# Statistical Methodology in the Pharmaceutical Sciences

*Edited by*

DONALD A. BERRY

*School of Statistics*
*University of Minnesota*
*Minneapolis, Minnesota*

MARCEL DEKKER, INC. New York and Basel

# 13

# Categorical Data Analysis

**GARY G. KOCH and GREGORY J. CARR** University of North Carolina, Chapel Hill, North Carolina

**INGRID A. AMARA** Quintiles, Inc., Chapel Hill, North Carolina

**MAURA E. STOKES** SAS Institute, Inc., Cary, North Carolina

**THOMAS J. URYNIAK** Fisons Corporation, Bedford, Massachusetts

## 1 INTRODUCTION

Many studies in the pharmaceutical sciences are concerned with the relationships between categorical response variables which describe favorable or unfavorable outcome and one or more explanatory variables. The explanatory variables include experimental factors such as treatment in clinical trials and background characteristics such as age, sex, and baseline status of subjects; additional factors that often need to be taken into account are "center" in multicenter studies and "visit" in multivisit studies. Since the explanatory variables which are considered can have either a categorical (e.g., sex) or continuous (e.g., age) nature, the distinguishing feature of situations which require categorical data analyses is that the response variables are categorical.

The statistical questions of interest for categorical data analysis are addressed in this chapter through a specific example from a multicenter, multivisit clinical trial for patients with a respiratory disorder. The data for this example are displayed as case records in Table 1. There are two centers

**Table 1** Data from a Multicenter, Multivisit Clinical Trial to Compare Two Treatments for Patients with a Respiratory Disorder[a]

| Center | Patient | Drug | Sex | Age | Base | Visit 1 | Visit 2 | Visit 3 | Visit 4 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 53 | A | F | 32 | 1 | 2 | 2 | 4 | 2 |
| | 18 | A | F | 47 | 2 | 2 | 3 | 4 | 4 |
| | 54 | A | M | 11 | 4 | 4 | 4 | 4 | 2 |
| | 12 | A | M | 14 | 2 | 3 | 3 | 3 | 2 |
| | 51 | A | M | 15 | 0 | 2 | 3 | 3 | 3 |
| | 20 | A | M | 20 | 3 | 3 | 2 | 3 | 1 |
| | 16 | A | M | 22 | 1 | 2 | 2 | 2 | 3 |
| | 50 | A | M | 22 | 2 | 1 | 3 | 4 | 4 |
| | 03 | A | M | 23 | 3 | 3 | 4 | 4 | 3 |
| | 32 | A | M | 23 | 2 | 3 | 4 | 4 | 4 |
| | 56 | A | M | 25 | 2 | 3 | 3 | 2 | 3 |
| | 35 | A | M | 26 | 1 | 2 | 2 | 3 | 2 |
| | 26 | A | M | 26 | 2 | 2 | 2 | 2 | 2 |
| | 21 | A | M | 26 | 2 | 4 | 1 | 4 | 2 |
| | 08 | A | M | 28 | 1 | 2 | 2 | 1 | 2 |
| | 30 | A | M | 28 | 0 | 0 | 1 | 2 | 1 |
| | 33 | A | M | 30 | 3 | 3 | 4 | 4 | 2 |
| | 11 | A | M | 30 | 3 | 4 | 4 | 4 | 3 |
| | 42 | A | M | 31 | 1 | 2 | 3 | 1 | 1 |
| | 09 | A | M | 31 | 3 | 3 | 4 | 4 | 4 |
| | 37 | A | M | 31 | 0 | 2 | 3 | 2 | 1 |
| | 23 | A | M | 32 | 3 | 4 | 4 | 3 | 3 |
| | 06 | A | M | 34 | 1 | 1 | 2 | 1 | 1 |
| | 22 | A | M | 46 | 4 | 3 | 4 | 3 | 4 |
| | 24 | A | M | 48 | 2 | 3 | 2 | 0 | 2 |
| | 38 | A | M | 50 | 2 | 2 | 2 | 2 | 2 |
| | 48 | A | M | 57 | 3 | 3 | 4 | 3 | 4 |
| | 05 | P | F | 13 | 4 | 4 | 4 | 4 | 4 |
| | 19 | P | F | 31 | 2 | 1 | 0 | 2 | 2 |
| | 25 | P | F | 35 | 1 | 0 | 0 | 0 | 0 |
| | 28 | P | F | 36 | 2 | 3 | 3 | 2 | 2 |
| | 36 | P | F | 45 | 2 | 2 | 2 | 2 | 1 |
| | 43 | P | M | 13 | 3 | 4 | 4 | 4 | 4 |
| | 41 | P | M | 14 | 2 | 2 | 1 | 2 | 3 |
| | 34 | P | M | 15 | 2 | 2 | 3 | 3 | 2 |
| | 29 | P | M | 19 | 2 | 3 | 3 | 0 | 0 |
| | 15 | P | M | 20 | 4 | 4 | 4 | 4 | 4 |
| | 13 | P | M | 23 | 3 | 3 | 1 | 1 | 1 |

**Table 1** (*Continued*)

| Center | Patient | Drug | Sex | Age | Base | Visit 1 | Visit 2 | Visit 3 | Visit 4 |
|---|---|---|---|---|---|---|---|---|---|
| | 27 | P | M | 23 | 4 | 4 | 2 | 4 | 4 |
| | 55 | P | M | 24 | 3 | 4 | 4 | 4 | 3 |
| | 17 | P | M | 25 | 1 | 1 | 2 | 2 | 2 |
| | 45 | P | M | 26 | 2 | 4 | 2 | 4 | 3 |
| | 40 | P | M | 26 | 1 | 2 | 1 | 2 | 2 |
| | 44 | P | M | 27 | 1 | 2 | 2 | 1 | 2 |
| | 49 | P | M | 27 | 3 | 3 | 4 | 3 | 3 |
| | 39 | P | M | 28 | 2 | 1 | 1 | 1 | 1 |
| | 02 | P | M | 28 | 2 | 0 | 0 | 0 | 0 |
| | 14 | P | M | 30 | 1 | 0 | 0 | 0 | 0 |
| | 31 | P | M | 37 | 1 | 0 | 0 | 0 | 0 |
| | 10 | P | M | 37 | 3 | 2 | 3 | 3 | 2 |
| | 07 | P | M | 43 | 2 | 3 | 2 | 4 | 4 |
| | 52 | P | M | 43 | 1 | 1 | 1 | 3 | 2 |
| | 04 | P | M | 44 | 3 | 4 | 3 | 4 | 2 |
| | 01 | P | M | 46 | 2 | 2 | 2 | 2 | 2 |
| | 46 | P | M | 49 | 2 | 2 | 2 | 2 | 2 |
| | 47 | P | M | 63 | 2 | 2 | 2 | 2 | 2 |
| 2 | 30 | A | F | 37 | 1 | 3 | 4 | 4 | 4 |
| | 52 | A | F | 39 | 2 | 3 | 4 | 4 | 4 |
| | 23 | A | F | 60 | 4 | 4 | 3 | 3 | 4 |
| | 54 | A | F | 63 | 4 | 4 | 4 | 4 | 4 |
| | 12 | A | M | 13 | 4 | 4 | 4 | 4 | 4 |
| | 10 | A | M | 14 | 1 | 4 | 4 | 4 | 4 |
| | 27 | A | M | 19 | 3 | 3 | 2 | 3 | 3 |
| | 47 | A | M | 20 | 2 | 4 | 4 | 4 | 3 |
| | 16 | A | M | 20 | 2 | 1 | 1 | 0 | 0 |
| | 29 | A | M | 21 | 3 | 3 | 4 | 4 | 4 |
| | 20 | A | M | 24 | 4 | 4 | 4 | 4 | 4 |
| | 25 | A | M | 25 | 3 | 4 | 3 | 3 | 1 |
| | 15 | A | M | 25 | 3 | 4 | 4 | 3 | 3 |
| | 02 | A | M | 25 | 2 | 2 | 4 | 4 | 4 |
| | 09 | A | M | 26 | 2 | 3 | 4 | 4 | 4 |
| | 49 | A | M | 28 | 2 | 3 | 2 | 2 | 1 |
| | 55 | A | M | 31 | 4 | 4 | 4 | 4 | 4 |
| | 43 | A | M | 34 | 2 | 4 | 4 | 2 | 4 |
| | 26 | A | M | 35 | 4 | 4 | 4 | 4 | 4 |
| | 14 | A | M | 37 | 4 | 3 | 2 | 2 | 4 |
| | 36 | A | M | 41 | 3 | 4 | 4 | 3 | 4 |

**Table 1** (*Continued*)

| Center | Patient | Drug | Sex | Age | Base | Visit 1 | Visit 2 | Visit 3 | Visit 4 |
|---|---|---|---|---|---|---|---|---|---|
| | 51 | A | M | 43 | 3 | 3 | 4 | 4 | 2 |
| | 37 | A | M | 52 | 1 | 2 | 1 | 2 | 2 |
| | 19 | A | M | 55 | 4 | 4 | 4 | 4 | 4 |
| | 32 | A | M | 55 | 2 | 2 | 3 | 3 | 1 |
| | 03 | A | M | 58 | 4 | 4 | 4 | 4 | 4 |
| | 53 | A | M | 68 | 2 | 3 | 3 | 3 | 4 |
| | 28 | P | F | 31 | 3 | 4 | 4 | 4 | 4 |
| | 05 | P | F | 32 | 3 | 2 | 2 | 3 | 4 |
| | 21 | P | F | 36 | 3 | 3 | 2 | 1 | 3 |
| | 50 | P | F | 38 | 1 | 2 | 0 | 0 | 0 |
| | 01 | P | F | 39 | 1 | 2 | 1 | 1 | 2 |
| | 48 | P | F | 39 | 3 | 2 | 3 | 0 | 0 |
| | 07 | P | F | 44 | 3 | 4 | 4 | 4 | 4 |
| | 38 | P | F | 47 | 2 | 3 | 3 | 2 | 3 |
| | 08 | P | F | 48 | 2 | 2 | 1 | 0 | 0 |
| | 11 | P | F | 48 | 2 | 2 | 2 | 2 | 2 |
| | 04 | P | F | 51 | 3 | 4 | 2 | 4 | 4 |
| | 17 | P | F | 58 | 1 | 4 | 2 | 2 | 0 |
| | 39 | P | M | 11 | 3 | 4 | 4 | 4 | 4 |
| | 40 | P | M | 14 | 2 | 1 | 2 | 3 | 2 |
| | 24 | P | M | 15 | 3 | 2 | 2 | 3 | 3 |
| | 41 | P | M | 15 | 4 | 3 | 3 | 3 | 4 |
| | 33 | P | M | 19 | 4 | 2 | 2 | 3 | 3 |
| | 34 | P | M | 20 | 3 | 2 | 4 | 4 | 4 |
| | 13 | P | M | 20 | 1 | 4 | 4 | 4 | 4 |
| | 45 | P | M | 33 | 3 | 3 | 3 | 2 | 3 |
| | 22 | P | M | 36 | 2 | 4 | 3 | 3 | 4 |
| | 18 | P | M | 38 | 4 | 3 | 0 | 0 | 0 |
| | 35 | P | M | 42 | 3 | 2 | 2 | 2 | 2 |
| | 44 | P | M | 43 | 2 | 1 | 0 | 0 | 0 |
| | 06 | P | M | 45 | 3 | 4 | 2 | 1 | 2 |
| | 46 | P | M | 48 | 4 | 4 | 0 | 0 | 0 |
| | 31 | P | M | 52 | 2 | 3 | 4 | 3 | 4 |
| | 42 | P | M | 66 | 3 | 3 | 3 | 4 | 4 |

[a] 0, terrible; 1, poor; 2, fair; 3, good; 4, excellent.

(center 1, center 2) within which patients were randomly assigned to two treatments (active, placebo) in successive blocks of 6. The status of each patient was classified relative to a set of five ordinal categories (0 = terrible, 1 = poor, 2 = fair, 3 = good, 4 = excellent) at baseline and at each of four visits (visit 1, visit 2, visit 3, visit 4) during the time period over which the treatments were administered. Data for the age and sex of all patients were obtained at the time of entry to the study.

The categorical response variables for the example are the classifications of the status of patients at the four visits. These response variables have an ordinal measurement scale which expresses an ordering of possible outcomes from most unfavorable (i.e., terrible), to most favorable (i.e., excellent). The simplification of the observed classifications to the dichotomous measurement scale of (terrible, poor, or fair) versus (good or excellent) provides another set of response variables of interest. Dichotomous classifications and ordinal classifications are the most prominent types of categorical response variables in the pharmaceutical sciences, so this chapter focuses primary attention on methods for their analysis. Other measurement scales for categorical data and references for their discussion are as follows:

i. Discrete counts: for example, number of infected quarters of the udder of dairy cows receiving treatment for mastitis (Koch et al., 1978) or number of hours with little or no pain for women receiving treatment for obstetrical-related pain (Koch et al., 1985a); methods for this type of data are similar to those for ordinal classifications.

ii. Grouped survival data: for example, time interval between periodic evaluations for healing of duodenal ulcer (Koch and Edwards, 1987) or time interval between periodic evaluations for death or recurrence of duodenal ulcer (Koch et al., 1986); methods for this type of data are analogous to those for dichotomous data; they include the Mantel–Haenszel statistic (Mantel and Haenszel, 1959) as extended by Mantel (1966) to a life-table format, Poisson regression for piecewise exponential models (Holford, 1980), and extensions of logistic models (Chapter 11).

iii. Nominal classifications: for example, type or site of pain or infection with there being no ordering for three or more outcomes; one strategy for this type of data is the application of methods for dichotomous classifications to the presence or absence of each outcome separately or to specified combinations of outcomes; extensions of the Mantel–Haenszel statistic (Landis et al., 1978; Koch et al., 1985a) and log-linear models (Bishop et al., 1975; Fienberg, 1980; Imrey et al., 1981, 1982) are useful for the joint analysis of all outcomes.

The explanatory variables for the example further illustrate the different types of measurement scales for data from studies in the pharmaceutical sciences. Center, sex, and treatment are dichotomous classifications; baseline status is an ordinal classification; visit is a nominal classification (since the ordering of visits in time may not be relevant to their relationship with the response variables); and age has a continuous distribution. Thus categorical data analyses for the response variables from the example need to include consideration of both categorical and continuous explanatory variables.

The subsequent sections of this chapter address the statistical questions of interest for the data in Table 1. Section 2 is concerned with comparisons between treatment groups for the background variables of age, sex, and baseline status. Such analysis is of interest because any background variable that does not have equivalent distributions for the two treatments might partly explain the observed differences between treatments for response variables if it is also strongly associated with the response variables. For this reason, potentially important associations between background variables and response variables are evaluated to identify those background variables for which the framework for treatment comparisons for response variables requires statistical adjustment; these are background variables that are both strongly associated with response variables and have different distributions for the two treatments. Two methods for adjustment of treatment comparisons for such background variables are stratification and covariance analysis. Adjustment by one of these methods is necessary for observed differences between treatment groups to be interpretable as due to treatment (as opposed to a random lack of equivalence of the treatment groups for a background variable with strong association with the response variables). Stratification and covariance analysis are useful for other purposes besides adjustment for differences between treatment groups for background variables. One of these is a capability for more powerful statistical tests through the relatively smaller variances which are provided for the estimates of treatment differences. Another is that their scope enables questions concerning homogeneity of treatment differences across subgroups based on background variables to be addressed through tests of treatment × background variable interaction. More specific aspects of stratification and covariance analysis are discussed subsequently through the roles they serve for the analysis of the data in Table 1.

Univariate methods for the analysis of the response variables at each visit separately are presented in Section 3. Attention is initially given to nonparametric statistical tests for comparisons between treatments under minimal assumptions. These include Fisher's exact test for $(2 \times 2)$ contingency tables and rank tests (with adjustments for ties) for ordinal data within

each center separately, and the Mantel–Haenszel test and its extensions for the combined centers. An alternative analysis strategy that is more useful for descriptive purposes is based on statistical models for the relationship between response variables and explanatory variables for treatment, center, and background variables. Methods include logistic regression for dichotomous data, the proportional odds model for ordinal data, and weighted least squares for mean scores or other relevant functions of response distributions. Their range of application enables the evaluation of the similarity of treatment differences across centers through tests of treatment × center interaction and across subgroups based on background variables through tests of treatment × background variable interaction. Also, background variables without equivalent distributions for treatments need to be included in the statistical model for covariance adjustment purposes. A noteworthy limitation of model-based methods is their requirement of assumptions about how data from a clinical study are representative of a general population.

In Section 4, multivariate methods for the analysis of the response variables at all four visits jointly are discussed. They include both nonparametric statistical tests of comparisons between treatments under minimal assumptions, and statistical models which account for variation of response across visits as well as treatment and center. An important feature of the statistical models is that they enable the evaluation of the similarity of treatment differences across visits through tests of treatment × visit interaction.

The roles of the alternative methods in Sections 3 and 4 are summarized in Section 5. Relative strengths and limitations are also discussed there. Throughout Sections 2 to 5, concepts, analysis procedures, results, and interpretation are emphasized. The reader is assumed to have general familiarity with statistical issues for studies in the pharmaceutical sciences and basic methods for the statistical analysis of categorical data and contingency tables. Some background references for these topics are: Everitt (1977), Fienberg (1980), Fleiss (1981, 1986), Friedman et al. (1981), Koch and Sollecito (1984), Shapiro and Louis (1983), and Tygstrup et al. (1982). Also, the scope of this chapter does not include a substantial discussion for the statistical theory and technical structure of available methods for the analysis of categorical data; some references for these topics are: Agresti (1984), Bishop et al. (1975), Cox (1970), Forthofer and Lehnen (1981), Freeman (1987), Imrey et al. (1981, 1982), Koch et al. (1985a), and McCullagh and Nelder (1983). All computations to illustrate the application of the methods in this chapter were performed with the SAS System (1985); the types of statements that were used are described in the appendix to this chapter.

## 2 EVALUATION OF BACKGROUND VARIABLES

The case record data in Table 1 are from $n = 111$ patients of whom $n_1 = 56$ participated at center 1 and $n_2 = 55$ participated at center 2. The numbers of patients $n_{hi}$ who received active ($A$) and placebo ($P$) treatment were $n_{1A} = 27$ and $n_{1P} = 29$ at center 1 and $n_{2A} = 27$ and $n_{2P} = 28$ at center 2 (where $h = 1, 2$ indexes center and $i = A, P$ indexes treatment). The total numbers of patients $n_{+i} = (n_{1i} + n_{2i})$ for the two treatment groups were $n_{+A} = 54$ for active and $n_{+P} = 57$ for placebo.

Three background variables that characterized patients prior to treatment in the clinical trial were age, sex, and baseline status. For most clinical trials, patients are selected for study by convenience mechanisms related to their need or eligibility for treatment at a particular time as opposed to a probabilistic sampling process. Similarly, centers are selected according to judgmental criteria for their qualifications and willingness to conduct the study. These considerations imply that the patients in a clinical trial might not represent a general target population in a formal statistical way. Thus distributions of background variables such as age, sex, and baseline status in the study population (which the patients in a clinical trial constitute) might be different from those in the target population (to which conclusions concerning treatments are to be generalized). For example, the study population for a clinical trial might have relatively more patients who are younger, male, or have more favorable baseline status than the target population.

One way to evaluate treatments in a clinical trial relative to the issue about how a study population represents a target population is for the statistical analysis to have the following two parts:

I. Usage of nonparametric statistical tests to compare treatment groups with respect to background variables and response variables in the study population under minimal assumptions that involve only study design considerations

II. Usage of statistical models to describe the relationships between response variables and treatment, center, and background variables for the target population under the (not provable) assumption that the conditional distributions of response variables given treatment, center, and background variables for patients in the study reasonably represent those in the target population

The objective for part I is the determination of the existence of a treatment difference for the study population; whereas the objective for part II is the evaluation of the extent to which a treatment difference is generalizable throughout a target population. Since the nonparametric methods

for part I are based only on study design considerations, the conclusions from them are often called design-based inferences (for the study population). In a similar spirit, the conclusions from analysis in part II are often called model-based inferences (for the target population). Both parts I and II serve important roles. The primary advantage of part I is its applicability without assumptions external to the study design; but this is counterbalanced by the limitation of its scope of inference to the study population. In contrast, part II has the advantage of providing conclusions about the target population, but its applicability has the limitation of requiring potentially debatable assumptions that express how patients in the study population are conceptually representative of their counterparts in the target population. Also, any need for assumptions about model structure would be another concern. Since the advantages and limitations of parts I and II are in some sense complementary, the combined usage of both is recommended in this chapter. Other references that discuss this strategy are Koch et al. (1980b, 1982), Koch and Gillings (1983), Koch and Sollecito (1984), and Koch and Edwards (1987). The remainder of this section is concerned with aspects of the application of nonparametric statistical tests to address the comparisons in part I for the background variables and related questions concerning the association between background variables and response variables. Both parts I and II are discussed for the response variables in Sections 3 and 4.

## 2.1 Comparisons Between Treatment Groups Within Centers

For the specific example in this chapter, the patients at each center were separately assigned to the two treatments through random partitions of successive blocks of six patients (i.e., within each block, three patients received active treatment and three patients received placebo). These blocks are ignored henceforth on the basis of the assumption that the order of entry of patients to the study was sufficiently random for them to have no association with either background variables or response variables. The randomization process in the study design and the ignorability of blocks jointly imply that the patients in each treatment group at each center are a simple random sample of the finite subpopulation of all patients in the study at the corresponding center (if blocks needed to be taken into account, each treatment group would be a stratified simple random sample and the methods in Section 2.2 would apply).

The statistical properties of sums or means of observations from simple random samples [as discussed in Cochran (1977)] then enables the construction of nonparametric statistical tests for comparisons between treatment

groups. To be specific, let $x_{hik}$ denote the value of a background variable for the $k$th patient with the $i$th treatment at the $h$th center. Since treatment has no effect on background variables (as a consequence of their determination prior to treatment), the subpopulation mean $\mu_h$ and variance $v_h$ for all patients at each of the centers are known constants, that is,

$$\mu_h = \frac{\sum_{i=A}^{P} \sum_{k=1}^{n_{hi}} x_{hik}}{n_h}, \qquad v_h = \frac{\sum_{i=A}^{P} \sum_{k=1}^{n_{hi}} (x_{hik} - \mu_h)^2}{n_h}. \tag{1}$$

The status of each treatment group as a simple random sample from the finite subpopulation at the corresponding center implies that the expected values and covariance structure for their observed means $\bar{x}_{hi} = \{\sum_{k=1}^{n_{hi}} x_{hik}/n_{hi}\}$ are

$$E\{\bar{x}_{hi}\} = \mu_h, \qquad \text{var}\{\bar{x}_{hi}\} = \frac{(n_h - n_{hi})v_h}{n_{hi}(n_h - 1)},$$

$$\text{cov}\{\bar{x}_{hA}, \bar{x}_{hP}\} = \frac{-v_h}{n_h - 1}. \tag{2}$$

When the sample sizes for the two treatment groups are sufficiently large (e.g., the $n_{hi} \geq 15$), the mean score statistic relative to either treatment $i = A, P$,

$$Q_{S,h} = \frac{(\bar{x}_{hi} - \mu_h)^2}{\text{var}\{\bar{x}_{hi}\}} = \frac{n_h - 1}{n_h}(\bar{x}_{hA} - \bar{x}_{hP})^2 \Big/ \left(\frac{1}{n_{hA}} + \frac{1}{n_{hP}}\right) v_h, \tag{3}$$

approximately has the chi-square distribution with one degree of freedom (i.e., $d.f. = 1$). Thus the $p$-value from this chi-square approximation for outcomes $\geq Q_{S,h}$ enables evaluation of the extent to which randomization has provided the treatment groups with similar distributions for the background variable. In this regard, small $p$-values (e.g., $p \leq 0.05$) correspond to large values of $Q_{S,h}$ and hence identify background variables with noteworthy differences between treatment groups due to chance (since treatment should have no effect on background variables). The determination of $p$-values for situations with small samples requires consideration of the exact distribution of the $\bar{x}_{hi}$ across all possible randomizations of patients to treatments at each center.

The mean score statistic $Q_{S,h}$ in (3) has forms that encompass well-known methods for particular types of background variables. Dichotomous variables such as sex are handled with indicator variables such as $x_{hik} = 1$ if male or $x_{hik} = 0$ if female. It then follows that $\bar{x}_{hi}$ is the proportion of males with the $i$th treatment at the $h$th center and $n_{hi}\bar{x}_{hi} = n_{hiM}$ is the number of males with the $i$th treatment at the $h$th center. Similarly, $\mu_h$ is the proportion of males at the $h$th center and $n_h\mu_h = n_{h+M}$ is the number

of males at the $h$th center. Also, $v_h = (n_{h+F}n_{h+M})/n_h^2$, where $n_{h+F}$ is the number of females at the $h$th center. It then follows that

$$Q_{S,h} = \left(n_{hiM} - \frac{n_{hi}n_{h+M}}{n_h}\right)^2 \Big/ \left[\frac{n_{hA}n_{hP}n_{h+M}n_{h+F}}{n_h^2(n_h-1)}\right]$$
$$= \frac{n_h - 1}{n_h}\sum_i\sum_j \frac{(n_{hij} - m_{hij})^2}{m_{hij}} = \frac{(n_h-1)Q_{P,h}}{n_h}; \quad (4)$$

here $i = A, P$ and $j = F, M$ index treatment and sex; $n_{hij}$ and $m_{hij} = (n_{hi}n_{h+j}/n_h)$ denote the observed and expected numbers of patients with the $j$th sex in the $i$th treatment group under randomization at the $h$th center; and $Q_{P,h}$ denotes the well-known Pearson chi-square statistic for the $(2\times2)$ contingency table of frequencies $n_{hij}$ from the cross-classification of treatment and sex for each center. For the data in Table 1, the $(2 \times 2)$ contingency tables for treatment $\times$ sex are:

| Center 1 | Female | Male | Center 2 | Female | Male |
|---|---|---|---|---|---|
| Active (A) | 2 | 25 | Active (A) | 4 | 23 |
| Placebo (P) | 5 | 24 | Placebo (P) | 12 | 16 |

(5)

Their corresponding mean score statistics are $Q_{S,1} = 1.21$ with $p = 0.271$ and $Q_{S,2} = 5.15$ with $p = 0.023$. However, for $(2 \times 2)$ contingency tables, the determination of exact $p$-values through Fisher's exact test is straightforward and usually desirable unless sample sizes are clearly large enough for a chi-square approximation (e.g., all $m_{hij} \geq 10$).

The $p$-values for Fisher's exact test is determined by identification of all possible $(2 \times 2)$ contingency tables with the same row and column sums as the observed table, computation of the probability of occurrence of each with respect to the randomization-induced hypergeometric distribution, and the summation of those probabilities that are less than or equal to the probability of the observed table. One-sided $p$-values for a specified alternative are obtained by summation of all probabilities for tables in the corresponding direction for differences $(\bar{x}_{hA} - \bar{x}_{hP})$ between treatments at least as large as that for the observed table. The two-sided $p$-values from Fisher's exact test for sex are $p_1 = 0.42$ at center 1 and $p_2 = 0.037$ at center 2. These results suggest the presence of an imbalance in the sex distributions for the treatment groups at center 2 with a substantially larger percentage of males being randomly assigned to $A$ (85.2%) than to P (57.1%). Also, at center 1, somewhat more males were assigned to $A$ (92.6%) than to $P$ (82.8%).

