Journal of Biopharmaceutical Statistics, 4(2), 233–265 (1994)

STATISTICAL EVALUATION OF RELATIONSHIPS BETWEEN ANALGESIC DOSE AND ORDERED RATINGS OF PAIN RELIEF OVER AN EIGHT-HOUR PERIOD

Stuart A. Gansky,[1] Gary G. Koch,[1] and Jerome Wilson[2]

[1]Department of Biostatistics
School of Public Health
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina 27599-7400

[2]Roberts Pharmaceutical Corporation
Meridian Center III
6 Industrial Way West
Eatontown, New Jersey 07724

Key words. Longitudinal design; Relative potency; Weighted least squares; Ordinal response; Multivariate response; Marginal models

Abstract

Statistical considerations are discussed for the application of alternative methods to a clinical trial involving repeated ordinal ratings and multiple dosage levels of active drugs. Analyses included summary measures traditionally employed in studies of acute pain: sum of pain intensity differences from baseline, total pain relief, and total pain half gone. Estimators and confidence intervals of relative potency are developed for univariate and multivariate situations, using weighted least squares analysis with mean response and variances from Taylor series linearizations. The estimates from these

methods are compared to those from traditional methods, such as ordinary least squares regression and Fieller's method for confidence intervals, as well as those from more recent developments, such as generalized estimating equations and sample survey data regression. A double-blind, two-center, randomized clinical trial of acute pain relief comparing placebo with two analgesics, each at two dosage levels, over an 8-hr period serves as an illustrative example for these techniques and comparisons.

Introduction

Biopharmaceutical research studies frequently involve ordered responses evaluated at several time points after a drug is taken [e.g., Bredfeldt *et al.* (1) and Mehlisch *et al.* (2)]. Since ordinal repeated measures data are not necessarily compatible with the assumptions of normality and homogeneous variances, methods utilizing normal theory may not be valid; thus categorical data analyses might be needed.

Often in drug trials, the experimental compound is compared to a standard agent (active control) and/or a placebo to evaluate the new drug's efficacy. In most trials of this nature, the primary goal is to evaluate the test medication's efficacy relative to the standard, as well as placebo. When investigators wish to compare the new drug's strength to the standard's, more than one dosage level for each is studied. Various statistical methods using marginal (population-averaged) models to assess the relationships between drug dose and repeated ordinal responses will be presented and illustrated.

The example for this paper included the following elements: multiple treatments and doses, repeated measures, ordinal response, multiple end-points, missing data, and dose-response via relative potency. We describe this clinical trial and the statistical considerations for the application of alternative methods to its analysis in the next section. A brief overview of methods commonly used for such situations, as well as techniques and their extensions that may be more appropriate, is provided along with results in the following section. Further details of the methods can be found in the Appendices and in the original articles. We provide discussion and our conclusions in the final sections.

Example Study Design

In a two-center, parallel-groups, randomized, double-blind study of analgesics, 258 patients with a dental pain condition provided written informed consent. Following oral surgery, these otherwise healthy males and females aged

	Treatment Group					
Center	P	SL	SH	TL	TH	Total
1	27	27	26	26	27	133
2	25	25	25	25	25	125
Both	52	52	51	51	52	258

Figure 1. Sample sizes for treatments at each center.

10–54 years old were observed until local surgical anesthesia subsided; those experiencing at least moderate pain on a none, slight/mild, moderate, or severe scale were given one oral administration of one of five randomly assigned, similar-appearing study medication capsules: placebo (P), low dose (200 mg) standard drug (SL), high dose (400 mg) standard drug (SH), low dose (50 mg) test drug (TL), or high dose (100 mg) test drug (TH). Prenumbered treatments were assigned sequentially in blocks of 20 to promote balance, separately for each center, resulting in a randomized block design. Subjects were split nearly equally between the treatment groups and two centers (Fig. 1).

Patients recorded ordinally scaled observations of pain intensity and pain relief in analgesia diaries at 0.5, 1, 2, 3, 4, 5, 6, 7, and 8 hr after taking the study medication. Data were not, in Rubin's terminology (3), missing completely at random (MCAR) since missingness was related to response; the missingness mechanism may have depended solely on the observed responses (i.e., missing at random) since patients with either very poor or very good responses tended to have incomplete diaries. Intent-to-treat (ITT) methods (4) were used to impute data; thus even when data were not MCAR, but due to rescue or remedication (R/R), they were replaced conservatively. Those patients not receiving adequate relief could have remedicated with their assigned study drug or taken a rescue medication to alleviate pain; such persons were considered treatment failures.

Missing Data and Imputation

The P and SL treatment groups had slightly higher, but not significantly so, percentages of R/R (chi-square test of homogeneity). Hence, the 15 patients with R/R, 14 from center 1, had their responses modified accordingly for the time of R/R and all subsequent time points, similar to Siegel *et al.* (5): pain intensity was set to the worse of baseline pain intensity and pain intensity at the time of R/R, pain relief was set to "none," and pain half relieved was set to "no."

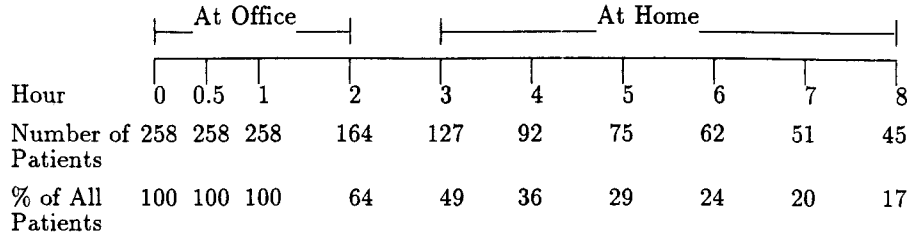

Figure 2. Number and percent of patients making entries at each time point.

Many patients did not complete their analgesia diaries for all 10 time points, as Figure 2 illustrates. Entries for the first three time points (through 2 hr postmedicating) were completed in the physicians' offices, but reports for the remaining time points were logged in the patients' homes. All 258 patients filled in their diaries for the first two time points postmedication, and 164 (64%) completed the 2 hr postmedicating entry. Upon leaving the offices, patients had poorer compliance with diary entries as only 127 (49%) provided entries at 3 hr postmedication. Due to extensive missing data, time point by time point analyses were performed only through 3 hr postmedication. Since summary measures and some analysis methods required completely observed data vectors, imputation was used for these missing times. In an ITT rationale (4,6), the last recorded entry was carried forward to the missing times.

Summary Measures

Typically in trials concerning acute pain [e.g., Mehlisch *et al.* (2), Powell *et al.* (6), Kempf *et al.* (7), and Schachtel *et al.* (8)], the efficacy measures include the sum of pain intensity differences (SPID), total pain relief (TOTPAR), and the total extent of half relieved pain (TOTGONE). These measures assume totally observed response vectors; thus their formation requires imputation for times with missing values. These measures are calculated as follows:

$$\text{SPID} = \sum_{k=1}^{\tau} \{(\text{pain intensity at } t_0 - \text{pain intensity at } t_k) \times (t_k - t_{k-1})\} \quad (1)$$

$$\text{TOTPAR} = \sum_{k=1}^{\tau} \{(\text{pain relief at } t_k) \times (t_k - t_{k-1})\} \quad (2)$$

$$\text{TOTGONE} = \sum_{k=1}^{\tau} \{(\text{pain half relieved at } t_k) \times (t_k - t_{k-1})\} \quad (3)$$

where k indexes the τ time points, t_k is the time in hours of the kth time point; $t_0 = 0$; pain is scored 0 (none), 1 (slight/mild), 2 (moderate), or 3 (severe); pain relief 0 (none), 1 (a little), 2 (some), 3 (a lot), or 4 (complete); and pain half relieved 0 (no) or 1 (yes). Thus all three are weighted sums across the τ time points, with the weights being the length of time (interval) between entries. Summaries through $\tau^* < \tau$ time points with the suffix τ^* are sometimes used; e.g., SPID3 is the sum of the first three times' pain intensity differences.

Relative Potency

Establishing the relationship between dose and the response produced for each drug is frequently important to estimate the comparably effective doses of the drugs (9,10). Relative potency is the ratio of doses that achieve the same effect; in other words, it is the ratio of the estimated amount (dose) of one drug needed to produce the same level of response as a specified dosage of the other drug. Studies often seek to find the amount of standard medication that achieves the same effect as a certain test drug dose.

Analysis and Results

First, nonparametric, minimal assumption tests of treatment efficacy were performed using Mantel-Haenszel tests (11,12) to provide confirmatory inferences for standard and test drugs being better than placebo. Then marginal (population-averaged) model approaches were used to assess treatment differences and to develop relative potency estimates and confidence intervals (CIs) for test drug relative to standard. Ordinary least squares (OLS) analyses assuming homogeneous variances and normal model errors were performed, as is traditional in studies of this nature (7,10). For relative potency, Taylor series linearizations (TSL) were used to produce CIs (9,13,14). The proportional odds assumption of cumulative logit models (15) was inappropriate, so weighted least squares (WLS) methods (12,16,17), which allow for general covariance structures, were utilized with mean response using TSL to obtain CIs for both univariate and multivariate data structures. In addition, Fieller's formula (18) was used with WLS to verify relative potency CIs. WLS methodology was used with the repeated imputed ordered pain relief in conjunction with TSL to provide estimates and CIs of relative potency. Generalized estimating equation (GEE) (19,20) and survey data regression (SDR) methods (21,22) were used to evaluate relative potency estimates and CIs for repeated responses in a way that allowed greater scope for inclusion of explanatory variables in relationships.

Extended Mantel-Haenszel Tests

Results of the extended Mantel-Haenszel tests, outlined in Appendix I, are exhibited in Table 1. In these inferential nonparametric efficacy tests, responses for each active drug dose were compared to placebo, while adjusting for center and baseline pain intensity, with a statistic directed at differences between mean scores. A hierarchical closed procedure (23) adjusted for multiple comparisons for each drug versus placebo for each measure; i.e. the procedure was closed in that if the composite null hypotheses of equivalence to placebo could not be rejected for either drug for a particular measure, the pre-specified order of comparisons was halted. First the standard (i.e., active control) was tested versus the placebo; given significance (i.e. rejection of the composite null hypothesis that both doses were equal to placebo) in the first stage, the test drug was compared to the placebo (in terms of the composite hypothesis that both doses were equal to placebo). To adjust for multiplicity of doses in the composite null hypothesis versus placebo for each of the standard and test drugs, a Bonferroni-Holm method (which Holm (24) called a sequentially rejective Bonferroni procedure) was applied by testing the dose with smaller p-value at $\alpha = .05/2 = .025$ and the one with a higher p-value at $\alpha = .05$, given significance at $\alpha = .025$ for the other dose; only rejection with the smaller p-value was required in order to support the conclusion that at least one dose was different from placebo. For the previously described testing procedure, TOTPAR was evaluated first as the primary efficacy measure; then SPID, TOTGONE, and the pain relief ratings at 3 hr, 2 hr, 1 hr and 0.5 hr were evaluated. The summary measures through the first three hours were considered for primary inference, due to the extensive missing data described earlier; the summary measures across all time points were used supportively.

