since we have repeated measurements, i.e., measurements on different days involving the same subjects, dependencies arise which need to be taken into account. This can be done naturally using marginal modeling techniques.

3.1 Definition of the marginal proportions

Let the variable G denote the group a respondent is in ($G = 1$ for the active group and $G = 2$ for the control group), let S denote whether or not the solicited symptom occurred ($S = 1$ if the symptom occurred, $S = 2$ if it didn't), and let T denote the time after intervention in days ($T = 1, 2, 3, 4$). The proportion of respondents who are in group $G = g$ with solicited symptom $S = s$ given time $T = t$ is denoted by $\pi_{s\,g\,t}^{SG|T}$. We should note that these $\pi_{s\,g\,t}^{SG|T}$ are not proportions of an ordinary contingency table, but *marginal* proportions of a larger contingency table. Since for each solicited symptom, each subject in each group has four

Table 3: *Differences between groups in percentage of subjects reporting a specified solicited local symptom during the 4-day post vaccination period.*

		Control (N=499)		Active (N=1381)		Control - Active 95% CI			p-value	
Symptoms	Intensity	n	%	n	%	%	LL	UL	Raw	B-H
					Day 1					
Pain	All	333	66.7	929	67.3	-0.54	-5.8	4.39	0.824	1.000
	2 or 3	143	28.7	300	21.7	6.93	2.25	12.02	0.002	0.021
	3	33	6.6	31	2.2	4.37	1.83	7.63	<0.001	0.001
Redness	All	310	62.1	797	57.7	4.41	-0.7	9.59	0.090	0.538
	2 or 3	39	7.8	80	5.8	2.02	-0.93	5.69	0.133	0.626
	3	2	0.4	18	1.3	-0.9	-3.22	1	0.125	0.626
Irritability	All	221	44.3	570	41.3	3.01	-2.1	8.32	0.25	1.000
	2 or 3	31	6.2	88	6.4	-0.16	-3.86	3.17	1.000	1.000
	3	4	0.8	19	1.4	-0.57	-3.07	1.49	0.476	1.000
					Day 2					
Pain	All	252	50.5	613	44.4	6.11	1	11.41	0.021	0.169
	2 or 3	93	18.6	153	11.1	7.56	3.55	12.11	<0.001	0.001
	3	14	2.8	14	1	1.79	-0.03	4.42	0.008	0.075
Redness	All	312	62.5	769	55.7	6.84	1.73	12.01	0.008	0.091
	2 or 3	91	18.2	189	13.7	4.55	0.48	9.17	0.016	0.125
	3	16	3.2	42	3	0.17	-2.01	3.13	0.880	1.000
Irritability	All	218	43.7	587	42.5	1.18	-3.93	6.5	0.673	1.000
	2 or 3	67	13.4	149	10.8	2.64	-1.03	6.93	0.120	0.958
	3	15	3	34	2.5	0.54	-1.53	3.41	0.514	1.000
					Day 3					
Pain	All	116	23.2	264	19.1	4.13	-0.31	9.05	0.051	0.358
	2 or 3	25	5	43	3.1	1.9	-0.53	5.08	0.068	0.407
	3	2	0.4	7	0.5	-0.11	-2.22	1.74	1.000	1.000
Redness	All	214	42.9	504	36.5	6.39	1.3	11.69	0.013	0.121
	2 or 3	54	10.8	102	7.4	3.44	0.11	7.42	0.023	0.159
	3	11	2.2	19	1.4	0.83	-0.95	3.43	0.214	0.641
Irritability	All	164	32.9	413	29.9	2.96	-1.94	8.2	0.234	1.000
	2 or 3	54	10.8	115	8.3	2.49	-0.88	6.52	0.101	0.905
	3	14	2.8	21	1.5	1.28	-0.62	4	0.081	0.813
					Day 4					
Pain	All	49	9.8	102	7.4	2.43	-0.8	6.35	0.102	0.509
	2 or 3	10	2	14	1	0.99	-0.68	3.49	0.104	0.509
	3	2	0.4	4	0.3	0.11	-0.98	2.08	0.659	1.000
Redness	All	115	23	236	17.1	5.96	1.57	10.83	0.004	0.047
	2 or 3	23	4.6	31	2.2	2.36	0.09	5.41	0.011	0.114
	3	0	0	2	0.1	-0.14	-1.49	1.42	1.000	1.000
Irritability	All	115	23	246	17.8	5.23	0.83	10.12	0.012	0.142
	2 or 3	38	7.6	69	5	2.62	-0.26	6.21	0.042	0.459
	3	4	0.8	8	0.6	0.22	-1.09	2.39	0.530	1.000