The percentages of males for $A$ and $P$ at each center and the two-sided $p$-values from Fisher's exact test are displayed in Table 2. These

**Table 2** Descriptive Statistics and $p$-Values from Treatment Comparisons for Background Variables[a]

| Background variable | Treatment | Statistic | Center 1 | Center 2 | Combined centers[b] |
|---|---|---|---|---|---|
| Number of patients | Active (A) | $n$ | 27 | 27 | 54 |
| | Placebo (P) | $n$ | 29 | 28 | 57 |
| | Total | $n$ | 56 | 55 | 111 |
| Age | Active (A) | Mean | 29.93 | 35.85 | 32.89 |
| | | SE | 2.16 | 3.07 | 1.88 |
| | Placebo (P) | Mean | 30.69 | 36.71 | 33.65 |
| | | SE | 2.25 | 2.70 | 1.75 |
| | A vs. P | $p$-value | 0.93 | 0.69 | 0.73 |
| Sex | Active (A) | % male | 92.6 | 85.2 | 88.9 |
| | | SE | 5.1 | 7.0 | 4.3 |
| | Placebo (P) | % male | 82.8 | 57.1 | 70.2 |
| | | SE | 7.1 | 9.5 | 5.9 |
| | A vs. P | $p$-value | 0.42 | 0.037[c] | 0.013[c] |
| Baseline status | Active (A) | Mean | 1.96 | 2.78 | 2.37 |
| | | SE | 0.22 | 0.20 | 0.15 |
| | Placebo (P) | Mean | 2.17 | 2.61 | 2.39 |
| | | SE | 0.17 | 0.17 | 0.12 |
| | A vs. P | $p$-value | 0.57 | 0.55 | 0.98 |
| Dichotomous baseline status | Active (A) | % $\geq$ good | 33.3 | 55.6 | 44.4 |
| | | SE | 9.2 | 9.7 | 6.7 |
| | Placebo (P) | % $\geq$ good | 31.0 | 60.7 | 45.6 |
| | | SE | 8.7 | 9.4 | 6.4 |
| | A vs. P | $p$-value | 1.00 | 0.79 | 0.88 |

[a]The $p$-values for age and baseline status are based on the Wilcoxon rank sum statistic via (3) for centers 1 and 2 separately and on the extended Mantel–Haenszel statistic (12) relative to within-center standardized ranks (or the van Elteren statistic) for the combined centers; midranks were used to account for ties via (16). The $p$-values for sex and dichotomous baseline status are based on Fisher's exact test for centers 1 and 2 separately [as described relative to (5)] and the Mantel–Haenszel statistic (13) for the combined centers. Computations were performed with the FREQ Procedure in the SAS System (1985).

[b]Means for the combined centers are based on the pooled data for the two centers; their standard errors are adjusted for center via (21). The $p$-values for the combined centers are adjusted for centers by the stratification for the extended Mantel–Haenszel statistic (12).

[c]Treatment comparisons with $p \leq 0.05$.

types of results are also given there for the dichotomous classification of baseline status as (good or excellent) or not. For both sex and dichotomous baseline status, standard errors (SE) are shown in Table 2 for the reported percentages. They were obtained via

$$(\mathrm{SE})_{hi} = 100\left[\frac{p_{hij}(1-p_{hij})}{n_{hi}-1}\right]^{1/2} \tag{6}$$

where $p_{hij} = (n_{hij}/n_{hi})$ denotes the proportion of patients with the $j$th category of the background variable for the $i$th treatment group at the $h$th center; the multiplication by 100 accounts for percentages being $100p_{hij}$. The purpose of the SE's is to describe the variability of the reported percentages relative to the general setting of simple random sampling from corresponding infinite populations, so their computation from (6) presumes this framework (by involving no finite population correction and being based only on the data for the $i$th treatment at the $h$th center).

For age and baseline status, the means $\bar{x}_{hi}$ and their standard errors

$$(\mathrm{SE})_{hi} = \left[\frac{\sum_{k=1}^{n_{hi}}(x_{hik}-\bar{x}_{hi})^2}{n_{hi}(n_{hi}-1)}\right]^{1/2} \tag{7}$$

are given in Table 2 for each treatment group at each center; the SE's in (7) apply to infinite populations for the same reasons stated for (6). Although the actual values of age and baseline status can be used in the mean score statistic $Q_{S,h}$ for the comparison of the two treatment groups at each center, their transformation to ranks for this purpose is often preferable. An advantage of ranks for an ordinal variable such as baseline status (for which midranks are used to account for ties) is that they basically express the relative ordering of the observed categories as opposed to an explicit scaling such as 0, 1, 2, 3, 4, which might be debatable (the issue of using 0, 1, 2, 3, 4 for means and standard errors is discussed further in Section 3.1). For continuous variables such as age with distributions that may be nonsymmetric and may have a wide range, use of ranks can enhance the applicability of chi-square approximations to $Q_{S,h}$ for the available sample sizes. Also, when the $\{x_{hik}\}$ are ranks, the sum $n_{hi}\bar{x}_{hi}$ of observed values for the $i$th treatment is the Wilcoxon rank sum statistic, so analysis for small samples can be undertaken with exact $p$-values. An algorithm due to Mehta et al. (1984) can be used to obtain such exact $p$-values for ordinal variables with a small number of categories and possibly many ties; for continuous variables with only a few ties, tables like those in Owen (1962) for the Wilcoxon rank sum statistic (or its Mann–Whitney counterpart) are applicable.

In Table 2, the $p$-values for the comparison of the two treatment groups with respect to age and baseline status are based on the Wilcoxon rank-sum statistic. They were obtained through the approximate chi-square distribution of the mean score statistics $Q_{S,h}$ for which the $\{x_{hik}\}$ were the ranks of the $k$th patient relative to all patients at the $h$th center. This method is considered to be suitably supported by the available sample size (i.e., all $n_{hi} \geq 25$).

The statistical comparisons in Table 2 for the separate centers generally confirm that randomization has provided similar distributions of background variables for the two treatment groups. Only the $p$-value for sex at center 2 suggests a noteworthy imbalance by its $\leq 0.05$ status; the other seven $p$-values for within center comparisons had $\geq 0.25$ status and hence were clearly compatible with what might be expected from randomization.

## 2.2 Comparisons Between Treatment Groups for the Combined Centers

The patients from the combined centers constitute a stratified population (with centers as the strata), and the patients in the two treatment groups are stratified simple random samples of this population (since the random assignment of patients to treatments was undertaken separately at the two centers). Thus the question of whether the distribution of background variables for the two treatment groups is similar is equivalent to the question of whether the distribution of background variables for each treatment group is similar to what would be expected from the structure of its corresponding stratified simple random sample. The latter question can be addressed for the $i$th treatment group by comparing the across-center sum

$$x_{+i+} = \sum_{h=1}^{2} \sum_{k=1}^{n_{hi}} x_{hik} = \sum_{h=1}^{2} n_{hi} \bar{x}_{hi} \tag{8}$$

of its observations to their expected value

$$E\{x_{+i+}\} = \sum_{h=1}^{2} n_{hi} \mu_h. \tag{9}$$

The stratified structure of the randomization of patients to treatments implies that the means $\bar{x}_{hi}$ from different centers are independent of one another; so

$$\mathbf{cov}\{\bar{x}_{1i}, \bar{x}_{2i}\} = 0 \qquad \text{for } i = A, P. \tag{10}$$

Also, for (10), it is implicitly assumed that the $x_{hik}$ are obtained either without any measurement error or with mutually independent measurement errors (which are ignorable by restriction of attention to the data as given). From (2) and (10), the variance of $x_{+i+}$ is

$$\text{var}(x_{+i+}) = \frac{\sum_{h=1}^{2} n_{hi}(n_h - n_{hi})v_h}{n_h - 1}. \tag{11}$$

Thus an approximate test statistic for the comparison of the two treatment groups for the combined centers is

$$\begin{aligned} Q_{EMH} &= \frac{(x_{+i+} - E\{x_{+i+}\})^2}{\text{var}\{x_{+i+}\}} \\ &= \frac{[\sum_{h=1}^{q}(n_{hA}n_{hP}/n_h)(\bar{x}_{hA} - \bar{x}_{hP})]^2}{\sum_{h=1}^{q}(n_{hA}n_{hP}/n_h)^2\tilde{v}_h}, \end{aligned} \tag{12}$$

where $q = 2$ is the number of centers (or strata), $\tilde{v}_h = \text{var}\{\bar{x}_{hA} - \bar{x}_{hP}\}$ and $x_{+i+}$ can refer to either $i = A$ or $i = P$. This criterion is often called the extended Mantel–Haenszel statistic. It approximately has the chi-square distribution with d.f. = 1 when the two treatment groups have sufficiently large sample sizes $\{n_{+i}\}$ for the combined centers (*e.g.*, $n_{+i} \geq 20$). Small $p$-values (e.g., $p \leq 0.05$) relative to this chi-square approximation indicate background variables for which atypically large differences between the two treatment groups for the combined centers resulted from randomization. For situations with small samples, the determination of $p$-values for $Q_{EMH}$ needs to be based on the exact distribution of $x_{+i+}$ across all possible randomizations of patients to treatments at the two centers. Regardless of whether sample sizes are large or small, the $p$-values for $Q_{EMH}$ are adjusted for centers in the sense of being based on the within-center differences of the means $(\bar{x}_{hA} - \bar{x}_{hP})$ for the two treatments.

The definition of $Q_{EMH}$ in (12) for dichotomous background variables with possible categories $j = 1, 2$ provides the usual Mantel–Haenszel (Mantel and Haenszel, 1959) statistic,

$$\begin{aligned} Q_{MH} &= \frac{[\sum_{h=1}^{q}(n_{hA}n_{hP}/n_h)(p_{hA1} - p_{hP1})]^2}{\sum_{h=1}^{q}[n_{hA}n_{hP}n_{h+1}n_{h+2}/n_h^2(n_h - 1)]} \\ &= \frac{[\sum_{h=1}^{q}(n_{hA1} - m_{hA1})]^2}{\sum_{h=1}^{q}[n_{hA}n_{hP}n_{h+1}n_{h+2}/n_h^2(n_h - 1)]}, \end{aligned} \tag{13}$$

where $q = 2$, $n_{h+j} = (n_{hAj} + n_{hPj})$ and $m_{hij} = (n_{hi}n_{h+j}/n_h)$. In (13), $Q_{MH}$ is specified relative to the $p_{hi1}$, but the same result would be obtained relative to the $p_{hi2}$; also, $Q_{MH}$ would remain the same if the $(n_{hA1} -$

$m_{hA1}$) were replaced by the $(n_{hij} - m_{hij})$ for some other $i$, $j$. Through its definition in terms of frequencies $\{n_{hij}\}$, $Q_{MH}$ is a method for the combined analysis of a set of $(2 \times 2)$ contingency tables like those shown in (5) for treatment × sex at each center. A criterion due to Mantel and Fleiss (1980) is available for confirming the appropriateness of the chi-square approximation to the distribution of $Q_{MH}$ for this data structure. It is that the difference between the across-center sum of expected values $\{\sum_{h=1}^{2} m_{hij}\}$ for any $i$, $j$ and both the minimum possible value and the maximum possible value for the corresponding sums of the observed values exceed 5. For the situations where the Mantel and Fleiss (1980) criterion is not satisfied, the methods reviewed in Gart (1971) can be used to determine an exact $p$-value for $Q_{MH}$. Algorithms for the computation of such exact $p$-values or other types of exact results for sets of $(2 \times 2)$ contingency tables are discussed in Thomas (1975) and Mehta et al. (1985).

The extended Mantel–Haenszel statistic in (12) was proposed by Mantel (1963) for the comparison of two groups with respect to an ordinal variable in a way that adjusts for a set of strata. Although (12) expressed how $Q_{EMH}$ is obtained from the case record data $\{x_{hik}\}$ for patients, a common framework for its usage is a set of $(2 \times r)$ contingency tables for group × ordinal variable. In this setting, computation of $Q_{EMH}$ would also be based on (12), but with the modifications that the

$$\bar{x}_{hi} = \frac{\sum_{j=1}^{r} a_{hj} n_{hij}}{n_{hi}} \tag{14}$$

become means with respect to scores $\{a_{hj}\}$ at the $h$th center for the numbers of patients $\{n_{hij}\}$ with the $j$th category of the observed variable in the $i$th treatment group at the $h$th center and

$$\begin{aligned} \tilde{v}_h &= \frac{n_h}{n_h - 1}\left(\frac{1}{n_{hA}} + \frac{1}{n_{hP}}\right) v_h \\ &= \left(\frac{1}{n_{hA}} + \frac{1}{n_{hP}}\right) \frac{\sum_{j=1}^{r} (a_{hj} - \mu_h)^2 n_{h+j}}{n_h - 1} \end{aligned} \tag{15}$$

where $n_{h+j} = n_{hAj} + n_{hPj}$ and $\mu_h = \sum_{j=1}^{r} a_{hj} n_{h+j}/n_h$. As discussed in Section 2.1 for the within-center mean score statistic $Q_{S,h}$ in (3), analyses based on ranks have important advantages. A nonparametric rank procedure with a locally most powerful property was proposed by van Elteren (1960) and is discussed by Lehmann (1975). It has essentially the same structure as $Q_{EMH}$ relative to scores that are called within-center, standardized midranks here [and modified ridits in the documentation for the **FREQ** Procedure in the SAS System (1985)]. The definition of these scores

with respect to a set of ordinal categories is

$$a_{hj} = \begin{cases} 0 & \text{for all } j \text{ such that } \sum_{j'=1}^{j} n_{h+j'} = 0 \\ \dfrac{\sum_{j'=1}^{j} n_{h+j'} - (1/2)n_{h+j} + 1/2}{n_h + 1} & \text{for all } j \text{ such that } \sum_{j'=1}^{j} n_{h+j'} \geq 0. \end{cases} \tag{16}$$

It also applies to variables with continuous distributions through the special case where all $n_{h+j} = 1$. Thus the extended Mantel–Haenszel statistic for within-center, standardized midranks is the categorical data counterpart to the van Elteren statistic.

Consistent tendencies for one treatment group to have larger means than the other treatment group across the strata (centers) are the types of imbalance which the Mantel–Haenszel statistic most effectively detects for the distribution of a background variable. This aspect of its performance is a consequence of the differences $(\bar{x}_{hA} - \bar{x}_{hP})$ reinforcing one another in the numerator of $Q_{EMH}$ when they predominantly have the same sign (i.e., nearly all are positive or nearly all are negative). Accordingly, $Q_{EMH}$ is often said to provide a test of average partial association [see Landis et al. (1978)]. Aside from the previous considerations about the optimal setting for its performance, $Q_{EMH}$ is a valid test statistic for the comparison of treatment groups for a combined set of strata regardless of the pattern of differences across them. Thus use of $Q_{EMH}$ has the important advantage of being unconditionally specifiable in the protocol for a study (i.e., at a time prior to data collection or data analysis).

When the directions of the differences between treatments for a background variable conflict to the extent that positive ones offset negative ones, a potential limitation of $Q_{EMH}$ is an inability to detect the extent to which their absolute magnitudes for the respective strata might collectively suggest an atypical distribution from randomization. Two more effective methods for evaluating this pattern for the means $\bar{x}_{hi}$ are the total association, mean score statistic

$$Q_{S,T} = \sum_{h=1}^{q} Q_{S,h}, \tag{17}$$

where $q = 2$, and the pseudohomogeneity statistic

$$Q_{S,PH} = Q_{S,T} - Q_{EMH}. \tag{18}$$

When the sample sizes within each center are sufficiently large (e.g., all $n_{hi} \geq 15$) for the $Q_{S,h}$ to have approximate chi-square distributions with d.f. = 1, then $Q_{S,T}$ and $Q_{S,PH}$ approximately have the chi-square distributions with d.f. = $q = 2$ and d.f. = $(q-1) = 1$, respectively. This sample size requirement is more stringent than that for $Q_{EMH}$ relative to the sample sizes $\{n_{+i}\}$ for the combined strata, so $Q_{EMH}$ has the advantage that its chi-square approximation is applicable in a broader range of situations than those for $Q_{S,T}$ or $Q_{S,PH}$; these include the case of matched pairs where there is $n_{hi} = 1$ observation per treatment and for which $Q_{EMH}$ is analogous to the paired $t$-test for continuous data and the sign test (or McNemar's test) for dichotomous data.

Another reason why $Q_{EMH}$ is used more extensively than $Q_{S,T}$ or $Q_{S,PH}$ is that the consistent patterns of imbalance which it is better able to detect are typically of greater interest than the more general patterns which it may be less able to detect. Regardless of their somewhat different capabilities, there is merit in evaluating the results for $Q_{EMH}$, $Q_{S,T}$, and $Q_{S,PH}$ together when within-center sample sizes are sufficiently large for such analysis. For these situations, $Q_{S,T}$ is directed at general patterns of treatment differences which may or may not have consistent direction; $Q_{EMH}$ is directed at the specific pattern of differences with consistent direction; and $Q_{S,PH}$ is directed at patterns of differences which are not encompassed by $Q_{EMH}$ in the sense of its definition in (18). However, $Q_{S,PH}$ needs to be interpreted cautiously because it is not a test of homogeneity of treatment differences across the strata (or centers), although it can often shed light on whether such homogeneity seems to apply. Appropriate methods for evaluating homogeneity of treatment differences are described in Section 3 through tests of treatment × center interaction in statistical models. Other discussion of the roles of $Q_{EMH}$, and $Q_{S,PH}$ is given in Koch et al. (1985a).

Statistical results pertaining to comparisons between treatment groups for the combined centers are summarized for the background variables in the last column of Table 2. For sex, application of the Mantel–Haenszel statistic in (13) to the set of $(2 \times 2)$ contingency tables shown in (5) yielded $Q_{MH} = 6.15$; so $p = 0.013$ relative to the chi-square distribution with d.f. = 1. A chi-square approximation was considered reasonable here because the Mantel and Fleiss (1980) criterion was satisfied. More specifically, the minimum and maximum possible values for $(n_{1AF} + n_{2AF})$ are 0 and 23, and both are different from $(m_{1AF} + m_{2AF}) = 11.23$ by at least 5. If consideration is given to the exact distribution for (5) through the algorithm of Thomas (1975), one-sided $p = 0.011$ is obtained. Both this result and that from the chi-square approximation agree in indicating that an atypically larger percentage of males were randomly assigned to active treatment than to placebo for the combined centers.

When the one-sided exact $p$-value requires a close approximation (and is not available in its own right), some references, such as Breslow and Day (1980) and Fleiss (1981), recommend usage of a continuity correction; that is, (13) is modified to

$$Q_{MH,C} = \frac{[|\sum_{h=1}^{q}(n_{hA1} - m_{hA1})| - 0.5]^2}{\sum_{h=1}^{q}[n_{hA}n_{hP}n_{h+1}n_{h+2}/n_h^2(n_h - 1)]}, \tag{19}$$

and the one-sided $p$-value from its chi-square approximation with d.f. = 1 is determined.

The total association statistic in (17) for the comparison of the sex distributions for the two treatments is $Q_{S,T} = 6.36$ with d.f. = 2; so the pseudohomogeneity statistic in (18) is $Q_{S,PH} = (6.36 - 6.15) = 0.21$ with d.f. = 1. Since $Q_{S,PH}$ is a relatively small component of $Q_{S,T}$, these results indicate that the imbalance in sex distributions for the two treatments is due primarily to a consistent pattern of treatment differences (i.e., a larger percentage of males for $A$ than for $P$ for both centers). Chi-square approximations are not used here to determine $p$-values for $Q_{S,T}$ and $Q_{S,PH}$ because the sample sizes at center 1 are not considered large enough for this purpose (e.g., $m_{1iF} \leq 5$ for $i = A, P$).

The distribution of baseline status for the two treatment groups at the two centers is described by the following set of $(2 \times 5)$ contingency tables:

| Center 1 | Terrible (0) | Poor (1) | Fair (2) | Good (3) | Excellent (4) |
|---|---|---|---|---|---|
| Active (A) | 3 | 6 | 9 | 7 | 2 |
| Placebo (P) | 0 | 7 | 13 | 6 | 3 |

(20)

| Center 2 | Terrible (0) | Poor (1) | Fair (2) | Good (3) | Excellent (4) |
|---|---|---|---|---|---|
| Active (A) | 0 | 3 | 9 | 6 | 9 |
| Placebo (P) | 0 | 4 | 7 | 13 | 4 |

Since baseline status is an ordinal variable, its comparison for the two treatments is based on the extended Mantel–Haenszel statistic in (12) for the within-center, standardized midrank scores in (16). This method corresponds to the van Elteren statistic. For center 1 the standardized midrank scores are $a_{1j} = (2/57), (10/57), (27.5/57), (45/57), (54/57)$; and for center 2, they are $a_{2j} = (0), (4/56), (15.5/56), (33/56), (49/56)$. The result of this analysis was $Q_{EMH} = 0.001$, for which $p = 0.98$. It clearly confirms the fact that randomization has provided similar distributions of baseline status for the two treatment groups at the two centers. Such similarity was also evident for dichotomous baseline status on the basis of $Q_{MH} = 0.023$

with $p = 0.88$, and for age on the basis of $Q_{EMH} = 0.122$ with $p = 0.73$. Thus, among the four background variables for which statistical comparisons between treatments were evaluated for the combined centers, only sex exhibited a noteworthy imbalance. Although such a finding is not a substantial departure from what might be expected from randomization (relative to the framework of four statistical tests), it does merit some concern by identifying the possibility that any treatment difference for the response variables might be due partly to differences in the sex distributions for the two treatment groups. Analyses to address this issue are undertaken in Sections 2.3, 3.2, and 3.3.

A description is given in Table 2 for the distributions of background variables for all patients in each treatment group through the means $\bar{x}_{+i+}$ for the pooled data from the two centers and their corresponding standard errors $(\text{S.E.})_{+i}$ relative to stratified sampling from an infinite population. These results were obtained via

$$\bar{x}_{+i+} = \frac{\sum_{h=1}^{2} n_{hi}\bar{x}_{hi}}{n_{+i}}, \qquad (\text{SE})_{+i} = \left[\frac{\sum_{h=1}^{2} n_{hi}^2(\text{SE})_{hi}^2}{n_{+i}^2}\right]^{1/2}, \tag{21}$$

where the $(\text{SE})_{hi}$ are defined by (6) for dichotomous variables and by (7) for variables with reasonably meaningful scores for ordinal categories or measured numerical values on a continuum. The $\bar{x}_{+i+}$ are the quantities at which $Q_{EMH}$ in (12) is directed; but the $(\text{SE})_{+i}$ are not based on (11), but rather are stratified counterparts to (6) and (7) in the sense of being based on the data for the $i$th treatment within each of the respective centers and not involving any finite population corrections.

## 2.3 Association Between Background Variables and Response Variables

An atypically large difference in the sex distributions for the two treatment groups was identified through the analyses in Sections 2.1 and 2.2. As discussed in Section 1, comparisons between treatments for response variables would need to be adjusted for sex if important associations between sex and the response variables existed in the study population (to avoid potential bias from the imbalance in sex distributions favoring one of the treatments). Since sex is a dichotomous variable, one way to evaluate its effects under minimal assumptions is through nonparametric statistical tests like the mean score statistic in (3) for the separate centers and the extended Mantel–Haenszel statistic in (12) for the combined centers. For such analysis, sex defines the groups, and comparisons are directed

at distributions of the response variables at each of the four visits. However, sex is a known characteristic of patients at the time of entry to a study rather than a randomly allocated condition, so the application of (3) and (12) to it requires somewhat different justification than that which applies to treatment comparisons. One strategy here is to argue that no association between sex and a response variable is hypothetically equivalent to sex categories being perceivable as randomly assigned labels for response distributions. In other words, randomization is invoked by hypothesis; and this provides the framework for sex comparisons with the same statistical properties as that discussed for treatment comparisons in Sections 2.1 and 2.2. Thus under the hypothesis of no association between response variables and groups based on sex (or those based on any other background variable), the counterparts to the mean score statistic in (3) for the separate centers and the extended Mantel–Haenszel statistic in (12) for the combined centers approximately have chi-square distributions when sample sizes are sufficiently large; also, exact methods are applicable for small-sample situations. The occurrence of small $p$-values for these methods is interpreted as contradicting the hypothesis of no association which provided the basis for their determination, thereby indicating the presence of an association. Additional discussion of this analysis strategy is given in Landis, et al. (1978) and Koch et al. (1980b).

Approximate $p$-values from the analyses of the association between sex and the response variables at visits 1 to 4 are shown in the upper part of Table 3. For the separate centers, these results are obtained from the Wilcoxon rank sum statistic through its specification in (3) as a mean score statistic; and for the combined centers, they are from the extended Mantel–Haenszel statistic in (12) for within-center, standardized midranks (i.e., the van Elteren statistic). The $\geq 0.10$ status of all $p$-values in Table 3 for sex comparisons suggests that there is little or no association between sex and response variables. On this basis, the imbalance in sex distributions for the two treatments is interpreted as ignorable in the sense of only inducing negligible bias on treatment comparisons for response variables. Thus analyses of treatment effects are not considered to need any adjustment for sex. Nevertheless, the role of sex is evaluated further in Section 3.4 and found to have a possible interaction with treatment at visits 3 and 4; for this reason, the discussion there does include analyses that are adjusted for sex.