While this approach is reasonable, it does not fully control the familywise error rate in the strongest or strictest sense. An alternative multiple comparison procedure (which would control the familywise error rate at 0.05) begins testing with the dose of standard versus placebo with the smaller p-value at $\alpha = .025$, and given its significance, continues with assessing the dose of test versus placebo with the smaller p-value at $\alpha = .025$; given significance of the former two, the smaller of the p-values of the two remaining comparisons would be assessed at $\alpha = .025$, while the final one would be assessed at $\alpha = .05$. For the example discussed in this paper, this alternative method supported the same conclusions as the one involving the Bonferroni-Holm method; however the alternative is more awkward to specify and may not identify all doses which differ from placebo as well. Analyses were performed with modified ridits and midranks to break ties using SAS's FREQ procedure (25).

Table 1. *p*-Values from Mantel-Haenszel Statistics Adjusting for Center and Baseline Severity[a]

	Drug			
	Active control (standard)		Test	
Response[d]	Low dose	High dose	Low dose	High dose
TOTPAR3	.001**[c]	<.001**[b]	.001**[c]	<.001**[b]
TOTPAR	.002**[c]	<.001**[b]	.001**[c]	<.001**[b]
SPID3	<.001**[c]	<.001**[b]	.021*[c]	<.001**[b]
SPID	.001**[c]	<.001**[b]	.021*[c]	<.001**[b]
TOTGONE3	.001**[c]	<.001**[b]	.002**[c]	<.001**[b]
TOTGONE	.001**[c]	<.001**[b]	.002**[c]	<.001**[b]
Relief 3 hr	.020*[c]	.009*[b]	.003**[c]	.003**[b]
Relief 2 hr	<.001**[c]	<.001**[b]	<.001**[c]	<.001**[b]
Relief 1 hr	<.001**[c]	<.001**[b]	.001**[c]	<.001**[b]
Relief 0.5 hr[e]	.024*[b]	.069	.323	.012*[b]

[a]Testing equivalence of active drugs with placebo using a hierarchical closed procedure with Bonferroni-Holm multiple dose adjustment.
[b]The more significant dose is tested at $\alpha = .05/2 = .025$ (or 0.005) with the conventions: *[b] if $0.005 < p \leq 0.025$ (or **[b] if $p \leq 0.005$).
[c]The less significant dose is tested at $\alpha = .05$ (or 0.01) with the conventions: *[c] if $0.01 < p \leq 0.05$ (or **[c] if $p \leq 0.01$).
[d]Summary variables with the suffix "3" are only through the first three time points.
[e]For pain relief at 0.5 hr, $p < .025$ for standard's low dose demonstrates standard's statistical significance and justifies evaluating test; $p < .025$ for test's high dose demonstrates test's statistical significance.

Differences between placebo and the active drug doses were clearly significant ($p < .01$) for all summary measures through both the first three and all nine time points. Pain relief differences from placebo were clearly significant for both drugs at 1 and 2 hr postmedication and for the test drug at 3 hr postmedication. Differences were moderately significant ($p < .05$) for the standard drug's pain relief at 3 hr postmedication, as well as for the standard drug's low dose and the test drug's high dose at 0.5 hr postmedication.

Mantel-Haenszel procedures are minimal assumption methods (11,12) in that they only assume randomization of patients to treatment groups and they do not require random sampling from a general population or a particular underlying distribution. Moreover, Mantel-Haenszel procedures only require similar direction with respect to homogeneity of effect across strata; varying directions of effect only reduce efficiency and weaken power. Mantel-Haenszel tests are appropriate with small within-stratum sample sizes and moderate sample sizes across strata. Statistically, conclusions apply only to the study sample; nonstatistical arguments about the subjects representing a target population must be used for generalizability.

Inferential comparisons between doses of standard and test drug were not made since the doses of these treatments were specified in the study design to have similar ranges of expected response so as to facilitate estimation of relative potency. In other words, relative potency estimation sufficiently quantifies comparisons between standard and test drug as discussed subsequently; thus direct comparisons between doses of standard and test were not necessary.

Ordinary Least Squares Regression

As in similar studies, analyses that assume normality and homogeneous variances were employed using the summary measures (through all nine time points) along with imputed ordered pain relief through 3 hr postmedication, using a weighted mean of 0.5 and 1 hr relief (TOTPAR2). Well-known OLS regression analyses using the placebo group as the reference were performed. Models included an indicator variable for each of the two drugs, a natural log transform of dose level, and a drug by $\log_e$(dose) interaction to parameterize the treatment groups, along with center and baseline pain intensity indicators and an intercept term. The intercept was a reference for placebo patients in center 1 with moderate baseline pain. These models were applied using SAS's PROC GLM (26).

For the summary measures and imputed pain relief through the first 3 hr as univariate responses, parallelism of the $\log_e$(dose)-response relationships for each drug was supported, as well as a tendency for baseline to be unimportant; SPID was the only measure for which importance of baseline pain intensity was suggested. Since SPID is the sum of pain differences from baseline, it is already adjusted for baseline pain intensity to some extent. Thus further consideration was only for models including center and parallel dose relationships for standard and test drugs. Results from analyses with these models indicated significant differences existed between the two centers' responses. Descriptive statistics for each treatment group and center are presented in Tables 2 and 3.

Univariate relative potency estimates and 95% CIs, obtained using OLS and SAS's Data Step (27) as explained in Appendix I, are shown in Table 4. Calculations, assuming parallelism (justified above), were performed adjusting for center. Relative potency estimates ranged between 2.84 and 4.31 depending on the response used; the corresponding CIs were reasonably compatible with one another. A noteworthy consideration for these relative potency estimates and their CIs is that the corresponding $\log_e$(dose) effects be reasonably nonnull, and preferably significant, since relative potency is not a meaningful concept when a dose-response relationship does not exist;

Table 2. Descriptive Statistics of Summary Measures for each Treatment Group and Center (N = 258)

		SPID		TOTPAR		TOTGONE		(SPID, TOTPAR)
Treatment	Center	Mean	SE	Mean	SE	Mean	SE	Covariance
P	1	−3.89	0.86	0.76	0.45	0.04	0.04	0.21
	2	0.84	1.45	10.14	1.91	2.90	0.75	2.38
SL	1	0.76	1.31	6.81	1.92	1.76	0.58	2.12
	2	6.52	1.43	17.54	2.11	5.26	0.68	2.72
SH	1	2.04	1.29	11.60	2.01	3.52	0.66	2.30
	2	7.92	1.10	18.76	1.82	5.40	0.65	1.54
TL	1	−1.10	1.36	6.00	1.60	1.81	0.57	1.91
	2	5.54	1.01	16.84	1.85	4.84	0.72	1.65
TH	1	2.67	1.41	10.17	1.91	3.65	0.66	2.33
	2	6.40	0.95	19.04	1.31	6.16	0.49	0.94

moreover when the $\log_e$(dose) effect is clearly nonsignificant, CIs for relative potency may be nonestimable. In Table 4, $\log_e$(dose) was significant ($p < .05$) for all measures except the weighted mean relief for 0.5 and 1 hr ($p = .1356$), but even for this measure a relatively reasonable, albeit wider, CI was available. However, a general issue of interest is that these CIs may be too narrow (anticonservative) since OLS' homogeneous variance assumption appears unrealistic.

Table 3. Descriptive Statistics of Imputed Pain Relief at the First Three Hours for Each Treatment Group and Center (N = 258)

		Pain relief hour					
		1[a]		2		3	
Treatment	Center	Mean	SE	Mean	SE	Mean	SE
P	1	0.17	0.06	0.15	0.11	0.07	0.05
	2	1.18	0.21	1.28	0.24	1.28	0.24
SL	1	0.74	0.19	1.04	0.27	1.07	0.29
	2	2.02	0.27	2.48	0.28	2.32	0.27
SH	1	0.90	0.19	1.65	0.29	1.62	0.30
	2	1.92	0.20	2.48	0.23	2.52	0.23
TL	1	0.73	0.18	0.96	0.26	0.96	0.28
	2	1.64	0.20	2.44	0.25	2.16	0.24
TH	1	1.06	0.19	1.59	0.30	1.59	0.30
	2	2.08	0.19	2.84	0.15	2.60	0.17

[a]Weighted average of the 0.5 and 1 hr measures.

Table 4. Univariate Relative Potency Estimates Adjusting for Center Using Ordinary Least Squares ($N = 258$)

Measure	Point est.	95% CI (TSL) Lower	95% CI (TSL) Upper	p-value of $\log_e$(dose)
SPID	2.84	2.26	3.80	.0380*
TOTPAR	3.45	3.06	3.95	.0132*
TOTGONE	4.30	3.99	4.65	.0036**
Relief 1 hr[a,b]	3.79	2.86	5.42	.1356
Relief 2 hr[a]	4.31	3.85	4.87	.0193*
Relief 3 hr[a]	3.70	3.32	4.15	.0112*

[a]Imputed pain relief.
[b]Weighted mean of pain relief at 0.5 and 1 hr.
*$.01 < p \leq .05$.
**$p \leq .01$.

Univariate Weighted Least Squares with Mean Response

Analyses employing WLS used mean response for the summary measures and pain relief. Since cumulative logit models were not appropriate according to the proportional odds assumption (score tests from the LOGISTIC procedure of SAS for imputed pain relief at each of the first four time points: $\chi^2 \geq 26$, d.f. $= 15$, $p \leq .04$), WLS with mean response was utilized. For the ordinal outcomes, the perspective of equally distant categories—reflecting their labeling—supports uniformly spaced integer scores for which mean scores can be used (17, p. 264). [Additional rationales for these scores are given in Koch and Edwards (11), Graubard and Korn (28), and Koch *et al.* (29).] Since the summary measures are weighted sums across the time points, mean response can be used for them, too. More specifically, TOTPAR can be calculated using the sum of binary indicators for each cut point for pain relief at each time point and summing over cut points and time, lending further support for integer scores.

The models in these analyses had similar predictors as OLS: an intercept term, two indicators for drug, $\log_e$(dose), drug by $\log_e$(dose) interaction, and an indicator for center. Due to the small percentage of patients with severe baseline pain intensity, baseline pain could not be fit as a separate effect. However, this should not arouse undue concern considering the relative unimpatance of baseline pain intensity in the OLS analyses. The data from the 10 center $\times$ treatment groups were considered in these WLS analyses to be conceptually similar to a stratified simple random sample of the 10 populations they represent; sample sizes were sufficient for the mean responses to

Table 5. Univariate Relative Potency Estimates Adjusting for Center Using Weighted Least Squares with Mean Response ($N = 258$)

Measure	Point est.	TSL 95% CI		Fieller's Formula 95% CI		p-value of $\log_e$(dose)
		Lower	Upper	Lower	Upper	
SPID	2.64	1.50	10.25	Nonestimable		.0535
TOTPAR	3.44	2.18	7.08	1.01	7.20	.0145*
TOTGONE	4.29	2.89	7.38	2.54	8.83	.0030**
Relief 1 hr[a,b]	4.15	2.25	12.15	Nonestimable		.0901
Relief 2 hr[a]	4.74	2.90	9.75	2.43	24.92	.0178*
Relief 3 hr[a]	3.72	2.35	7.52	1.32	9.43	.0166*

[a]Imputed pain relief.
[b]Weighted mean of pain relief at 0.5 and 1 hr.
*$.01 < p \leq .05$.
**$p \leq .01$.

be approximately multivariate normal. The CATMOD procedure of SAS (25) was used to obtain estimates of the model parameters and their associated covariance matrix. Relative potency estimates and CIs were formed using the estimates, obtained from PROC CATMOD, along with a TSL method in the SAS/IML software (30). The test of the $\log_e$(dose) effect was assessed with a Wald chi-square statistic. Fieller's formula was employed in the SAS/IML environment with PROC CATMOD's parameter and covariance estimates to calculate alternative CIs.