measurements taken on the four days, the full contingency table for a solicited symptom involves five variables: G, the group the subject is in, S_1, whether a solicited symptom occurred on day 1, S_2, whether a solicited symptom occurred on day 2, and so on. Denote by $\pi^{GS_1S_2S_3S_4}_{g\ s_1\,s_2\,s_3\,s_4}$ the proportion of subjects in group $G = g$ with symptom $S_i = s_i$ on day i ($S_i = 1$ if the symptom occurred on day i and $S_i = 0$ if the symptom did not occur on day i). Then, with a '+' in the subscript denoting summation over that subscript,

$$\begin{aligned}
\pi^{SG|T}_{s\,g\ 1} &= \pi^{GS_1S_2S_3S_4}_{g\ s\ +\,+\,+} \\
\pi^{SG|T}_{s\,g\ 2} &= \pi^{GS_1S_2S_3S_4}_{g\ +\ s\ +\,+} \\
\pi^{SG|T}_{s\,g\ 3} &= \pi^{GS_1S_2S_3S_4}_{g\ +\,+\ s\ +} \\
\pi^{SG|T}_{s\,g\ 4} &= \pi^{GS_1S_2S_3S_4}_{g\ +\,+\,+\ s}
\end{aligned}$$

That is, the $\pi^{SG|T}_{s\,g\ t}$ are marginal proportions. We next discuss some models for these marginal proportions.

3.2 Modeling the differences in marginal proportions

Various questions can be asked about the data in Table 2 concerning changes in the response patterns over the four days. The conditional probability that $S = s$ given $G = g$ and $T = t$ is denoted

$$\pi^{S|GT}_{s\ g\,t} = \frac{\pi^{SG|T}_{s\,g\ t}}{\pi^{S|T}_{s\ t}}$$

The differences in the marginal proportions for active and control group at time t are denoted

$$\delta^T_t = \pi^{S|GT}_{1\ 1\,t} - \pi^{S|GT}_{1\ 2\,t}$$

and can be estimated by different models. The saturated model, later called *varying difference model*, which does not impose any restrictions but whose parameters can be useful for interpretation, is denoted by

$$\delta^T_t = \alpha + \beta_t \qquad \text{for all } t, \tag{1}$$

for some unknown parameters α and β_t. Here, the β parameters are not identified but can be identified by imposing a restriction such as $\sum_t \beta_t = 0$ (cf. effect coding in ANOVA). Various other models of interest are obtained by imposing restrictions on the α and β parameters.

The most parsimonious model asserting no differences between active and control is obtained by setting $\alpha = 0$ and $\beta_t = 0$, i.e.,

$$\delta^T_t = 0 \qquad \text{for all } t. \tag{2}$$

We will refer to this model as the *no difference model*. The presence of a difference but one which does not change over time is

$$\delta^T_t = \alpha \qquad \text{for all } t. \tag{3}$$

We will refer to this model as the *constant difference model*. Finally, a difference between active and control which changes linearly over time is formulated as

$$\delta^T_t = \alpha + \beta.t \qquad \text{for all } t. \tag{4}$$

We will refer to this model as the *linear difference model*.

Since the $\pi^{S|GT}_{s\ g\,t}$ are marginal proportions, and the aforementioned models are linear in these, we will call them *Linear Categorical Marginal Models* (LCMMs)

3.3 Fitting Linear Categorical Marginal Models

Before we describe the fitting procedure, we first formulate the model in matrix notation. Denote the vector of proportions for the full table, i.e., the $\pi^{GS_1S_2S_3S_4}_{g\,s_1\,s_2\,s_3\,s_4}$, by $\boldsymbol{\pi}$. The vector of marginal proportions of interest are a linear combination of the elements of $\boldsymbol{\pi}$ and can thus be written as

$$\mathbf{M}\boldsymbol{\pi}$$

where $\mathbf{M}$ is an appropriate matrix of zeroes and ones (for more details see Bergsma, Croon, & Hagenaars, 2009). Let $\boldsymbol{\delta}$ be the vector of δ_t^T. We can use the generalized exp-log notation of Kritzer (1977) and Bergsma et al. (2009) to represent $\boldsymbol{\delta}$, which we denote $\boldsymbol{\delta}(\mathbf{M}\boldsymbol{\pi})$ to indicate the dependence on the marginal proportions:

$$\boldsymbol{\delta}(\mathbf{M}\boldsymbol{\pi}) = \mathbf{C}' \exp \mathbf{B}' \log \mathbf{A}'\mathbf{M}\boldsymbol{\pi}$$