Since baseline status reflects the condition of patients prior to treatment, its association with response variables is of natural interest. A relevant consideration for the analysis of such association is that both baseline status and the response variables have ordinal measurement scales. A method that effectively accounts for this data structure within each center sepa-

**Table 3** $p$-Values for Association of Sex and Baseline Status with Response Variables at Visits 1, 2, 3, and 4[a]

| Background variable | Response variable | Degrees of freedom | Center 1 | Center 2 | Combined centers |
|---|---|---|---|---|---|
| Sex | Visit 1 | 1 | 0.36 | 0.61 | 0.34 |
| | Visit 2 | 1 | 0.47 | 0.24 | 0.17 |
| | Visit 3 | 1 | 0.87 | 0.30 | 0.47 |
| | Visit 4 | 1 | 0.77 | 0.59 | 0.54 |
| Baseline status | Visit 1 | 1 | $< 0.001$[b] | 0.012[c] | $< 0.001$[b] |
| | Visit 2 | 1 | $< 0.001$[b] | 0.29 | $< 0.001$[b] |
| | Visit 3 | 1 | $< 0.001$[b] | 0.15 | $< 0.001$[b] |
| | Visit 4 | 1 | $< 0.001$[b] | 0.059 | $< 0.001$[b] |

[a]The $p$-values for sex are based on the Wilcoxon rank sum statistic via (3) for centers 1 and 2 separately and on the extended Mantel–Haenszel statistic (12) relative to within-center standardized ranks (or the van Elteren statistic) for the combined centers; midranks were used to account for ties via (16). The $p$-values for baseline status are based on the Spearman rank correlation statistic (25) for centers 1 and 2 separately and on its Mantel–Haenszel extension (26) with standardized ranks for the combined centers. Computations were performed with the FREQ Procedure in the SAS System (1985).
[b]Relationships with $p \leq 0.01$.
[c]Relationships with $p \leq 0.05$.

rately is the Spearman rank correlation statistic (with midranks applied to ties). The relevant quantities for the expression of this criterion for categorical data are frequencies $\{n_{hh'j}\}$ for the frequency of the $h'$th category of baseline status and the $j$th category for a response variable at the $h$th center, standardized midranks $\{c_{hh'}\}$ as scores for baseline status, and standardized midranks $\{a_{hj}\}$ as scores for the response variable. Then let

$$f_h = \frac{\sum_{h'=0}^{4} \sum_{j=0}^{4} c_{hh'} a_{hj} n_{hh'j}}{n_h}. \tag{22}$$

Relative to the perspective that the hypothesis $H_0$ of no association between baseline status and a response variable for each center is equivalent

to each being randomly distributed relative to the other, it follows that

$$\begin{aligned} E\{f_h \mid H_0\} &= \frac{\sum_{h'=0}^{4}\sum_{j=0}^{4} c_{hh'}a_{hj}n_{hh'+}n_{h+j}}{n_h^2} \\ &= \frac{\sum_{h'=0}^{4} c_{hh'}n_{hh'+}}{n_h}\,\frac{\sum_{j=0}^{4} a_{hj}n_{h+j}}{n_h} \\ &= \mu_{hc}\mu_{ha}, \end{aligned} \tag{23}$$

$$\begin{aligned} \operatorname{var}\{f_h \mid H_0\} &= \frac{\sum_{h'=0}^{4}(c_{hh'}-\mu_{hc})^2 n_{hh'+}}{n_h}\,\frac{\sum_{j=0}^{4}(a_{hj}-\mu_{ha})^2 n_{h+j}}{n_h(n_h-1)} \\ &= \frac{v_{hc}v_{ha}}{n_h(n_h-1)}, \end{aligned} \tag{24}$$

where $n_{hh'+} = \sum_{j=0}^{4} n_{hh'j}$ and $n_{h+j} = \sum_{h'=0}^{4} n_{hh'j}$. The test statistic that emerges from this framework is

$$\begin{aligned} Q_{CS,h} &= \frac{(f_h - E\{f_h \mid H_0\})^2}{\operatorname{var}\{f_h \mid H_0\}} \\ &= \frac{(n_h-1)[\sum_{h'=0}^{4}\sum_{j=0}^{4}(c_{hh'}-\mu_{hc})(a_{hj}-\mu_{ha})n_{hh'j}]^2}{n_h^2 v_{hc}v_{ha}} \\ &= (n_h-1)r_{ca,h}^2, \end{aligned} \tag{25}$$

where $r_{ca,h}$ denotes the correlation coefficient between the baseline status scores and the response variable scores for the $h$th center. When standardized midranks are used for baseline status and the response variable, $r_{ca,h}$ becomes the Spearman rank correlation coefficient. Since $Q_{CS,h}$ approximately has the chi-square distribution with d.f. = 1 when the sample size for the corresponding center is large (e.g., $n_h \geq 30$), it is often called a correlation chi-square statistic.

A synthesis of the principles underlying the $\{Q_{CS,h}\}$ in (25) and the extended Mantel–Haenszel statistic in (12) provides the basis for a test statistic suggested by Mantel (1963) for the association between two ordinal variables for a combined set of strata (e.g., centers). It is given by

$$\begin{aligned} Q_{CSMH} &= \frac{[\sum_{h=1}^{q} n_h(f_h - E\{f_h \mid H_0\})]^2}{\sum_{h=1}^{q} n_h^2 \operatorname{var}\{f_h \mid H_0\}} \\ &= \frac{[\sum_{h=1}^{q} n_h(v_{hc}v_{ha})^{1/2}r_{ca,h}]^2}{\sum_{h=1}^{q}[n_h^2 v_{hc}v_{ha}/(n_h-1)]}, \end{aligned} \tag{26}$$

where $q = 2$ is the number of strata. Sufficiently large sample size for the combined strata (e.g., $\sum_{h=1}^{q} n_h \geq 40$) supports usage of an approximate chi-square distribution with d.f. = 1 for $Q_{CSMH}$.

The lower part of Table 3 displays approximate $p$-values for statistical tests of the association between baseline status and the response variables at visits 1 to 4. The results for the separate centers are based on the Spearman rank correlation chi-square statistic in (25), and those for the combined centers are based on its extended Mantel–Haenszel counterpart in (26) for within-center, standardized midranks. For the combined centers, all $p$-values have $\leq 0.01$ status, so the hypothesis of no association between baseline status and response variables is clearly contradicted. This conclusion applies equally strongly to center 1, but it has somewhat weaker support in center 2. Further evaluation of differences between centers for the relationship of response variables to baseline status requires methods analogous to the pseudohomogeneity statistic in (18) or usage of appropriate statistical models along the lines described in Section 3. Aside from this issue, the strong relationship between baseline status and the response variables for the combined centers suggests that treatment comparisons for response variables would be more effective (in the sense of involving relatively smaller variances) if they were adjusted for baseline status.

Other nonparametric strategies are available for the analysis of the association between background variables and response variables. Comparisons among more than two groups can be based on the Kruskal–Wallis statistic (Kruskal and Wallis, 1953) for the separate centers and its extended Mantel–Haenszel counterpart for the combined centers. Also, scores other than within-center, standardized midranks can be used, and strata can be based on treatment × center or other cross-classifications. For further discussion of extended Mantel–Haenszel statistics and related nonparametric methods based on randomization principles, see Landis et al. (1978, 1979), Koch and Bhapkar (1982), Koch et al. (1980b, 1982, 1985a), and Koch and Edwards (1987).

## 2.4 Computations

The $p$-values in Tables 2 and 3 were obtained with the FREQ Procedure in the SAS System (1985). The MEANS Procedure was used to determine descriptive statistics for the separate centers, and algorithms in the IML Procedure were used to determine descriptive statistics for the combined centers. Additional documentation for computations is given in the footnotes of Tables 2 and 3 and in the appendix.

## 3 UNIVARIATE METHODS FOR RESPONSE VARIABLES

As discussed at the beginning of Section 2, the analysis of a categorical response variable in a clinical trial often requires the use of more than one method in order to address the questions of statistical interest. These questions include:

a. Is there a difference between treatments for a response variable under minimal assumptions and no (or only necessary) adjustment for background variables?
b. Is there a difference between treatments for a response variable after adjustments for appropriate background variables and under minimal assumptions?
c. Is any difference between treatments for a response variable homogeneous across centers?
d. Is any difference between treatments for a response variable homogeneous across background variables?

The rationale for question a is that randomization usually is successful in providing the treatment groups with equivalent distributions of background variables. Moreover, the imbalances that may occasionally occur are often ignorable because the corresponding background variables have little or no association with the response variables (e.g., sex for the data in Table 1) or have offsetting associations (i.e., one may favor active treatment while another may favor placebo). Thus, for most situations, no adjustment for any background variable is necessary to avoid bias in treatment comparisons from imbalances, and any that is applied may have the potential limitation of seeming judgmental and thereby debatable unless its role was well justified (e.g., by inclusion in the protocol for a study or by the argument that substantial bias from no adjustment will be misleading, etc.). In Section 3.1, analysis of the direct comparisons in question a is undertaken with minimal assumptions for each response variable by the same nonparametric statistical tests that were discussed in Section 2. These methods and certain refinements are also applicable to the adjusted comparisons between treatments in question b. The objective of the latter analyses, which are discussed in Section 3.2, is the confirmation that conclusions are as well supported when adjustments for relevant background variables are applied as when they are not. For the data in Table 1, adjustment for sex is evaluated because of the imbalance in its distribution in the two treatment groups; and adjustment for baseline is evaluated because of the strong association between baseline and the response variable.

Statistical models are used to address questions c and d. For these analyses, the patients in the study are assumed to represent a general target population along the lines indicated in part II at the beginning of Section 2. The homogeneity of treatment effects across centers in question c is evaluated in Section 3.3 by determining whether a model with components for treatment and center needs to be expended to include the treatment × center interaction. Such analysis is applied to the dichotomous response variables of (good or excellent) or not through maximum likelihood procedures for fitting logistic regression models. For the scaling of the ordinal classification of response with successive integers, it is applied through weighted least squares procedures for fitting linear models. The evaluation of the homogeneity of treatment differences across background variables is undertaken in Section 3.4 with similar strategies. Attention is given there to whether relatively comprehensive models with components for treatment, center, and background variables also need to include treatment × background variable interactions. Logistic regression is used for such analysis of the dichotomous response variable of (good or excellent) or not, and its extension to the proportional odds model [as discussed in McCullagh (1980) and Harrell (1986)] is used for the ordinal expression of each response variable.

## 3.1 Direct Comparisons Between Treatments Under Minimal Assumptions

The primary conceptual difference between response variables and background variables is that the response variables for a patient may be substantially influenced by the treatment that is received. However, under the hypothesis $H_0$ of equivalence of treatment effects for each patient (i.e., the responses of each patient to the assigned treatment are identical to what would occur for the other treatment), the statistical framework provided by the randomization process in the study design is essentially the same for both response variables and background variables. In other words, the observed values of a response variable for all patients under $H_0$ constitute a finite population (which would hypothetically remain the same relative to all possible randomizations), and those for the two treatment groups are corresponding stratified simple random samples. Thus the nonparametric statistical tests described in Sections 2.1 and 2.2 for comparisons between treatment groups for background variables are also appropriate for such analyses of response variables. However, the role they serve for response variables involves somewhat different considerations. One is that maintenance of study design integrity is presumed; the basis for this includes double blinding and other mechanisms for avoiding bias in treatment com-

parisons from possible associations between response variables and inconsistencies in study management or measurement procedures. Another is that a small $p$-value $\leq 0.05$ is interpreted as a contradiction to the equivalence of treatment groups for the distribution of a response variable. Both together enable the contradiction to apply to the hypothesis $H_0$ of equivalence of treatment effects, and thereby support the evaluation of $p$-values $\leq 0.05$ as indicative of a significant difference in treatment effects.

Results for the comparison of the two treatment groups are given in Table 4 for the dichotomous response variables of (good or excellent) or not at visits 1 to 4. For each center × treatment × visit, the percentages of patients with good or excellent response are reported with their corresponding standard errors. These descriptive statistics were obtained from the frequencies $\{n_{ghij}\}$ of the $j$th response category at the $g$th visit for patients in the $i$th treatment group at the $h$th center via

$$\begin{aligned}(\%)_{ghi} &= 100p_{ghi} = 100\left(\frac{\sum_{j=3}^{4} n_{ghij}}{n_{hi}}\right), \\ (\mathrm{SE})_{ghi} &= 100\left[\frac{p_{ghi}(1-p_{ghi})}{n_{hi}-1}\right]^{1/2}\end{aligned} \tag{27}$$

As was the case with (6) for dichotomous background variables, the standard errors in (27) apply to random samples for infinite populations. Within each center, the $p$-values for each visit are based on Fisher's exact test. Those at visits 1 to 3 for center 2 are $\leq 0.05$, and thus indicate that the percentages of good or excellent response for $A$ in these settings are significantly larger than those for $P$. A directional tendency for good or excellent response to be more prevalent for $A$ than for $P$ also is apparent for each of the other center × visit combinations. The comparisons of the treatment groups for the combined centers are based on the Mantel–Haenszel statistic (13). The approximate $p$-values from this method are significant ($p \leq 0.05$) at visits 1 to 3 and nearly so at visit 4 ($p = 0.063$). These results provide the principal basis for the conclusion that the percentages of good or excellent response are significantly greater for $A$ than $P$ since they are based on all patients in the study population. They also do this more effectively than their within-center counterparts because of the consistency of the pattern of treatment differences across centers. A descriptive statistic for the treatment difference at which the Mantel–Haenszel statistic is directed for the $g$th visit response is

$$d_g = 100\left[\frac{\sum_{h=1}^{2} w_h(p_{ghA} - p_{ghP})}{\sum_{h=1}^{2} w_h}\right] = 100(p_{g*A} - p_{g*P}), \tag{28}$$

**Table 4** Descriptive Statistics and $p$-Values from Treatment Comparisons for Percentages of Patients with Good or Excellent Response at Visits 1, 2, 3, and 4

| Response variable | Treatment | Statistic[a] | Center 1 | Center 2 | Combined centers[b] |
|---|---|---|---|---|---|
| Visit 1 | Active (A) | % ≥ good | 51.9 | 85.2 | 68.4 |
| | | SE | 9.8 | 7.0 | 6.0 |
| | Placebo (P) | % ≥ good | 41.4 | 57.1 | 49.2 |
| | | SE | 9.3 | 9.5 | 6.7 |
| | A vs. P | Difference | 10.5 | 28.0 | 19.2 |
| | | SE | 13.5 | 11.8 | 9.0 |
| | | $p$-value | 0.59 | 0.037[c] | 0.036[c] |
| Visit 2 | Active (A) | % ≥ good | 59.3 | 81.5 | 70.3 |
| | | SE | 9.6 | 7.6 | 6.2 |
| | Placebo (P) | % ≥ good | 34.5 | 42.9 | 38.6 |
| | | SE | 9.0 | 9.5 | 6.5 |
| | A vs. P | Difference | 24.8 | 38.6 | 31.6 |
| | | SE | 13.2 | 12.2 | 9.0 |
| | | $p$-value | 0.11 | 0.005[d] | 0.001[d] |
| Visit 3 | Active (A) | % ≥ good | 63.0 | 81.5 | 72.1 |
| | | SE | 9.5 | 7.6 | 6.1 |
| | Placebo (P) | % ≥ good | 41.4 | 50.0 | 45.7 |
| | | SE | 9.3 | 9.6 | 6.7 |
| | A vs. P | Difference | 21.6 | 31.5 | 26.5 |
| | | SE | 13.3 | 12.3 | 9.0 |
| | | $p$-value | 0.12 | 0.023[c] | 0.005[d] |
| Visit 4 | Active (A) | % ≥ good | 44.4 | 77.8 | 61.0 |
| | | SE | 9.7 | 8.2 | 6.4 |
| | Placebo (P) | % ≥ good | 31.0 | 57.1 | 44.0 |
| | | SE | 8.7 | 9.5 | 6.5 |
| | A vs. P | Difference | 13.4 | 20.6 | 17.0 |
| | | SE | 13.1 | 12.5 | 9.1 |
| | | $p$-value | 0.41 | 0.15 | 0.063 |

[a]The $p$-values are based on Fisher's exact tests [as described relative to (5)] for centers 1 and 2 separately and on the Mantel–Haenszel statistic (13) for the combined centers. Computations were performed with the FREQ Procedure in the SAS System (1985).

[b]Descriptive statistics for the combined centers are based on weighted averages (28)–(30) for which the weights reflect adjustment for centers through their relative contributions to the Mantel–Haenszel statistic; that is, the weight for the $h$th center is $w_h/(w_1 + w_2)$ with $w_h = n_{h1}n_{h2}/(n_{h1} + n_{h2})$, where the $n_{hi}$ denote the numbers of patients in the $i$th treatment group at the $h$th center.

[c]Treatment comparisons with $p \leq 0.05$.

[d]Treatment comparisons with $p \leq 0.01$.

where $w_h = n_{h1}n_{h2}/(n_{h1} + n_{h2})$; its standard error relative to stratified random sampling from an infinite population is

$$\mathrm{SE}(d_g) = 100\left(\frac{\left\{\sum_{h=1}^{2} w_h^2[(\mathrm{SE})_{ghA}^2 + (\mathrm{SE})_{ghP}^2]\right\}^{1/2}}{\sum_{h=1}^{2} w_h}\right); \tag{29}$$

corresponding statistics that describe each treatment group's percentage of good or excellent response for the combined centers are

$$(\%)_{g*i} = 100(p_{g*i}), \qquad (\mathrm{SE})_{g*i} = \frac{[\sum_{h=1}^{2} w_h^2(\mathrm{SE})_{ghi}^2]^{1/2}}{\sum_{h=1}^{2} w_h}, \tag{30}$$

where the $p_{g*i}$ are defined in (28). The $(\%)_{g*i}$, the $d_g$, and their standard errors are given in Table 4. The range for the percentage of good or excellent response over the four visits is about 40 to 50% for $P$ and about 60 to 70% for $A$. The difference between $A$ and $P$ has a range of about 20 to 30% and a standard error of about 10%.

The analysis of treatment comparisons for the ordinal expression of the response variables at visits 1 to 4 is similar to that which was described for baseline status in Sections 2.1 and 2.2. Means with respect to the integer scores 0, 1, 2, 3, 4 for the ordinal categories of terrible, poor, fair, good, and excellent, respectively, and their standard errors are used to describe the response distribution for each center × treatment × visit. These statistics were obtained from the frequencies $\{n_{ghij}\}$ in contingency tables such as (20) via

$$\bar{y}_{ghi} = \frac{\sum_{j=0}^{4} j n_{ghij}}{n_{hi}}, \qquad (\mathrm{SE})_{ghi} = \left[\frac{\sum_{j=0}^{4}(j - \bar{y}_{ghi})^2 n_{ghij}}{n_{hi}(n_{hi} - 1)}\right]^{1/2}. \tag{31}$$

Alternatively, if $y_{ghik}$ denotes the response at the $g$th visit for the $k$th patient in the $i$th treatment group at the $h$th center, they could have been computed directly via

$$\bar{y}_{ghi} = \frac{\sum_{k=1}^{n_{hi}} y_{ghik}}{n_{hi}}, \qquad (\mathrm{SE})_{ghi} = \left[\frac{\sum_{k=1}^{n_{hi}}(y_{ghik} - \bar{y}_{ghi})^2}{n_{hi}(n_{hi} - 1)}\right]^{1/2}. \tag{32}$$

In accordance with (6), (7), and (27), the standard errors in (31) and (32) apply to random samples from infinite populations in order to describe the observed experience of the patients for this general setting.

A potential concern for the interpretation of the means and standard errors in (31) or (32) is that the integer scores on which they are based are not necessarily clinically meaningful values. One way to address this

issue is to rewrite $\bar{y}_{ghi}$ in (31) as

$$\bar{y}_{ghi} = \frac{\sum_{j=0}^{4} j n_{ghij}}{n_{hi}} = \frac{\sum_{j=1}^{4}\sum_{j'=1}^{j} n_{ghij}}{n_{hi}}$$
$$= \frac{\sum_{j'=1}^{4}\sum_{j=j'}^{4} n_{ghij}}{n_{hi}} = \frac{\sum_{j'=1}^{4} N_{ghij'}}{n_{hi}} = \sum_{j'=1}^{4} P_{ghij'}, \tag{33}$$

where the $P_{ghij'}$ are the proportions of patients in the $i$th treatment group at the $h$th center with response at the $g$th visit at least as favorable as the $j'$th category. In view of (33), the difference between the treatment groups for the $g$th visit and $h$th center is

$$d_{gh} = (\bar{y}_{ghA} - \bar{y}_{ghP}) = \sum_{j=1}^{4}(P_{ghAj} - P_{ghPj}). \tag{34}$$

Through (34), the $\{d_{gh}\}$ have a meaningful interpretation as measures of the consistency with which the differences $\{(P_{ghAj} - P_{ghPj})\}$ across $j = 1$, 2, 3, 4, favor active treatment (or placebo). This rationale for the use of integer scores in the determination of the $d_{gh}$ is considered to provide reasonable support for their usage in the means $\bar{y}_{ghi}$. As a consequence of (31) and (34), the standard errors for the $d_{gh}$ are given by

$$(\mathrm{SE})_{gh*} = [(\mathrm{SE})^2_{ghA} + (\mathrm{SE})^2_{ghP}]^{1/2}. \tag{35}$$

For each center, the means $\bar{y}_{ghi}$, the differences $d_{gh}$, and their standard errors are given in Table 5 for the responses at each visit. Approximate $p$-values are also given there for the comparison between treatments through two specifications of the mean score statistic in (3). One of these is based on the integer scores 0, 1, 2, 3, 4 for the ordinal categories and hence has the advantage of being directed at the treatment difference that is described by the $d_{gh}$. The other is based on standardized midranks and hence is equivalent to the Wilcoxon rank-sum statistic; it has the advantage of only making use of the ordering of response categories rather than involving an explicit scaling. As often occurs in practice, the results from these methods support essentially the same conclusions. In Table 5, both methods indicate significantly ($p \leq 0.05$) more favorable response for $A$ than $P$ for all visits at center 2 and visit 2 at center 1; for the other visits at center 1, the positive values of the $d_{gh}$ indicate directional tendencies in favor of active treatment.

For the combined centers, the statistical comparisons between the treatments for the ordinal response variables are assessed with the extended Mantel–Haenszel statistic in (12). Results were obtained for both inte-

**Table 5** Descriptive Statistics and $p$-Values from Treatment Comparisons for Ordinal Response at Visits 1, 2, 3, and 4

| Response variable | Treatment | Statistic[a] | Center 1 | Center 2 | Combined centers[b] |
|---|---|---|---|---|---|
| Visit 1 | Active (A) | Mean | 2.52 | 3.33 | 2.92 |
| | | SE | 0.19 | 0.16 | 0.12 |
| | Placebo (P) | Mean | 2.24 | 2.82 | 2.53 |
| | | SE | 0.25 | 0.19 | 0.16 |
| | A vs. P | Difference | 0.28 | 0.51 | 0.39 |
| | | SE | 0.31 | 0.25 | 0.20 |
| | | $p$-value(int) | 0.38 | 0.043[c] | 0.053 |
| | | $p$-value(sr) | 0.45 | 0.045[c] | 0.052 |
| Visit 2 | Active (A) | Mean | 2.85 | 3.41 | 3.13 |
| | | SE | 0.19 | 0.19 | 0.13 |
| | Placebo (P) | Mean | 2.00 | 2.29 | 2.14 |
| | | SE | 0.25 | 0.25 | 0.17 |
| | A vs. P | Difference | 0.85 | 1.12 | 0.99 |
| | | SE | 0.31 | 0.31 | 0.22 |
| | | $p$-value(int) | 0.011[c] | 0.001[d] | $< 0.001$[d] |
| | | $p$-value(sr) | 0.016[c] | 0.001[d] | $< 0.001$[d] |
| Visit 3 | Active (A) | Mean | 2.81 | 3.30 | 3.05 |
| | | SE | 0.23 | 0.19 | 0.15 |
| | Placebo (P) | Mean | 2.24 | 2.21 | 2.23 |
| | | SE | 0.27 | 0.28 | 0.19 |
| | A vs. P | Difference | 0.57 | 1.08 | 0.83 |
| | | SE | 0.35 | 0.34 | 0.24 |
| | | $p$-value(int) | 0.11 | 0.004[d] | 0.001[d] |
| | | $p$-value(sr) | 0.13 | 0.005[d] | 0.002[d] |
| Visit 4 | Active (A) | Mean | 2.48 | 3.26 | 2.87 |
| | | SE | 0.20 | 0.24 | 0.16 |
| | Placebo (P) | Mean | 2.03 | 2.46 | 2.25 |
| | | SE | 0.24 | 0.31 | 0.19 |
| | A vs. P | Difference | 0.45 | 0.79 | 0.62 |
| | | SE | 0.31 | 0.39 | 0.25 |
| | | $p$-value(int) | 0.16 | 0.047[c] | 0.015[c] |
| | | $p$-value(sr) | 0.22 | 0.037[c] | 0.020[c] |

[a]The $p$-values are based on randomization chi-square statistics (3) to compare mean scores for treatments within centers 1 and 2 separately and on

ger scores and within-center, standardized midrank scores; use of the latter corresponds to the van Elteren statistic. Approximate $p$-values from these analyses are given in the last column of Table 5. For visits 2 to 4, they are significant ($p \leq 0.05$); and for visit 1, they are nearly significant ($p = 0.053, 0.052$). The extent to which the responses for active treatment are more favorable than those for placebo are described in a spirit similar to (28)–(29) with weighted linear combinations of the $d_{gh}$ and their corresponding standard errors. These statistics have the form

$$d_g = \frac{\sum_{h=1}^{2} w_h(\bar{y}_{ghA} - \bar{y}_{ghP})}{\sum_{h=1}^{2} w_h},$$

$$\mathrm{SE}(d_g) = \frac{\left\{\sum_{h=1}^{2} w_h^2[(\mathrm{SE})_{ghA}^2 + (\mathrm{SE})_{ghP}^2]\right\}^{1/2}}{\sum_{h=1}^{2} w_h}, \tag{36}$$

where $w_h = n_{h1}n_{h2}/(n_{h1} + n_{h2})$ and the $\bar{y}_{ghi}$ and the $(\mathrm{SE})_{ghi}$ are given in (32). The counterparts to (36) for describing each treatment group's response distributions for the combined centers are

$$\bar{y}_{g*i} = \frac{\sum_{h=1}^{2} w_h \bar{y}_{ghi}}{\sum_{h=1}^{2} w_h}, \qquad (\mathrm{SE})_{g*i} = \frac{[\sum_{h=1}^{2} w_h^2 (\mathrm{SE})_{ghi}^2]^{1/2}}{\sum_{h=1}^{2} w_h}. \tag{37}$$

The $\bar{y}_{g*i}$, the $d_g$, and their standard errors are given in Table 5. The difference between the two treatments over the four visits is indicated there to have a range of about 0.4 to 1.0 and a standard error of about 0.2.

The results in Table 4 and Table 5 have the attractive feature of being based on methods with minimal assumptions for which justification usually is reasonably supported by study design considerations. These assumptions were randomization, ignorability of blocks in the randomization, and maintenance of study design integrity. Of particular importance, no

---

the extended Mantel–Haenszel statistic (12) for such analysis of the combined centers. Results for both integer scores (int) and within center, standardized ranks (sr) via (16) are presented; the latter correspond to the Wilcoxon rank sum statistic for the separate centers and the van Elteren statistic for the combined centers. Computations were performed with the FREQ Procedure in the SAS System (1985).

[b] Descriptive statistics for the combined centers are based on weighted averages (36)–(37) for which the weights reflect adjustment for centers through their relative contributions to the extended Mantel–Haenszel statistic; that is, the weight for the $h$th center is $w_h/(w_1 + w_2)$ with $w_h = n_{h1}n_{h2}/(n_{h1} + n_{h2})$, where the $n_{hi}$ denote the numbers of patients in the $i$th treatment group at the $h$th center.

[c] Treatment comparisons with $p \leq 0.05$.