Parallelism of the two dose-response lines was not rejected for any response, but the two centers were significantly different for all responses. After the interaction term was dropped from the models, as supported above, univariate relative potency estimates and 95% CIs were generated, as described in Appendix II, with both TSL and Fieller's methods. These results are shown in Table 5. In situations with insufficient evidence of a $\log_e$(dose) effect, no dose-response relationships could be concluded with this method; hence CIs were unobtainable in these cases. Estimates of relative potency, not controlling for baseline pain, ranged from 2.64 to 4.74, agreeing rather well with the OLS results. WLS allows for heterogeneous variances, as opposed to OLS, which assumes homoscedasticity; so the WLS-based CIs, which are wider, are probably more reasonable. The CIs calculated with Fieller's formula, though slightly wider, agree fairly well with the TSL intervals; Fieller's formula could not provide bounds for marginally significant $\log_e$(dose)-response relationships.

Table 6. Bivariate Relative Potency Estimates Adjusting for Center Using Weighted Least Squares with Mean Response (N = 258)

Measure	Point est.	TSL 95% CI Lower	TSL 95% CI Upper	Fieller's Formula 95% CI Lower	Fieller's Formula 95% CI Upper	p-value of $\log_e$(dose)
SPID	2.66	1.49	10.94	Nonestimable		.0596
TOTPAR	3.31	2.06	7.22	0.69	6.78	.0173*
Average[a]	2.95	1.74	8.24	—	—	—

[a]Relative potency homogeneity test had p = 0.5144.
*$.01 < p \leq .05$

Bivariate WLS with Mean Response for Summary Measures

Estimates and 95% CIs of relative potency were computed using the output from PROC CATMOD in the SAS/IML software with a TSL method, as well as with Fieller's method for comparison, similar to the univariate case. The two primary efficacy measures, SPID and TOTPAR, were analyzed bivariately, adjusting for center. Relative potency estimates for each measure $\hat{\rho}_m$ were tested for homogeneity with a 1 d.f. χ^2 test:

$$(\hat{R}_1 - \hat{R}_2)^2/\{\mathrm{Var}(\hat{R}_1) - 2\mathrm{Cov}(\hat{R}_1, \hat{R}_2) + \mathrm{Var}(\hat{R}_2)\},$$

where $\hat{R}_m = \log_e\log_e\hat{\rho}_m$ with m = 1 for SPID or 2 for TOTPAR. After verification of homogeneity, estimates for each measure were averaged for an overall bivariate estimate and TSL-based variance for determination of a 95% CI of relative potency:

$$\hat{\bar{R}} = \tfrac{1}{2}(\hat{R}_1 + \hat{R}_2) \text{ and } \mathrm{Var}(\hat{\bar{R}}) = \tfrac{1}{4}\mathrm{Var}(\hat{R}_1) + \tfrac{1}{2}\mathrm{Cov}(\hat{R}_1, \hat{R}_2) + \tfrac{1}{4}\mathrm{Var}(\hat{R}_2).$$

Further details are provided in Appendix II. Results shown in Table 6 agree quite well with the univariate WLS results. The findings using Fieller's formula also agree with the univariate ones: the Fieller's CIs are a bit wider. The findings for the "average relative potency" across SPID and TOTPAR are appealing in that consideration of such a single measure is sufficient to encompass both the primary response measures.

Repeated Measures with Ordinal Response

Imputed pain relief scores through the first 3 hr (with TOTPAR2 as the first hour's response) were used in analyses accounting for repeated measures through mean response and WLS, as described in Appendix III. The center effect

Table 7. Descriptive Statistics of Imputed Pain Relief through the First Three Hours for Each Treatment Group ($N = 258$)

Treatment	Time	Mean	Time covariance matrix		
			1[a]	2	3
P	1[a]	0.654	0.017	0.018	0.018
	2	0.692		0.023	0.022
	3	0.654			0.021
SL	1[a]	1.356	0.034	0.036	0.033
	2	1.731		0.049	0.045
	3	1.673			0.046
SH	1[a]	1.402	0.025	0.020	0.019
	2	2.059		0.037	0.035
	3	2.059			0.041
TL	1[a]	1.177	0.022	0.027	0.024
	2	1.686		0.044	0.040
	3	1.549			0.042
TH	1[a]	1.548	0.024	0.020	0.017
	2	2.192		0.036	0.032
	3	2.077			0.036

[a]Weighted average of the 0.5 and 1 hr measures.

could not be fit due to a singularity in the covariance matrix for the separate estimates from the two centers; thus results presented here do not adjust for center. Center could have been adjusted in a reweighting scheme (17) or with smoothed covariance matrices (31), but these methods were not used in this paper. Models were fit using PROC CATMOD in SAS; relative potency estimates and CIs were constructed in SAS/IML with a TSL method. Model effects were tested with Wald chi-square statistics.

Table 7 provides pain relief means and covariance matrices through the first 3 hr of each treatment group. Results from an initial model with an intercept, an indicator for active drug, an indicator for test drug (i.e., the difference between test and standard), a natural log transform of dose, two indicators for the second and third hours, and interactions with time are shown in Table 8. The nonsignificant residual goodness of fit (GOF) chi-square statistic supported the model's fit. Factors not varying over time can be simplified to a common effect; e.g., homogeneous dose-response over time is modeled with a common slope. Thus $\log_e$(dose) and (test − standard) interactions with time were removed to simplify the model. Moreover, the time effect was reduced to a 1 d.f. indicator of the second or third hour; its interaction with active drug also was simplified to a similar 1 d.f. indicator, since these effects do not seem to vary over those two times. This reduced model, shown in Table 9, had adequate fit as assessed by the residual GOF

Table 8. Analysis of Variance of Repeated Measures for Imputed Pain Relief with Time Interaction Effects Without Adjusting for Center Using Weighted Least Squares with Mean Response (N = 258)

Parameter	d.f.	χ^2	p-value
Intercept	1	25.82	<.0001**
Active drug	1	0.78	.3762
Test-standard drug	1	1.43	.2316
Log_e(dose)	1	2.11	.1467
Time	2	1.02	.6006
Active drug × time	2	1.67	.4344
Test-standard drug × time	2	3.40	.1831
Log_e(dose) × time	2	2.88	.2365
Residual	3	1.32	.7252

**$p \leq .01$.

statistic, which had a value of 10.99, 9 d.f. and a p-value of .2764. The marginally significant $\log_e$(dose) effect suggests the existence of a dose-response relationship for this type of analysis.

For the reduced model, estimates of the parameter vector and the covariance matrix were used with a TSL method to develop an estimate and 95% CI of relative potency as shown in Table 10.

This model was also fit, for comparison purposes, to imputed pain relief scores using GEE and survey data regression methods, as detailed in Appendix III. Although GEE methods are quite flexible and can be extended to ordinal models, they have yet to be fully developed and software is not readily available. The comment of Kenward and Jones (32) regarding Clayton's suggestion (33) of analyzing ordered response as a series of correlated dichotomies to model proportional odds does not help in this

Table 9. Reduced Analysis of Variance of Repeated Measures for Imputed Pain Relief Using Weighted Least Squares with Mean Response (N = 258)

Parameter	d.f.	χ^2	p-value
Intercept	1	25.36	<.0001**
Active drug	1	1.42	.2333
Test-standard drug	1	2.06	.1514
Log_e(dose)	1	3.18	.0747
2 or 3 hour	1	0.00	1.000
Active drug × 2 or 3 hour	1	41.43	<.0001**
Residual	9	10.99	.2764

**$p \leq .01$.

Table 10. Multivariate Relative Potency Estimates for Imputed Pain Relief Allowing for Time Effects (N = 258)

Method	Point est.	TSL 95% CI		p-value of $\log_e$(dose)
		Lower	Upper	
WLS	3.62	2.03	10.42	.075
GEE	3.94	2.32	9.32	.042*
SDR	3.92	2.31	9.32	.043*
GEE[a]	3.82	2.39	7.88	.020*
SDR[a]	3.82	2.38	7.90	.021*

[a]Adjusted for baseline pain severity and center.
*$.01 < p \leq .05$.

case since the proportional odds assumption does not hold. Thus the GEE linear model for pain relief ratings, fit with the SPIDA software package (34), assumed an identity link with normal (constant) variance. Although the specified working correlation matrix had an independence structure, other correlation structures produced very similar results. (Misspecification of the working correlation matrix in GEE may affect the parameter estimates' efficiency, but not their consistency.) Some of GEE's flexibility was demonstrated by considering an additional model that adjusts for severe baseline pain and center.

Regularly timed repeated measures data from a clinical trial can be interpreted as similar to a single-stage cluster sample with subject as a primary sampling unit (PSU) and each time point as a within-cluster observation. Using SUDAAN software (35), survey data linear regression was performed for this example as a simple random sample with replacement of PSUs (subjects) and sampling weights of ones. SDR uses TSL approximations for covariance matrix estimations. SDR was used to fit the same models as GEE, both unadjusted and adjusted for baseline pain severity and center.

GEE and SDR produced nearly identical relative potency estimates and CIs, as other comparisons of SDR and GEE with an independence correlation matrix have shown; in variance computations GEE uses the number of subjects in the denominator and SDR uses the number of subjects minus one (36). The unadjusted and adjusted analyses produced very similar findings, though the adjusted CIs were slightly smaller. These results agree reasonably well with the WLS results; moreover, they are similar to WLS TSL-based results for SPID and TOTPAR (univariately and averaged bivariately), as well as for hourly pain relief ratings.

Discussion

In general, weighted least squares methods perform well with sufficient sample sizes and a limited number of categorical predictors. For these data consisting of repeated ordinal measures and weighted summary measures in univariate and bivariate contexts, WLS analyses produce appropriate results. Relative potency estimates are attainable in most situations (as long as the slope and the difference between intercepts are both positive), and their confidence intervals using TSL approximations appear reasonable compared to other methods; moreover, test statistics are provided for assessing GOF for the models applied.

WLS methods are particularly useful in situations where many other methods, such as proportional odds models (15), analyses of variance, repeated measures analyses of variance, multivariate normal bioassay methods (37–42), and multivariate analyses of variance, may not have their assumptions met. Furthermore, these WLS methods can be employed with standard statistical computing packages and matrix language programming software.

It may not always be appropriate to use equally spaced integer scales for the ordered responses. Some researchers (43) have identified problems with such scales, particularly in parametric analyses. When the strength between levels of response is unknown, alternative scores could be used with these methods instead of integer scores. Rapid pain relief in some situations may decrease exponentially, not linearly. This area should be explored in future investigations. Visual analog scales, which are gaining popularity, also could be used with these methods when other methods' assumptions are not met. However, Graubard and Korn (28) have recommended using meaningful integer scores (which they call column scores), developed *a priori*; in cases either without reasonable natural integer scores or without uniformly distributed column marginals, they suggest equally spaced integer scores.

Other researchers have explored the analysis of ordered pain ratings differently. For example, Cox and Chuang (44) and Chuang and Agresti (45) reduce a similar trial's data from longitudinal measurements to a univariate summary assessment of efficacy. Cox and Chuang explored several types of logits, drawing similar conclusions from them. Chuang and Agresti analyzed one type of logit with equally spaced, monotone scores for easier interpretation. SPID and TOTPAR could be analyzed using these methods as long as their assumptions (e.g., proportional odds) seemed reasonable.