A linear model for this vector of coefficients, i.e., a LCMM, can then be denoted as

$$\boldsymbol{\delta}(\mathbf{M}\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta} \tag{5}$$

for an appropriate design matrix $\mathbf{X}$ and a parameter vector $\boldsymbol{\beta}$. With the columns of $\mathbf{U}$ spanning the orthogonal complement of the space spanned by the columns of $\mathbf{X}$, we can give the equivalent representation

$$\mathbf{U}'\boldsymbol{\delta}(\mathbf{M}\boldsymbol{\pi}) = \mathbf{0} \tag{6}$$

With $\mathbf{n}$ a vector of frequencies, the kernel of the multinomial log likelihood is given as

$$L(\boldsymbol{\pi}|\mathbf{n}) = \mathbf{n}' \log \boldsymbol{\pi} - N\mathbf{1}'\boldsymbol{\pi} \tag{7}$$

where N is the sample size. The problem now is to find an estimator of $\boldsymbol{\pi}$ subject to the constraint (6), or of $\boldsymbol{\beta}$ subject to (5), when the data vector $\mathbf{n}$ follows a multinomial likelihood. Two different estimation procedures have been developed for models of this type and more general models: the weighted least squares (WLS) method (Grizzle, Starmer, & Koch, 1969) and the maximum likelihood (ML) method (Lang & Agresti, 1994; Bergsma, 1997; Lang, 2004; Bergsma et al., 2009).

WLS is based on the the asymptotic covariance matrix of the sample value of $\boldsymbol{\theta}(\mathbf{M}\boldsymbol{\pi})$. Using the delta method this leads to the WLS estimator

$$\tilde{\boldsymbol{\beta}} = \left(\mathbf{X}'\left(\mathbf{JMD_pM}'\mathbf{J}'\right)^{-\mathbf{1}}\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{JMD_pM}'\mathbf{J}'\right)^{-\mathbf{1}}\mathbf{JMp}.$$

where $\mathbf{J}$ is the Jacobian of $\boldsymbol{\delta}$, $\mathbf{p}$ is the vector of observed probabilities, and $\mathbf{D_p}$ is the diagonal matrix with $\mathbf{p}$ on the main diagonal (see also, e.g., Koch, Landis, Freeman, & Lehnen, 1977).

The ML method is computationally more complex and based on maximizing the multinomial log likelihood (7) subject to the constraint (6). The constrained ML solution is a stationary point of the Lagrangian expression

$$L(\pi|\mathbf{n}) - \boldsymbol{\lambda}'\mathbf{U}'\boldsymbol{\delta}(\mathbf{M}\boldsymbol{\pi})$$

where $\boldsymbol{\lambda}$ is a vector of Lagrange multipliers. A scoring type algorithm which works well in practice is given in Bergsma et al. (2009) (see also Bergsma, 1997). The algorithm assumes the regularity conditions that $\mathbf{U}$ has full column rank and the Jacobian J has full row rank, which are normally satisfied in practice.

Once the estimates $\hat{\pi}$ have been obtained, marginal models can be tested by means of two well-known test statistics: the likelihood ratio test statistic

$$G^2 = -2N \sum_i p_i \log \frac{\hat{\pi}_i}{p_i}$$

and Pearson's chi-square test statistic

$$X^2 = N \sum_i \frac{(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i}.$$

If the postulated model is true, these test statistics have an asymptotic chi-square distribution with degrees of freedom (*df*) equal to the number of independent constraints on the cell probabilities. Assuming the aforementioned regularity conditions, *df* equals the row rank of $\mathbf{U}$.

Both ML and WLS share the same desirable asymptotic properties. The advantage of WLS is the ease of computation, in particular, closed form expressions for the estimators exist. However, the WLS method is very sensitive to sparseness in the data, while the ML method can be used for much smaller data sets (see Berkson, 1980, and the discussion of that paper). An alternative method is *Generalized Estimating Equations* (GEE) (Liang & Zeger, 1986). Here, unlike for the ML method, a predefined correlation structure has to be assumed, which may be arbitrary. For more details about this method, the reader is referred to Skrondal and Rabe-Hesketh (2004, Section 6.9), Molenberghs and Verbeke (2005), or Bergsma et al. (2009, Section 7.2.1).