[d] Treatment comparisons with $p \leq 0.01$.

assumption about specific structure for the distribution of response variables within the two centers nor the homogeneity of treatment differences across centers is required. Thus the conclusions from Table 4 and Table 5 are design-based inferences. Another important advantage of the methods used to obtain the descriptive statistics and $p$-values for the treatment comparisons in Tables 4 and 5 is that they are a priori specifiable in the protocol of a study. In this role, the extended Mantel–Haenszel statistic is a valid method for the analysis of all patients in a way that is particularly effective for detecting treatment differences with consistent direction across centers. As indicated in Section 2.2, it is not useful when treatment differences predominantly have conflicting direction, but the detection of such patterns of contradictory results is rarely an objective of a clinical trial. When necessary, these situations can be addressed with either methods like the total association statistic in (17) or separate tests for each center. Most clinical trials, however, are undertaken to detect a consistent pattern of treatment differences, and their analysis for this purpose is well served by the extended Mantel–Haenszel statistic. Since its usage is a priori specifiable, involves only minimal assumptions, and encompasses all patients, the extended Mantel–Haenszel statistic is often the most appropriate method for the inferential analysis of categorical data from multicenter clinical trials.

## 3.2 Adjusted Comparisons Between Treatments Under Minimal Assumptions

In Section 2, an imbalance in the sex distributions for the treatment groups was identified. This imbalance merits attention since it might partially account for any treatment differences for the response variables. The extent of such a confounding influence is determined by the strength of the association between sex and the response variable. Since analyses reported in Section 2.3 indicate little or no association between sex and the response variables, the imbalance in the sex distributions for the treatment groups was interpreted as ignorable. Thus direct comparisons between treatments in Section 3.1 are considered to provide appropriate conclusions for the response variables.

Alternatively, if there had been substantial association between sex and the response variables, treatment comparisons for response variables would need to be adjusted for sex to avoid any bias from the imbalance in its distribution. Such analyses are undertaken here to confirm that they yield the same conclusions as the direct comparisons in Section 3.1. A way of implementing adjustment for a categorical background variable like sex for a nonparametric statistical test is through its use for the post-stratification of patients (i.e., sex is "held constant" in subgroups of the

center × sex × treatment cross-classification). Then treatment comparisons for response variables are combined across center × sex strata through the extended Mantel–Haenszel statistic. Results from such analysis (which involves minimal assumptions) are given in Table 6 for both the dichotomous and ordinal expressions of the response variables at visits 1 to 4; also, those for the ordinal response variables are based on within center × sex standardized midranks. The conclusions from Table 6 agree very well with those in Tables 4 and 5 from direct comparisons between treatments for the combined centers. For the ordinal response variables, the stratification adjusted $p$-values in Table 6 indicate that the differences between treatments are significant ($p \leq 0.05$) at visits 2 to 4 and nearly significant ($p = 0.066$) at visit 1. Those for the dichotomous response variable are significant ($p \leq 0.05$) at visits 2 and 3 and nearly significant at visits 1 and 4 ($p = 0.063, 0.081$).

Analysis of baseline status in Section 2 supported the following two conclusions: the two treatment groups had equivalent distributions of baseline status (see Table 2); and baseline status had strong relationships with the response variables (see Table 3). As a consequence of the former, no adjustment for baseline status is necessary for the avoidance of bias in treatment comparisons for response variables. However, the latter suggests that adjustment can strengthen statistical tests for treatment comparisons; this

**Table 6** $p$-Values for Association of Treatment with Response Variables at Visits 1, 2, 3, and 4 under Stratification Adjustment for Center and Sex

| Response variable | Degrees of freedom | Good or excellent response[a] | | Ordinal response[b] | |
|---|---|---|---|---|---|
| | | $Q_{MH}$ | $p$-value | $Q_{EMH}$ | $p$-value |
| Visit 1 | 1 | 3.46 | 0.063 | 3.37 | 0.066 |
| Visit 2 | 1 | 10.07 | 0.002[c] | 15.22 | $< 0.001$[c] |
| Visit 3 | 1 | 6.45 | 0.011[d] | 7.76 | 0.005[c] |
| Visit 4 | 1 | 3.04 | 0.081 | 4.57 | 0.032[d] |

[a]The $p$-values for (good or excellent response) are based on the Mantel–Haenszel statistic (13) with adjustment for center × sex strata.
[b]The $p$-values for the ordinal response are based on the extended Mantel–Haenszel statistic (12) relative to within center, standardized ranks via (16) and adjustment for center × sex strata (i.e., the van Elteren statistic).
[c]Treatment comparisons with $p \leq 0.01$.
[d]Treatment comparisons with $p \leq 0.05$.

occurs through a reduction in the relative variability of the applicable distributions for estimates of treatment differences.

Poststratification could be used to adjust for baseline status along the lines discussed previously in this section for sex. This strategy has two limitations: it does not account for the ordinality of baseline status, and the center × baseline status cross-classification produces 10 strata within many of which the numbers of patients may be too small to contribute effectively to the detection of a treatment difference. In this regard, any stratum for which all patients correspond to one treatment or one response category merits particular concern because of the zero values for their components in the numerator and denominator of the extended Mantel–Haenszel statistic; that is, they provide no information, and this essentially implies the exclusion of such patients from the analysis. The issue here is that when results are not effectively based on all patients, their ability to support conclusions may become debatable.

The previously stated limitations of poststratification provide the rationale for the consideration of another method for adjustment. It involves the synthesis of nonparametric covariance analysis principles [described in Quade (1967) for ranks] with the randomization framework of the extended Mantel–Haenszel test. For each center separately, covariance analysis is applied by the construction of the residuals from the ordinary least squares prediction of a response variable as a linear function of baseline status. These residuals for the response variable at the $g$th visit are given by

$$z_{ghik} = (y_{ghik} - \bar{y}_{gh*}) - \lambda_{gh}(x_{hik} - \mu_h); \tag{38}$$

here $\bar{y}_{gh*} = \sum_{i=A}^{P} \sum_{k=1}^{n_{hi}} y_{ghik}/n_h$ and $\mu_h = \sum_{i=A}^{P} \sum_{k=1}^{n_{hi}} x_{hik}/n_h$ denote the means for the response levels $y_{ghik}$ and baseline levels $x_{hik}$ for all patients in the finite study population at the $h$th center; and

$$\lambda_{gh} = \frac{\sum_{i=A}^{P} \sum_{k=1}^{n_{hi}} (y_{ghik} - \bar{y}_{gh*})(x_{hik} - \mu_h)}{\sum_{i=A}^{P} \sum_{k=1}^{n_{hi}} (x_{hik} - \mu_h)^2} \tag{39}$$

denotes the least squares slope for the linear prediction of the response at the $g$th visit by baseline status for the finite population of patients at the $h$th center. Under the hypothesis $H_0$ of equivalence of treatment effects for each patient (as discussed in Section 3.1), $\bar{y}_{gh*}$, $\mu_h$, and $\lambda_{gh}$ are constants that apply to all possible randomizations. Thus the observed values of the residuals $z_{ghik}$ for all patients constitute a finite population under $H_0$, and those for the two treatment groups within each center are simple random samples. These considerations provide the basis for the usage of the mean score statistic in (3) for the comparison of the two treatment groups with respect to the distribution of the response residuals $z_{ghik}$. This statistic

has the form

$$Q_{z,gh} = \frac{(\bar{z}_{ghA} - \bar{z}_{ghP})^2}{\text{var}\{(\bar{z}_{ghA} - \bar{z}_{ghP}) \mid H_0\}} = \frac{[(n_h - 1)/n_h][(\bar{y}_{ghA} - \bar{y}_{ghP}) - \lambda_{gh}(\bar{x}_{hA} - \bar{x}_{hP})]^2}{(1/n_{hA} + 1/n_{hP})(1 - r^2_{xg,h})v_{g,h}}, \tag{40}$$

where $v_{g,h} = \sum_{i=A}^{P} \sum_{k=1}^{n_{hi}} (y_{ghik} - \bar{y}_{gh*})^2/n_h$ and $r^2_{xg,h} = \lambda^2_{gh} v_h / v_{g,h}$, with $v_k$ given by (1). The test statistic $Q_{z,gh}$ in (40) approximately has the chi-square distribution with d.f. = 1 when the sample sizes for the two treatment groups are sufficiently large (e.g., the $n_{hi} \geq 15$). For the situation where the $y_{ghik}$ and the $x_{hik}$ are ranks, it is equivalent to the rank analysis of covariance statistic discussed in Quade (1967). Another noteworthy feature of $Q_{z,gh}$ is that its addition to the mean score statistic $Q_{x,h}$ from (3) for baseline status yields the bivariate mean score statistic $Q_{xg,h}$ for the comparison of the two treatment groups with respect to the response variable at the $g$th visit and baseline status simultaneously; that is,

$$Q_{xg,h} = Q_{x,h} + Q_{z,gh} = \mathbf{f}' \mathbf{V}_f^{-1} \mathbf{f},$$

where

$$\mathbf{f} = \begin{bmatrix} (\bar{y}_{ghA} - \bar{y}_{ghP}) \\ (\bar{x}_{hA} - \bar{x}_{hP}) \end{bmatrix}, \qquad \mathbf{V_f} = \text{var}(\mathbf{f} \mid H_0). \tag{41}$$

Under $H_0$, $Q_{xg,h}$ approximately has the chi-square distribution with d.f. = 2 for large-sample situations. The partition (41) is of interest because it enables $Q_{z,gh}$ to be interpreted as applying to the comparison of adjusted means of the response variables for the prediction setting where the treatment groups have the same mean for baseline status [see Koch et al. (1982) for further explanation].

The counterpart to the $Q_{z,gh}$ for the combined centers is formulated by applying the principles underlying the extended Mantel–Haenszel statistic in (12) to the $z_{ghik}$. It is given by

$$Q_{z,g} = \frac{[\sum_{h=1}^{q} (n_{hA} n_{hP}/n_h)(\bar{z}_{ghA} - \bar{z}_{ghP})]^2}{\sum_{h=1}^{q} [n_{hA} n_{hP}/(n_h - 1)](1 - r^2_{xg,h})v_{g,h}}, \tag{42}$$

where $q = 2$ is the number of strata (e.g., centers). An approximate chi-square distribution with d.f. = 1 applies to $Q_{z,g}$ under $H_0$ when the two treatment groups have sufficiently large sample sizes $n_{+i}$ for the combined centers (e.g., $n_{+i} \geq 20$). As summarized in this discussion, $Q_{z,g}$ is based on randomization, accounts for centers through stratification, and provides covariance adjustment for baseline status through the residuals $z_{ghik}$; so its

usage is describable as stratified randomization covariance analysis. Another important feature of this nonparametric method is its applicability under the same minimal assumptions as the extended Mantel–Haenszel statistic (see Section 3.1). In this regard, no assumption about the relationship between the $y_{ghik}$ and the $x_{hik}$ is required even though the definition of the $z_{ghik}$ involves a linear structure.

Results from randomization covariance analysis with baseline adjustment of treatment comparisons are shown in Table 7 for the ordinal re-

**Table 7** $p$-Values from Randomization Covariance Analysis Relative to Baseline Status for Ordinal Response at Visits 1, 2, 3, and 4

| Response variable | Statistic[a] | Center 1 | Center 2 | Combined centers | Combined for center × sex |
|---|---|---|---|---|---|
| Visit 1 | $Q_Z$ $(d.f. = 1)$ | 3.03 | 3.67 | 6.64 | 6.65 |
| | $p$-value | 0.082 | 0.055 | 0.010[c] | 0.010[c] |
| Visit 2 | $Q_Z$ $(d.f. = 1)$ | 11.63 | 11.17 | 22.59 | 20.95 |
| | $p$-value | 0.001[c] | 0.001[c] | < 0.001[c] | < 0.001[c] |
| Visit 3 | $Q_Z$ $(d.f. = 1)$ | 4.69 | 7.68 | 12.33 | 10.50 |
| | $p$-value | 0.030[b] | 0.006[c] | < 0.001[c] | 0.001[c] |
| Visit 4 | $Q_Z$ $(d.f. = 1)$ | 2.99 | 4.00 | 6.97 | 6.19 |
| | $p$-value | 0.084 | 0.046[b] | 0.008[c] | 0.013[b] |
| All visits | $Q_Z$ $(d.f. = 4)$ | 12.34 | 11.74 | 23.41 | 21.58 |
| | $p$-value | 0.015[b] | 0.019[b] | < 0.001[c] | < 0.001[c] |
| Average over 4 visits | $Q_Z$ $(d.f. = 1)$ | 8.76 | 8.97 | 17.47 | 16.09 |
| | $p$-value | 0.003[c] | 0.003[c] | < 0.001[c] | < 0.001[c] |

[a]The $p$-values for each center are based on randomization chi-square statistics (40) to compare mean scores for treatment with covariance adjustment for baseline status. For both response variables and baseline status, within center, standardized midranks are used as framework for analysis. Stratification adjustment in a spirit similar to that for the Mantel–Haenszel statistic is used via (42) to determine the $p$-values for treatment comparisons for the combined centers and the combined (center × sex) strata with covariance adjustment for baseline status. Computations were performed with the procedures documented in Amara and Koch (1980).

[b]Treatment comparisons with $p \leq 0.05$.

[c]Treatment comparisons with $p \leq 0.01$.

sponse variables at each visit. The approximate $p$-values for each center are based on the mean score statistics $Q_{z,gh}$ in (40) for within center, standardized midranks of both the response variable and baseline status; i.e., they correspond to rank analysis of covariance. For center 1, these $p$-values are significant ($p \leq 0.05$) at visits 2 and 3 and nearly significant at visits 1 and 4 ($p = 0.082, 0.084$); and for center 2, they are significant at visits 2 to 4 and nearly significant at visit 1 ($p = 0.055$). These patterns of results agree well with their counterparts in Table 5 from direct comparisons between treatments, although those for center 1 are slightly stronger and those for center 2 are slightly weaker; the reason for these minor variations is the association of more favorable baseline status with more favorable response status, so adjustment offsets the somewhat less favorable baseline status of active treatment in center 1 and its somewhat more favorable status in center 2. Adjusted treatment comparisons for the combined centers are evaluated with the stratified randomization covariance statistic $Q_{z,g}$ in (42); this method was also applied with stratification adjustment for both center and sex. The approximate $p$-values from these analyses indicated significant ($p \leq 0.05$) differences between treatments at all visits. Thus they supported somewhat stronger conclusions than their counterparts in Table 5 from direct comparisons between treatments and in Table 6 from stratification-adjusted comparisons with respect to sex.

More general versions of randomization covariance statistics such as the $Q_{z,gh}$ and the $Q_{z,g}$ are available. They encompass comparisons among more than two groups, stratification adjustment for cross-classifications of center with one or more background variables, covariance adjustment for more than one background variable, and usage of scores other than ranks. Also, as described in Section 4.1, multivariate analyses of more than one response variable can be undertaken. For some background variables, such as sex, either stratification adjustment or covariance adjustment could be applied, so the choice between them requires attention. The basic consideration here is that stratification provides a more explicit adjustment by "holding the background variable constant" at the individual patient level, whereas covariance analysis involves adjustment for which equivalence of means of background variables is induced for groups of patients. The usefulness of each method is greater when adjustment is for background variables which are strongly associated with response variables. Further discussion of nonparametric methods for covariance analysis is given in Quade (1967, 1982), Puri and Sen (1971), Amara and Koch (1980), Huitema (1980), and Koch et al. (1982).

### 3.3 Evaluation of Treatment × Center Interaction

At the beginning of Section 2 the conduct of the statistical analysis of a clinical trial in two parts was identified as an effective strategy for having the evaluation of treatments account for how the study population was selected and how it conceptually represents some target population. Part I is concerned with the existence of a treatment difference for the study population under minimal assumptions. It was addressed with the nonparametric methods in Sections 2, 3.1, and 3.2. The results of these analyses indicated the existence of a significant difference between treatments for the patients in the study population (by the contradiction of the hypothesis of no difference with $p$-values $\leq 0.05$). This significant difference was due to the tendency for patients with active treatment to have more favorable responses than placebo patients.

After the determination of the existence of a difference between treatments for a study population in part I of the statistical analysis of a clinical trial, the objective for part II is the evaluation of whether this difference is homogeneous across study factors such as center or subgroups based on background variables such as age, sex, and baseline status. When homogeneity applies, a treatment difference is interpretable as being independent of study factors and background variables; in this sense, it is generalizable throughout some large target population that the patients conceptually represent (i.e., it applies to younger persons, older persons, females, males, etc.). An important consideration here is that generalizability is an issue for the target population, so its evaluation requires assumptions about how patients in the study population represent those in the target population. Since the target population can be viewed as containing patients like those in the study population, a reasonable (but not provable) assumption is that representation is provided by a process equivalent to stratified simple random sampling where the strata are based on the cross-classification of center × treatment (i.e., representation is assumed for the conditional distributions of response variables given center and treatment). Through this assumption, statistical models are formulated in this section for the relationship of response variables to treatment and center. Then, homogeneity of treatment effects across centers is evaluated by the determination of whether such statistical models need to be expanded to include the treatment × center interaction. In Section 3.4 attention is given to homogeneity of treatment effects across subgroups based on background variables through statistical models which include components for treatment, center, and background variables. For these analyses, the assumed sampling process for the rep-

resentation of the target population has its stratification structure based on the cross-classification of center, treatment, and the background variables.

On the basis of the assumption that the patients within each center × treatment group represent a target population in a sense equivalent to stratified simple random sampling, the frequencies $n_{ghij}$ for the distribution of the response variable at the $g$th visit have the product multinomial distribution

$$\Pr(\{n_{ghij}\}) = \prod_{h=1}^{2}\prod_{i=A}^{P} n_{hi}! \frac{\prod_{j=0}^{4} \pi_{ghij}^{n_{ghij}}}{n_{ghij}!}; \tag{43}$$

here the $\pi_{ghij}$ denote the probabilities with which a randomly selected patient from the stratum corresponding to the $h$th center and $i$th treatment is observed to have the $j$th response category at the $g$th visit. From the structure in (43), it follows that the frequencies $N_{ghi*} = n_{ghi3} + n_{ghi4}$ for good or excellent response have the product binomial distribution

$$\Pr(\{N_{ghi*}\}) = \frac{\prod_{h=1}^{2}\prod_{i=A}^{p} n_{hi}! \theta_{ghi}^{N_{ghi*}} (1-\theta_{ghi})^{(n_{hi}-N_{ghi*})}}{N_{ghi*}!(n_{hi}-N_{ghi*})!}, \tag{44}$$

for which $\theta_{ghi} = \pi_{ghi3} + \pi_{ghi4}$; that is, the $\theta_{ghi}$ denote the probabilities of good or excellent response at the $g$th visit for a randomly selected patient from the stratum corresponding to the $h$th center and $i$th treatment.

A useful analysis of the relationship between a dichotomous response variable and a set of explanatory variables (e.g., center, treatment) is provided by logistic regression. It is based on the fitting of a logistic model to the probabilities of favorable (i.e., good or excellent) response. For the $\theta_{ghi}$ in (44), this model has the specification

$$\theta_{ghi} = \left[1 + \exp\left(-\xi_g - \sum_{G=1}^{u} \beta_{gG} x_{Ghi}\right)\right]^{-1}, \tag{45}$$

where the $x_{Ghi}$ are the values of $u$ explanatory variables for the respective center × treatment strata, the $\{\beta_{gG}\}$ are corresponding regression parameters, and $\xi_g$ is an intercept parameter. The parameters $\xi_g$ and $\{\beta_{gG}\}$ can have any value in $(-\infty, \infty)$ since values of $(\xi_g + \sum_{G=1}^{u} \beta_{gG} x_{Ghi})$ in $(-\infty, \infty)$ correspond to values of the $\theta_{ghi}$ in (0,1). Additional insights concerning the parameters $\xi_g$ and $\{\beta_{gG}\}$ are provided by consideration of the logistic transformation of (45) to the linear model

$$\text{logit}(\theta_{ghi}) = \log_e \frac{\theta_{ghi}}{(1-\theta_{ghi})} = \xi_g + \sum_{G=1}^{u} \beta_{gG} x_{Ghi}. \tag{46}$$

In (46), the $\phi_{ghi} = \theta_{ghi}/(1-\theta_{ghi})$ represent the "odds" of favorable versus unfavorable response. Thus the $\{\exp(\beta_{gG})\}$ are interpretable as multipliers for the "odds" per unit change in the $x_{Ghi}$.

Estimates of the parameters $\xi_g$ and $\{\beta_{gG}\}$ are usually obtained by maximum likelihood methods. More specifically, the likelihood (44) is expressed as a function of $\xi_g$ and the $\{\beta_{gG}\}$ by replacement of the $\theta_{ghi}$ with their model counterparts from (45); then its natural logarithm is differentiated with respect to $\xi_g$ and the $\{\beta_{gG}\}$. The nonlinear equations that result from setting these derivatives equal to 0 are then solved for the maximum likelihood estimates $\hat{\xi}_g$ and $\{\hat{\beta}_{gG}\}$ by an iterative procedure such as the Newton–Raphson method. When the sample sizes $n_{hi}$ are sufficiently large, $\hat{\xi}_g$ and the $\{\hat{\beta}_{gG}\}$ approximately have a multivariate normal distribution for which the covariance matrix can be consistently estimated by

$$\hat{V}_{g,\mathbf{x}} = \mathbf{V}(\hat{\xi}_g, \{\hat{\beta}_{gG}\}) = \left[\sum_{h=1}^{2}\sum_{i=A}^{P} n_{ni}\hat{\theta}_{ghi}(1-\hat{\theta}_{ghi})\mathbf{x}_{*hi}\mathbf{x}'_{*hi}\right]^{-1}, \qquad (47)$$

where $\mathbf{x}_{*hi} = (1, x_{1hi}, \ldots, x_{uhi})'$ and

$$\hat{\theta}_{ghi} = \left[1 + \exp\left(-\hat{\xi}_g - \sum_{G=1}^{u}\hat{\beta}_{gG}x_{Ghi}\right)\right]^{-1};$$

a supportive condition for approximate normality of $\hat{\xi}_g$ and the $\{\hat{\beta}_{gG}\}$ for situations with a small number of strata (e.g., $\leq 30$) is that nearly all of the $N_{ghi*}$ and their complements $(n_{hi} - N_{ghi*})$ are $\geq 5$; more generally, the relevant consideration is whether the sample size is sufficiently large to support approximate normality for the linear statistics $\sum_{h=1}^{2}\sum_{i=A}^{P} N_{ghi*}x_{Ghi}$ [see Cox (1970), Imrey et al. (1981, 1982), McCullagh and Nelder (1983), and Koch and Edwards (1985) for further discussion].

Two asymptotically equivalent criteria for evaluating the goodness of fit of the logistic model (45) for situations where the sample sizes within the strata are not excessively small (e.g., nearly all $n_{hi} \geq 5$) are the log-likelihood ratio chi-square statistic $Q_{L,g}$ and the Pearson chi-square statistic $Q_{P,g}$. Their definitions are

$$Q_{L,g} = \sum_{h=1}^{2}\sum_{i=A}^{P}\left[N_{ghi*}\log_e\frac{N_{ghi*}}{\hat{M}_{ghi*}} + (n_{hi} - N_{ghi*})\log_e\frac{(n_{hi} - N_{ghi*})}{(n_{hi} - \hat{M}_{ghi*})}\right], \qquad (48)$$

$$Q_{P,g} = \sum_{h=1}^{2}\sum_{i=A}^{P}\left[\frac{(N_{ghi*} - \hat{M}_{ghi*})^2}{\hat{M}_{ghi*}} + \frac{(N_{ghi*} - \hat{M}_{ghi*})^2}{n_{hi} - \hat{M}_{ghi*}}\right]$$
$$= \sum_{h=1}^{2}\sum_{i=A}^{P}\frac{n_{hi}(N_{ghi*} - \hat{M}_{ghi*})^2}{\hat{M}_{ghi*}(n_{hi} - \hat{M}_{ghi*})}, \tag{49}$$

where $\hat{M}_{ghi*} = n_{hi}\hat{\theta}_{ghi}$ is the model predicted frequency of good or excellent response for the $h$th center and $i$th treatment (also, for $Q_{L,g}$, $0\{\log_e(0)\}$ is defined to be 0). The test statistics $Q_{L,g}$ and $Q_{P,g}$ approximately have chi-square distributions with d.f. = [(number of strata)-(number of parameters)] = $4 - (u + 1)$ when the sample sizes within the strata are sufficiently large [e.g., all $\hat{M}_{ghi*} \geq 5$ and all $(n_{hi} - \hat{M}_{ghi*}) \geq 5$; or equivalently, $5 \leq \hat{M}_{ghi*} \leq (n_{hi} - 5)$ for all strata]. Applicability of chi-square approximations to $Q_{P,g}$ have also been found through numerical studies [e.g., Larntz (1978)] to be reasonable for many situations with small or moderate sample sizes (e.g., $2 < \hat{M}_{ghi*} < n_{hi} - 2$ for most strata and only a few strata with $\hat{M}_{ghi*} < 1$ or $\hat{M}_{ghi*} > n_{hi} - 1$).

For the dichotomous response variables at visits 1 to 4, results for a logistic regression model with explanatory variables for treatment and center are given in Table 8. The model specification in these analyses for the probability of good or excellent response at the $g$th visit is

$$\theta_{ghi} = [1 + \exp(-\xi_g - \beta_{g1}x_{1hi} - \beta_{g2}x_{2hi})]^{-1}, \tag{50}$$

where $x_{1hi} = 1$ if $i = A$ and $x_{1hi} = 0$ if $i = P$ indicates treatment and $x_{2hi} = 1$ if $h = 2$ and $x_{2hi} = 0$ if $h = 1$ indicates center; the parameters $\xi_g$, $\beta_{g1}$, and $\beta_{g2}$ are the reference value (of the logit) for placebo in center 1, the effect for active treatment, and the effect for center 2, respectively. The maximum likelihood estimates of these parameters and their standard errors [from the square roots of the diagonal elements of (47)] are shown in Table 8. Since the sample sizes for the data in Table 1 are considered large enough to support approximately normal distributions for the maximum likelihood estimates $\hat{\xi}_g$, $\hat{\beta}_{g1}$, and $\hat{\beta}_{g2}$, statistical tests for whether their corresponding parameters are equal to 0 can be undertaken with Wald statistics. These criteria have the form

$$Q_W = \frac{(\text{maximum likelihood estimate})^2}{(\text{estimated standard error})^2}; \tag{51}$$

approximate $p$-values for $Q_W$ are based on the chi-square distribution with d.f. = 1. These $p$-values are shown in Table 8 with the estimates to which they apply. For the estimated treatment effects $\hat{\beta}_{g1}$, they are significant ($p \leq 0.05$) at visits 1 to 3 and nearly significant at visit 4 ($p = 0.063$). These

**Table 8** Maximum Likelihood Estimates, Standard Errors, and $p$-Values for Parameters in Logistic Regression Models for Probability of Good or Excellent Response at Visits 1, 2, 3, and 4 and Corresponding Goodness-of-Fit Statistics[a]

| Parameter | Statistic | Visit 1 | Visit 2 | Visit 3 | Visit 4 |
|---|---|---|---|---|---|
| Reference value for placebo in center 1 | Estimate | −0.564 | −0.814 | −0.476 | −0.894 |
| | SE | 0.341 | 0.351 | 0.338 | 0.354 |
| | $p$-value | 0.099 | 0.020[b] | 0.16 | 0.012[b] |
| Effect for active treatment | Estimate | 0.859 | 1.361 | 1.150 | 0.758 |
| | SE | 0.410 | 0.411 | 0.409 | 0.407 |
| | $p$-value | 0.036[b] | 0.001[c] | 0.005[c] | 0.063 |
| Effect for center 2 | Estimate | 1.072 | 0.685 | 0.603 | 1.268 |
| | SE | 0.410 | 0.409 | 0.406 | 0.407 |
| | $p$-value | 0.009[c] | 0.094 | 0.14 | 0.002[c] |
| Goodness of fit statistic (treatment × center interaction) | $Q_L$ (d.f. = 1) | 1.52 | 0.82 | 0.53 | 0.23 |
| | $p$-value | 0.22 | 0.36 | 0.47 | 0.63 |
| | $Q_P$ (d.f. = 1) | 1.50 | 0.82 | 0.53 | 0.23 |
| | $p$-value | 0.22 | 0.37 | 0.47 | 0.63 |

[a]The logistic model has the specification in (50); that is, (probability of good or excellent response) $= [1+\exp(-\xi-\beta_1 x_1-\beta_2 x_2)]^{-1}$, where $x_1 = 1$ if active or $x_1 = 0$ if placebo indicates treatment, $x_2 = 1$ if center 2 or $x_2 = 0$ if center 1 indicates center, and $\xi$, $\beta_1$, $\beta_2$ are unknown parameters corresponding to reference value for placebo in center 1, effect for active treatment, and effect for center 2. Estimates for $\xi$, $\beta_1$, $\beta_2$ are from maximum likelihood methods; their standard errors are from square roots of the diagonal elements of (47); and their $p$-values are from Wald statistics $Q_W$ in (51); goodness-of-fit statistics are the log-likelihood ratio statistic $Q_L$ in (48) and the Pearson statistic $Q_P$ in (49). Both have one degree of freedom and pertain to treatment × center interaction. Computations were performed with the CATMOD and LOGIST Procedures in the SAS System (1985).