In the repeated measures analyses, Kenward and Jones (32) noted that as within-subject correlation increases, WLS's sample size requirement increases, too. Moreover, GEE and survey data regression methods have ad-

vantages of fitting more predictors than WLS, as well as continuous covariates, although computing packages incorporating them may not be as widely available as those with WLS capabilities. The GEE models in this example used a normal (constant) variance, whereas WLS, which allows general covariances, may enable somewhat better efficiency when applicable. Also, Zeger and Liang (19) warn variances are robust only for data with a "diminishing fraction of missing" values or that is missing completely at random (MCAR) (3). Although GEE methods can be flexible, the software available does not fully take advantage of the method's capabilities. The SUDAAN package provides user-friendly, flexible software for survey data regression methods. SDR also assumes data are MCAR. Although SDR methods validly account for within-subject correlations, they may not be the most efficient, since the parameters are not estimated using the correlation structure (36).

Conclusions

Extended Mantel-Haenszel tests with hierarchical closed procedures are straightforward to plan and perform. In this case they clearly indicate both doses of both active drugs are better than placebo.

Moreover, relative potency, a useful concept for comparing doses of different agents, can be estimated via several methods and models. Relative potency estimates can be combined across multiple end-points (responses or times) to produce a single estimate.

Survey data regression via SUDAAN software is a versatile method for modeling complicated data structures, such as clustered or correlated measures. In addition, SUDAAN can model irregularly timed data and missing values. The software performs computationally efficient (36), friendly survey data regression with straightforward syntax as in SAS, and can produce vector and matrix estimates to be utilized in other packages such as SAS/IML. Thus SUDAAN readily provides relative potency estimates. GEE methods have these same capabilities, but user-friendly software for large datasets is not readily available.

Weighted least squares methods involve a tradeoff. Basic WLS analyses are easily attained and provide measures of GOF; however, WLS analyses limit the models available and require moderate sample sizes. Extensions to WLS methods, such as smoothed covariance or weighted averages over strata, increase flexibility but also complexity and are not very user-friendly.

Acknowledgments

The activity of Jerome Wilson in the research this paper reports took place at: Warner-Lambert Company, CP R&D Division, 170 Tabor Road, Morris Plains, NJ 07950-2598. Partial support for this research was provided to the

Department of Biostatistics at the University of North Carolina by the Warner-Lambert Company and by National Institute of Environmental Health Sciences Training Grant 5T32-ES07018-15. The authors thank Professor Lloyd Edwards, Glenn M. Davies, and John S. Preisser for their comments on earlier versions of this paper, as well as the reviewers and editor for their helpful suggestions. Any ambiguities or errors that remain are our own.

Appendices

In general, studies with repeated measures gather data from one or more groups during at least two different times. These data consist of response(s) and possibly covariates that may be predictive of the response(s).

Generally, the vector of r responses for the ith subject from the jth group at the kth time point is denoted as $\underset{\sim}{\mathbf{y}}_{ijk} = [y_{ijk1}, y_{ijk2}, \ldots, y_{ijkr}]'$, where $i = 1, 2, \ldots, n_j$ indexes the subjects in the jth group, $j = 1, 2, \ldots, s$ indexes the number of groups, and $k = 1, 2, \ldots, \tau$ indexes the time points. Similarly, the vector of c covariates measured for the ith subject from the jth group at the kth time point is denoted $\underset{\sim}{\mathbf{x}}_{ijk} = [x_{ijk1}, x_{ijk2}, \ldots, x_{ijkc}]'$.

In the example presented in this paper, the number of groups, s, is usually equal to 10 (five treatments at two centers). The number of time points, τ, is equal to nine, although many analyses described use τ^* equal to three or four, due to extensive missing values. Moreover, some analyses use the weighted mean of the first two times (0.5 hr and 1 hr) as the 1 hr response for pain relief; this corresponds to TOTPAR2 in previous notation. The number of subjects in each group, n_j, ranges from 25 to 27. The number of responses, r, is equal to three, and the response vector, $\underset{\sim}{\mathbf{y}}_{ijk}$, is comprised of the pain intensity, the pain relief and whether the pain was half relieved. The covariate vector is actually a scalar, x_{ij}, composed merely of baseline pain intensity.

Appendix I. Traditional Methods

The $\underset{\sim}{\mathbf{y}}_{ijk}$ vectors can be rewritten as $\underset{\sim}{\mathbf{y}}_{kl}$ vectors of the lth response at the kth time point with dimension $(N \times 1)$, so

$$\underset{\sim}{\mathbf{y}}_{kl} = [y_{11kl}, y_{21kl}, \ldots, y_{n_1 1kl}, y_{12kl}, y_{22kl}, \ldots, y_{n_2 2kl}, \ldots, y_{1skl}, y_{2skl}, \ldots, y_{n_s skl}]' \quad \text{(A.1)}$$

where $N = \sum_{j=1}^{s} n_j$ is the total sample size for the study.

		Response Category				
		1	2	⋯	v	
	1	n_{h11}	n_{h12}	⋯	n_{h1v}	n_{h1+}
Factor	2	n_{h21}	n_{h22}	⋯	n_{h2v}	n_{h2+}
Category	⋮	⋮	⋮		⋮	⋮
	u	n_{hu1}	n_{hu2}	⋯	n_{huv}	n_{hu+}
		n_{h+1}	n_{h+2}	⋯	n_{h+v}	n_{h++}

Figure A1. Cross-tabulation of the hth stratum.

Extended Mantel-Haenszel Test

For randomized trials, minimal assumption methods (11,12) can be used as primary significance tests of efficacy, but cannot estimate those effects with confidence intervals. Cross-tabulations for assessing the relationship of a u-level factor with a v-level ordinal response, while adjusting across the q strata, formed by other factors, can be constructed. The adjustment factors form q strata, where q is equal to the product of the number of levels of each factor; e.g., with two dichotomous factors, such as center and baseline severity, four strata are formed. Figure A1 depicts the cross-tabulation of the hth such stratum, where n_{hpg} is the number of subjects in the hth stratum having the pth level of the factor for comparison and the gth level of the response. (A subscripted "+" indicates a summation over that particular term, e.g., n_{h++} is the total number of subjects in the hth stratum.) Moreover, under the null hypothesis, H_0, of no factor difference in response for each patient, a product multivariate hypergeometric distribution applies:

$$\Pr\{n_{hpg} \mid H_0\} = \prod_h \frac{\prod_{p=1}^{u} n_{hp+}! \prod_{g=1}^{v} n_{h+g}!}{n_{h++}! \prod_{p=1}^{u} \prod_{g=1}^{v} n_{hpg}!}$$

Since the responses in this situation are ordinal, modified ridit (standardized midrank) scores, which only scale the responses according to their

relative ordering, are used in the spirit of minimal assumptions. Such scores, calculated as

$$a_{hg} = \frac{2\sum_{g'=1}^{g} n_{h+g'} - n_{h+g} + 1}{2(n_{h++} + 1)},$$

lie in the (0,1) interval and use midranks to manage ties. Thus, the sum of across-strata scores for the factor's pth level is $f_{+p+} = \Sigma_{h=1}^{q} \Sigma_{g=1}^{v} a_{hg} n_{hpg}$. Stacking any $(u - 1)$ of the u across-strata sums creates the $((u - 1) \times 1)$ vector $\underset{\sim}{\mathbf{f}}_{+}$. Under H_0, each sum has expected value $E\{f_{+p+}|H_0\} = \Sigma_{h=1}^{q} n_{hp+}\mu_h$ and covariance structure

$$\mathrm{Cov}\{f_{+p+}, f_{+p'+} \mid H_0\} = \sum_{h=1}^{q} \frac{n_{hp'+}(n_{h++}\,\Delta_{pp'} - n_{hp+})}{n_{h++} - 1} v_h,$$

where

$$\Delta_{pp'} = \begin{cases} 0, \text{ if } p \neq p' \\ 1, \text{ if } p = p' \end{cases}, \quad \mu_h = \sum_{g=1}^{v} \frac{a_{hg} n_{h+g}}{n_{h++}}, \text{ and } v_h = \sum_{g=1}^{v} \frac{n_{h+g}}{n_{h++}} (a_{hg} - \mu_h)^2.$$

(When $p = p'$, the covariance term is the variance.) The $(u - 1)$ of the across-strata expected values can be arranged in a $((u - 1) \times 1)$ vector $\underset{\sim}{\mathbf{E}}$ like the observed sums. Similarly, variances can be arranged on the diagonal and covariances off the diagonal to form a $((u - 1) \times (u - 1))$ covariance matrix $\underset{\sim}{\mathbf{V}}_{\underset{\sim}{\mathbf{f}}_{+}}$. For large n_{+p+}, by the central limit theorem, $\underset{\sim}{\mathbf{f}}_{+}$ is distributed approximately multivariate normal, so the extended Mantel-Haenszel statistic, $Q_{\mathrm{EMH}} = (\underset{\sim}{\mathbf{f}}_{+} - \underset{\sim}{\mathbf{E}})' \underset{\sim}{\mathbf{V}}_{\underset{\sim}{\mathbf{f}}_{+}}^{-1} (\underset{\sim}{\mathbf{f}}_{+} - \underset{\sim}{\mathbf{E}})$, has an approximate chi-square distribution with $(u - 1)$ d.f. This statistic (also called the row mean score difference statistic in the FREQ procedure of the SAS system) is an extension of both the Kruskal-Wallis (46) and Friedman rank (47) tests and is invariant to the choice of the $(u - 1)$ of u functions among the $\{f_{+p+}\}$.

In this study, differences in response are assessed for each active treatment dose versus placebo with adjustment for four strata based on center and baseline pain intensity. A Bonferroni-Holm hierarchical closed procedure (23) adjusted for multiple comparisons for each drug versus placebo for each measure. Following a significant difference between the standard and placebo (i.e. rejecting the composite null hypothesis that both doses were equal to placebo) in the first stage, the test drug was compared to placebo (in terms of the composite hypothesis). To adjust for multiple doses in the composite null hypothsis versus placebo for each of the standard and test drugs, a Bonferroni-Holm method (sequentially rejective Bonferroni procedure (24)) was applied by testing the dose with smaller p-value at $\alpha/2$ and the one with

higher p-value at α, given significance at $\alpha/2$ for the other dose; only rejection with the smaller p-value is needed to conclude at least one dose is different from placebo. For the previously described testing procedure, measures are evaluated in a set order: TOTPAR, SPID, TOTGONE, and pain relief ratings at 3, 2, 1 and 0.5 hr.

Ordinary Least Squares Regression

Well-known OLS regression for the lth response at the kth time point can be expressed using a general linear univariate model [e.g., Searle (48)] as

$$\underset{\sim}{\mathbf{y}}_{kl} = \underset{\sim}{\mathbf{X}}\underset{\sim}{\boldsymbol{\beta}}_{kl} + \underset{\sim}{\boldsymbol{\epsilon}}_{kl}, \tag{A.2}$$

where $\underset{\sim}{\mathbf{X}}$ is a matrix of fixed known constants, $\underset{\sim}{\boldsymbol{\beta}}_{kl}$ is a vector of w_{kl} fixed unknown constant parameters, and $\underset{\sim}{\boldsymbol{\epsilon}}_{kl}$ is a vector of unobserved errors for the kth time point and lth response. The error vector is assumed to be distributed multivariate normal with mean $\underset{\sim}{\mathbf{0}}$ and covariance matrix $\sigma^2_{kl}\underset{\sim}{\mathbf{I}}$, where $\underset{\sim}{\mathbf{I}}$ is an identity matrix. These assumptions include independence of observations, finite second-moment existence, and homoscedasticity, as well as parameter linearity.

