The use of bootstrap methods for estimating sample size and analysing health-related quality of life outcomes (particularly the SF-36)

Stephen John Walters

Degree of PhD

School of Health and Related Research

Sheffield University

September 2003

To My Father

Frederick Walters

And To the Memory of

My Mother

Frances Elizabeth Walters

Contents

Acknowledgements

I would like to thank my supervisor Professor Michael Campbell for all his unstinting help, advice, support and encouragement throughout the six years it has taken me to produce this thesis.

I am therefore particularly grateful for Mike Campbell's detailed comments on the manuscript.

I would also like to thank the anonymous reviewers who commented on earlier drafts of the four published papers.

I am also grateful to Myfanwy Lloyd-Jones and Marion Stobbs for helpful comments on various chapters.

Any remaining errors are my own.

Stephen J Walters

Eckington, Sheffield

September 2003

Stephen John Walters

The use of bootstrap methods for estimating sample size and analysing health-related quality of life outcomes (particularly the SF-36)

Summary of PhD thesis

Health-Related Quality of Life (HRQoL) measures are becoming increasingly used in clinical trials and health services research, both as primary and secondary outcome measures. Investigators are now asking statisticians for advice on how to plan (e.g. sample size) and analyse studies using HRQoL outcomes.

HRQoL outcomes like the SF-36 are usually measured on an ordinal scale. However, most investigators assume that there exists an underlying continuous latent variable that measures HRQoL, and that the actual measured outcomes (the ordered categories), reflect contiguous intervals along this continuum.

The ordinal scaling of HRQoL measures means they tend to generate data that have discrete, bounded and skewed distributions. Thus, standard methods of analysis such as the *t*-test and linear regression that assume Normality and constant variance may not be appropriate. For this reason, non-parametric methods are often used to analyse HRQoL data. The bootstrap is one such computer intensive non-parametric method for estimating sample sizes and analysing data.

From a review of the literature, I found five methods of estimating sample sizes for two-group cross-sectional comparisons of HRQoL outcomes. All five methods (amongst other factors) require the specification of an *effect size*, which varies according to the method of sample size estimation. The empirical effect sizes calculated from the various datasets suggested that large differences in HRQoL (as measured by the SF-36) between groups are unlikely, particularly from the RCT comparisons. Most of the observed effect sizes were mainly in the 'small' to 'moderate' range (0.2 to 0.5) using Cohen's (1988) criteria.

I compared the power of various methods of sample size estimation for two-group cross-sectional study designs via bootstrap simulation. The results showed that under the location shift alternative hypothesis, conventional methods of sample size estimation performed well, particularly Whitehead's (1993) method. Whitehead's method is recommended if the HRQoL outcome has a limited number of discrete values (< 7) and/or the expected proportion of cases at either of the bounds is high. If a pilot dataset is readily available (to estimate the shape of the distribution) then bootstrap simulation may provide a more accurate and reliable estimate, than conventional methods.

Finally, I used the bootstrap for hypothesis testing and the estimation of standard errors and confidence intervals for parameters, in four datasets (which illustrate the different aspects of study design). I then compared and contrasted the bootstrap with standard methods of analysing HRQoL outcomes as described in Fayers and Machin (2000).

Overall, in the datasets studied with the SF-36 outcome the use of the bootstrap for estimating sample sizes and analysing HRQoL data appears to produce results similar to conventional statistical methods. Therefore, the results of this thesis suggest that bootstrap methods are not more appropriate for analysing HRQoL outcome data than standard methods. This result requires confirmation with other HRQoL outcome measures, interventions and populations.

Chapter 1: Introduction

Health Related Quality of Life (HRQoL) measures are becoming more frequently used in clinical trials and health services research (HSR), as both primary and secondary endpoints. Investigators are now asking statisticians for advice on how to plan (e.g. sample size) and analyse studies using HRQoL measures.

The analysis of data from quality of life (QoL) measurements requires some basic assumptions. We will assume that (Olschewski and Schumacher, 1990):

(1) QoL is a subjective construct which is not directly observable and measurable;

(2) QoL is a multi-dimensional construct consisting of different aspects of physical and psychological well being;

(3) QoL is a time-dependent construct reflecting a person's experiences and perceptions over their life history.