[b]Results with $p \leq 0.05$.

[c]Results with $p \leq 0.01$.

results agree very well with those from the Mantel–Haenszel statistic in Table 4; indeed, they appear to be virtually the same. The principal reason for this is that the Mantel–Haenszel statistic is asymptotically equivalent to test statistics for treatment effects in a logistic regression model like (50) with additive effects for center and treatment for the $\{\text{logit}(\theta_{ghi})\}$ [see Birch (1964, 1965) and Breslow and Day (1980) for further discussion]. Moreover, the compatibility of the $\theta_{ghi}$ with this type of model is supported for the dichotomous response variables at visits 1 to 4 by the nonsignificance of the goodness-of-fit statistics $Q_{L,g}$ and $Q_{P,g}$. The approximate $p$-values for these criteria are shown in Table 8; they are based on the chi-square distribution with d.f. $= 4-(u+1) = 1$, and all of them have $\geq 0.10$ status. The remaining results in Table 8 that merit a comment are the $p$-values for the center effects $\hat{\beta}_{g2}$; those at visits 1 and 4 are significant ($p \leq 0.05$) and that at visit 2 is suggestive ($p = 0.094$). Thus center effects need to be maintained in the logistic model (50) for the $\theta_{ghi}$ so that variation across centers as well as between treatments is described appropriately.

Another logistic model that is of potential interest for the $\theta_{ghi}$ is the expansion of (50) to include a component for treatment × center interaction; the specification for this expanded model is

$$\theta_{ghi} = [1 + \exp(-\xi_g - \beta_{g1}x_{1hi} - \beta_{g2}x_{2hi} - \beta_{g3}x_{3hi})]^{-1}, \tag{52}$$

where $x_{3hi} = x_{1hi}x_{2hi}$ and $\beta_{g3}$ is the effect for treatment × center interaction. For the expanded model (52), the extent to which the odds of favorable versus unfavorable response is greater for active treatment than for placebo at the $h$th center is expressed through the "odds ratios"

$$\psi_{gh} = \frac{\phi_{ghA}}{\phi_{ghP}} = \frac{\theta_{ghA}(1-\theta_{ghP})}{\theta_{ghP}(1-\theta_{ghA})} = \exp(\beta_{g1} + \beta_{g3}x_{*h*}), \tag{53}$$

where $x_{*h*} = 1$ if $h = 2$ and $x_{*h*} = 0$ if $h = 1$ (i.e., the "odds ratios" are measures of treatment effects within the respective centers). When $\beta_{g3} = 0$, the "odds ratios" for the respective centers are all equal to $\exp(\beta_{g1})$, so treatment effects are homogeneous across centers in this sense. On this basis, a statistical test of whether $\beta_{g3}$ is 0 is a test of homogeneity of treatment effects across centers. However, when $\beta_{g3} = 0$, the model (52) simplifies to the model (50). For this reason, the goodness-of-fit statistics $Q_{L,g}$ and $Q_{P,g}$ in Table 8 for the model (50) are also test statistics for $\beta_{g3} = 0$ in the model (52); as tests of treatment × center interaction, they are also tests of homogeneity of treatment effects across centers. Thus the nonsignificance of the results of the goodness-of-fit tests in Table 8 enables treatment effects to be interpreted as homogeneous across centers, and this supports their generalizability for the target population.

The estimated treatment effect that corresponds to the homogeneous "odds ratios" for the respective centers is $\hat{\psi}_{g**} = \exp(\hat{\beta}_{g1})$. At visit 1, $\hat{\psi}_{1**} = \exp(0.859) = 2.36$; so the odds of favorable versus unfavorable response is 2.36 times greater for active than for placebo. Also, since $\{\hat{\beta}_{g1} \pm 1.96[\mathrm{SE}(\hat{\beta}_{g1})]\}$ is an approximately 95% confidence interval for $\beta_{g1}$, its exponential transformation is an approximately 95% confidence interval for $\psi_{g**}$. On this basis, $\exp(0.055, 1.663) = (1.06, 5.28)$ is an approximately 95% confidence interval for $\psi_{1**}$. When sample sizes are not considered large enough to support the approximate normality of $\hat{\beta}_{g1}$ in a logistic model with a no-interaction structure such as (50), methods are available for determining an exact confidence interval for the homogeneous odds ratio $\psi_{g**}$ that applies across a set of strata [for their discussion, see Gart (1971), Thomas (1975), Breslow and Day (1980), and Mehta et al. (1985)].

Another noteworthy feature of logistic regression analysis is the description that its predicted values $\hat{\theta}_{ghi}$ provide for the variation of the probabilities of favorable response across treatments and centers. These predicted values [which are defined following (47)] and their standard errors are shown in Table 9; the standard errors were obtained via

$$\begin{aligned}\mathrm{SE}(\hat{\theta}_{ghi}) &= \hat{\theta}_{ghi}(1-\hat{\theta}_{ghi})[\mathrm{SE}(\hat{\xi}_g + \sum_{G=1}^{u} \hat{\beta}_{gG} x_{Ghi})] \\ &= \hat{\theta}_{ghi}(1-\hat{\theta}_{ghi})(\mathbf{x}'_{*hi}\hat{V}_{g,\mathbf{x}}\mathbf{x}_{*hi})^{1/2}, \end{aligned} \tag{54}$$

where $\hat{\mathbf{V}}_{g,x}$ and the $\mathbf{x}_{*hi}$ are defined via (47). There is agreement between the $\hat{\theta}_{ghi}$ in Table 9 and their counterparts from Table 4 for the actual proportions of patients with good or excellent responses for the respective center × treatment × visit combinations [i.e., the $p_{ghi} = (\%)_{ghi}/100$, where the $(\%)_{ghi}$ are defined in (27)]. This finding is compatible with the previously stated support provided for the model by the nonsignificance of the goodness-of-fit statistics $Q_{L,g}$ and $Q_{P,g}$. An important descriptive advantage of the predictive values $\hat{\theta}_{ghi}$ is that they tend to have smaller estimated standard errors than the actual proportions $p_{ghi}$. This property is a consequence of their structure not involving the extraneous variability for factors not included in the model (i.e., treatment × center interaction).

As discussed in Section 3.1, the integer score means $\bar{y}_{ghi}$ in (31) describe the distributions of the ordinal response variables for each center × treatment × visit. These statistics are unbiased estimates for the subpopulation means $\eta_{ghi} = \sum_{j=0}^{4} j\pi_{ghij}$ of the multinomial distribution (43) for the response at the $g$th visit by patients receiving the $i$th treatment at the $h$th center. Relative to the means $\eta_{ghi}$, the difference between active and placebo treatments for the $h$th center is $\delta_{gh} = (\eta_{ghA} - \eta_{ghP})$; so ho-

**Table 9** Predicted Values and Standard Errors from Logistic Regression Model for Probability of Good or Excellent Response and Linear Model for Integer Score Mean Response

| Response variable | Treatment | Statistic | Probability good or excellent response[a] | | Integer score mean response[b] | |
|---|---|---|---|---|---|---|
| | | | Center 1 | Center 2 | Center 1 | Center 2 |
| Visit 1 | Active | Pred. val. | 0.573 | 0.797 | 2.57 | 3.30 |
| | | SE | 0.084 | 0.063 | 0.16 | 0.14 |
| | Placebo | Pred. val. | 0.363 | 0.624 | 2.15 | 2.87 |
| | | SE | 0.079 | 0.081 | 0.19 | 0.16 |
| Visit 2 | Active | Pred. val. | 0.633 | 0.774 | 2.90 | 3.36 |
| | | SE | 0.081 | 0.067 | 0.17 | 0.17 |
| | Placebo | Pred. val. | 0.307 | 0.468 | 1.91 | 2.37 |
| | | SE | 0.075 | 0.084 | 0.20 | 0.20 |
| Visit 3 | Active | Pred. val. | 0.662 | 0.782 | 2.92 | 3.22 |
| | | SE | 0.079 | 0.065 | 0.20 | 0.17 |
| | Placebo | Pred. val. | 0.383 | 0.532 | 2.09 | 2.39 |
| | | SE | 0.080 | 0.084 | 0.22 | 0.23 |
| Visit 4 | Active | Pred. val. | 0.466 | 0.756 | 2.54 | 3.18 |
| | | SE | 0.085 | 0.069 | 0.18 | 0.20 |
| | Placebo | Pred. val. | 0.290 | 0.592 | 1.95 | 2.60 |
| | | SE | 0.073 | 0.082 | 0.21 | 0.24 |

[a]Results are based on (47) and (54) for the logistic regression model with specification given in Table 8.
[b]Results are based on (61) for the linear model with specification given in Table 10.

mogeneity of treatment differences across centers applies when there is no treatment × center interaction in the sense that

$$\delta_{g1} - \delta_{g2} = \eta_{g1A} - \eta_{g1P} - \eta_{g2A} + \eta_{g2P} = 0. \tag{55}$$

However, the constraint (55) for the $\eta_{ghi}$ corresponds to the compatibility of the $\eta_{ghi}$ with the linear model

$$\eta_{ghi} = \xi_g + \beta_{g1} x_{1hi} + \beta_{g2} x_{2hi}, \tag{56}$$

where $\xi_g$, $\beta_{g1}$, $\beta_{g2}$, the $x_{1hi}$, and the $x_{2hi}$ have definitions like those stated for (50). As a consequence of this consideration, statistical tests for the

goodness of fit of the linear model (56) are also tests for the homogeneity of treatment effects across centers for the $\eta_{ghi}$.

Since the sample sizes $n_{hi}$ are sufficiently large (e.g., $n_{hi} \geq 25$) to support approximately normal distributions for the $\bar{y}_{ghi}$, an effective way to assess the goodness of fit of the linear model in (56) is through the weighted least squares methods discussed in Grizzle et al. (1969) and Koch et al. (1977, 1985a). For such analysis, estimates $\hat{\xi}_g$, $\hat{\beta}_{g1}$, and $\hat{\beta}_{g2}$ are determined so as to minimize the weighted residual sum of squares

$$Q_{W,g} = \sum_{h=1}^{2} \sum_{i=A}^{P} \left\{ (\bar{y}_{ghi} - \hat{\xi}_g - \hat{\beta}_{g1} x_{1hi} - \hat{\beta}_{g2} x_{2hi})^2 / v_{ghi,y} \right\}, \tag{57}$$

where the $v_{ghi,y} = \left\{ \sum_{k=1}^{n_{hi}} (y_{ghik} - \bar{y}_{ghi})^2 / n_{hi}^2 \right\}$ are estimated variances for the $\bar{y}_{ghi}$. This process involves the solution of a set of linear equations, so matrix notation provides a concise expression for the estimates that result from it. More specifically, let $\hat{\beta}_{g*} = (\hat{\xi}_g, \hat{\beta}_{g1}, \hat{\beta}_{g2})'$ denote the vector of weighted least squares estimates, let $\bar{y}_{g**} = (\bar{y}_{g1A}, \bar{y}_{g1P}, \bar{y}_{g2A}, \bar{y}_{g2P})'$ denote the vector of mean scores, let $\mathbf{V}_{g,y}$ denote the estimated covariance matrix for $\bar{y}_{g**}$ with diagonal elements equal to the $v_{ghi,y}$ and all off-diagonal elements equal to 0, and let

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} \tag{58}$$

denote the specification matrix for the model (56); then the matrix expression for the weighted least squares estimates is

$$\hat{\beta}_{g*} = (\mathbf{X}'\mathbf{V}_{g,y}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_{g,y}^{-1}\bar{y}_{g**}; \tag{59}$$

also, a consistent estimate for their corresponding covariance matrix is

$$\hat{\mathbf{V}}_{g,\beta} = (\mathbf{X}'\mathbf{V}_{g,y}^{-1}\mathbf{X})^{-1}. \tag{60}$$

The minimized weighted residual sum of squares $Q_{W,g}$ from the substitution of the weighted least squares estimates $\hat{\beta}_{g*}$ from (59) into (57) is a goodness-of-fit statistic for the linear model (56). This test statistic approximately has the chi-square distribution with d.f. = (number of strata) − (number of parameters) = 4 − 3 = 1 when the sample sizes are sufficiently large for the $\bar{y}_{ghi}$ to have approximately normal distributions. Also, for such large-sample situations, $\hat{\xi}_g$ and the $\{\hat{\beta}_{gG}\}$ approximately have normal distributions, so $p$-values for tests of whether their corresponding parameters are equal to 0 can be based on chi-square approximations for Wald statistics such as (51).

**Table 10** Weighted Least Squares Estimates, Standard Errors, and $p$-Values for Parameters in Linear Models for Integer Scores Mean Response at Visit 1, 2, 3, and 4 and Corresponding Goodness-of-Fit Statistics[a]

| Parameter | Statistic | Visit 1 | Visit 2 | Visit 3 | Visit 4 |
|---|---|---|---|---|---|
| Reference value for placebo in center 1 | Estimate | 2.15 | 1.91 | 2.09 | 1.95 |
| | SE | 0.19 | 0.20 | 0.22 | 0.21 |
| | $p$-value | $< 0.001$[b] | $< 0.001$[b] | $< 0.001$[b] | $< 0.001$[b] |
| Effect for active treatment | Estimate | 0.42 | 0.99 | 0.83 | 0.59 |
| | SE | 0.19 | 0.22 | 0.24 | 0.24 |
| | $p$-value | 0.026[c] | $< 0.001$[b] | 0.001[b] | 0.014[c] |
| Effect for center 2 | Estimate | 0.72 | 0.46 | 0.29 | 0.64 |
| | SE | 0.19 | 0.21 | 0.23 | 0.24 |
| | $p$-value | $< 0.001$[b] | 0.028[c] | 0.20 | 0.007[b] |
| Goodness of fit statistic (treatment × center interaction) | $Q_W$ (d.f. = 1) | 0.36 | 0.39 | 1.12 | 0.51 |
| | $p$-value | 0.55 | 0.53 | 0.29 | 0.48 |

[a]The linear model has the specification in (56); that is, (integer score mean response) $= \xi + \beta_1 x_1 + \beta_2 x_2$, where $x_1 = 1$ if active or $x_1 = 0$ if placebo indicates treatment, $x_2 = 1$ if center 2 or $x_2 = 0$ if center 1 indicates center, and $\xi, \beta_1, \beta_2$ are unknown parameters corresponding to reference value for placebo in center 1, effect for active treatment, and effect for center 2. Estimates for $\xi$, $\beta_1$, $\beta_2$ are from weighted least squares methods via (59); their standard errors are square roots of the diagonal elements of (60); and their $p$-values are from Wald statistics such as (51). The goodness-of-fit statistic is the weighted sum of squares (57) due to the residuals from the model's predicted values for (integer score mean response); it has one degree of freedom and is the same as the Wald statistic (62) for the treatment × center interaction. Computations were performed with the CATMOD Procedure in the SAS System (1985).
[b]Results with $p \leq 0.01$.
[c]Results with $p \leq 0.05$.

Results from the weighted least squares analysis for the linear model (56) are given in Table 10 for the ordinal responses at visits 1 to 4. They include estimates from (59) for the predicted mean response for placebo at center 1, the effect for active treatment, and the effect for center 2 together with the standard errors of these estimates [from square roots of the diagonal elements of the estimated covariance matrix in (60)] and $p$-values for tests of zero values from Wald statistics.

For the differences between treatments, the estimates, standard errors, and $p$-values in Table 10 for weighted least squares analyses are very similar to their counterparts in Table 5 from analyses through the extended Mantel–Haenszel statistic. A reason for this is the across-center homogeneity of the differences between the integer mean scores for the treatments. In this regard, such homogeneity is supported by the nonsignificance of the goodness-of-fit statistics $Q_{W,g}$ in (57) for the model (56); the $p$-values from these tests are shown in Table 10, and all have $\geq 0.25$ status relative to their approximate chi-square distribution with d.f. = 1.

Some other results from weighted least squares analysis merit attention. For visits 1, 2, and 4, the $p$-values for center effects $\hat{\beta}_{g2}$ are significant ($p \leq 0.05$); so center effects need to be maintained in the linear model. Predicted values $\hat{\eta}_{ghi}$ from the linear model (56) and their standard errors are given for the integer mean scores $\eta_{ghi}$ in Table 9. These results were obtained via

$$\hat{\eta}_{ghi} = \hat{\xi}_g + \hat{\beta}_{g1}x_{1hi} + \hat{\beta}_{g2}x_{2hi} = \mathbf{x}'_{*hi}\hat{\boldsymbol{\beta}}_{g*},$$
$$\mathrm{SE}(\hat{\eta}_{ghi}) = \mathbf{x}'_{*hi}\hat{\mathbf{V}}_{g,\beta}\mathbf{x}_{*hi}. \tag{61}$$

Since usage of the linear model (56) was supported by the nonsignificance of the goodness-of-fit statistics $Q_{W,g}$ in (57), the predicted values $\hat{\eta}_{ghi}$ in Table 9 agree well with the actual mean scores $\bar{y}_{ghi}$ for the respective center × treatment × visit combinations. Also, by having a structure that does not involve sources of extraneous variation, the $\hat{\eta}_{ghi}$ tend to have smaller standard errors than the $\bar{y}_{ghi}$.

The weighted least squares analyses for the linear model (56) have some noteworthy theoretical properties. The goodness-of-fit statistic $Q_{W,g}$ in (57) is identical to the Wald statistic for the constraint (55) which corresponds to the model; that is,

$$Q_{W,g} = \frac{(\bar{y}_{g1A} - \bar{y}_{g1P} - \bar{y}_{g2A} + \bar{y}_{g2P})^2}{\sum_{h=1}^{2}\sum_{i=A}^{P} v_{ghi,y}}. \tag{62}$$

For models with linear constraints such as (55) for the parameters $\pi_{ghij}$, **Bhapkar** (1966) showed generally that the Wald statistic (Wald, 1943) was identical to the Neyman minimum modified chi-square statistic for good-

ness of fit. Thus, from Neyman (1949), $Q_{W,g}$ is asymptotically equivalent to the log-likelihood ratio statistic for the goodness of fit of the model (56) in the sense that the probability that the two methods contradict each other tends to zero as the sample sizes $n_{hi}$ become large. Also, the weighted least squares estimates $\hat{\beta}_{g*}$ belong to the class of *best asymptotic normal* (BAN) estimates; so they are asymptotically unbiased, efficient, and equivalent to maximum likelihood estimates. For additional discussion of theoretical properties of weighted least squares methods for categorical data, see Koch et al. (1985a).

In summary, homogeneity of treatment effects across centers was evaluated in this section through the goodness of fit of statistical models which included explanatory variables for treatment and center. For the dichotomous variables of good or excellent response at each visit, logistic regression models were used; and for the ordinal response variables at each visit, linear models for integer mean scores were used. The goodness-of-fit statistics for these models corresponded to test statistics for treatment × center interaction. Although attention here was focused on an example with two centers and two treatments, the same principles and methods are applicable to studies with $q$ centers and $s$ treatments. The models for these situations would include $(s-1)$ indicator variables for treatment and $(q-1)$ indicator variables for center in addition to the reference value; and the statistical test for treatment × center interaction would be based on chi-square approximations with d.f. $= (q-1)(s-1)$ for the appropriate extensions of $Q_{L,g}$ in (48), $Q_{P,g}$ in (49), and $Q_{W,g}$ in (57). For the example, the results of the statistical tests for treatment × center interaction were nonsignificant; so they supported the conclusion that treatment effects were homogeneous across centers and, in this sense, were generalizable.

## 3.4 Evaluation of Treatment × Background Variable Interaction

Conclusions concerning treatment effects in a clinical trial are generalizable throughout a target population if they apply to all subgroups based on background variables such as age, sex, and baseline status. Generalizability in this sense requires homogeneity of treatment effects across these subgroups. Such homogeneity is evaluated in this section through analyses which are similar in spirit to those in Section 3.3 for homogeneity of treatment effects across centers. For these analyses, statistical models are used to describe the relationship between the response variables at each visit and center, treatment, and background variables. Confirmation that these models do not need to include components for treatment × background

variable interaction or for treatment × center interaction supports the interpretation that treatment effects are homogeneous and hence generalizable throughout the target population.

Statistical models for the analysis of the homogeneity of treatment effects across subgroups based on background variables and centers for a multicenter clinical trial require an assumption for how the patients in the study population represent their counterparts in the target population. As discussed in Section 3.3, the perspective that the target population contains patients like those in the study population (e.g., younger persons, older persons, females, males, etc.) supports its representation in terms of conditional distributions of response variables given center, treatments, and background variables. This reasonable (but not provable) assumption is more formally expressed as the equivalence of the study population to a stratified simple random sample from the target population where the strata correspond to the cross-classification of center, treatment, and background variables. A lenient feature of this assumption is that the background variables themselves can possibly have different distributions in the study population than in the target population (i.e., the sampling rates can vary across the totality of possible strata).

The assumption that the patients in the study population represent those in the target population in a sense equivalent to stratified simple random sampling implies that their responses at the $g$th visit have the product multinomial distribution

$$\Pr(\{y_{ghijk}\}) = \prod_{h=1}^{2} \prod_{i=A}^{P} \prod_{k=1}^{n_{hi}} \left( \prod_{j=0}^{4} \pi_{ghijk}^{y_{ghijk}} \right), \tag{63}$$

where $y_{ghijk} = 1$ if the $k$th patient with the $i$th treatment at the $h$th center is classified into the $j$th response category at the $g$th visit and $y_{ghijk} = 0$ if otherwise; also, $\pi_{ghijk} = E\{y_{ghijk}\}$ denotes the probability that a randomly selected patient from the stratum corresponding to the $h$th center, $i$th treatment, and the same profile of background variables as the $(hik)$th patient (e.g., the same age, sex, and baseline status) is observed to have the $j$th response category at the $g$th visit; finally, $\sum_{j=0}^{4} y_{ghijk} = 1$ since the response at the $g$th visit is classified into one category.

As a consequence of the structure in (63), the dichotomous response variables $Y_{ghi*k} = y_{ghi3k} + y_{ghi4k}$ for good or excellent classifications have the product binomial distribution

$$\Pr(\{Y_{ghi*k}\}) = \prod_{h=1}^{2} \prod_{i=A}^{P} \prod_{k=1}^{n_{hi}} \theta_{ghik}^{Y_{ghi*k}} (1 - \theta_{ghik})^{1 - Y_{ghi*k}}; \tag{64}$$

here $\theta_{ghik} = \pi_{ghi3k} + \pi_{ghi4k}$ denotes the probability of good or excellent response at the $g$th visit for a randomly selected patient from the stratum corresponding to the $h$th center, $i$th treatment, and the same profile of background variables as the $(hik)$th patient. In accordance with the discussion in Section 3.3, logistic regression is a useful method for analyzing the relationship between the response variable at the $g$th visit and explanatory variables for center, treatment, and background characteristics such as age, sex, and baseline status. A specific model that is of interest for this purpose is

$$\theta_{ghik} = \left[1 + \exp\left(-\xi_g - \sum_{G=1}^{5} \beta_{gG} x_{Ghik}\right)\right]^{-1}; \tag{65}$$

for this model, $x_{1hik} = 1$ if $i = A$ and $x_{1hik} = 0$ if $i = P$ indicates treatment; $x_{2hik} = 1$ if $h = 2$ and $x_{2hik} = 0$ if $h = 1$ indicates center, $x_{3hik} = 1$ if the $(hik)$th patient is male and $x_{3hik} = 0$ if female, $x_{4hik}$ = baseline status for $(hik)$th patient, and $x_{5hik}$ = age/10 for $(hik)$th patient; the parameters $\xi_g$, $\beta_{g1}$, $\beta_{g2}$, $\beta_{g3}$, $\beta_{g4}$, and $\beta_{g5}$ are unknown parameters corresponding to a reference value, the effect for active treatment, the effect for center 2, the effect for males, the rate of change per category of baseline status, and the rate of change per 10 years of age, respectively. For models such as (65), the $x_{Ghik}$ need to be nonredundant in the sense that none of them is expressible as a linear function of the others and a constant.

A noteworthy feature of the logistic model (65) is that it applies to the responses of individual patients and that it includes both categorical and continuous explanatory variables. In contrast, the logistic models (50) and (52) were specified for strata with a moderately large number of patients and included only categorical explanatory variables. However, even though the logistic model in (65) has a relatively general structure, the principles and procedures for estimating its parameters through maximum likelihood methods are essentially the same as those discussed in Section 3.3 for the much simpler models (50) and (52). The maximum likelihood estimates $\hat{\xi}_g$ and the $\{\hat{\beta}_{gG}\}$ approximately have a multivariate normal distribution when the sample sizes $n_{hi}$ are sufficiently large to support approximately normal distributions for the linear statistics $\sum_{h=1}^{2} \sum_{i=A}^{P} \sum_{k=1}^{n_{hi}} x_{Ghik} Y_{ghi*k}$. A consistent estimate of the covariance matrix for these estimates is

$$\begin{aligned} \hat{\mathbf{V}}_{g,\mathbf{x}} &= \mathbf{V}(\hat{\xi}_g, \{\hat{\beta}_{gG}\}) \\ &= \left[\sum_{h=1}^{2} \sum_{i=A}^{P} \sum_{k=1}^{n_{hi}} \hat{\theta}_{ghik}(1 - \hat{\theta}_{ghik}) \mathbf{x}_{*hik} \mathbf{x}'_{*hik}\right]^{-1}, \end{aligned} \tag{66}$$

where $x_{*hik} = (1, x_{1hik}, x_{2hik}, x_{3hik}, x_{4hik}, x_{5hik})'$ is the vector of explanatory variables for the $(hik)$th patient and $\hat{\theta}_{ghik} = [1 + \exp(-\hat{\xi}_g - \sum_{G=1}^{5} \hat{\beta}_{gG} x_{Ghik})]^{-1}$ is the predicted probability of good or excellent response at the $g$th visit for the $(hik)$th patient.