3.4 Application to the Case Study

The models described in previous section were applied to the data from our case study. Results of several different LCMMs for pain, redness and irritability are presented in Table 4. For all three symptoms, the no difference model fit the data poorly, indicating that there could be a difference between groups, either constant across all 4 days or not.

For redness (resp. irritability), the hypothesis of a constant difference is acceptable as the estimated values from the constant difference model are not statistically significantly different from the observed ones ($p = 0.849$ and $p = 0.344$, respectively). The estimated value for the difference overall per day is 5.60% (resp. 4.21%) and it is statistically significant ($p = 0.001$ and $p = 0.022$, respectively).

For pain, even if the data does not show strong evidence that the differences vary among time (fit of the constant difference model is $p = 0.106$), we could consider to evaluate differences between group by day. Considering the varying difference model, we see that the active group has a statistically lower incidence of pain than the control group only for day 2 if not controlled for multiplicity ($p = 0.019$, $p_{B-H} = 0.076$). For the other days, no statistically significant difference was found. Hence, looking per day here, does not seems to bring much additional information to the evaluation of the difference between group. In addition, there does not seem to be a strong evidence for an overall difference between the groups (overall effect p-value is $p = 0.140$).

Comparing the results produced here and the ones obtained by standard analysis techniques, several remarks can be made. For pain, using the LCMM allows to see that although there might be some variability of the difference between groups across days, this variability is rather small, and the overall difference between the group also seems negligible. For redness, using the LCMM provides similar information as we would have found considering differences

Table 4: *Fit and effect estimates of different Linear Categorical Marginal Models with ML Estimation for the solicited symptoms observed during the 4-day post vaccination period.*

		Model Fit				Expected difference Control - Active		Model-based p-value	
Model	Symptoms	G^2	df	p-value	Day	Diff	se	Unadjusted	B-H
No difference	Pain	8.46	4	0.076	1, 2, 3 or 4	0	-	-	-
	Redness	8.82	4	0.066	1, 2, 3 or 4	0	-	-	-
	Irritability	11.46	4	0.022	1, 2, 3 or 4	0	-	-	-
Constant difference	Pain	6.16	3	0.106	1, 2, 3 or 4	1.95	1.32	0.140	-
	Redness	0.84	3	0.849	1, 2, 3 or 4	5.60	1.74	0.001	-
	Irritability	3.32	3	0.344	1, 2, 3 or 4	4.21	1.85	0.022	-
Varying difference	Pain	0.00	0	1.000	1	-0.54	2.46	0.827	0.827
			-		2	6.11	2.61	0.019	0.076
			-		3	4.13	2.17	0.057	0.171
			-		4	2.43	1.51	0.106	0.212
	Redness	0.00	0	1.000	1	4.41	2.55	0.083	0.083
			-		2	6.84	2.55	0.007	0.021
			-		3	6.39	2.57	0.013	0.026
			-		4	5.96	2.14	0.005	0.020
	Irritability	0.00	0	1.000	1	3.01	2.59	0.244	0.675
			-		2	1.18	2.59	0.648	0.675
			-		3	2.96	2.43	0.224	0.675
			-		4	5.23	2.14	0.015	0.060

overall: a statistically significant difference is found overall between the two groups, and this difference seems to be constant across time. Further this difference is very close using the 2 approaches: 5.7% for the differences and 5.6% for the LCMM. However, using the LCMM allows us to show that this is a correct strategy to describe the result overall as the *constant difference model* fits the data well. For irritability, using the LCMM, a statistically significant difference is found between the two groups (4.2%), and seems to be constant along time between the two groups (Constant difference model $p = 0.344$)). However, this difference is not really seen when considering the differences overall per subject (0.7%), or within each day independently except for day 4 where the difference is statistically significant even after correction for multiplicity. In this case, it might make more sense to conclude that there is a small but constant difference between the two groups, rather than no difference until day 3 and a difference at day 4.

We also fitted the same model using WLS and found little differences in p-values obtained compared to the ML procedure. For example, the only difference, in terms of unadjusted p-values, were for pain ($p = 0.138$ instead of 0.140 for the constant difference model), and irritability ($p = 0.024$ instead of 0.022 for the constant difference model, and 0.225 instead of 0.224 for the difference on day 3 for the varying difference model).