HRQoL measures such as the Short Form (SF)-36, Nottingham Health Profile (NHP) and EORTC QLQ-C30 are described in Bowling (1995, 1997) and are usually measured on an ordered categorical (ordinal) scale. This means that responses to individual questions are usually classified into a small number of response categories, which can be ordered, for example, poor, moderate and good. The question responses are often analysed by assigning equally spaced numerical scores to the ordinal categories (e.g. 0 = 'poor', 1 = 'moderate' and 2 = 'good') and the scores across similar questions are then summed to generate a HRQoL measurement. These 'summated scores' are treated as if they were from a continuous distribution and were Normally distributed. We will also assume that there exists an underlying continuous latent variable that measures HRQoL (although not necessarily Normally distributed), and that the actual measured outcomes are ordered categories that reflect contiguous intervals along this continuum.

However, this scaling of HRQoL measures may lead to several problems in determining sample size and analysing the data (Walters *et al* 2001a, 2001b).

(1) The apparent continuum hides the fact that only a few discrete values are possible.
(2) The scale may not be linear.
(3) The scales are bounded and have range-limited values.
(4) Methods based on the Normal distribution (such as linear regression) assume that the outcome variable has a constant variance. The variance of HRQoL may not be constant.
(5) Normal approximations may not apply.
(6) Missing values are likely.
(7) It is difficult to quantify an effect size in advance.

The advantages in being able to treat HRQoL scales as continuous and Normally distributed are simplicity in sample size estimation and statistical analysis. Therefore, it is important to examine such simplifying assumptions for different instruments and their scales. Since HRQoL outcome measures may not meet the distributional requirements (usually that the data have a Normal distribution) of parametric methods of sample size estimation and analysis, non-parametric methods are often used to analyse HRQoL data.

The bootstrap is an important non-parametric method for estimating sample size and analysing data (including hypothesis testing, standard error and confidence interval estimation). The bootstrap is a data based simulation method for statistical inference, which involves repeatedly drawing random samples from the original data, with replacement. It seeks to mimic, in an appropriate manner, the way the sample is collected from the population in the bootstrap samples from the observed data. The 'with replacement' means that any observation can be sampled more than once. HRQoL outcome measures actually generate data with discrete, bounded and non-standard distributions. So, in theory, computer intensive methods such as the bootstrap that make no distributional assumptions may therefore be more appropriate for estimating sample size and analysing HRQoL data than conventional statistical methods.

Conventional methods of analysis of HRQoL outcomes are extensively discussed in Fayers and Machin (2000) and Fairclough (2002). However, neither of these texts used the bootstrap to analyse HRQoL outcomes. As a consequence of this omission, the aim of this thesis is to compare bootstrap computer simulation methods with standard methods of sample size determination and analysis of HRQoL measures (particularly the SF-36).

The SF-36 is the most commonly used health status measure in the world today and has many of the problems with HRQoL measures described above. To compare and contrast the bootstrap methods with standard methods, we use SF-36 HRQoL data from a variety of cross-sectional and longitudinal studies including randomised controlled trials (RCTs).

The remainder of this thesis is structured into the following chapters. Chapter 2 describes the SF-36 HRQoL measure and the potential problems in calculating sample sizes and analysing such an outcome. Chapter 2 also describes the various datasets that are going to be used throughout the rest of the thesis to illustrate the methods. Chapter 3 briefly describes the bootstrap method of computer simulation. The results of a literature review of methods of sample size estimation and analysis of HRQoL measures are summarised in Chapter 4. The observed effects and summary statistics from the example datasets are shown in Chapter 5. Chapter 6 compares the power of various methods of sample size estimation and test statistics for some simple two group cross-sectional comparisons via bootstrap simulation. Methods of analysing cross-sectional HRQoL data and HRQoL measured at two time-points (say baseline and a single follow-up) are discussed in Chapter 7. Chapter 8 discusses summary measures, such as the area under the curve (AUC) and the modelling of longitudinal data, for repeated HRQoL (three or more) measurements over time. The final two chapters (9 and 10) end with a discussion and conclusions.

Chapter 2: Description of the SF-36 and the datasets

The SF-36 Health Survey

The Medical Outcomes Study (MOS) Short Form (SF)-36 is the most commonly used health status measure in the world today (Kaplan, 1998). It originated in the USA (Ware *et al 1992*), but it has been validated for use in the UK (Brazier *et al* 1992). It contains 36 questions measuring health across eight dimensions: physical functioning (PF) 10 items; role limitation because of physical health (RP) 4 items; social functioning (SF) 2 items, vitality (VT) 4 items; bodily pain (BP) 2 items; mental health (MH) 5 items; role limitation because of emotional problems (RE) 3 items and general health (GH) 5 items. (The full version of the SF-36 is shown in Appendix 1).