The maximum likelihood estimates for the parameters in the model (65) and their estimated standard errors [from square roots of the diagonal elements of (66)] are given in Table 11. Approximate $p$-values from Wald statistics such as (51) are also given there for tests of whether model parameters are equal to 0. In this regard, the sample size is considered to be sufficiently large to support approximately normal distributions for $\hat{\xi}_g$ and the $\{\hat{\beta}_{gG}\}$ and hence approximately chi-square distributions (with d.f. = 1) for the corresponding Wald statistics $Q_W$. The principal conclusions from the statistical tests in Table 11 for the parameters in the logistic model (65) are that the probability of good or excellent response at visits 1 to 4 is significantly ($p \leq 0.05$) higher for active treatment than for placebo and significantly ($p \leq 0.01$) increases with the extent to which baseline status is favorable. Also, the effects of sex are clearly nonsignificant (all $p$'s $\geq 0.25$). For center and age, there are mixed results across visits; the $p$-values for the center effects at visits 1 to 3 are clearly nonsignificant (all $p$'s $\geq 0.25$) while that for visit 4 is suggestive ($p = 0.025$); and for age, the $p$-values at visits 1 and 2 are clearly nonsignificant ($p \geq 0.25$) while those at visits 3 and 4 are suggestive ($p = 0.040, 0.055$). An interesting aspect of the results in Table 11 is that they indicate significantly more favorable response for active treatment than placebo in a stronger way than their counterparts in Table 8 for the logistic model (50) or in Table 4 from Mantel–Haenszel analyses. A reason for this is that the logistic model (65) provides covariance adjustment for baseline status which has a strong association with the response variables at visits 1 to 4 (see Table 3). Moreover, since the descriptive results in Table 2 suggest that baseline status was somewhat less favorable for patients at center 1 than those at center 2, adjustment for it might partly account for center effects; this consideration provides a reason why the results in Table 11 did not indicate differences between centers as strongly as those in Table 8.

Since the strata for a cross-classification of center, treatment, and background variables usually have very small sample sizes (e.g., essentially all can be 0 or 1 when the model includes continuous background variables), chi-square approximations are not applicable to goodness-of-fit statistics such as the log-likelihood ratio chi-square statistic $Q_{L,g}$ in (48) or the Pearson chi-square statistic $Q_{P,g}$ in (49). This consideration implies that these criteria can at most serve a descriptive role. Thus, other methods are necessary for the evaluation of the goodness of fit of models like (65). A

**Table 11** Maximum Likelihood Estimates, Standard Errors, and $p$-Values for Parameters in Logistic Regression Model for Relationship Between Probability of Good or Excellent Response and Treatment, Center, Sex, Baseline Status, and Age and $p$-Values for Goodness of Fit Through Pairwise Interactions Not Included in Models[a]

| Parameter | Statistic | Visit 1 | Visit 2 | Visit 3 | Visit 4 |
|---|---|---|---|---|---|
| Reference value | Estimate | −5.57 | −2.37 | −2.51 | −2.38 |
| | SE | 1.43 | 1.09 | 1.14 | 1.12 |
| | $p$-value | < 0.001[b] | 0.029[c] | 0.028[c] | 0.033[c] |
| Effect for active treatment | Estimate | 1.31 | 1.58 | 1.46 | 0.99 |
| | SE | 0.53 | 0.46 | 0.49 | 0.47 |
| | $p$-value | 0.013[c] | < 0.001[b] | 0.003[b] | 0.035[c] |
| Effect for center 2 | Estimate | 0.53 | 0.38 | 0.29 | 1.08 |
| | SE | 0.51 | 0.47 | 0.50 | 0.48 |
| | $p$-value | 0.30 | 0.42 | 0.56 | 0.025[c] |
| Effect for males | Estimate | −0.02 | −0.27 | −0.14 | −0.48 |
| | SE | 0.64 | 0.57 | 0.59 | 0.60 |
| | $p$-value | 0.98 | 0.64 | 0.82 | 0.42 |
| Baseline status | Estimate | 1.54 | 0.72 | 1.04 | 0.90 |
| | SE | 0.33 | 0.23 | 0.26 | 0.25 |
| | $p$-value | < 0.001[b] | 0.002[b] | < 0.001[b] | < 0.001[b] |
| Age/10 | Estimate | 0.01 | −0.18 | −0.38 | −0.35 |
| | SE | 0.19 | 0.17 | 0.18 | 0.18 |
| | $p$-value | 0.94 | 0.29 | 0.040[c] | 0.055 |

| Goodness-of-fit Tests | Statistic | Visit 1 | Visit 2 | Visit 3 | Visit 4 |
|---|---|---|---|---|---|
| Treatment × center | $Q_S$ (d.f. = 1) | 1.20 | 0.49 | 0.17 | 0.05 |
| | $p$-value | 0.27 | 0.49 | 0.68 | 0.82 |
| Treatment × sex | $Q_S$ (d.f. = 1) | 0.02 | 0.76 | 5.80 | 2.09 |
| | $p$-value | 0.89 | 0.38 | 0.016[c] | 0.15 |
| Treatment × baseline | $Q_S$ (d.f. = 1) | 2.99 | 0.01 | 0.22 | 0.38 |
| | $p$-value | 0.084 | 0.94 | 0.64 | 0.54 |
| Treatment × age | $Q_S$ (d.f. = 1) | 1.08 | 0.96 | 0.67 | 2.36 |
| | $p$-value | 0.30 | 0.33 | 0.41 | 0.12 |
| **All pairwise interactions with treatment** | $Q_S$ (d.f. = 4) | 5.35 | 1.54 | 6.18 | 3.85 |
| | $p$-value | 0.25 | 0.82 | 0.19 | 0.43 |
| **All pairwise interactions** | $Q_S$ (d.f. = 10) | 18.20 | 11.10 | 13.27 | 9.28 |
| | $p$-value | 0.052 | 0.35 | 0.21 | 0.51 |

reasonable principle on which to base such methods is a correspondence of lack of fit of a model to the need for the model to include one or more variables. Conversely, if statistical tests for the contribution of these additional variables are non-significant, then the goodness of fit of the model is supported. Evaluation of whether a logistic model needs to include additional explanatory variables can be undertaken effectively with the Rao score statistic. This criterion is directed at the extent to which the residuals $(Y_{ghi*k} - \hat{\theta}_{ghik})$ from the model are linearly associated with the additional explanatory variables. More specifically, let $w_{1hik}, w_{2hik}, \ldots, w_{whik}$ denote a set of $w$ additional explanatory variables; let

$$f_{G'} = \sum_{h=1}^{2} \sum_{i=1}^{2} \sum_{k=1}^{n_{hi}} w_{G'hik}(Y_{ghi*k} - \hat{\theta}_{ghik}), \tag{67}$$

where $G' = 1, 2, \ldots, w$; and let $\mathbf{f} = (f_1, f_2, \ldots, f_w)'$. Then the Rao score statistic has the general form

$$Q_S = \mathbf{f}'\mathbf{V}_f^{-1}\mathbf{f} = \mathbf{f}'\{\mathbf{W}'[\mathbf{D}_{\hat{V}} - \mathbf{D}_{\hat{V}}\mathbf{X}\hat{\mathbf{V}}_{g,x}\mathbf{X}'\mathbf{D}_{\hat{V}}]\mathbf{W}\}^{-1}\mathbf{f}, \tag{68}$$

for which $\mathbf{V}_f$ is the estimated covariance matrix for the linear functions of residuals $\mathbf{f}$. Also, $\mathbf{X}$ is the model specification matrix and has rows $\{\mathbf{x}'_{*hik}\}$, $\mathbf{W}$ is the specification matrix for the additional variables and has rows $\{\mathbf{w}'_{*hik} = (w_{1hik}, w_{2hik}, \ldots, w_{whik})\}$, $\mathbf{D}_{\hat{V}}$ is a diagonal matrix with the $\{\hat{v}_{ghik} = \hat{\theta}_{ghik}(1-\hat{\theta}_{ghik})\}$ on the diagonal, and $\hat{\mathbf{V}}_{g,x}$ is the estimated covariance matrix for $\hat{\xi}_g$ and the $\{\hat{\beta}_{g,G}\}$ from (66). The Rao score statistic $Q_S$ in (67) approximately has the chi-square distribution with d.f. $= w$ when the sample sizes are sufficiently large; that is, the $n_{hi}$ are large enough to support approximately normal distributions for the maximum likelihood

---

[a]The logistic model has the specification in (65); that is, probability of good or excellent response $= [1 + \exp(-\xi - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3 - \beta_4 x_4 - \beta_5 x_5)]^{-1}$, where $x_1 = 1$ if active or $x_1 = 0$ if placebo indicates treatment, $x_2 = 1$ if center 2 or $x_2 = 0$ if center 1 indicates center, $x_3 = 1$ if male or $x_3 = 0$ if female indicates sex, $x_4 =$ baseline status, $x_5 =$ (age/10), and $\xi$, $\beta_1$, $\beta_2$, $\beta_3$, $\beta_4$, $\beta_5$ are unknown parameters corresponding to reference value, effect for active treatment, effect for center 2, effect for males, rate of change per category of baseline status, and rate of change per 10 years of age. Estimates for $\xi$, $\beta_1$, $\beta_2$, $\beta_3$, $\beta_4$, $\beta_5$ are from maximum likelihood methods; their standard errors are from square roots of the diagonal elements of (66); and their $p$-values are from Wald statistics like (51). The $p$-values for goodness of fit through pairwise interactions not in model are based on score statistics via (68). Computations were performed with the LOGIST Procedure in the SAS System (1985).

[b]Results with $p \leq 0.01$.

[c]Results with $p \leq 0.05$.

estimates of the parameters in the expanded model which includes both the specific variables of interest $x_{Ghik}$ and the additional variables $w_{G'hik}$. Also, the $w_{G'hik}$ need to be nonredundant in the sense that the specification matrix $[\mathbf{X}, \mathbf{W}]$ for the expanded model has full rank $(1 + u + w)$. The Rao score statistic in (68) has two noteworthy properties. One is its asymptotic equivalence to the log-likelihood ratio chi-square statistic for testing whether the parameters for the $w_{G'hik}$ in the expanded model are zero. Since the latter criterion is obtained as the difference between the $Q_{L,g}$ in (48) for the model of interest and that for the expanded model, its determination requires the fitting of both models, whereas the determination of $Q_S$ is more convenient in the sense of only involving results for the model of interest. This computational advantage is the other noteworthy property of the Rao score statistic.

Approximate $p$-values for $Q_S$ are given in Table 11 for several types of additional variables that might be included in the logistic model (65). These correspond to the treatment × center interaction via $w_{1hik} = x_{1hik}x_{2hik}$, the treatment × sex interaction via $w_{2hik} = x_{1hik}x_{3hik}$, the treatment × baseline status interaction via $w_{3hik} = x_{1hik}x_{4hik}$, the treatment × age interaction via $w_{4hik} = x_{1hik}x_{5hik}$, all pairwise interactions with treatment via $(w_{1hik}, w_{2hik}, w_{3hik}, w_{4hik})$, and all pairwise interactions via $(w_{1hik}, \ldots, w_{10,hik})$, where the $w_{G'hik}$ encompass all pairwise products of the $x_{Ghik}$. Among the 16 statistical tests in Table 11 for the separate interactions of treatment with center and each of the background variables, only that for treatment × sex at visit 3 had $\leq 0.05$ status, and only that for treatment × baseline status at visit 1 had $0.05 \leq p \leq 0.10$ status; all of the others were nonsignificant with $p \geq 0.10$. Moreover, the overall statistical tests for all pairwise interactions with treatment were nonsignificant $(p \geq 0.10)$ at all four visits. Since this pattern of results is compatible with chance for the situation of no association between the residuals of the model (65) and the $w_{1hik}$, $w_{2hik}$, $w_{3hik}$, and $w_{4hik}$, it is interpreted as indicating that the model does not need to include the interactions of treatment with center and background variables. Thus treatment effects are concluded to be homogeneous.

The goodness of fit of the model (65) is also reasonably supported by the results of the statistical tests for all pairwise interactions. At visits 2 to 4, the $p$-values for these tests are nonsignificant $(p \geq 0.10)$; but at visit 1, some departure from the model is suggested by $p = 0.052$. By further analysis, this departure was identified as due largely to center × baseline status interaction in the sense of a stronger association between the response at visit 1 and baseline status for center 1 than for center 2. Such interaction is considered here to be ignorable since it seems to be more an atypical feature of the study population than a meaningful

source of variation. Thus, for each visit, the model (65) is concluded to provide a satisfactory description of the relationship of the probabilities of good or excellent response to treatment, center, sex, baseline status, and age. Moreover, it does this in a way that expresses the generalizability of treatment effects in terms of odds ratios. The estimates for these odds ratios are $\exp(\hat{\beta}_{g1}) = 3.71, 4.85, 4.31, 2.69$ for visits 1 to 4, respectively. As discussed in Section 3.3, they reflect the extent to which the odds of more favorable response is greater for active treatment than for placebo; and they apply in a homogeneous way to all subgroups of the target population (with respect to center, sex, age, and baseline status).

The probabilities of good or excellent response for different types of patients can be described further with predicted values from (65); in their determination, the range of a continuous variable such as age should be restricted to that for the study population. In other words, prediction from a model should not be extrapolated beyond the types of patients from whom the estimates for its parameters were obtained.

For purposes of completeness, a few additional comments about the goodness of fit of the model (65) merit attention. First, even though the $p = 0.016$ result for treatment × sex at visit 3 was interpreted as compatible with chance, consideration of its implications to generalizability is still of interest. The issue here is that the objective of analyses of interactions involving treatment effects is usually to confirm their absence because their presence might suggest that the generalizability of treatment effects is limited to a particular subgroup. For this reason, these analyses need to be reasonably comprehensive in exploring possibilities for interactions even though many of them may be for chance patterns of variation. The findings from such evaluation of the treatment × sex interaction at visit 3 indicated substantially more favorable responses for active treatment for both sexes and a stronger tendency of this type for the 23 females than the 88 males. This pattern of variation is interpreted as not suggesting any limitation of generalizability.

Another issue for the treatment × sex interaction at visit 3 is that 100% of the females on active treatment had good or excellent response; this aspect of its structure is contrary to its analysis in an expansion of the model (65) because it leads to computational anomalies (i.e., infinite estimates for one or more parameters may be forlornly sought by iterative estimation procedures). Also, it can undermine the large-sample properties which are principal reasons for the use of logistic regression. A rough rule that usually enables the avoidance of this awkward practical problem is the requirement that no linear combination of the columns of the model specification matrix $\mathbf{X}$ correspond to a set of strata for which all patients have the same response status (i.e., either 0% or 100% applies to an outcome such as good

or excellent response). For the treatment × sex interaction at visit 3, this rule is not satisfied, so logistic regression analysis is not applicable to an expansion of the model (65) with this variable. A more formal discussion of conditions for the applicability of logistic regression is given in Silvapulle (1981). A third issue for the model (65) is that it only includes the linear components for age and baseline status. The need to include higher-order components (e.g., quadratic) also could be evaluated through Rao score statistics. Such analysis was not undertaken here because the role of the model (65) was more to provide a reasonable framework for the evaluation of generalizability than to seek an as complete as possible specification for the relationships between the dichotomous response variables and center, treatment, and background variables. The latter objective is also worthwhile and can identify additional models of interest. Since many aspects of the search for such models are subjective, use of a straightforward model such as (65) can be argued as appropriate on practical grounds.

For the ordinal response variables at each visit, the evaluation of treatment × background variable interaction cannot be based on weighted least squares methods (like those discussed for treatment × center interaction in Section 3.3) because the sample sizes for the strata in a cross-classification of center, treatment, and background variables are too small. An alternative strategy for the analysis of the relationships between an ordinal response variable and center, treatment, and background variables is based on maximum likelihood methods for the extension of the logistic regression model to what is often called the proportional odds model. Its applicability has the less stringent sample size requirement of sufficiently large numbers of patients for all treatments at all centers (e.g., $\sum_{h=1}^{2}\sum_{i=A}^{P} n_{hi} \geq 40$). The structure of the proportional odds model for the parameters $\pi_{ghijk}$ in the product multinomial distribution (63) is applied through the probabilities $\theta_{ghijk} = \sum_{j'=j}^{4} \pi_{ghij'k}$ of response at least as favorable as the $j$th category for the $g$th visit and the $(hik)$th patient. For $j = 1, 2, 3, 4$, its specification consists of a parallel set of logistic models with the form

$$\theta_{ghijk} = \left[1 + \exp\left(-\xi_{gj} - \sum_{G=1}^{u} \beta_{gG} x_{Ghik}\right)\right]^{-1}, \tag{69}$$

where the $x_{Ghik}$ are the values of $u$ explanatory variables for the $(hik)$th patient, the $\{\beta_{gG}\}$ are corresponding regression parameters, and the $\xi_{gj}$ are intercept parameters. The $x_{Ghik}$ can correspond to either categorical or continuous variables; the parameters $\{\xi_{gj}\}$ and $\{\beta_{gG}\}$ can have any value in $(-\infty, \infty)$ since the $\theta_{ghijk}$ and the $\pi_{ghijk}$ determined from them are always in (0,1). The proportional odds model (69) implies that the

odds ratios for the $(hik)$th patient versus the $(h'i'k')$th patient have the same value

$$\frac{\theta_{ghijk}(1-\theta_{gh'i'jk'})}{(1-\theta_{ghijk})\theta_{gh'i'jk'}} = \exp\left[\sum_{G=1}^{u}\beta_{gG}(x_{Ghik}-x_{Gh'i'k'})\right] \tag{70}$$

for $j = 1, 2, 3, 4$. Thus the $\exp(\beta_{gG})$ are interpretable as multipliers for the odds of more favorable response versus less favorable response per unit change in the $x_{Ghik}$ for each of the four partitions of the five response categories into unfavorable and favorable subsets. Also, by expressing the extent to which more favorable responses are more likely for the $(hik)$th patient than the $(h'i'k')$th patient, the $\exp(\beta_{gG})$ are indicative of location shifts for the corresponding distributions.

Aspects of statistical analysis for the proportional odds model are similar to those for a general logistic model like (65). Maximum likelihood methods can be used to obtain estimates of its parameters and an estimate of their covariance matrix. Statistical tests for whether model parameters are equal to zero can then be undertaken with Wald statistics like (51). Evaluation of whether additional explanatory variables like those for treatment × background variable interaction need to be included in the model can be based on Rao score statistics which are analogous to (67). Discussion of these methods for the proportional odds model is given in Harrell (1986), McCullagh (1980), McCullagh and Nelder (1983), and Walker and Duncan (1967).

Another consideration for the analysis of the goodness of fit of the proportional odds model is the compatibility of the data with its underlying assumption that the odds ratios in (70) for the four partitions of the five response categories are equal to one another for any pair of patients. Some insight for addressing this question about the appropriateness of the proportional odds model can be obtained by fitting separate logistic regression models to each response partition $j = 1, 2, 3, 4$. Similarity of the four estimated parameters for these separate analyses for the coefficients of each explanatory variable would tend to support the goodness of fit of the proportional odds model (69). Other methods for evaluating the compatibility of ordinal data with the proportional odds assumption are discussed in Genter and Farewell (1985), Harrell (1986), Koch et al. (1985b), McCullagh and Nelder (1983), and Peterson (1986). Since the proportional odds model provides an effective way to describe the relationship between an ordinal response variable and a set of explanatory variables, its use for exploratory analysis of treatment × center interaction and treatment × background variable interaction may be reasonable even when its goodness of fit may seem questionable. However, when usage of the proportional odds model seems

clearly inappropriate, analyses of homogeneity of treatment effects across centers and subgroups based on background variables would need to be based on other methods for ordinal data. Some references concerning such methods are Agresti (1984), Clogg (1982), Cox and Chuang (1984), Koch and Edwards (1987), and McCullagh and Nelder (1983).

For the example considered in this chapter, the proportional odds model that was used for analysis of the relationship between the ordinal responses at each visit and center, treatment, sex, baseline status, and age had the specification

$$\theta_{ghijk} = \left[1 + \exp\left(-\xi_{gj} - \sum_{G=1}^{5} \beta_{gG} x_{Ghik}\right)\right]^{-1}; \tag{71}$$

the $x_{Ghik}$ and the $\{\beta_{gG}\}$ have the same definitions as for the analogous logistic model in (65); and the $\{\xi_{gj}\}$ are intercept parameters for the respective partitions of the five response categories into unfavorable and favorable; also, $\xi_{g3}$ is analogous to $\xi_g$ in (65) since both correspond to good or excellent response. The maximum likelihood estimates for the parameters in the model (71), their standard errors, and approximate $p$-values for tests of 0 values from Wald statistics such as (51) are given in Table 12. The conclusions from these results are similar to those from Table 11 for the logistic model (65). At visits 1 to 4, the odds of more favorable response is significantly ($p \leq 0.05$) higher for active treatment than placebo and significantly increases with the extent to which baseline status is favorable. The $p$-values from the statistical tests for center and sex are nonsignificant at all visits (all $p$'s $\geq 0.10$), and those for age have a mixed nature; at visits 1, 2, and 4, the $p$-values for age are nonsignificant (all $p$'s $\geq 0.10$) while that at visit 3 is suggestive. For the comparisons between treatments, the results in Table 12 [from analyses of ordinal response variables at each visit with the proportional odds model (71)] indicate significantly more favorable response for active treatment than placebo in a similar way to those in Table 7 (from randomization covariance analyses), and in a stronger way than those in Table 10 for the analyses of the integer score means with the linear model (56) or in Table 5 from extended Mantel–Haenszel analyses. As discussed for Table 11 relative to Tables 4 and 8, the methods for Tables 7 and 12 have the advantage of involving covariance adjustment for baseline status and hence provide a more effective analysis in the sense of accounting for the strong association between baseline status and the response variables at visit 1 to 4.

The need for the proportional odds model (71) to include additional variables for the interaction of treatment with center, sex, baseline status, and age or for all pairwise interactions of its explanatory variables is eval-

**Table 12** Maximum Likelihood Estimates, Standard Errors, and $p$-Values for Parameters in Proportional Odds Model for Relationship Between Ordinal Response and Treatment, Center, Sex, Baseline Status, and Age and $p$-Values for Goodness of Fit Through Pairwise Interactions Not Included in Models[a]

| Parameter | Statistic | Visit 1 | Visit 2 | Visit 3 | Visit 4 |
|---|---|---|---|---|---|
| Reference | Estimate | −0.73 | −0.22 | 0.44 | −0.43 |
| value for | SE | 0.97 | 0.92 | 0.88 | 0.91 |
| ≥ poor | $p$-value | 0.45 | 0.81 | 0.61 | 0.64 |
| Reference | Estimate | −2.04 | −1.28 | −0.24 | −1.23 |
| value for | SE | 0.92 | 0.91 | 0.87 | 0.90 |
| ≥ fair | $p$-value | 0.027[b] | 0.16 | 0.79 | 0.17 |
| Reference | Estimate | −4.22 | −3.00 | −1.52 | −2.69 |
| value for | SE | 0.98 | 0.95 | 0.89 | 0.93 |
| ≥ good | $p$-value | < 0.001[c] | 0.002[c] | 0.085 | 0.004[c] |
| Reference | Estimate | −5.89 | −4.10 | −2.71 | −3.56 |
| value for | SE | 1.05 | 0.98 | 0.91 | 0.95 |
| excellent | $p$-value | < 0.001[c] | < 0.001[c] | 0.003[c] | < 0.001[c] |
| Effect for | Estimate | 0.98 | 1.75 | 1.30 | 0.98 |
| active | SE | 0.39 | 0.39 | 0.38 | 0.37 |
| treatment | $p$-value | 0.011[b] | < 0.001[c] | 0.001[c] | 0.009[c] |
| Effect for | Estimate | 0.60 | 0.32 | 0.02 | 0.61 |
| Center 2 | SE | 0.41 | 0.41 | 0.39 | 0.40 |
| | $p$-value | 0.14 | 0.42 | 0.97 | 0.12 |
| Effect for | Estimate | −0.33 | −0.13 | −0.45 | −0.33 |
| males | SE | 0.49 | 0.48 | 0.49 | 0.49 |
| | $p$-value | 0.50 | 0.79 | 0.35 | 0.51 |
| Baseline | Estimate | 1.29 | 0.89 | 0.76 | 0.81 |
| status | SE | 0.22 | 0.21 | 0.19 | 0.20 |
| | $p$-value | < 0.001[c] | < 0.001[c] | < 0.001[c] | < 0.001[c] |
| Age/10 | Estimate | −0.06 | −0.19 | −0.27 | −0.11 |
| | SE | 0.14 | 0.14 | 0.14 | 0.14 |
| | $p$-value | 0.67 | 0.18 | 0.051 | 0.45 |

**Table 12** (*continued*)

| Goodness-of-fit tests | Statistic | Visit 1 | Visit 2 | Visit 3 | Visit 4 |
|---|---|---|---|---|---|
| Treatment × center | $Q_S$ (d.f. = 1) | 0.62 | 0.58 | 1.15 | 0.73 |
| | $p$-value | 0.43 | 0.44 | 0.28 | 0.39 |
| Treatment × sex | $Q_S$ (d.f. = 1) | 0.01 | 0.20 | 5.44 | 4.74 |
| | $p$-value | 0.94 | 0.66 | 0.020[b] | 0.029[b] |
| Treatment × baseline | $Q_S$ (d.f. = 1) | 0.11 | 0.06 | 0.10 | 0.38 |
| | $p$-value | 0.74 | 0.81 | 0.75 | 0.54 |
| Treatment × age | $Q_S$ (d.f. = 1) | 0.86 | 0.02 | 0.01 | 3.75 |
| | $p$-value | 0.35 | 0.89 | 0.99 | 0.053 |
| All pairwise interactions with treatment | $Q_S$ (d.f. = 4) | 2.48 | 0.70 | 6.90 | 7.13 |
| | $p$-value | 0.65 | 0.95 | 0.14 | 0.13 |
| All pairwise interactions | $Q_S$ (d.f. = 10) | 17.73 | 12.51 | 14.23 | 16.95 |
| | $p$-value | 0.060 | 0.25 | 0.16 | 0.076 |

[a]The proportional odds model has the specification in (71); that is, probability of response $\geq j = [1 + \exp(-\xi_j - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3 - \beta_4 x_4 - \beta_5 x_5]^{-1}$, where $x_1 = 1$ if active or $x_1 = 0$ if placebo indicates treatment, $x_2 = 1$ if center 2 or $x_2 = 0$ if center 1 indicates center, $x_3 = 1$ if male or $x_3 = 0$ if female indicates sex, $x_4$ = baseline status, $x_5$ = age/10. The $\xi_j$ are unknown parameters corresponding to the reference distribution, $\beta_1$ is the effect for active treatment, $\beta_2$ is the effect for center 2, $\beta_3$ is the effect for male sex, $\beta_4$ is the rate of change per category for baseline status, $\beta_5$ is the rate of change per 10 years of age. Estimates for $\xi_1$, $\xi_2$, $\xi_3$, $\xi_4$, $\beta_1$, $\beta_2$, $\beta_3$, $\beta_4$, $\beta_5$ are from maximum likelihood methods. The $p$-value for parameters in model are based on Wald statistics; the $p$-values for goodness of fit through pairwise interactions not in model are based on score statistics. Computations were performed with the LOGIST Procedure in the SAS System (1985).