Using a full rank $\underset{\sim}{\mathbf{X}}$ matrix with "reference" or "regression" coding, parameters can be estimated uniquely:

$$\hat{\underset{\sim}{\boldsymbol{\beta}}}_{kl} = (\underset{\sim}{\mathbf{X}}'\,\underset{\sim}{\mathbf{X}})^{-1}\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{y}}_{kl}, \tag{A.3}$$

since the inverse exists, and $\hat{\underset{\sim}{\boldsymbol{\beta}}}_{kl}$ is distributed multivariate normal with mean vector $\underset{\sim}{\boldsymbol{\beta}}_{kl}$ and covariance matrix $\sigma^2_{kl}(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{X}})^{-1}$, while the predicted response vector given the factors can be computed as $\hat{\underset{\sim}{\mathbf{y}}}_{kl} = \underset{\sim}{\mathbf{X}}\hat{\underset{\sim}{\boldsymbol{\beta}}}_{kl}$. Estimable functions, a class of linear combinations of the parameter estimates ($\hat{\underset{\sim}{\boldsymbol{\beta}}}_{kl}$), can be devised to test certain aspects of the grouping classifications using a contrast matrix $\underset{\sim}{\mathbf{C}}$. The estimator for σ^2_{kl} is $s^2_{kl} = (\underset{\sim}{\mathbf{y}}_{kl} - \hat{\underset{\sim}{\mathbf{y}}}_{kl})'(\underset{\sim}{\mathbf{y}}_{kl} - \hat{\underset{\sim}{\mathbf{y}}}_{kl})/(N - w_{kl})$, where w_{kl} is the number of parameters.

In this example the placebo group in center 1 with moderate baseline pain intensity is used as the reference. Indicators for each of the two drugs are used along with a natural logarithmic transformation of the dose to characterize the four active treatment groups. Since the doses (50 mg and 100 mg for test drug and 200 mg and 400 mg for standard) are equally spaced on the natural logarithmic scale, transforming the dose seems reasonable. Dichotomies for center and baseline pain are incorporated; an interaction between the standard drug indicator and the $\log_e$(dose) effect is used to assess the similarity of the $\log_e$(dose) effect for the two active drugs; this leads to a (258 × 7) design matrix with the following essence (unique rows):

$$
\begin{bmatrix}
1 & & & & & 0 \\
1 & & & & & 1 \\
1 & 1 & & \log_e 200 & \log_e 200 & 0 \\
1 & 1 & & \log_e 200 & \log_e 200 & 1 \\
1 & 1 & & \log_e 400 & \log_e 400 & 0 \\
1 & 1 & & \log_e 400 & \log_e 400 & 1 \\
1 & & 1 & \log_e 50 & & 0 \\
1 & & 1 & \log_e 50 & & 1 \\
1 & & 1 & \log_e 100 & & 0 \\
1 & & 1 & \log_e 100 & & 1 \\
1 & & & & & 1 \quad 0 \\
1 & & & & & 1 \quad 1 \\
1 & 1 & & \log_e 200 & \log_e 200 & 1 \quad 0 \\
1 & 1 & & \log_e 200 & \log_e 200 & 1 \quad 1 \\
1 & 1 & & \log_e 400 & \log_e 400 & 1 \quad 0 \\
1 & 1 & & \log_e 400 & \log_e 400 & 1 \quad 1 \\
1 & & 1 & \log_e 50 & & 1 \quad 0 \\
1 & & 1 & \log_e 50 & & 1 \quad 1 \\
1 & & 1 & \log_e 100 & & 1 \quad 0 \\
1 & & 1 & \log_e 100 & & 1 \quad 1
\end{bmatrix} \tag{A.4}
$$

and the following parameter vector

$$\underset{\sim}{\beta}_O = [\alpha_P, \alpha_S, \alpha_T, \beta, \alpha_S\beta, \psi, \nu]', \tag{A.5}$$

where α_P is the intercept for the reference group, α_S and α_T are the increments due to one milligram of standard and test drugs, respectively, β is the slope from the $\log_e$(dose) effect, $\alpha_S\beta$ is the increase in slope from the interaction between the standard drug and the $\log_e$(dose), ψ is the increase from being in center 2, and ν is the increase due to severe initial pain.

Removing $\alpha_S\beta$ from this model provides a simplified model for which parallelism of slopes is assumed. Relative potency estimates and confidence intervals can be generated (9,13,14) for this simplified model. The equivalent response, y_{ijkl}, for the two active drugs, can be achieved by setting their predicted values equal to each other:

$$\begin{aligned}
&\alpha_P + \alpha_S + [\beta \times \log_e(\text{dose Standard})] + \psi + \nu \\
&\quad = \alpha_p + \alpha_T + [\beta \times \log_e(\text{dose Test})] + \psi + \nu,
\end{aligned}$$

which implies $\log_e[\text{dose Standard}/\text{dose Test}] = (\alpha_T - \alpha_S)/\beta$. Exponentiating both sides yields the relative potency:

$$\rho = \frac{\text{dose Standard}}{\text{dose Test}} = \exp\left[\frac{\alpha_T - \alpha_S}{\beta}\right], \tag{A.6}$$

which means that ρ units of the standard drug produce the same expected response as one unit of the test drug. Let

$$R = \exp\left[\frac{D}{B}\right] \tag{A.7}$$

be an estimator for ρ, where $D = \hat{\alpha}_T - \hat{\alpha}_S$ and $B = \hat{\beta}$. Taking the natural logarithm twice, the variance can be approximated with TSL as

$$\mathrm{Var}\{\log_e(\log_e R)\} = \frac{\mathrm{Var}\, D}{D^2} - \frac{2\mathrm{Cov}(D, B)}{DB} + \frac{\mathrm{Var}\, B}{B^2},$$

which is meaningful when $\log_e R$ is positive. A $100(1 - \alpha)\%$ CI for R can be formed, where α is the type I error rate. Confidence bounds for $\log_e(\log_e R)$ are calculated in the usual way: estimate $\log_e(\log_e R)$; then subtract and add the product of the standard error of $\log_e(\log_e R)$ and the $(1 - \alpha/2)$th percentile of the standard normal distribution. Double exponentiating the bounds produces the CI for R:

$$\exp\left[\exp\left[\log_e(\log_e R) \pm \left\{z_{1-(\alpha/2)}\sqrt{\mathrm{Var}\{\log_e(\log_e R)\}}\right\}\right]\right]. \tag{A.8}$$

Appendix II. Weighted Least Squares Methods with Mean Response

Univariate Measures

OLS, explained in Appendix I, can be generalized to WLS to allow heteroscedasticity (12,16), when the number of categorical predictors is small. WLS methods assume only cross-tabulation sample sizes sufficient for the mean responses to be asymptotically multivariate normal and the classification group data to be considered representative of the populations they portray, similar to stratified simple random samples.

Altering Figure A1 slightly to form a contingency table with $u^* \equiv (q \times u)$ subpopulations on the vertical axis and v response levels on the horizontal one and letting h^* index the subpopulations gives the n_{h^*g} with product multinomial distributions in this situation.

$$\Pr\{n_{h^*g}\} = \prod_{h^*=1}^{u^*}\left[n_{h^*+}!\prod_{g=1}^{v}\left\{\frac{\{\pi_{h^*g}\}^{n_{h^*g}}}{n_{h^*g}!}\right\}\right],$$

where $\pi_{h^*g} = E\{n_{h^*g}/n_{h^*+}\}$ is the probability that a randomly selected patient from the h^*th subpopulation has the gth response and $\Sigma_{g=1}^{v} \pi_{h^*g} = 1$.

Now using the conventions from Landis and Koch (49), let $\underset{\sim}{\mathbf{m}}_{\mathbf{h}^*} = \underset{\sim}{\mathbf{n}}_{\mathbf{h}^*}/n_{h^*+}$,a $(v \times 1)$ vector, be the observed sample proportions from the h^*th subpopulation. Further, stack the $\underset{\sim}{\mathbf{m}}_{\mathbf{h}^*}$ to form the $(u^*v \times 1)$ compound vector $\underset{\sim}{\mathbf{m}}$, which is the unbiased maximum likelihood estimator for the compound parameter vector $\underset{\sim}{\boldsymbol{\pi}}$, shaped in the same manner. The covariance matrix can be consistently estimated by the $(u^*v \times u^*v)$ block diagonal matrix $\underset{\sim}{\mathbf{V}}(\underset{\sim}{\mathbf{m}})$, which has the $(v \times v)$ matrices $\underset{\sim}{\mathbf{V}}_{\mathbf{h}^*}(\underset{\sim}{\mathbf{m}}_{\mathbf{h}^*}) = (\underset{\sim}{\mathbf{D}}_{\underset{\sim}{\mathbf{m}}_{\mathbf{h}^*}} - \underset{\sim}{\mathbf{m}}_{\mathbf{h}^*}\underset{\sim}{\mathbf{m}}'_{\mathbf{h}^*})/n_{h^*+}$ on the main diagonal, where $\underset{\sim}{\mathbf{D}}_{\underset{\sim}{\mathbf{m}}_{\mathbf{h}^*}}$ is a $(v \times v)$ diagonal matrix with $\underset{\sim}{\mathbf{m}}_{\mathbf{h}^*}$'s elements on the diagonal. Each $\underset{\sim}{\mathbf{V}}_{\mathbf{h}^*}(\underset{\sim}{\mathbf{m}}_{\mathbf{h}^*})$ is the multinomial covariance matrix for the h^*th stratum. Then let $\underset{\sim}{\mathbf{F}} \equiv \underset{\sim}{\mathbf{F}}(\underset{\sim}{\mathbf{m}})$ be a matrix of f functions of $\underset{\sim}{\mathbf{m}}$ of interest, with a consistent, asymptotically nonsingular, $(f \times f)$ covariance matrix estimator $\underset{\sim}{\mathbf{V}}_{\underset{\sim}{\mathbf{F}}} = \underset{\sim}{\mathbf{H}}\underset{\sim}{\mathbf{V}}(\underset{\sim}{\mathbf{m}})\underset{\sim}{\mathbf{H}}'$, where $\underset{\sim}{\mathbf{H}} = \partial\underset{\sim}{\mathbf{F}}(\underset{\sim}{\mathbf{x}})/\partial\underset{\sim}{\mathbf{x}}|_{\underset{\sim}{\mathbf{x}}=\underset{\sim}{\mathbf{m}}}$ has dimensions $(f \times u^*v)$. The asymptotic expected value of $\underset{\sim}{\mathbf{F}}$ is $\underset{\sim}{\mathbf{E}}_{\underset{\sim}{\mathbf{A}}}\{\underset{\sim}{\mathbf{F}}(\underset{\sim}{\mathbf{m}})\} = \underset{\sim}{\mathbf{F}}(\underset{\sim}{\boldsymbol{\pi}}) = \underset{\sim}{\mathbf{X}}\underset{\sim}{\boldsymbol{\beta}}$, where $\underset{\sim}{\mathbf{X}}$ is a full rank design matrix with rank w. The parameter estimate vector

$$\hat{\underset{\sim}{\boldsymbol{\beta}}} = (\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{V}}_{\underset{\sim}{\mathbf{F}}}^{-1}\underset{\sim}{\mathbf{X}})^{-1}\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{V}}_{\underset{\sim}{\mathbf{F}}}^{-1}\underset{\sim}{\mathbf{F}}, \tag{A.9}$$

a best asymptotically normal estimator distributed approximately multivariate normal, has its covariance matrix consistently estimated as:

$$\underset{\sim}{\mathbf{V}}_{\hat{\underset{\sim}{\boldsymbol{\beta}}}} = (\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{V}}_{\underset{\sim}{\mathbf{F}}}^{-1}\underset{\sim}{\mathbf{X}})^{-1}. \tag{A.10}$$

The residual Wald statistic

$$Q = (\underset{\sim}{\mathbf{F}} - \underset{\sim}{\mathbf{X}}\hat{\underset{\sim}{\boldsymbol{\beta}}})'\,\underset{\sim}{\mathbf{V}}_{\underset{\sim}{\mathbf{F}}}^{-1}\,(\underset{\sim}{\mathbf{F}} - \underset{\sim}{\mathbf{X}}\hat{\underset{\sim}{\boldsymbol{\beta}}}), \tag{A.11}$$

which has the approximate chi-square distribution with $(f - w)$ d.f., is used to assess GOF.