4 Use of Linear Categorical Marginal Models for the Analysis of Several Intensities

In Section 3, we have tested for differences between the two groups regarding the occurrence of symptoms at a specific intensity. In this section, we extend this approach by formulating models that take into account the full scale at which the intensities were measured, rather than the simple dichotomy used in the previous section. We will treat the scale of intensity as ordinal; subjects are considered to have answered three successive dichotomous questions: were there any symptoms? were they least of moderate intensity? were they of severe intensity? Thus, we treat the data as truly categorical rather than as a realization of an underlying continuum (see Hagenaars, 2010, for an extensive discussion of the history and philosophy of this approach). Therefore, in the present setup we not only have the dependencies over time which we already encountered in Section 3, but also, at each time point, we have several dependent dichotomous variables. Below, we show how marginal models can also be used to handle these more complex dependence relations.

4.1 Modeling the difference of multivariate marginal proportions

In Section 3 we have not taken the fact that symptoms were measured on a four point scale into account, that is we only compared proportions of having no symptoms ($S = 0$) versus having a symptom with a certain intensity ($S = 1, 2, 3$). To make use of the intensity information, we can also look at the differences in proportions of those who had $S = 0, 1$ vs. $S = 2, 3$ and those who had $S = 0, 1, 2$ vs $S = 3$. Introduce the new variable R, which takes the following values: $R = 1$ if $S \leq 1$, $R = 2$ if $S \leq 2$, and $R = 3$ if $S = 3$. Let

$$\pi_{r\ g\,t}^{R|GT}$$

be the proportion of respondents from group g given time t who experience the symptom with an intensity of at most r. Then we can define the two-way table of differences in proportions for active and control by

$$\delta_{t\,r}^{TR} = \pi_{r\ 1\,t}^{R|GT} - \pi_{r\ 2\,t}^{R|GT}$$

The saturated model, which does not impose any restrictions, is denoted as

$$\delta_{t\ r}^{TR} = \alpha + \beta_t^T + \beta_r^R + \beta_{t\ r}^{TR} \qquad \text{for all } r \text{ and } t$$

for some unknown α and β parameters. The β parameters can be identified by imposing restrictions such as $\sum_t \beta_t^T = 0$, $\sum_r \beta_r^R = 0$, and $\sum_t \beta_{t\ r}^{TR} = \sum_r \beta_{t\ r}^{TR} = 0$.
The most parsimonious model is that all differences are zero,

$$\delta_{t\ r}^{TR} = 0$$

We will refer to this model as the *no difference model*. It has here twelve independent restrictions on the probability distribution, so there are 12 degrees of freedom. The model asserting constant differences between active and control groups across time and intensities is

$$\delta_{t\ r}^{TR} = \alpha$$

We will refer to this model as the *constant difference model*. It has 11 degrees of freedom. The model asserting constant differences between active and control groups for each level of intensity R is

$$\delta_{t\ r}^{TR} = \alpha + \beta_r^R$$

We will refer to this model as the *constant difference model by intensity*. It has 9 degrees of freedom. The model asserting an independent effect of both time and intensity between active and control groups is

$$\delta_{t\ r}^{TR} = \alpha + \beta_r^R + \beta_t^T$$

We will refer to this model as the *independent intensity and time effect model*. For this model, the difference between active and control group for each level of intensity R is different, but the effect of time on this difference is the same for all intensities. It has 6 degrees of freedom.
Here also, the likelihood ratio test statistic G^2 and X^2 have an asymptotic chi-square distribution if the postulated model is true, with degrees of freedom equal to the number of independent constraints on the cell probabilities.
Several other models for modeling repeated ordered categorical variables have been developed but will not be discussed here (see, e.g. Molenberghs & Verbeke, 2005).

4.2 Application to the Case Study

Considering results by day from Table 3, significant differences not taking into account multiplicity, are noted for pain at day 1 (for intensity higher than 2 or intensity 3), at day 2 (for all 3 categories), for redness at days 2, 3 and 4 (for any intensity or intensity higher than 2), and for irritability at day 4 (for any intensity or intensity higher than 2). However, when correcting for multiplicity via the Bonferroni-Holm procedure, only 4 significant differences are found: for pain at day 1 (for intensity higher than 2 or intensity 3), at day 2 (for intensity higher than 2), and for redness at day 4 (for any intensity)
Analysis of this data by LCMMs may provide more insight in the data. Results of the fit of LCMMs taking into account the several intensities are shown in Table 5 (intensity 0 vs. 1, 2 and 3, or 0 and 1 vs. 2 and 3, or 0, 1 and 2 vs. 3 are simultaneously tested).
For irritability, as the no difference model yields a good fit to the data ($p = 0.221$), we find no evidence for a difference between the 2 groups in terms of any intensity at any time