[b]Results with $p \leq 0.05$.

[c]Results with $p \leq 0.01$.

uated with Rao score statistics in a manner similar to that discussed for the logistic model (65). Approximate $p$-values from these statistical tests are given in Table 12. As was the case for their counterparts in Table 11, they tend to support the conclusion that treatment effects are homogeneous across centers and subgroups based on sex, baseline status, and age. In this regard the statistical tests for all pairwise interactions with treatment were nonsignificant ($p \geq 0.10$) at all four visits; also 13 of the 16 separate tests for such interactions were nonsignificant. The departures from this pattern were $\leq 0.05$ $p$-values for treatment $\times$ sex interaction at visits 3 and 4 and $p = 0.053$ for treatment $\times$ age at visit 4. Since both females and males as well as patients of all ages tended to have more favorable responses for active treatment than placebo at each visit, these possible interactions are interpreted as not suggesting any limitation to the generalizability of treatment effects for the target population.

There is also reasonable support for the model (71) from the statistical tests for all pairwise interactions. The $p$-values for these tests are nonsignificant ($p \geq 0.10$) for visits 2 and 3, but suggest some departure from the model at visits 1 and 4 ($p = 0.060, 0.076$). This departure was found by additional analysis to be due to center $\times$ baseline status interaction, so it was considered to be an ignorable, atypical feature of the study population. On the basis of the interpretation given here for the Rao score statistics in Table 12, the proportional odds model (71) is concluded to provide a satisfactory description of the relationship between the distributions of the ordinal response variable at each visit and the effects of treatment, center, sex, baseline status, and age. Through this model, the extent to which the odds of more favorable response is greater for active treatment than for placebo is expressed by the estimated odds ratios $\exp(\hat{\beta}_{g1}) = 2.66$, 5.75, 3.67, 2.66 for visits 1 to 4, respectively. Also, treatment effects are homogeneous in this sense across subgroups of the target population (with respect to center, sex, age, and baseline status).

In summary, statistical models provide a useful framework for the analysis of the relationship between response variables in a clinical trial and treatment, center, and background variables. They also enable the evaluation of generalizability of treatment effects through statistical tests for whether a model needs to include components for treatment $\times$ center interaction or treatment $\times$ background variable interaction. The logistic regression model is of interest for such analyses of dichotomous response variables and its extension to the proportional odds model is of interest for ordinal response variables. Estimates of parameters and statistical tests of goodness of fit can be obtained for both of these types of models with maximum likelihood methods.

### 3.5 Computations

The $p$-values in Tables 4 to 6 were obtained with the FREQ Procedure in the SAS System (1985). The MEANS Procedure was used to determine the descriptive statistics for the separate centers in Tables 4 and 5, and algorithms in the IML Procedure were used to determine descriptive statistics for the combined centers. The computations for the $p$-values in Table 7 from randomization covariance analyses were performed with the procedures documented in Amara and Koch (1980); they could also be obtained with algorithms in the IML Procedure. The CATMOD Procedure in the SAS System (1985) and the LOGIST Procedure of Harrell (1986) were used to obtain the results in Tables 8 to 10; also the analyses for Table 11 and 12 were undertaken with the LOGIST Procedure.

## 4 MULTIVARIATE METHODS FOR RESPONSE VARIABLES

For studies with more than one visit, there is often interest in analyses that encompass the data for all visits simultaneously as well as those for each of the separate visits. The need for such multivariate analyses arises from the tendency for the differences between treatments to vary across visits. Such variation is of particular concern when it corresponds to results of statistical tests for treatment comparisons seeming inconsistent in the sense of being significant at some visits but not at others. Moreover, the multiplicity of statistical tests across visits can make such concern become greater by increasing the probability that observed patterns of apparently significant differences for the respective visits are due to chance. This issue can be addressed for some studies by formally indicating (in the protocol) that the response at one particular visit (e.g., the last visit) is the primary criterion for analysis and those at all others are supportive. For other situations, the data from all visits are considered informative, so alternative strategies are needed. One of these involves methods through which treatment differences are combined across visits in a similar spirit to how they are combined across centers for each visit separately. Aspects of its application are discussed in Section 4.1 in the context of multivariate extensions of the randomization covariance analyses in Section 3.2. The other strategy involves statistical models that describe the variation of response distributions across treatments, centers, and visits. Confirmation that such models do not need to include components for treatment $\times$ visit interaction supports the conclusion that treatment effects are homogeneous across visits and hence are interpretable in a unified way for all visits. Analyses

along these lines are discussed in Section 4.2 in the context of multivariate extensions of the weighted least squares methods for fitting linear models in Section 3.3.

## 4.1 Randomization Covariance Analysis

Multivariate extensions of nonparametric methods such as the randomization covariance statistics in Section 3.2 enable comparisons between treatments to encompass the responses at all visits in a study under minimal assumptions. One family of these test statistics is directed at all visits simultaneously. When covariance adjustment is applied to a background variable such as baseline status and stratification adjustment is applied to center, the multivariate consideration of the residuals $z_{ghik}$ in (38) in an extended Mantel–Haenszel sense such as (12) for the combined centers leads to the multivariate randomization covariance statistic

$$Q_{\mathbf{z}} = \left(\sum_{h=1}^{q}\sum_{k=1}^{n_{hA}} \mathbf{z}_{*hAk}\right)' \left[\text{var}\left(\sum_{h=1}^{q}\sum_{k=1}^{n_{hA}} \mathbf{z}_{*hAk}\right)\right]^{-1} \left(\sum_{h=1}^{q}\sum_{k=1}^{n_{hA}} \mathbf{z}_{*hAk}\right); \tag{72}$$

here $\mathbf{z}_{*hik} = (z_{1hik}, z_{2hik}, z_{3hik}, z_{4hik})'$ is the vector of residuals from the least squares prediction of the response at each of the respective visits as a linear function of baseline status,

$$\text{var}\left\{\sum_{h=1}^{q}\sum_{k=1}^{n_{hA}} \mathbf{z}_{*hAk}\right\} = \sum_{h=1}^{q} \frac{n_{hi}(n_h - n_{hi})}{n_h(n_h - 1)} \left(\sum_{i=A}^{P}\sum_{k=1}^{n_{hi}} \mathbf{z}_{*hik}\mathbf{z}'_{*hik}\right) \tag{73}$$

is the randomization based covariance matrix for the sum

$$\sum_{h=1}^{q}\sum_{k=1}^{n_{hA}} \mathbf{z}_{*hAK}$$

of all the residual vectors for active treatment, and $q = 2$ is the number of centers. The test statistic $Q_{\mathbf{z}}$ approximately has the chi-square distribution with d.f. = (number of visits) = 4 for situations where the two treatment groups have sufficiently large sample sizes $n_{+i}$ for the combined centers (e.g., $n_{+i} \geq 40$ if number of visits $\leq 6$).

From (72), the structure of multivariate test statistics that do not involve covariance adjustment or stratification adjustment is reasonably apparent. If there is no covariance adjustment, the $z_{ghik}$ in (38) are simplified to $y_{ghik} - \bar{y}_{gh*}$ in (72); and if there is no stratification adjustment, summation over $h$ is not needed since only $h = 1 = q$ applies. Alternatively, more general versions of $Q_{\mathbf{z}}$ can be specified for situations with more than two

treatments and covariance adjustment for more than one background variable; for their discussion, see Amara and Koch (1980), Koch and Bhapkar (1982), and Koch et al. (1982).

Another strategy for the usage of responses at all visits in treatment comparisons is based on the univariate analysis of the average responses $\bar{y}_{*hik} = (\sum_{g=1}^{4} y_{ghik}/4)$ over visits. The nonparametric methods that are applicable to this summary measure are the same as those discussed in Sections 3.1 and 3.2 for the responses at each visit separately. In particular, the counterpart to (72) for the covariance and stratification adjusted comparison between the two treatment groups with respect to the average responses over visits has the same form as (42), but with the $g$th visit residuals $z_{ghik}$ from (38) replaced by the average residuals $\mathbf{z}_{*hik} = (\sum_{g=1}^{4} z_{ghik}/4)$. The resulting test statistic $Q_{\bar{z}}$ approximately has the chi-square distribution with d.f. = 1 when the two treatment groups have sufficiently large sample sizes $n_{+i}$ for the combined centers (e.g., $n_{+i} \geq 20$). An important property of $Q_{\bar{z}}$ and other methods based on the average responses over visits is their greater effectiveness than their multivariate counterparts such as (72) for detecting consistent tendencies for one treatment group to have more favorable responses than the other across visits. The reason for this is that such patterns of treatment differences for the respective visits are reinforced in their average. In this sense, the advantage of $Q_{\bar{z}}$ over $Q_{\blacksquare}$ is analogous to that discussed for the extended Mantel–Haenszel statistic $Q_{EMH}$ relative to the total association statistic $Q_{S,T}$ in (17). Moreover, $Q_{\bar{z}}$ has the noteworthy virtue of being specifiable as the primary method of analysis at the time the protocol for a study is prepared. This analysis strategy enables the issue of multiple comparisons (over center or visits) to be avoided in the sense that its basis is one test statistic that encompasses all visits by all patients at all centers. In this setting, the test statistics for the separate centers or separate visits would tend to serve a supportive and descriptive role.

Results from randomization covariance analysis for the responses at all visits of the study in the example are shown in the last two rows of Table 7. The responses in these analyses were expressed in terms of within-center, standardized midranks, and covariance adjustment was based on within-center, standardized midranks for baseline. The strata corresponded to the centers for the analyses in columns 3 to 5 and the center×sex groups for that in column 6. For the multivariate statistic (72), all $p$-values in the next-to-last row of Table 7 are significant ($p \leq 0.05$). The difference between treatment groups is more strongly indicated by the significance ($p \leq 0.005$) of the test statistics $Q_{\bar{z}}$ for the average of the within-center, standardized

midranks over visits. Thus, these multivisit analyses clearly provide overall support for the conclusion that the responses to active treatment are more favorable than those to placebo.

## 4.2 Weighted Least Squares Analysis

Methods for fitting statistical models to describe the relationship between the response variables at each of the respective visits of a clinical trial and treatment, center, and background variables were discussed in Sections 3.3 and 3.4. For any of these models, the variation of the corresponding parameters across visits expresses the effects of visits. Such variation for the treatment effect constitutes treatment × visit interaction; so confirmation that treatment × visit interaction is negligible enables treatment effects to be interpretable as homogeneous across visits. Similar considerations apply to interactions between visits and other components of within-visit models.

A general strategy for analyzing the across-visits variation of the parameters of within-visit models has the following three stages:

i. Univariate methods such as those in Sections 3.3 and 3.4 are applied to fit models for each visit separately.
ii. A consistent estimate for the covariance matrix for the estimated parameters corresponding to all visits is constructed in a manner that accounts appropriately for the multivariate structure of the responses at the respective visits.
iii. Hypotheses concerning model parameters from stage (i) for all visits are tested with Wald statistics such as (51) or (62) or equivalent methods; also, simplified linear models can be fit to the parameters of within-visit models by weighted least squares methods in order to describe the effects of visits and the interactions between visits and treatments, centers, and background variables.

Aspects of the strategy in (i)–(iii) are described by Stram et al. (1988) for situations where maximum likelihood methods are used to fit logistic models to dichotomous response variables at the respective visits or proportional odds models to ordinal response variables. However, the application of such analysis to the data in Table 1 is beyond the scope of this chapter since the construction of an appropriate estimated covariance matrix for stage (ii) is conceptually and computationally complicated.

A relatively straightforward framework for the application of the strategy in (i)–(iii) is based on the multivisit extension of the weighted least squares methods outlined in (56)–(61) for the fitting of linear models. For such analysis, a noteworthy requirement is the availability of moderately

large sample sizes for the subpopulations that correspond to the cross-classification of the explanatory variables in the model. Thus it has the limitation of only being able to account for a small number of categorical explanatory variables. In view of this consideration, its illustration here is directed primarily at the effects of treatment and center on the integer score means $\eta_{ghi}$ through linear models like (56); such analysis is also provided for the probabilities of good or excellent response $\theta_{ghi}$ in (44).

Weighted least squares analysis of linear models for the $\eta_{ghi}$ is applied to the corresponding estimates $\bar{y}_{ghi}$ in (32). Let $\bar{\mathbf{y}}_{*hi} = (\bar{y}_{1hi}, \bar{y}_{2hi}, \bar{y}_{3hi}, \bar{y}_{4hi})'$ denote the vector of estimated means at the respective visits for the patients who received the $i$th treatment at the $h$th center. A consistent estimate for the covariance matrix of $\bar{\mathbf{y}}_{*hi}$ is

$$\mathbf{V}_{hi,y} = \frac{\sum_{k=1}^{n_{hi}} (\mathbf{y}_{*hik} - \bar{\mathbf{y}}_{*hi})(\mathbf{y}_{*hik} - \bar{\mathbf{y}}_{*hi})'}{n_{hi}^2}, \tag{74}$$

where $\mathbf{y}_{*hik} = (y_{1hik}, y_{2hik}, y_{3hik}, y_{4hik})'$ denotes the vector of responses at the respective visits by the $k$th patient with the $i$th treatment at the $h$th center. The compound vector $\bar{\mathbf{y}} = (\bar{\mathbf{y}}'_{*1A}, \bar{\mathbf{y}}'_{*1P}, \bar{\mathbf{y}}'_{*2A}, \bar{\mathbf{y}}'_{*2P})'$ concisely expresses all means in the $(2 \times 2 \times 4)$ cross-classification of center, treatment, and visit. A consistent estimate for its covariance matrix is the block diagonal matrix $\mathbf{V}_y$ with the $\mathbf{V}_{hi,y}$ in (74) as the diagonal blocks; that is,

$$\mathbf{V}_y = \begin{bmatrix} \mathbf{V}_{1A,y} & \mathbf{0}_{44} & \mathbf{0}_{44} & \mathbf{0}_{44} \\ \mathbf{0}_{44} & \mathbf{V}_{1P,y} & \mathbf{0}_{44} & \mathbf{0}_{44} \\ \mathbf{0}_{44} & \mathbf{0}_{44} & \mathbf{V}_{2A,y} & \mathbf{0}_{44} \\ \mathbf{0}_{44} & \mathbf{0}_{44} & \mathbf{0}_{44} & \mathbf{V}_{2P,y} \end{bmatrix} \tag{75}$$

where $\mathbf{0}_{44}$ denotes a $(4 \times 4)$ matrix of 0's. A linear model that describes the variation among the $\eta_{ghi} = E(\bar{y}_{ghi})$ across centers, treatments, and visits, can be concisely expressed as

$$E\{\bar{\mathbf{y}}\} = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}, \tag{76}$$

where $\mathbf{X}$ is the specification matrix with full rank $u$, and $\boldsymbol{\beta}$ is the $(u \times 1)$ vector of unknown coefficients. The weighted least squares estimates $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ are given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}_y^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_y^{-1}\bar{\mathbf{y}}. \tag{77}$$

Since the sample sizes $n_{hi}$ are considered sufficiently large for $\bar{\mathbf{y}}$ to have an approximately multivariate normal distribution, $\hat{\boldsymbol{\beta}}$ has an approximately multivariate normal distribution for which the covariance matrix is consis-

tently estimated by

$$\hat{\mathbf{V}}_{\beta} = (\mathbf{X}'\mathbf{V}_{y}^{-1}X)^{-1}. \tag{78}$$

A goodness-of-fit statistic for the model (76) is the weighted residual sum of squares

$$Q_W = (\bar{\mathbf{y}} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{V}_{y}^{-1}(\bar{\mathbf{y}} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{79}$$

When $\bar{\mathbf{y}}$ is compatible with (76), $Q_W$ approximately has the chi-square distribution with d.f. = (dimension $\bar{\mathbf{y}}$) − (dimension $\boldsymbol{\beta}$) = $16 - u$. Further analysis of models with satisfactory goodness of fit is often undertaken through tests of linear hypotheses $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, where $\mathbf{C}$ is a corresponding $(c \times u)$ specification matrix. For such hypotheses, the Wald statistic

$$Q_C = \hat{\boldsymbol{\beta}}'\mathbf{C}'\{\mathbf{C}\hat{\mathbf{V}}_{\beta}\mathbf{C}'\}^{-1}\mathbf{C}\hat{\boldsymbol{\beta}} \tag{80}$$

approximately has the chi-square distribution with d.f. = $c$.

The weighted least squares methods described in (74)–(80) for the integer score means $\bar{y}_{ghi}$ in (32) are also applicable to the analysis of the linear models for the proportions of good or excellent response $p_{ghi}$ in (27). In fact, the $p_{ghi}$ are means of indicator variables that have the value 1 if the response of a subject is good or excellent (i.e., $y_{ghik} = 3, 4$) and the value 0 if otherwise. For both the $p_{ghi}$ and the $\bar{y}_{ghi}$, the weighted least squares methods in (74)–(80) have the same advantageous theoretical properties which were discussed in Section 3.3 for their counterparts (55)–(62) for the separate visits. These include the BAN property for the estimates $\hat{\boldsymbol{\beta}}$ and asymptotic equivalence of $Q_W$ and $Q_C$ to log-likelihood ratio test statistics.

A convenient model for the preliminary evaluation of sources of variation for the means $\bar{y}_{ghi}$ or the proportions of good or excellent response $p_{ghi}$ has $\mathbf{X} = \mathbf{I}_{16}$, where $\mathbf{I}_{16}$ denotes the $(16 \times 16)$ identity matrix. It is usually called the cell means (or identity) model. Since $\boldsymbol{\beta} = \boldsymbol{\eta}$, $\hat{\boldsymbol{\beta}} = \bar{\mathbf{y}}$, and $\hat{\mathbf{V}}_{\beta} = \mathbf{V}_{y}$ for this model, the Wald statistic $Q_C$ in (80) applies to tests of hypotheses $H_0$: $\mathbf{C}\boldsymbol{\eta} = \mathbf{0}$. Results from tests of this type for the effects of treatment, center, visit, and their interactions are shown in Table 13 for the model O heading. The corresponding $\mathbf{C}$ matrix for some of these tests are

$$\mathbf{C}_T = [\mathbf{1}_4', -\mathbf{1}_4', \mathbf{1}_4', -\mathbf{1}_4'] \text{ for treatment,} \tag{81}$$

$$\mathbf{C}_V = [\mathbf{I}_3, -\mathbf{1}_3, \mathbf{I}_3, -\mathbf{1}_3, \mathbf{I}_3, -\mathbf{1}_3, \mathbf{I}_3, -\mathbf{1}_3] \text{ for visit,} \tag{82}$$

$$\mathbf{C}_{TV} = [\mathbf{I}_3, -\mathbf{1}_3, -\mathbf{I}_3, \mathbf{1}_3, \mathbf{I}_3, -\mathbf{1}_3, -\mathbf{I}_3, \mathbf{1}_3] \text{ for treatment} \times \text{visit,} \tag{83}$$

where $\mathbf{I}_3$ denotes the $(3 \times 3)$ identity matrix and $\mathbf{1}_3 = (1, 1, 1)'$. For both the $p_{ghi}$ and the $\bar{y}_{ghi}$, the tests for treatment × center interaction and treatment × center × visit interaction had clearly nonsignificant $p$-values

**Table 13** $p$-Values for Effects of Treatment, Center, Visit, and Their Interactions in Linear Models to Describe the Variation of Probability of Good or Excellent Response and Integer Score Mean Response at Visit 1, 2, 3, and 4[a]

| Source of variation | Statistic | Probability good or excellent response | | Integer score mean response | |
|---|---|---|---|---|---|
| | | Model 0 | Model 1 | Model 0 | Model 1 |
| Treatment | $Q_W$ (d.f. = 1) | 11.72 | 13.18 | 13.76 | 15.73 |
| | $p$-value | 0.001[b] | < 0.001[b] | < 0.001[b] | < 0.001[b] |
| Center | $Q_W$ (d.f. = 1) | 9.07 | 9.80 | 6.53 | 9.20 |
| | $p$-value | 0.003[b] | 0.002[b] | 0.011[c] | 0.002[b] |
| Treatment × center | $Q_W$ (d.f. = 1) | 0.77 | Not in | 0.80 | Not in |
| | $p$-value | 0.38 | model | 0.37 | model |
| Visit | $Q_W$ (d.f. = 3) | 3.52 | 4.67 | 2.40 | 2.57 |
| | $p$-value | 0.32 | 0.20 | 0.49 | 0.46 |
| Treatment × visit | $Q_W$ (d.f. = 3) | 3.34 | 3.79 | 12.60 | 12.01 |
| | $p$-value | 0.34 | 0.29 | 0.006[b] | 0.007[b] |
| Center × visit | $Q_W$ (d.f. = 3) | 4.49 | 5.30 | 8.13 | 10.16 |
| | $p$-value | 0.21 | 0.15 | 0.043[c] | 0.017[c] |
| Treatment × center × visit | $Q_W$ (d.f. = 3) | 0.34 | Not in | 0.58 | Not in |
| | $p$-value | 0.95 | model | 0.90 | model |
| Goodness of fit (i.e., sources not in model) | $Q_W$ (d.f. = 4) | Does not | 1.18 | Does not | 1.18 |
| | $p$-value | apply | 0.88 | apply | 0.88 |

[a]Model 0 is the cell mean model with $\mathbf{X} = \mathbf{I}_{16}$, where $\mathbf{I}_{16}$ is the 16 × 16 identity matrix; model 1 is the reduced model in (84) with the treatment × center interaction and the treatment × center × visit interaction excluded. The $p$-values are based on Wald statistics from (80) with specifications such as (81)–(83) for model 0. Computations were performed with the CATMOD Procedure in the SAS System (1985).

[b]Results with $p \leq 0.01$.

[c]Results with $p \leq 0.05$.

(i.e., $p > 0.25$). This finding suggested that a reduced model which did not include these sources of variation might be appropriate. This reduced model has the specification

$$
\begin{aligned}
\mathbf{E}\{\bar{\mathbf{y}}\} &= \boldsymbol{\eta} \\
&= \begin{bmatrix}
1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\
1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\
1 & -1 & 1 & 1 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 \\
1 & -1 & 1 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 1 & 0 \\
1 & -1 & 1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 1 \\
1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 \\
1 & 1 & -1 & 1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 \\
1 & 1 & -1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\
1 & 1 & -1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & -1 \\
1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 \\
1 & -1 & -1 & 1 & 0 & 0 & -1 & 0 & 0 & -1 & 0 & 0 \\
1 & -1 & -1 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & -1 & 0 \\
1 & -1 & -1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & -1 \\
1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & 1
\end{bmatrix}
\begin{bmatrix} \xi \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \\ \beta_8 \\ \beta_9 \\ \beta_{10} \\ \beta_{11} \end{bmatrix} \\
&= \mathbf{X}_1\boldsymbol{\beta}, \qquad (84)
\end{aligned}
$$

where $\xi$ represents the average of the $\eta_{ghi}$, $\beta_1$ corresponds to treatment, $\beta_2$ corresponds to center, $\beta_3$, $\beta_4$, $\beta_5$ correspond to visits, $\beta_6$, $\beta_7$, $\beta_8$ correspond to treatment × visit and $\beta_9$, $\beta_{10}$, $\beta_{11}$ correspond to center × visit. Results from statistical tests concerning the model in (84) are shown in Table 13 under the model 1 heading. The ones for goodness of fit supported use of the model (84) by their nonsignificance. Among the tests concerning model parameters, those for treatment × visit and center × visit were significant ($p \leq 0.050$) for the integer mean scores $\bar{y}_{ghi}$ but nonsignificant for the $p_{ghi}$. On the basis of these findings, final descriptive models for the $p_{ghi}$ and the $\bar{y}_{ghi}$ were formulated. The specification matrices for these models were represented as follows:

$$
\mathbf{X}_{2P} = \begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{bmatrix}' \qquad (85)
$$

$$\mathbf{X}_{2y} = \begin{bmatrix} 1\,1\,1\,1 & 1\,1\,1\,1 & 1\,1\,1\,1 & 1\,1\,1\,1 \\ 1\,1\,1\,1 & 0\,0\,0\,0 & 1\,1\,1\,1 & 0\,0\,0\,0 \\ 0\,0\,0\,0 & 0\,0\,0\,0 & 1\,1\,1\,1 & 1\,1\,1\,1 \\ 0\,0\,0\,0 & 0\,0\,0\,0 & 0\,1\,1\,0 & 0\,1\,1\,0 \\ 0\,1\,1\,0 & 0\,0\,0\,0 & 0\,1\,1\,0 & 0\,0\,0\,0 \end{bmatrix}. \tag{86}$$

For the model $\mathbf{X}_{2y}$, the fourth column accounts for the center × visit interaction's correspondence to somewhat smaller differences between centers at visits 2 and 3 than at visits 1 and 4; and the fifth column accounts for the treatment × visit interaction's correspondence to somewhat larger differences between treatments at visits 2 and 3 than at visits 1 and 4. Estimated parameters, standard errors, and $p$-values from statistical tests are shown in Table 14 for the final models in (85) and (86). For the $p_{ghi}$, these results indicate that treatment effects are significant ($p \leq 0.001$) and that the proportion of patients with good or excellent response for active treatment homogeneously exceeds that for placebo by 0.268 at all visits for each center. The interpretation of treatment effects for the $\bar{y}_{ghi}$ is somewhat more complicated because their variation involved significant treatment × visit interaction. One way to address this matter is through consideration of the predicted values $\hat{\eta} = \mathbf{X}_{2y}\hat{\boldsymbol{\beta}}$ from the model (86). These predicted values and their corresponding standard errors (from square roots of the diagonal elements of $\hat{\mathbf{V}}_{\eta} = \mathbf{X}_{2y}\hat{\mathbf{V}}_{\beta}\mathbf{X}'_{2y}$) are given in Table 15; their counterparts for the $p_{ghi}$ are also given there. The pattern of variation among the predicted values $\hat{\eta}_{ghi}$ for both centers indicates that the integer mean score for active treatment exceeds that for placebo by 0.54 at visits 1 and 4 and by 0.89 at visits 2 and 3. Also, in view of the results in Table 14, each is significant ($p < 0.01$) in its own right, and the difference between them is significant ($p < 0.01$). Thus the conclusion of more favorable response for active treatment than placebo is generalizable across both centers and visits even though the extent of treatment differences is heterogeneous across visits.