Linear combinations $\underset{\sim}{\mathbf{C}}\hat{\underset{\sim}{\boldsymbol{\beta}}}$ can be formed with the result being asymptotically normally distributed with mean $\underset{\sim}{\mathbf{0}}$ and covariance matrix $\underset{\sim}{\mathbf{C}}(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{V}}_{\underset{\sim}{\mathbf{F}}}^{-1}\underset{\sim}{\mathbf{X}})^{-1}\underset{\sim}{\mathbf{C}}'$. Testing $\underset{\sim}{\mathbf{C}}\hat{\underset{\sim}{\boldsymbol{\beta}}} = \underset{\sim}{\mathbf{0}}$ is accomplished with a Wald statistic

$$(\underset{\sim}{\mathbf{C}}\hat{\underset{\sim}{\boldsymbol{\beta}}})'[\underset{\sim}{\mathbf{C}}(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{V}}_{\underset{\sim}{\mathbf{F}}}^{-1}\underset{\sim}{\mathbf{X}})^{-1}\underset{\sim}{\mathbf{C}}']^{-1}(\underset{\sim}{\mathbf{C}}\hat{\underset{\sim}{\boldsymbol{\beta}}}), \tag{A.12}$$

which is distributed approximately chi-square with rank of $\underset{\sim}{\mathbf{C}}$ d.f. Predicted values can be generated using

$$\hat{\underset{\sim}{\mathbf{F}}} = \underset{\sim}{\mathbf{X}}\hat{\underset{\sim}{\boldsymbol{\beta}}} = \underset{\sim}{\mathbf{X}}(\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{V}}_{\underset{\sim}{\mathbf{F}}}^{-1}\underset{\sim}{\mathbf{X}})^{-1}\underset{\sim}{\mathbf{X}}'\underset{\sim}{\mathbf{V}}_{\underset{\sim}{\mathbf{F}}}^{-1}\underset{\sim}{\mathbf{F}} \tag{A.13}$$

and its covariance matrix can be estimated consistently by

$$\mathbf{V}_{\hat{\mathbf{F}}} = \mathbf{X}\mathbf{V}_{\hat{\boldsymbol{\beta}}}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{V}_{\mathbf{F}}^{-1}\mathbf{X})^{-1}\mathbf{X}'. \tag{A.14}$$

When the model fits well and subpopulation sample sizes are adequate, $\hat{\mathbf{F}}$ and its estimated covariance matrix $\mathbf{V}_{\hat{\mathbf{F}}}$ are better estimates than $\mathbf{F}$ and $\mathbf{V}_{\mathbf{F}}$, the sample estimates.

Of particular interest is the linear transform $\mathbf{F}(\mathbf{m}) = \mathbf{L}\mathbf{m}$, where $\mathbf{L}$ is the $(u^* \times u^*v)$ block diagonal matrix of $[1\ 2 \cdots v]$ row vectors. This produces $\bar{\mathbf{y}}$, the vector of means of integer scores for responses in each subpopulation, with covariance matrix $\mathbf{V}_{\bar{\mathbf{y}}} = \mathbf{L}\mathbf{V}(\mathbf{m})\mathbf{L}'$. Estimates of the parameters $(\hat{\boldsymbol{\beta}}_*)$, its covariance matrix $(\mathbf{V}_{\hat{\boldsymbol{\beta}}*})$, predicted values $(\hat{\bar{\mathbf{y}}})$, and their covariance matrix $(\mathbf{V}_{\hat{\bar{\mathbf{y}}}})$ can be found as described previously.

In this study, WLS uses a design matrix $\mathbf{X}$ similar to OLS's essence matrix (A.4), but the last column is omitted, since the cross-tabulations form some sparse subpopulations when baseline pain intensity is included. WLS uses a similar parameter vector as OLS does in Eq. (A.5), but the last element is omitted.

Assuming parallelism, estimates and CIs for relative potency can be generated by performing matrix operations on the parameter and covariance estimates. The point estimates for relative potency can be found as

$$\hat{\rho} = \exp[\exp\{\mathbf{A}_2 \log_e(\mathbf{A}_1\hat{\boldsymbol{\beta}}_*)\}], \tag{A.15}$$

where $\mathbf{A}_1 = \begin{bmatrix} 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$ and $\mathbf{A}_2 = [1\ -1]$. This estimate has variance:

$$\mathrm{Var}\{\hat{\rho}\} = \mathbf{H}_*\mathbf{V}_{\hat{\boldsymbol{\beta}}}\mathbf{H}_*', \tag{A.16}$$

where $\mathbf{H}_* = \partial\boldsymbol{\rho}/\partial\boldsymbol{\beta}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_*} = a_3a_2\mathbf{A}_2\mathbf{D}_{\mathbf{a}_1}^{-1}\mathbf{A}_1$, $\mathbf{a}_1 = \mathbf{A}_1\hat{\boldsymbol{\beta}}_*$, $a_2 = \exp(\mathbf{A}_2\log_e\mathbf{a}_1)$, $a_3 = \exp(a_2)$, and a diagonal matrix of elements of $\mathbf{a}_1$ forms $\mathbf{D}_{\mathbf{a}_1}$. To find the CIs, calculate the bounds on the doubly logged quantities and then double-exponentiate, as was done for OLS. Transforming the point estimate with the natural logarithm twice gives:

$$\log_e\{\log_e(\hat{\rho})\} = \mathbf{A}_2\log_e(\mathbf{A}_1\hat{\boldsymbol{\beta}}_*), \tag{A.17}$$

which has its variance as

$$\text{Var}[\log_e\{\log_e(\hat{\rho})\}] = \tilde{\mathbf{H}}_* \mathbf{V}_{\hat{\boldsymbol{\beta}}_*} \tilde{\mathbf{H}}_*', \tag{A.18}$$

where $\tilde{\mathbf{H}}_* = \partial \log_e\{\log_e(\hat{\rho})\}/\partial \boldsymbol{\beta}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_*} = \mathbf{A}_2 \mathbf{D}_{\mathbf{a}_1}^{-1} \mathbf{A}_1$. This produces a $100(1 - \alpha)\%$ CI like (A.8):

$$\exp\left[\exp\left[\log_e(\log_e \hat{\rho}) \pm \left\{z_{1-(\alpha/2)} \sqrt{\text{Var}\{\log_e(\log_e \hat{\rho})\}}\right\}\right]\right]. \tag{A.19}$$

Fieller's method for finding confidence intervals (18) was used to judge the TSL-derived CIs for relative potency. Letting ξ be the logarithm of Eq. (A.6) relative to Eq. (A.7), the square of the standard normal percentile can be set equal to $(D - \xi B)^2/(V_D - 2\xi V_{D,B} + \xi^2 V_B)$ and rewritten as a quadratic equation: $a\xi^2 + b\xi + c = 0$, where the coefficients are $a = B^2 - (z_{1-(\alpha/2)})^2 V_B$, $b = 2\{(z_{1-(\alpha/2)})^2 V_{D,B} - DB\}$, and $c = D^2 - (z_{1-(\alpha/2)})^2 V_D$. Then the quadratic formula gives the solution's lower and upper bounds for ξ as $(-b - \sqrt{b^2 - 4ac})/2a$ and $(-b + \sqrt{b^2 - 4ac})/2a$, respectively. Exponentiating ξ's bounds provides a $100(1 - \alpha)\%$ CI for $\hat{\rho}$ from Fieller's formula.

Multivariate Summary Measures

In general, univariate WLS methods, as reviewed in the previous section, can be extended to multivariate situations. If the $\mathbf{y}_{\mathbf{kl}}$ are written as a compound vector, multivariate WLS methods can be used. For example, to model the r responses at all τ times, a compound vector of the $(r \times \tau)$ $\mathbf{y}_{\mathbf{kl}}$ vectors is used. The remaining methods in this section are shown for the r responses at the kth time point, so the compound vector $[\mathbf{y}_{\mathbf{k1}}', \mathbf{y}_{\mathbf{k2}}', \ldots, \mathbf{y}_{\mathbf{kr}}']'$ contains the responses to be modeled. Then using the multivariate extensions of $\mathbf{m}$ and $\mathbf{V}(\mathbf{m})$, along with the functions thereof ($\mathbf{F}$ and $\mathbf{V}_{\mathbf{F}}$), a compound parameter vector $\boldsymbol{\beta}_{*\mathbf{k}} = [\boldsymbol{\beta}_{*1}', \boldsymbol{\beta}_{*2}', \ldots, \boldsymbol{\beta}_{*\mathbf{r}}']'$, and an expanded design matrix defined with a left Kronecker product as $[\mathbf{X} \otimes \mathbf{I}_\mathbf{r}]$, WLS can be performed as in the univariate case by substituting those quantities into Eqs. (A.9) and (A.10) for estimation, (A.11) for GOF, (A.12) for hypothesis testing, and (A.13) and (A.14) for prediction. Relative potency estimates and CIs are obtained with Eqs. (A.15), (A.16), and (A.19). In multivariate situations, however, relative potency is estimated for each response, so $\hat{\boldsymbol{\rho}}$ is a vector with the same rank as the compound response vector. Appropriate premultiplication matrices for compound functions in the multivariate case are $[\mathbf{A}_1 \otimes \mathbf{I}_\mathbf{r}]$ and $[\mathbf{A}_2 \otimes \mathbf{I}_\mathbf{r}]$. Moreover, $\mathbf{H}_*$ is now defined using diagonal matrices $\mathbf{D}_{\mathbf{a}_2}$ and $\mathbf{D}_{\mathbf{a}_3}$ with the vectors $\mathbf{a}_2$ and $\mathbf{a}_3$, respectively, on the diagonal. The vector of the standard

errors (the square roots of the diagonals of the covariance matrix) of $\log_e(\log_e \hat{\boldsymbol{\rho}})$ is used to form the bounds. This vector of estimates, $\hat{\boldsymbol{\rho}}$, can be tested for homogeneity across responses and averaged to obtain a single multivariate relative potency estimate $\hat{\rho}$ and its own CI based on $\log_e (\log_e \hat{\rho})$ bounds.