The responses to the 36 individual questions are classified into a mixture of binary (yes/no) and three, five and six point ordered response categories. In planning and analysis, the question responses are often analysed by assigning equally spaced numerical scores to the ordinal categories (e.g. 0 = 'poor', 1 = 'moderate' and 2 = 'good'). The raw scores across similar questions are combined to generate a raw dimension score. Finally, these raw dimension scores are then transformed to generate a HRQoL score from 0 to 100, where 100 indicates "good health". Figure 2.1 shows the distribution of the eight main dimensions of the SF-36 from a general population survey of Sheffield residents (Brazier *et al* 1992).

Two further summary components, the Mental Component Summary (MCS) and Physical Component Summary (PCS) have also been derived from the eight dimensions using factor analysis (Ware *et al* 1994). The PCS and MCS scales of the SF-36 are standardized such that a mean score of 50 (standard deviation 10) reflects the mean score of a standard population. Thus, the SF-36 generates a profile of HRQoL outcomes (on up to 10 dimensions), which makes statistical analysis and interpretation difficult (Fayers and Machin, 2000). We will concentrate on the analysis of the eight main dimensions of the SF-36 rather than the two summary components.

The ordinal scaling of HRQoL measures such as the SF-36 may lead to several problems in determining sample size and analysing the data.

(1) The apparent continuum hides the fact that only a few discrete values are possible. For example, the Role Physical (RP) dimension of the SF-36 is scored on a 0 to 100 scale but there are only five possible categories/scores e.g. 0, 25, 50, 75 and 100 (see Figure 2.1c).

(2) The scale may not be linear. For example, using the SF-36 RP dimension, is a change of score from 0 to 25 the same as a change from 75 to 100?

(3) There is often a floor or ceiling effect. Patients cannot be worse than the worst category or better than the best category. (In the case of the SF-36 score either 0 or 100). For some populations the level is wrong and most people score on either the best category or the worst category. Floor and ceiling effects are more likely to be a problem in longitudinal studies because they limit the ability of the instrument to detect an improvement or deterioration in a patient's HRQoL over time. Figure 2.1c shows that for the RP dimension of the SF-36 over 72% (1000/1372) of the general population sample had scored 100 and were at the ceiling of the distribution.

(4) Methods based on the Normal distribution (such as linear regression) assume that the outcome variable has a constant variance. The variances of changes may depend on initial values. This is a common problem with range-limited values. Patients may enter the study with a wide variety of scores, but tend always to increase their scores. Thus patients who score lower at the start of the study have more range to improve than those who are already close to the maximum.

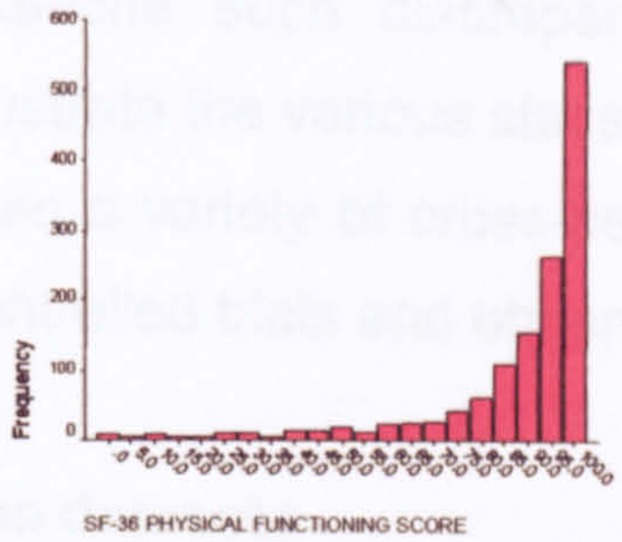
(5) Normal approximations may not apply. Since the data are in fact categorical, they may require different techniques of analysis. By definition, no ordinal variable can be Normally distributed, although in some cases a Normal approximation will suffice.

(6) Missing values are likely, for example in questionnaires that ask 'how far can you walk?' when the patient is in a wheel chair.