In summary, weighted least squares methods are useful for analyzing the variation of linear summary statistics such as the $p_{ghi}$ or the $\bar{y}_{ghi}$ across the treatments, centers, and visits of a multicenter, multivisit study. An important aspect of their application is the evaluation of the generalizability of conclusions concerning treatment effects across visits through statistical tests of treatment × visit interaction. Additional discussion of the usage of weighted least squares methods for the multivariate analysis of categorical data from multivisit studies is given in Koch et al. (1977, 1980a, 1983, 1987, 1989), and Carr et al. (1989). These references provide examples which are directed at the pattern of variation of certain types of nonlinear summary

**Table 14** Weighted Least Squares Estimates, Standard Errors, and $p$-Values for Parameters in Final Linear Models for Describing the Variation of Probability of Good or Excellent Response and Integer Score Mean Response in Terms of Effects of Treatment, Center, Visit, and Their Interactions and $p$-values for Goodness of Fit of Model[a]

| | Probability good or excellent response | | | Integer score mean response | | |
|---|---|---|---|---|---|---|
| Parameter | Estimate | SE | $p$-value | Estimate | SE | $p$-value |
| Reference value for placebo in center 1 at visit 1 | 0.300 | 0.060 | $< 0.001$[b] | 1.97 | 0.17 | $< 0.001$[b] |
| Effect for active treatment | 0.268 | 0.067 | $< 0.001$[b] | 0.54 | 0.17 | 0.002[b] |
| Effect for center 2 | 0.239 | 0.067 | $< 0.001$[b] | 0.80 | 0.17 | $< 0.001$[b] |
| Interaction effect for active treatment at visits 2 and 3 | Does not apply | Does not apply | Does not apply | 0.35 | 0.10 | $< 0.001$[b] |
| Interaction effect for visits 2 and 3 at Center 2 | Does not apply | Does not apply | Does not apply | −0.30 | 0.10 | 0.002[b] |
| Goodness of fit | $Q_W$ (d.f. = 13) = 15.35, $p = 0.29$ | | | $Q_W$ (d.f. = 11) = 8.79, $p = 0.64$ | | |

[a]The linear models have the specifications shown in (85) and (86); that is, probability of good or excellent response $= \xi + \beta_1 x_1 + \beta_2 x_2$ and integer score mean response $= \xi + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$. For these models, $x_1 = 1$ if active or $x_1 = 0$ if placebo indicates treatment, $x_2 = 1$ if center 2 or $x_2 = 0$ if center 1 indicates center, $x_3 = 1$ if visit 2 or visit 3 for center 2 or $x_3 = 0$ if otherwise indicates (center × visit) interaction, and $x_4 = 1$ if visit 2 or visit 3 and active or $x_4 = 0$ if otherwise indicates (treatment × visit) interaction, and $\xi, \beta_1, \beta_2, \beta_3, \beta_4$ are unknown parameters as applicable. Estimates for the unknown parameters are from weighted least squares methods via (77); their standard errors are square roots of the diagonal elements of (78); and their $p$-values are from Wald statistics such as (80). The goodness-of-fit statistic is the weighted sum of squares (79) due to residuals from the model's predicted values. Computations were performed with the CATMOD Procedure in the SAS System (1985).

[b]Results with $p \leq 0.01$.

**Table 15** Predicted Values and Standard Errors from Final Linear Models for Describing the Variation of Probability of Good or Excellent Response and Integer Score Mean Response in Terms of Effects of Treatment, Center, Visit, and Their Interactions[a]

| Response variable | Treatment | Statistic | Probability good or excellent response | | Integer score mean response | |
|---|---|---|---|---|---|---|
| | | | Center 1 | Center 2 | Center 1 | Center 2 |
| Visit 1 | Active | Estimate | 0.568 | 0.807 | 2.50 | 3.30 |
| | | SE | 0.061 | 0.051 | 0.14 | 0.14 |
| | Placebo | Estimate | 0.300 | 0.539 | 1.97 | 2.76 |
| | | SE | 0.060 | 0.059 | 0.17 | 0.15 |
| Visit 2 | Active | Estimate | 0.568 | 0.807 | 2.86 | 3.36 |
| | | SE | 0.061 | 0.051 | 0.15 | 0.15 |
| | Placebo | Estimate | 0.300 | 0.539 | 1.97 | 2.47 |
| | | SE | 0.060 | 0.059 | 0.17 | 0.16 |
| Visit 3 | Active | Estimate | 0.568 | 0.807 | 2.86 | 3.36 |
| | | SE | 0.061 | 0.051 | 0.15 | 0.15 |
| | Placebo | Estimate | 0.300 | 0.539 | 1.97 | 2.47 |
| | | SE | 0.060 | 0.059 | 0.17 | 0.16 |
| Visit 4 | Active | Estimate | 0.568 | 0.807 | 2.50 | 3.30 |
| | | SE | 0.061 | 0.051 | 0.14 | 0.14 |
| | Placebo | Estimate | 0.300 | 0.539 | 1.97 | 2.76 |
| | | SE | 0.060 | 0.059 | 0.17 | 0.15 |

[a]Results are based on the linear models with specifications given in Table 14.

statistics as well as linear ones; such nonlinear statistics include estimates for logits (e.g., the $\log_e[p_{ghi}/(1-p_{ghi})]$) and rank measures of association such as those considered in Semenya et al. (1983).

## 4.3 Computations

The computations for the $p$-values in Table 7 from randomization covariance analyses were performed with the procedures documented in Amara and Koch (1980); they could also be obtained with algorithms in the IML Procedure of the SAS System (1985). The CATMOD Procedure in the SAS System (1985) was used to obtain the results in Tables 13 to 15.

## 5 SUMMARY OF ROLES OF ALTERNATIVE METHODS

In this chapter we have methods for the analysis of categorical response variables from studies in the pharmaceutical sciences. These methods addressed the following underlying components of situations in statistical practice.

1. Measurement scale (dichotomous and ordinal response variables)
2. Data structure [case records as in Table 1 and contingency tables as in (5) and (20)]
3. Dimension (univariate consideration of the response at each visit separately and multivariate consideration of the responses at all visits simultaneously)
4. Extent of assumptions and their implications to the scope of inference (design-based inference for the study population and model-based inference for a conceptual target population)
5. Management of explanatory variables for center and background characteristics of patients (stratification adjustment and covariance adjustment)

Analyses that corresponded to appropriate combinations of the features encompassed by the components 1–5 were illustrated for data from a multicenter, multivisit clinical trial. These included nonparametric statistical tests for design-based inferences concerning the existence of differences between treatments and the fitting of statistical models for the description of the relationship between response variables and treatment, center, and background variables. Also, the statistical models provided a framework for the evaluation of the generalizability of treatment differences through statistical tests of treatment × center interaction and treatment × background variable interaction.

## APPENDIX: USAGE OF THE SAS SYSTEM (1985) TO GENERATE RESULTS FROM ANALYSIS

The following discussion assumes that the reader has a working knowledge of the SAS System, particularly the basic concepts of the data and proc steps.

## A.1 FREQ Procedure

The FREQ procedure was used to obtain the results displayed in Tables 1 to 5 and discussed in Sections 2, 3.1, 3.2. FREQ performs statistical tests and computes measures of association for contingency tables; it also calculates the extended Mantel–Haenszel statistic. The input for the procedure consists of observations with variables containing information pertaining to treatment, background, and response outcomes. FREQ produces the appropriate cross-tabulation or tables based on the specified variables, and then calculates the desired statistics. In the following illustrations, SEX refers to the variable for sex, CENTER indicates center 1 or center 2, TRTMENT indicates treatment category (active or placebo), AGE means age in years, BASE represents baseline status, and DBASE represents the dichotomous baseline response outcome. The following statements produce many of the results displayed in Table 2.

```
PROC FREQ;
TABLES CENTER*TRTMENT*(SEX BASE DBASE)/SCORES
    =MODRIDIT CHISQ CMH;
```

The TABLES statement request that three sets of two way tables be formed. Each set has two tables for the two centers, and each table has TRTMENT as the rows and SEX, BASE, or DBASE as the columns. The resulting sets of two-way tables have a three-way structure [e.g., a $(2\times2\times2)$ table applies to (CENTER × TRTMENT × SEX)]. The CHISQ option requests that Fisher's exact test, the Pearson chi-square, and other test statistics be calculated for each two-way table in each set. Measures of association based on chi-square are also printed. The CMH option requests that extended Mantel–Haenszel statistics be computed for the association of SEX with TRTMENT adjusting for CENTER. SCORES=MODRIDIT specifies that standardized ranks are to be used as scores. This is how the $p$-values for BASE in Table 2 for the combined centers were obtained. The default scores used by PROC FREQ are TABLE scores, which are the row and column heading values for numeric classification variables, and the integers for character-valued variables. They were used for SEX and DBASE.

The following statements allow one to produce Wilcoxon rank sum statistics for center 1 and center 2 for the association of treatment with age and also the extended Mantel–Haenszel statistic for the combined centers.

```
PROC FREQ;
BY CENTER;
TABLES TRTMENT*AGE/SCORES=MODRIDIT CMH NOPRINT;
```

```
PROC FREQ;
TABLES CENTER*TRTMENT*AGE/SCORES
    =MODRIDIT CMH NOPRINT:
```

The resulting mean score statistic printed for each center by the first invocation of PROC FREQ can be shown to be equivalent to Wilcoxon rank sum statistics. Note that since age is continuous, and has many levels, the NOPRINT option is employed to suppress printing of the tables. The second invocation of PROC FREQ requests the extended Mantel–Haenszel statistic for the combined centers.

An alternative way in which to obtain the Wilcoxon rank sum statistic for the separate centers would have been to employ the NPAR1WAY procedure of the SAS System. The following statements generate the appropriate results:

```
PROC NPAR1WAY WILCOXON;
    BY CENTER;
    CLASS TRTMENT;
    VAR AGE;
```

The chi-square approximation for the Kruskal–Wallis test is equivalent to the mean score statistic from PROC FREQ discussed previously.

Table 3 contains results of analyses to assess the association of SEX with the responses at visits 1 to 4. PROC FREQ was used to obtain the extended Mantel–Haenszel statistic for this association while adjusting for the effect of CENTER. The following statements illustrate how the FREQ procedure would be used for this purpose:

```
PROC FREQ;
  TABLES CENTER*SEX*(VISIT1 VISIT2 VISIT3 VISIT4)/CMH
      SCORES=MODRIDIT;
```

These statements produce the extended Mantel–Haenszel statistic with standardized ranks (the van Elteren statistic) for the association of sex with the response variables, adjusting for center. Also generated are the Spearman rank correlation chi-square statistic, and its extended Mantel–Haenszel counterpart. The CMH option produces three extended Mantel–Haenszel statistics: a mean score statistic, a correlation statistic, and a general association statistic; some of these may not be appropriate for the data under analysis. One can restrict the analysis to the correlation statistic by specifying the 'CMH1' option instead; similarly, the 'CMH2' option requests both the mean score and correlation statistic.

## A.2 CATMOD Procedure

The CATMOD procedure performs weighted least squares analysis for categorical data. It fits linear models to functions of response probabilities. It also can perform maximum likelihood analysis for logistic regression. Table 8 contains parameter estimates, standard errors, and $p$-values for the logistic regression models for good or excellent response. The following statements were used to obtain these results:

```
PROC CATMOD ORDER=DATA;
POPULATION CENTER TRTMENT;
RESPONSE LOGIT;
MODEL BWK1 = (1 1 0,
              1 0 0,
              1 1 1,
              1 0 1)/ML PRED NOGLS;
```

The input for PROC CATMOD consists of data observations containing the values (0,1) for BWK1, where 1 indicates good or excellent response at visit 1 and 0 other; the variable CENTER takes the value 1 or 2, and TRTMENT is the variable for treatment and takes the values 'active' or 'placebo.' The model specification matrix is direct input in this application of CATMOD since a reference cell model is desired. The default parameterization which CATMOD uses has a centerpoint structure such as (84), in which case the necessary MODEL statement is similar to that employed in the GLM procedure:

```
MODEL BWK1=CENTER TRTMENT/ML PRED NOGLS;
```

As noted previously, CATMOD does include the capacity for users to input their own model specification matrices for greater flexibility.

The POPULATION statement indicates which variables are to be used to determine the subpopulations under investigation. ML is the option used to request maximum likelihood analysis, PRED requests that the predicted and observed response functions be printed for each subpopulation, and NOGLS requests that the standard weighted least squares analysis be suppressed. Note the inclusion of the ORDER=DATA option in the PROC statement. Since it was desired to have center 1, placebo as the reference value, it was necessary to sort the data by CENTER and descending TRTMENT before PROC CATMOD was invoked and then to request CATMOD to create subpopulations based on the order in which it encountered variable values, rather than by the standard sort order. The

parameter estimates produced by CATMOD are the same as those that would be produced by a logistic regression procedure except that the signs are reversed. CATMOD normalizes on the '0' response rather than the '1' response.

Table 10 contains weighted least squares estimates and standard errors for a linear model for integer score mean response at visits 1 to 4. These estimates were also produced with the CATMOD procedure, by using the following statements:

```
PROC CATMOD ORDER= DATA;
 POPULATION CENTER TRTMENT;
 RESPONSE 0 1 2 3 4;
 MODEL WK1 = (1 1 0,
              1 0 0
              1 1 1,
              1 0 1)/PRED COV;
```

The RESPONSE statement serves to specify the type of response function desired. In this case, there will be four response functions, one for each combination of center and treatment. Each response function is the mean response generated by scoring the five possible response outcome categories by 0, 1, 2, 3, 4. WK1 is the variable corresponding to the response for week 1. Similar computer analyses were performed for WK2, WK3, and WK4.

## A.3 LOGIST Procedure

Table 12 contains results generated with the LOGIST procedure of Harrell (1986), also available with the SAS system. This procedure fits the logistic multiple regression model to either a binary response or an ordinal dependent variable. Proportional odds models can be fit with the LOGIST procedure. Variables may have to be coded beforehand in order to represent interaction terms. Following are the statements used to compute some of the results presented in Table 12:

```
PROC LOGIST K=4 PRINTC;
 MODEL WK2 =
     ITRT ISEX AGE ICLINIC BASE TRTSEX
         TRTCLIN TRTBASE AGETRT /
     STEPWISE INCLUDE=5 PRINTI PRINTQ
         SLENTRY=0.001 SLSTAY=0.10;
```

Many of the variables were created in a previous data step by statements such as AGEBASE=AGE*BASE in order to represent interaction effects. The dependent variable here is WK2, or the response outcome at the second week, with possible values from 0 to 4. $K = 4$ on the PROC statement specifies that the value '4' is the largest value allowable for the ordinal dependent variable. The stepwise mode of model building is specified by the STEPWISE option on the MODEL statement; INCLUDE= 5 indicates that the first five independent variables listed are to be included in every model. Other options used on the MODEL statement specify what type of criteria to use for a variable to be entered or retained in a model, as well as what types of parameter estimates and statistics to print out; in particular, PRINTQ indicates that score statistics analogous to (68) are to be printed.

This appendix is intended to give an overview of the types of computational strategies employed to generate the results of the analyses in this chapter. Additional details can be found in the appropriate documentation for the SAS software used.

*Acknowledgments.* The research for this chapter was supported in part by the U.S. Bureau of the Census through Joint Statistical Agreement JSA-84-5. The authors would like to thank Donald Berry, Myra Carpenter, Suzanne Edwards, Amy Goulson, James Grady, William Sollecito, and Kenneth Williams for helpful comments with respect to the preparation of the manuscript. They would also like to thank Ans Janssens for editorial assistance.

## REFERENCES

Agresti, A. (1984). *Analysis of Ordinal Categorical Data.* Wiley, New York.

Amara, I. A., and G. G. Koch (1980). A macro for multivariate randomization analyses of stratified sample data. *Proceedings of the Fifth Annual SAS Users Group International Conference*, pp. 134–144.

Bhapkar, V. P. (1966). A note on the equivalence of two test criteria for hypotheses in categorical data. *J. Amer. Statist. Assoc. 61*, 228–235.

Birch, M. W. (1964). The detection of partial association: I. The $2 \times 2$ case. *J. Roy. Statist. Soc. B 26*, 313–324.

Birch, M. W. (1965). The detection of partial association: II. The general case. *J. Roy. Statist. Soc. B 27*, 111–124.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland (1975). *Discrete Multivariate Analysis: Theory and Practice.* MIT Press, Cambridge, Mass.

Breslow, N. E., and N. E. Day (1980). *Statistical Methods in Cancer Research*, Vol. I: *The Analysis of Case-Control Studies*, International Agency for Research on Cancer, Lyon.

Carr, G. J., K. B. Hafner, and G. G. Koch (1989). Analysis of rank measures of association for ordinal data from longitudinal studies. *J. Amer. Statist. Assoc. 84*, in press.

Clogg, C. C. (1982). Some models for the analysis of association in multiway cross-classifications having ordered categories. *J. Amer. Statist. Assoc. 77*, 803–815.

Cochran, W. G. (1977). *Sampling Techniques*. Wiley, New York

Cox, C., and C. Chuang (1984). A comparison of chi-square partitioning and two logit analyses of ordinal pain data from a pharmaceutical study. *Statist. Med. 3*, 273–285.

Cox, D. R. (1970). *The Analysis of Binary Data*. Chapman & Hall, London.

Everitt, B. S. (1977). *The Analysis of Contingency Tables*. Chapman & Hall, London.

Fienberg, S. E. (1980). *The Analysis of Cross-Classified Categorical Data*. MIT Press, Cambridge, Mass.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*. Wiley, New York.

Fleiss, J. L. (1986). *The Design and Analysis of Clinical Experiments*. Wiley, New York.

Forthofer, R. N., and R. G. Lehnen (1981). *Public Program Analysis: A New Categorical Data Approach*, Wadsworth, Belmont, Calif.

Freeman, D. H., Jr. (1987). *Applied Categorical Data Analysis*, Marcel Dekker, New York.

Friedman, L. M., C. D. Furberg, and D. L. DeMets (1981). *Fundamentals of Clinical Trials*. John Wright-PSG, Littleton, Mass.

Gart, J. J. (1971). The comparison of proportions: a review of significance tests, confidence intervals, and adjustments for stratification. *Internat. Statist. Rev. 39*, 148–169.

Genter, F. C., and V. T. Farewell (1985). Goodness-of-link testing in ordinal regression models. *Canad. J. Statist. 13(1)*, 37–44.

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

Harrell, F. E. (1986). LOGIST. *SUGI Supplementary Library User's Guide*. SAS Institute, Cary, N. C. pp. 269–292.

Holford, T. R. (1980). The analysis of rates and survivorship using log-linear models. *Biometrics 36*, 299–305.

Huitema, B. E. (1980). *The Analysis of Covariance and Alternatives*. Wiley, New York.

Imrey, P. B., G. G. Koch, M. E. Stokes, J. N. Darroch, D. H. Freeman, Jr. and H. D. Tolley (1981). Categorical data analysis: some reflections on the log linear model and logistic regression: p. I. *Internat. Statist. Rev. 49*, 265–283.

Imrey, P. B., G. G. Koch, M. E. Stokes, J. N. Darroch, D. H. Freeman, Jr. and H. D. Tolley (1982). Categorical data analysis: some reflections on the log linear model and logistic regression: p. II. *Internat. Statist. Rev. 50*, 35–64.

Koch, G. G., and V. P. Bhapkar (1982). Chi-square tests. In *Encyclopedia of Statistical Sciences*, Vol. 1 (N. L. Johnson and S. Kotz, eds.). Wiley, New York, pp. 442–457.

Koch, G. G., and S. Edwards (1985). Logistic regression. In *Encyclopedia of Statistical Sciences*, Vol. 5 (N. L. Johnson and S. Kotz, eds.). Wiley, New York, pp. 128–132.

Koch, G. G., and S. Edwards (1987). Clinical efficacy trials with categorical data. In *Handbook of Biopharmaceutical Statistics in Human Drug Development* (Karl E. Peace, ed.). Marcel Dekker, New York, Chap. 9, pp. 403–457.

Koch, G. G., and D. B. Gillings (1983). Inference, design based vs. model based. In *Encyclopedia of Statistical Sciences*, Vol. 4 (N. L. Johnson and S. Kotz, eds.). Wiley, New York, pp. 84–88.

Koch, G. G., and W. A. Sollecito (1984). Statistical considerations in the design, analysis, and interpretation of comparative clinical studies: an academic perspective. *Drug Inform. J. 18*, 131–151.

Koch, G. G., J. R. Landis, J. L. Freeman, D. H. Freeman, Jr. and R. G. Lehnen (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics 33*, 133–158.

Koch, G. G., J. E. Grizzle, K. Semenya, and P. K. Sen (1978). Statistical methods for evaluation of mastitis treatment data. *J. Dairy Sci. 61*, 829–847.

Koch, G. G., I. A. Amara, M. E. Stokes, and D. B. Gillings (1980a). Some views on parametric and non-parametric analysis for repeated measurements and selected bibliography. *Internat. Statist. Rev. 48*, 249–265.

Koch, G. G., D. B. Gillings, and M. E. Stokes (1980b). Biostatistical implications of design, sampling, and measurement to health science data analysis. *Annual Rev. Public Health 1*, 163–225.

Koch, G. G., I. A. Amara, G. W. Davis, and D. B. Gillings (1982). A review of some statistical methods for covariance analysis of categorical data. *Biometrics 38*, 563–595.

Koch, G. G., S. L. Gitomer, L. Skalland, and M. E. Stokes (1983). Some non-parametric and categorical data analyses for a change-over design

study and discussion of apparent carry-over effects. *Statist. Med. 2*, 397–412.

Koch, G. G., P. B. Imrey, J. M. Singer, S. S. Atkinson, and M. E. Stokes (1985a). *Analysis of Categorical Data.* Les Presses de l'Université de Montréal, Montreal.

Koch, G. G., J. M. Singer, and I. A. Amara (1985b). A two-stage procedure for the analysis of ordinal categorical data. In *Biostatistics: Statistics in Biomedical, Public Health and Environmental Sciences*, The Bernard G. Greenberg Volume (P. K. Sen, ed.). North-Holland, New York, pp. 357–387.

Koch, G. G., S. S. Atkinson, and M. E. Stokes (1986). Poisson regression, In *Encyclopedia of Statistical Sciences*, Vol. 7 (N. L. Johnson and S. Kotz, eds.). Wiley, New York, pp. 32–42.

Koch, G. G., J. D. Elashoff, and I. A. Amara (1987). Repeated measurements studies, design and analysis. In *Encyclopedia of Statistical Sciences*, Vol. 8 (N. L. Johnson and S. Kotz, eds.). Wiley, New York, pp. 46–73.

Koch, G. G., J. M. Singer, M. E. Stokes, G. J. Carr, S. B. Cohen, and R. N. Forthofer (1989). Some aspects of weighted least squares analysis for longitudinal categorical data. In *Statistical Models for Longitudinal Studies of Health* (J. H. Dwyer, ed.). Oxford University Press, Oxford, in press.

Kruskal, W. H., and W. A. Wallis (1953). Use of ranks in one criterion variance analysis. *J. Amer. Statist. Assoc. 46*, 583–621.

Landis, J. R., E. R. Heyman, and G. G. Koch (1978). Average partial association in three-way contingency tables: a review and discussion of alternative tests. *Internat. Statist. Rev. 46*, 237–254.

Landis, J. R., M. M. Cooper, T. Kennedy, and G. G. Koch, (1979). A computer program for testing average partial association in three-way contingency tables (PARCAT). *Comput. Programs Biomed. 9*, 223–246.

Larntz, K. (1978). Small sample comparisons of exact levels for chi-squared goodness of fit statistics. *J. Amer. Statist. Assoc., 73*, 253–263.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* Holden–Day, San Francisco.

Mantel, N. (1963). Chi-square tests with one degree of freedom: extensions of the Mantel–Haenszel procedure. *J. Amer. Statist. Assoc. 58*, 690–700.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep. 50*, 163–170.

Mantel, N., and J. Fleiss (1980). Minimum expected cell size requirements for the Mantel–Haenszel one-degree-of-freedom chi-square test and a related rapid procedure. *Amer. J. Epidemiol. 112*, 129–134.

**Mantel**, N., and W. Haenszel (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Nat. Cancer. Inst. 22*, 719–748.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *J. Roy. Statist. Soc. B 42*, 109–142.

McCullagh, P., and J. A. Nelder (1983). *Generalized Linear Models*. Chapman & Hall, New York.

Mehta, C. R., N. R. Patel, and A. A. Tsiatis (1984). Exact significance testing to establish treatment equivalence with ordered categorical data. *Biometrics 40*, 819–825.

Mehta, C. R., N. R. Patel, and R. Gray (1985). Computing an exact confidence interval for the common odds ratio in several 2 × 2 contingency tables. *J. Amer. Statist. Assoc. 80*, 969–973.

Neyman, J. (1949). Contributions to the theory of the $\chi^2$-test. *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability* (J. Neyman ed.). University of California Press, Berkeley, pp. 239–273.

Owen, D. B. (1962). *Handbook of Statistical Tables*. Addison–Wesley, Reading, Mass.

Peterson, B. L. (1986). Proportional odds and partial proportional odds models for ordinal response variables. Dissertation submitted to the Department of Biostatistics, University of North Carolina, Chapel Hill.

Puri, M. L., and P. K. Sen, (1971). *Non-parametric Methods in Multivariate Analysis*. Wiley, New York.

Quade, D. (1967). Rank analysis of covariance. *J. Amer. Statist. Assoc. 62*, 1187–1200.

Quade, D. (1982). Nonparametric analysis of covariance by matching. *Biometrics 38*, 597–611.

SAS Institute, Inc. (1985). *SAS User's Guide: Statistics Version, 5th ed.* SAS Institute, Cary N. C.

Semenya, K. A., G. G. Koch, M. E. Stokes, R. N. Forthofer (1983). Linear models methods for some rank function analyses of ordinal categorical data. *Comm. Statist. 12*, 1277–1298.

Shapiro, S. H., and T. A. Louis, eds. (1983). *Clinical Trials: Issues and Approaches*. Marcel Dekker, New York.

Silvapulle, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *J. Roy. Statist. Soc. B 43*, 310–313.

Stram, D. O., L. J. Wei, and J. H. Ware (1988). Analysis of repeated ordered categorical outcomes with possibly missing observations and time dependent covariates. *J. Amer. Statist. Assoc. 83*, 631–637.

**Thomas**, D. G. (1975). Exact and asymptotic methods for the combination of 2 × 2 tables. *Comput. Biomed. Res. 8*, 423–446.

Tygstrup, N., J. M. Lachin, and E. Juhl (1982). *The Randomized Clinical Trial and Therapeutic Decisions*. Marcel Dekker, New York.

van Elteren, P. H. (1960). On the combination of independent two-sample tests of Wilcoxon. *Bull. Internat. Statist. Inst. 37*, 351–361.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc. 54*, 426–482.

Walker, S. H., and D. B. Duncan (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika 54*, 167–179.