In this particular example, bivariate WLS of the mean primary responses (SPID and TOTPAR) employs the parameter vector $\boldsymbol{\beta}_{*B} = [\boldsymbol{\beta}'_{*1}, \boldsymbol{\beta}'_{*2}]'$ and the design matrix, say $[\mathbf{X} \otimes \mathbf{I}_2]$, where each $\boldsymbol{\beta}_{*l} = [\alpha_{Pl}, \alpha_{Sl}, \alpha_{Tl}, \beta_l, \psi_l]'$ and $\mathbf{X}$ is similar to OLS' essence matrix (A.4) but omitting the fifth and last columns, representing $\log_e$(dose) $\times$ standard drug interaction and baseline severity, respectively. The elements of the $\boldsymbol{\beta}_{*l}$ vectors have interpretations as in univariate OLS, but these are specific for the lth response. Relative potency estimates and CIs are found using the compound functions with $[\mathbf{A}_1 \otimes \mathbf{I}_2]$ and $[\mathbf{A}_2 \otimes \mathbf{I}_2]$. A 1 d.f. chi-square test of the homogeneity of the elements of $\hat{\boldsymbol{\rho}}$ is $[\mathbf{A}_2 \text{ (A.17)}]^2 \, [\mathbf{A}_2 \text{ (A.18)} \mathbf{A}'_2]^{-1}$. Following a nonsignificant chi-square test result, a relative potency estimate can be calculated using $\mathbf{A}_3$ (A.17) and a CI formed by substituting $\mathbf{A}_3$ (A.18)$\mathbf{A}'_3$ into Eq. (A.19) instead of Eq. (A.18), where $\mathbf{A}_3 = \left[\frac{1}{2} \; \frac{1}{2}\right]$.

Fieller's method for obtaining CIs, described for the univariate situation, can be extended to multivariate ones as a verification method for the aforementioned TSL approximation methods. In multivariate settings, D and B, as well as their variances and covariances, become vectors with identical rank as the compound response vector. They can be substituted into the quadratic formula, shown in the univariate case; element-wise multiplication instead of matrix multiplication allows the quadratic formula's solution.

Appendix III. Repeated Measures Methods for Ordinal Response

Weighted Least Squares

One specific multivariate situation involves analyzing responses over the τ time points, using a compound response vector as described in Appendix II. In this case, the effect of time and the interactions involving time should be modeled. The parameter vector and design matrix mentioned in Appendix II should be augmented to incorporate these effects. So design matrices are of the form $[\mathbf{1}_\tau \otimes \mathbf{X}, \; \mathbf{T} \otimes \mathbf{1}_{u*}, \; \mathbf{Y}]$, where $\mathbf{T} = \begin{bmatrix} \mathbf{0}_{1,\tau-1} \\ \mathbf{I}_{\tau-1} \end{bmatrix}$ parameterizes the time effect with $\tau - 1$ indicators and $\mathbf{Y}$ contains the interactions between columns in $\mathbf{1}_\tau \otimes \mathbf{X}$ and columns in $\mathbf{T} \otimes \mathbf{1}_{u*}$. The parameter vector would be $[\boldsymbol{\beta}'_*, \boldsymbol{\gamma}', \boldsymbol{\eta}']'$, where $\boldsymbol{\beta}_*$ contains the group parameters as before, $\boldsymbol{\gamma}$ contains the $\tau - 1$ time parameters, and $\boldsymbol{\eta}$ contains the parameters involving inter-

actions with time. WLS as discussed in the multivariate section of Appendix II can be performed, and the model simplified as needed. Relative potency estimates and confidence intervals can be calculated in a similar manner with the appropriate compound functions.

Analyses of imputed pain relief through the first 3 hr (with TOTPAR2 as the 1 hr response) used a modified design matrix: in addition to an intercept, treatment was modeled with an indicator for active drug and an indicator for test drug, as well as the $\log_e$(dose) effect as before. The center effect had to be excluded due to insufficient sample size for adequate covariance matrix estimation for three functions from 10 subpopulations. So five subpopulations representing the treatment groups were used with the 2 d.f. time effect (indicators for the second and third time points) and three 2 d.f. interactions with time. The design matrix without interactions is

$$\begin{bmatrix}
1 & & & & & \\
1 & & & & 1 & \\
1 & & & & & 1 \\
1 & 1 & & \log_e 200 & & \\
1 & 1 & & \log_e 200 & 1 & \\
1 & 1 & & \log_e 200 & & 1 \\
1 & 1 & & \log_e 400 & & \\
1 & 1 & & \log_e 400 & 1 & \\
1 & 1 & & \log_e 400 & & 1 \\
1 & 1 & 1 & \log_e 50 & & \\
1 & 1 & 1 & \log_e 50 & 1 & \\
1 & 1 & 1 & \log_e 50 & & 1 \\
1 & 1 & 1 & \log_e 100 & & \\
1 & 1 & 1 & \log_e 100 & 1 & \\
1 & 1 & 1 & \log_e 100 & & 1
\end{bmatrix} \tag{A.20}$$

and the parameter vector is

$$\underset{\sim}{\beta}_W = [\alpha_P, \alpha_A, \delta, \beta, T_2, T_3]', \tag{A.21}$$

where the parameters have the same interpretation as in OLS (Appendix I) plus α_A is the increment due to either active drug, δ is the difference between test and standard drugs, and each T_k is the increment from the first time point to the kth.

Common effects are appropriate for factors that do not vary across time. A common slope would indicate a constant dose-response relationship over time. The methods explained in Appendix II can then be utilized. To obtain relative potency estimates and CIs with the no interaction model, the matrix $\underset{\sim}{A}_1 = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$ is used to form the premultiplier presented in Appendix II.

Generalized Estimating Equations

Quasi-likelihood theory (50) assumes a known transformation of the response's marginal expectation is a linear function of the covariates and the variance is a known function of its expected value, except for a possibly unknown scale parameter. In other words, the distribution of the response is not specified, just the relationship between the mean response and covariates, as well as between the mean and variance. Thus normal and nonnormal responses can be modeled. GEE methods (19, 20) use this quasi-likelihood approach along with a specified "working" correlation matrix in an iterative fashion to provide consistent regression parameter and variance estimates.

Using a compound response vector as described in Appendix II, let t^* index the observations for the $k = 1, 2, \ldots, \tau$ times and $l = 1, 2, \ldots, r$ responses of the ith subject belonging to the jth group, and let $i^* = 1, 2, \ldots, N$, where $N = \Sigma_j n_j$, index the $i = 1, 2, \ldots, n_j$ subjects from $j = 1, 2, \ldots, s$ groups. Then the expected value of the t^*th observation for the i^*th subject can be written as $E[y_{i^*t^*}] = \mu_{i^*t^*} = \hbar[\underset{\sim}{x}'_{i^*t^*}\underset{\sim}{\beta}_G]$ and the variance can be written as a function of the expected value $V[y_{i^*t^*}] = v_{i^*t^*} = \omega[\mu_{i^*t^*}]/\phi^{(g)}$, where the inverse of $\hbar$ is called the link function and $\phi^{(g)}$ is a scale parameter. (The superscript g merely reminds the reader of GEE parameters.)

Let $\underset{\sim}{P}(\underset{\sim}{\alpha}^{(g)})$ be the "working" correlation matrix, characterized by the unknown nuisance parameter vector $\underset{\sim}{\alpha}^{(g)}$. The working covariance matrix is $\underset{\sim}{V}_{i^*} = \underset{\sim}{D}_{i^*}^{1/2}\,\underset{\sim}{P}(\underset{\sim}{\alpha}^{(g)})\,\underset{\sim}{D}_{i^*}^{1/2}/\phi^{(g)}$, where $\underset{\sim}{D}_{i^*}$ is an $\tau r \times \tau r$ diagonal matrix with $\omega[\mu_{i^*t^*}]$ as the diagonal elements. The GEEs are $\Sigma_{i^*=1}^{N}\,\underset{\sim}{H}'_{i^*}\,\underset{\sim}{V}_{i^*}^{-1}\,\underset{\sim}{S}_{i^*} = 0$, where $\underset{\sim}{H}_{i^*} = \partial\underset{\sim}{\mu}_{i^*}/\partial\underset{\sim}{\beta}_G$ and $\underset{\sim}{S}_{i^*} = \underset{\sim}{y}_{i^*} - \underset{\sim}{\mu}_{i^*}$; $\underset{\sim}{\mu}_{i^*}$ is the vector of observations, $\mu_{i^*t^*}$, for the i^*th subject. The vector of consistent parameter estimates $\hat{\underset{\sim}{\beta}}_G$ is the solution of $\Sigma_{i^*=1}^{N}\,\underset{\sim}{F}_{i^*}\,[\underset{\sim}{\beta}_G, \hat{\alpha}\{\underset{\sim}{\beta}_G, \hat{\phi}^{(g)}(\underset{\sim}{\beta}_G)\}] = 0$, where $\hat{\alpha}^{(g)}$ and $\hat{\phi}^{(g)}$ are consistent estimates of $\alpha^{(g)}$ and $\phi^{(g)}$, respectively. Furthermore, $N^{1/2}(\hat{\underset{\sim}{\beta}}_G - \underset{\sim}{\beta}_G)$ is distributed asymptotically multivariate normal with a covariance matrix of

$$\underset{\sim}{V}_G = \lim_{N\to\infty} N\left(\sum_{i^*=1}^{N} \underset{\sim}{H}'_{i^*}\,\underset{\sim}{V}_{i^*}^{-1}\,\underset{\sim}{H}_{i^*}\right)^{-1}$$

$$\left\{\sum_{i^*=1}^{N} \underset{\sim}{H}'_{i^*}\,\underset{\sim}{V}_{i^*}^{-1}\,\mathrm{cov}(\underset{\sim}{y}_{i^*})\,\underset{\sim}{V}_{i^*}^{-1}\,\underset{\sim}{H}_{i^*}\right\}\left(\sum_{i^*=1}^{N} \underset{\sim}{H}'_{i^*}\,\underset{\sim}{V}_{i^*}^{-1}\,\underset{\sim}{H}_{i^*}\right)^{-1},$$

where $\mathrm{cov}(\underset{\sim}{y}_{i^*})$ is the true, not the assumed, covariance matrix. In practice, $\mathrm{cov}(\underset{\sim}{y}_{i^*})$ is replaced with $\underset{\sim}{S}_{i^*}\underset{\sim}{S}'_{i^*}$. These equations are solved using iteratively reweighted least squares and standardized residuals, iterated until convergence is achieved. A GEE analysis was applied with the (A.20) design matrix and the parameter estimate vector (A.21) as described in the preceding section. Also to illustrate a portion of GEE's versatility, a model adjusting for

severe baseline pain and center as covariates was examined. The adjusted model used $\left[(A.20) \otimes \underset{\sim}{1}_4, \underset{\sim}{1}_{15} \otimes \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}'\right]$ as the essence design matrix and $\underset{\sim}{\beta}_{GA} = [\alpha_P, \alpha_A, \delta, \beta, T_2, T_3, \psi, \nu]'$ as the parameter vector.

Survey Data Regression

SDR methods (21, 22, 51, 52) involve modifying usual OLS methods by changing the variance estimation to account for correlation of data from the same patient. Regularly- or irregularly-timed repeated measures data from a clinical trial can be interpreted as analogous to cluster samples with each person as a sampling unit and each time point as a within-cluster element. The covariance matrix of the parameter estimates is approximated with TSL.