Table 5: *Fit of different marginal models of the solicited symptoms of several intensities observed during the 4-day post vaccination period by either ML or WLS estimation procedures*

Estimation procedure	Symptom	No difference model			Constant difference model			Constant difference model by intensity			Independent intensity & time effect model		
ML		G^2	df	p-value	G^2	df	p-value	G^2	df	p-value	G^2	df	p-value
	Pain	33.8	12	<0.001	32.5	11	<0.001	29.9	9	<0.001	16.1	6	0.013
	Redness	22.5	12	0.032	21.7	11	0.027	7.11	9	0.625	1.8	6	0.938
	Irritability	15.3	12	0.221	15.4	11	0.166	10.75	9	0.293	7.6	6	0.269
WLS		W^2	df	p-value	W^2	df	p-value	W^2	df	p-value	W^2	df	p-value
	Pain	27.6	12	0.006	27.5	11	0.004	26.4	9	0.002	15.4	6	0.017
	Redness	22.7	12	0.031	20.2	11	0.043	7.34	9	0.602	1.7	6	0.941
	Irritability	14.3	12	0.283	14.1	11	0.225	10.17	9	0.337	7.6	6	0.267

Table 6: *Differences between the groups and the different intensities for the solicited symptom pain overall days post vaccination. Results of the constant difference model by intensity*

		Control - Active					
		0 vs 1,2,3		0,1 vs 2,3		0,1,2 vs 3	
Symptoms	Day	Diff (%)	p-value	Diff (%)	p-value	Diff (%)	p-value
Redness	1,2,3,4	4.93	0.003	2.18	0.009	-0.17	0.124
Pain	1,2,3,4	1.22	0.325	1.59	0.025	0.814	0.041

points. It can be observed that this result is not supported by the conclusions obtained in Section 3.4, in which a significant effect was found. This is may be attributed to the fact that extra comparisons are taken into account here that may mask a specific effect, which can also occur, e.g., when considering the effect of factors in ANOVA models.

For redness, the assumption of no difference cannot be sustained ($p = 0.032$) and it seems further that the difference between the groups is not the same for all intensities ($p = 0.027$), but when considering each intensity this difference seems constant along time ($p = 0.625$). In fact, it seems that for redness, the differences between the groups is much more pronounced for symptoms with low intensity than for symptoms with high intensity (see Table 6).

Contrary to the analyses of irritability or redness, the fit of the different models shown in Table 5 seems to indicate that the analysis of pain should be handled by intensity and time point. The independent intensity and time model does not fit the data very well either ($p = 0.013$) providing some evidence for an interaction between the two factors. From Table 3, we see indeed that for the highest intensities, a difference between the groups appears earlier and disappears also earlier, than when considering any intensity. Hence, for pain, summarizing the results across intensity and/or time may lead to substantial loss of information.

Using the LCMMs in this setting allows us to easily confirm that, there are no marked differences for irritability, that for redness differences seems to occur rather constantly per day for low to moderate intensity symptoms, and that for pain a more complex pattern has to be taken into account to analyze the differences as they are not only depending on the intensities, but also on the days and on their interaction.

The results obtained with ML and WLS estimation procedures can be compared. Results were similar when considering only one intensity comparison by solicited symptom, but are

somewhat more different though the differences remain quite small than when the intensity was not taken into account (see Table 5). Here, all p-values are different with a maximal difference of 0.06 ($p = 0.28$ for WLS versus $p = 0.22$ for ML for the no difference model for irritability). This is probably due to larger and more sparse tables used when considering all intensities at once. Different behaviors of the WLS and of the ML estimation procedure are indeed likely to occur on sparse tables. Note that here the estimation problem of the WLS procedure due to the sparseness of the table is more likely to happen, which makes the use of the ML procedure more attractive here.