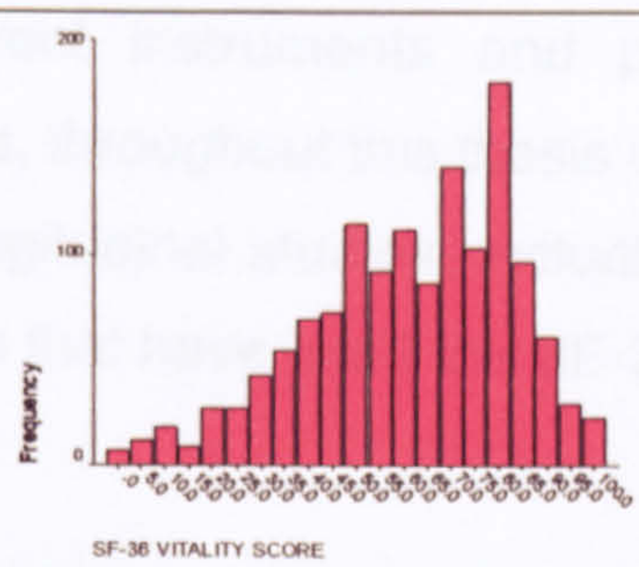
(7) It is difficult to quantify an effect size (e.g. a desirable difference in mean score between groups) in advance.

Figure 2.1: Distribution of the eight SF-36 Dimensions from a general population survey (n=1372)

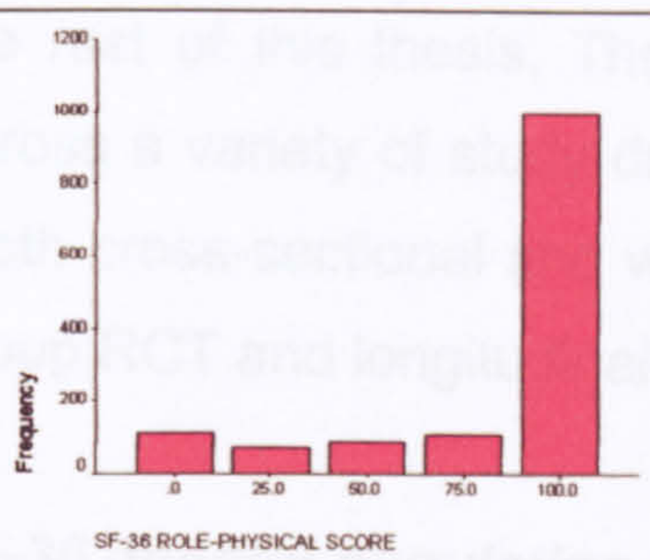
a)



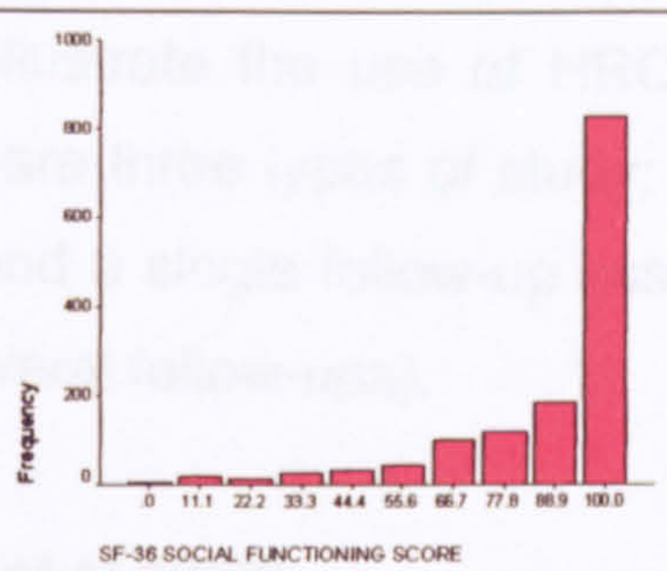
b)