The OLS regression model (A.2) and the solution of the normal equations (A.3) are altered, using weighted sums of squares and cross-products (Horvitz-Thompson estimators) for $\underset{\sim}{X}'\underset{\sim}{X}$ and $\underset{\sim}{X}'\underset{\sim}{y}_{kl}$. The counterpart to $\underset{\sim}{X}'\underset{\sim}{X}$ is $\Sigma_{i^*=1}^{N} \Sigma_{j=1}^{\tau} \underset{\sim}{x}_{i^*j}\underset{\sim}{x}'_{i^*j} w_{i^*} = \underset{\sim}{G}$, while the counterpart to $\underset{\sim}{X}'\underset{\sim}{y}_{kl}$ is $\Sigma_{i^*=1}^{N} \Sigma_{j=1}^{\tau} \underset{\sim}{x}_{i^*j} y_{i^*j} w_{i^*} = \underset{\sim}{K}$, where w_{i^*} is proportional to the inverse overall sampling selection probability for the i^*th subject; thus the SDR parameter estimator is $\hat{\underset{\sim}{\beta}}_S = \underset{\sim}{G}^{-1}\underset{\sim}{K}$. Let the linearized variate vector for the j^{th} observation of the i^{*th} subject be $\underset{\sim}{Z}_{i^*j} = \underset{\sim}{G}^{-1}\{\underset{\sim}{x}_{i^*j}\underset{\sim}{r}_{i^*j} w_{i^*}\}$, where $r_{i^*j} = (y_{i^*j} - \underset{\sim}{x}'_{i^*j}\hat{\underset{\sim}{\beta}}_S)$ is the adjusted residual; then summing them across observations in a within-subject manner provides $\underset{\sim}{Z}_{i^*} = \Sigma_{j^*=1}^{\tau} \underset{\sim}{Z}_{i^*j}$ and then averaging them across subjects provides $\bar{\underset{\sim}{Z}} = 1/N \Sigma_{i^*=1}^{N} \underset{\sim}{Z}_{i^*}$; the Taylor series variance estimator of $\hat{\underset{\sim}{\beta}}_S$ is written as $\hat{\underset{\sim}{V}}(\hat{\underset{\sim}{\beta}}_S) = N/(N-1) \Sigma_{i^*=1}^{N} (\underset{\sim}{Z}_{i^*} - \bar{\underset{\sim}{Z}})(\underset{\sim}{Z}_{i^*} - \bar{\underset{\sim}{Z}})'$, since the $\underset{\sim}{Z}_{i^*}$ vectors of within-subject residual aggregates are independent as long as subjects are independent. Although in this study each subject had an equal sampling weight (inverse probability of being sampled) so that the same parameter estimates as OLS were obtained, covariance computations involve within-subject sums of adjusted residuals which account for the aforementioned sampling design. Thus, each subject is viewed as a cluster (primary sampling unit) and each observation is treated as an observational unit within that cluster. As with the GEE analyses, the unadjusted and adjusted SDR analyses were applied with the (A.20) design matrix and the parameter estimate vector (A.21) as described earlier, along with the augmented design matrix and parameter estimate vector adjusting for center and baseline pain explained in the previous section.

References

1. Bredfeldt RC, Sutherland JE, Kruse JE: Efficacy of transdermal clonidine for headache prophylaxis and reduction of narcotic use in migraine patients: A randomized crossover trial. *J Family Pract* 29:153–158, 1989.
2. Mehlisch DR, Sollecito WA, Helfrick JF, Leibold DG, Markowitz R, Schow CE Jr, Schultz R, Waite DE: Multicenter clinical trial of ibuprofen and acetaminophen in the treatment of postoperative dental pain. *J Am Dent Assoc* 121:257–263, 1990.
3. Rubin DB: Inference and missing data. *Biometrika* 63:581–592, 1976.
4. Gillings D, Koch G: The application of the principle of intention-to-treat to the analysis of clinical trials. *Drug Inform J* 25:411–424, 1991.
5. Siegel C, Sunshine A, Richman H, Olson NZ, Robissa N, Cordone R, Estrada N, Laska E: Meptazinol and morphine in postoperative pain assessed with a new method for onset and duration. *J Clin Pharmacol* 29:1017–1025, 1989.
6. Powell H, Smallman JMB, Morgan M: Comparison of intramuscular ketorolac and morphine in pain control after laparotomy. *Anaesthesia* 45:538–542, 1990.
7. Kempf KK, Konzelman JL, Schultz RE, Turner JL: Comparison of meclofenamate sodium with buffered aspirin and placebo for the relief of postoperative dental pain. *Clin Ther* 9:594–601, 1987.
8. Schachtel BP, Thoden WR, Baybutt RI: Ibuprofen and acetaminophen in the relief of postpartum episiotomy pain. *J Clin Pharmacol* 29:550–553, 1989.
9. Tsutakawa RK: Bioassay, statistical methods in. In: *Encyclopedia of Statistical Sciences*, Vol 1 (Kotz S, Johnson NL, Eds). Wiley, New York, 1982, pp. 236–243.
10. Laska EM, Meisner M, Takeuchi K, Wanderling JA, Siegel C, Sunshine A: An analytic approach to quantifying pain scores. *Pharmacotherapy* 6:276–282, 1986.
11. Koch GG, Edwards SE: Clinical efficacy trials with categorical data. In: *Biopharmaceutical Statistics for Drug Development* (Peace KE, Ed). Marcel Dekker, New York, 1988, pp. 403–457.
12. Koch GG, Carr GJ, Amara IA, Stokes ME, Uryniak TJ: Categorical data analysis. In: *Statistical Methodology in the Pharmaceutical Sciences* (Berry DA, Ed). Marcel Dekker, New York, 1990, pp. 389–473.
13. Govindarajulu Z: *Statistical Techniques in Bioassay*. Karger, New York, 1988, pp. 3–27.
14. Finney DJ: *Statistical Methods in Biological Assay*, 3rd ed. Macmillan, New York, 1978, pp. 17–132.
15. McCullagh P: Regression models for ordinal data (with discussion). *J Roy Statist Soc B* 42:109–142, 1980.
16. Grizzle JE, Starmer CF, Koch GG: The analysis of categorical data by linear models. *Biometrics* 25:489–504, 1969.
17. Koch GG, Imrey PB, Singer JM, Atkinson SS, Stokes ME: *Analysis of Categorical Data*. Les Presses de l'Université de Montreal, Montreal, 1985.
18. Read CB: Fieller's theorem. In: *Encyclopedia of Statistical Sciences*, Vol 3 (Kotz S, Johnson NL, Eds). Wiley, New York, 1983, pp. 86–88.
19. Zeger SL, Liang K-Y: Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 42:121–130, 1986.
20. Liang K-Y, Zeger SL: Longitudinal data analysis using generalized linear models. *Biometrika* 73:13–22, 1986.
21. Binder DA: On the variances of asymptotically normal estimators from complex surveys. *Int Statist Rev* 51:279–292, 1983.

22. Shah BV, Holt MM, Folsom RE: Inference about regression models from sample survey data. *Bull Int Stat Inst* 47:43–57, 1977.
23. Bauer P: Multiple testing in clinical trials. *Statist Med* 10:871–890, 1991.
24. Holm S: A simple sequentially rejective multiple test procedure. *Scand J Statist* 6:65–70, 1979.
25. SAS Institute, Inc: *SAS/STAT User's Guide, Version 6*, Vol 1, 4th ed. SAS Institute Inc, Cary, NC, 1990, pp. 405–517, 851–889.
26. SAS Institute, Inc: *SAS/STAT User's Guide, Version 6*, Vol 2, 4th ed. SAS Institute, Inc, Cary, NC, 1990, pp. 891–996.
27. SAS Institute, Inc: The data step. In: *SAS Language Guide for Personal Computers, Release 6.03 ed.* SAS Institute, Inc, Cary, NC, 1988, pp. 17–278.
28. Graubard BI, Korn EL: Choice of column scores for testing independence in ordered $2 \times K$ contingency tables. *Biometrics* 43:471–476, 1987.
29. Koch GG, Amara IA, Davis GW, Gillings DB: A review of some statistical methods for covariance analysis of categorical data. *Biometrics* 38:563–595, 1982.
30. SAS Institute, Inc: *SAS/IML Software, Version 6*, 1st ed. SAS Institute, Inc, Cary, NC, 1990.
31. Turney EA, Amara IA, Koch GG, Stewart WH: Evaluation of alternative statistical methods for linear model analysis to compare two treatments for 24-hour blood pressure response. *Statist Med* 11:1843–1860, 1992.
32. Kenward MG, Jones B: Alternative approaches to the analysis of binary and categorical repeated measurements. *J Biopharm Statist* 2:137–170, 1992.
33. Clayton D: Repeated ordinal measurements: A generalized estimating equations approach. *Appl Statist* (submitted for publication 1992).
34. Gebski V, Leung O, McNeil D, Lunn D: *SPIDA User's Manual, Version 6*. Eastwood, New South Wales, Australia, 1992.
35. Shah BV, Barnwell BG, Hunt PN, LaVange LM: *SUDAAN User's Manual, Release 6.00*. Research Triangle Institute, Research Triangle Park, NC, 1992.
36. LaVange LM, Keyes LL, Koch GG, Margolis PA: Application of sample survey methods for modelling ratios to incidence densities. *Statist Med* 13:343–355, 1994.
37. Vølund A: Multivariate bioassay. *Biometrics* 36:255–236, 1980.
38. Vølund A: Combination of multivariate bioassay. *Biometrics* 38:181–190, 1982.
39. Vølund A: Response to "Multivariate bioassay." *Biometrics* 41:551–554, 1985.
40. Laska EM, Kushner HB, Meisner M: Multivariate bioassay. *Biometrics* 41:547–550, 1985.
41. Carter EM, Hubert JJ: Analysis of parallel-line assays with multivariate responses. *Biometrics* 41:703–710, 1985.
42. Williams DA: An exact confidence interval for the relative potency estimated from a multivariate bioassay. *Biometrics* 44:861–867, 1988.
43. Ekblom A, Hansson P: Pain intensity measurements in patients with acute pain receiving afferent stimulation. *J Neurol Neurosurg Psychiatry* 51:481–486, 1988.
44. Cox C, Chuang C: A comparison of chi-square partitioning and two logit analyses of ordinal pain data from a pharmaceutical trial. *Statist Med* 3:273–285, 1984.
45. Chuang C, Agresti A: A new model for ordinal pain data from a pharmaceutical trial. *Statist Med* 5:15–20, 1986.
46. Kruskal WH, Wallis WA: Use of ranks on one-criterion variance analysis. *J Am Statist Assoc* 47:583–621, 1952.
47. Friedman M: The use of ranks to avoid the assumptions of normality implicit in the analysis of variance. *J Am Statist Assoc* 32:675–701, 1937.

48. Searle SR: *Linear Models*. Wiley, New York, 1971.
49. Landis JR, Koch GG: The analysis of categorical data in longitudinal studies of behavioral development. In: *Longitudinal Methodology in the Study of Behavior and Development* (Nesselroade JR, Baltes PB, Eds). Academic Press, New York, 1979, pp. 233–261.
50. McCullagh P, Nelder JA: Quasi-likelihood functions. *Ann Statist* 11:59–67, 1983.
51. Shah BV, LaVange LM: Software for inference on linear models from survey data. *Proceedings of the American Statistical Association, Section on Survey Research Methods*. American Statistical Association, Washington, DC, 1982.
52. Shah BV, Folsom RE, LaVange LM, Wheeless SC, Boyle KE, Williams RL: *Statistical Methods and Mathematical Algorithms used in SUDAAN*. Research Triangle Institute, Research Triangle Park, NC, 1993.