5 Simulation studies

We now compare the performance of the marginal modeling approach with two classical approaches commonly used in clinical trials using a simulation study of 3 different types of models with different correlation between the days. The data are simulated as follows. We use $n = 300$ for both the control and the active groups and simulate from (i) the no difference model,

$$\delta_t^T = 0.00 \qquad \text{for } \; t = 1, \ldots, 4$$

(ii) the constant difference model

$$\delta_t^T = 0.025 \qquad \text{for } \; t = 1, \ldots, 4$$

(iii) the constant difference model

$$\delta_t^T = 0.05 \qquad \text{for } \; t = 1, \ldots, 4$$

(iv) the varying difference model

$$\delta_1^T = \delta_2^T = 0.05 \qquad \delta_3^T = \delta_4^T = 0 \tag{8}$$

(v) the varying difference model

$$\delta_1^T = \delta_2^T = 0.1 \qquad \delta_3^T = \delta_4^T = 0 \tag{9}$$

and (vi) the varying difference model

$$\delta_1^T = \delta_2^T = 0 \qquad \delta_3^T = \delta_4^T = 0.05 \quad . \tag{10}$$

In all cases, we set the proportions experiencing a symptom for the control group equal to 0.65, 0.50, 0.25 and 0.15.

To simulate correlated data we need a model for the nuisance parameters as well. For this we use two loglinear models. The first is a naive model that the occurrence of symptoms at the different time points are independent given treatment, denoted in loglinear model notation as

$$\{GS_1, GS_2, GS_3, GS_4\} \tag{11}$$

The second model adds conditional associations between symptom occurrence at the different time points, and is denoted in loglinear model terms as

$$\{GS_1, GS_2, GS_3, GS_4, S_1S_2, S_2S_3, S_3S_4\}$$

This model requires specification of an association parameter in marginal tables S_1S_2, S_2S_3, and S_3S_4. We take a constant difference in proportions,

$$\delta^S = \pi_{1\ 1}^{S_{t+1}|S_t} - \pi_{1\ 2}^{S_{t+1}|S_t} = 0.2$$

For the model (11), $\delta^S = 0$. For both approaches, and each case, 20,000 simulated samples are produced.

For each simulated sample, the first classical approach, called here the 'any day' approach, considers subjects who experienced the symptom at least once during any day of the follow up, and tests differences between groups using the Fisher's exact test. The second classical approach entails computing the Fisher's exact p-value for the hypotheses $H_{0t} : \delta_t^T = 0$ for all t with the Bonferroni-Holm (B-H) correction. For the marginal modeling approach, the strategy followed to detect an effect is similar to the one used in Section 3. First, the constant difference model is tested, the alternative being the varying difference model which is the saturated model. If the constant difference model is not rejected (model fit p-value ≥ 0.05), then the constant difference across all 4 days is tested for significance. If the constant difference model is rejected, the varying difference model is used and the estimated differences for each day are tested for significance using the B-H correction. The overall probability of obtaining significant differences between the groups by the marginal approach, denoted as *Pdiff* in the following, is thus obtained as the weighted average of the 2 types of difference tests previously estimated.

Table 7: *Simulation results: Type I error and power of the different models (H_1 being the constant difference model)*

		Classical approaches		LCMM approach			
Simulation model for $\delta_{t\ r}^{TR}$	Model for nuisance parameters	'any day'	B-H *Pdiff*	P(accept H_1)	H_1 accepted *Pdiff*	H_1 rejected *Pdiff*	Overall *Pdiff*
No difference ($\delta_t^T = 0$)	$\delta^S = 0$	0.032	0.038	0.948	0.050	0.314	0.063
	$\delta^S = .2$	0.058	0.038	0.948	0.051	0.261	0.063
Constant diff. ($\delta_t^T = 0.025$)	$\delta^S = 0$	0.323	0.196	0.951	0.306	0.703	0.326
	$\delta^S = .2$	0.535	0.304	0.947	0.401	0.765	0.420
Constant diff. ($\delta_t^T = 0.05$)	$\delta^S = 0$	0.831	0.587	0.948	0.837	0.948	0.844
	$\delta^S = .2$	0.967	0.835	0.938	0.946	0.989	0.948
Varying diff. (Eq. (8))	$\delta^S = 0$	0.325	0.273	0.818	0.190	0.726	0.287
	$\delta^S = .2$	0.488	0.354	0.800	0.245	0.789	0.354
Varying diff. (Eq. (9))	$\delta^S = 0$	0.823	0.810	0.371	0.584	0.950	0.815
	$\delta^S = .2$	0.929	0.903	0.346	0.717	0.982	0.890
Varying diff. (Eq. (10))	$\delta^S = 0$	0.322	0.452	0.799	0.478	0.803	0.544
	$\delta^S = .2$	0.619	0.659	0.762	0.629	0.911	0.696

The simulation results are shown in Table 7. We first see that the classical 'any day' approach can be either conservative or liberal according to the strength of the association between the different days. Then we see that the classical B-H approach is slightly conservative, with too small probabilities corresponding to the type I error, while the marginal modeling approach using LCMMs as applied with the above strategy is slightly liberal. However,

this shortcoming should not really penalize the marginal modeling approach in the context of analyses of solicited and unsolicited symptoms as the priority is often put on detecting differences rather than confirming them. In this respect, the simulations show that the marginal modeling approach is significantly more powerful than the B-H approach when simulations are done from the constant difference models. For simulations done from the varying difference models the two approaches yield comparable power. Compared to the 'any day' approach, the marginal modeling approach yields similar or better results (data corresponding to Equation (10)) in terms of power when there is no association between the days. Whenever there is some association between the days, the 'any day' approach can be more powerful than the marginal modeling one for the constant difference models, while for the varying effect models, which model is the most powerful seems to depend on the structure of the differences.

In addition, the marginal modeling approach is more flexible than the classical approaches, as a wider range of hypotheses can be tested and corresponding parameter estimates can be obtained. As shown in Table 7, using LCMMs the marginal modeling approach can also help to find the correct structure of the differences. For example, in the varying difference simulated data corresponding to Equation (9), in more than 60% of the cases it can be concluded that there is a significant effect and that this effect differs among days, and in the constant difference simulated data $\delta_t^T = 0.05$ it can be concluded in more than 80% of the cases that there is a significant effect and that this effect is constant among days, and only in less than 7% of the cases that there is a significant effect differing among days.

6 Conclusions

Without making unnecessary assumptions, LCMMs have been shown to take into account the dependencies that arise due to the repeated measurements of the solicited symptoms experienced after vaccination. They allow a better understanding of the relative safety profile of the several groups considered by testing correctly global hypotheses rather than looking at a list of p-values. In addition, interpretation is easy, the effect parameters derived from these models being expressed in terms of difference of percentages. The use of this method has the potential to improve the quality of global evaluation of the occurrence of solicited symptoms especially when several intensities, observation days, and/or doses are considered. Further, it is not limited to solicited symptoms and could also be applied to unsolicited symptoms, or even non safety data, provided that the occurrence of the event of interest is sufficiently frequent. Use of the ML estimation procedure will also bring some added value compared to the WLS procedure especially for complex models.

However, use of the ML estimation method presented above will not be straightforward in a clinical standard setting. Indeed, the estimation procedure, although available in an R package, is still not yet present in an ISO validated software. Furthermore, satisfactory strategies for the handling of missing data have not been yet been developed for this procedure. Hence, further developments in terms of handling missing data and accessibility of the method are still needed to be able to be used by a broader audience. Current work of the authors of this article is aimed at tackling these shortcomings.

Acknowledgements

The authors would like to thank the editor and the anonymous reviewer for their constructive comments which greatly helped to improve the quality of the final version of the manuscript.

The authors would also like to thank dr. Brigitte Cheuvart who has provided support and input along the entire project.

References

Bergsma, W. P. (1997). *Marginal models for categorical data.* Tilburg: Tilburg University Press.

Bergsma, W. P., Croon, M., & Hagenaars, J. A. P. (2009). *Marginal models for dependent, clustered and longitudinal categorical data.* NY: Springer.

Berkson, J. (1980). Minimum chi-square, not maximum likelihood! *The Annals of Statistics*, *8*(3), pp. 457-487.

Grizzle, J. E., Starmer, C. F., & Koch, G. G. (1969). Analysis of categorical data by linear models. *Biometrics*, *25*, 489-504.

Hagenaars, J. A. (2010). *Living in a categorical world.* Tilburg: Tilburg University, Valedictory Address.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Koch, G. G., Landis, J. R., Freeman, D. H., & Lehnen, R. G. (1977). A general methodology for the analysis of experiments with repeated measurements of categorical data. *Biometrics*, *33*, 133-158.

Kritzer, H. M. (1977). Analyzing measures of association derived from contingency tables. *Sociological Methods and Research*, *5*, 35-50.

Lang, J. B. (2004). Multinomial-Poisson homogeneous models for contingency tables. *Annals of Statistics*, *32*, 340-383.

Lang, J. B., & Agresti, A. (1994). Simultaneously modelling the joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association*, *89*, 625-632.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13-22.

Mehrotra, D. V., & Heyse, J. F. (2004). Use of the false discovery rate for evaluating clinical safety data. *Statistical Methods in Medical Research*, *13*, 227-238.

Molenberghs, G., & Verbeke, G. (2005). *Models for discrete longitudinal data.* New York: Springer-Verlag.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models.* Boca Raton, FL: Chapman and Hall.