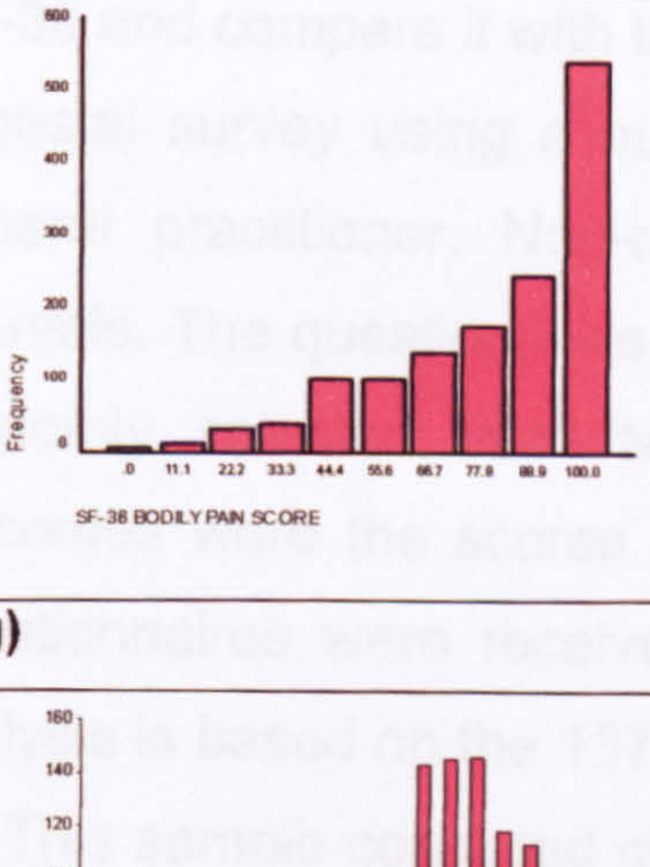
c)



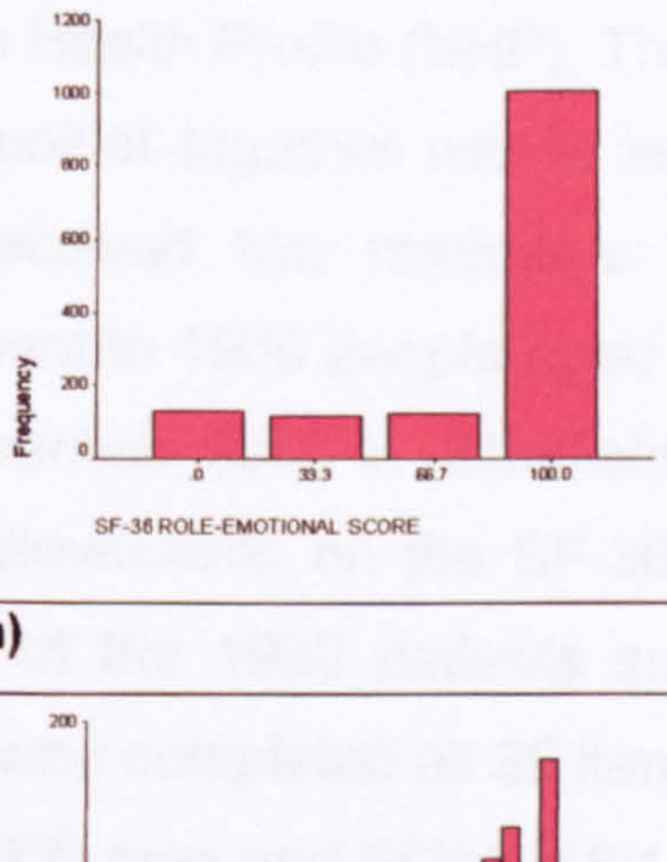
d)



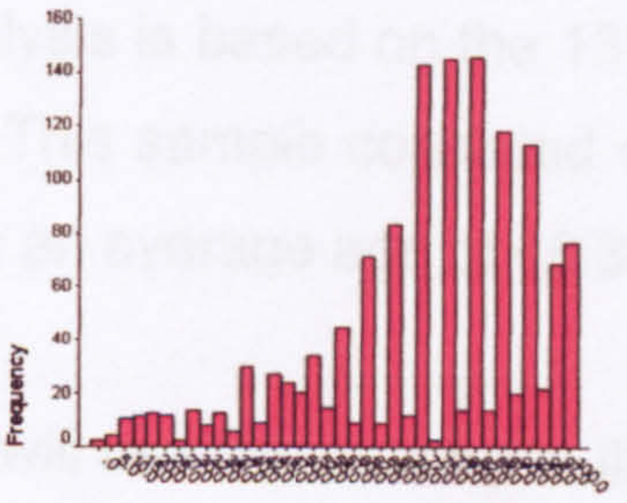
e)



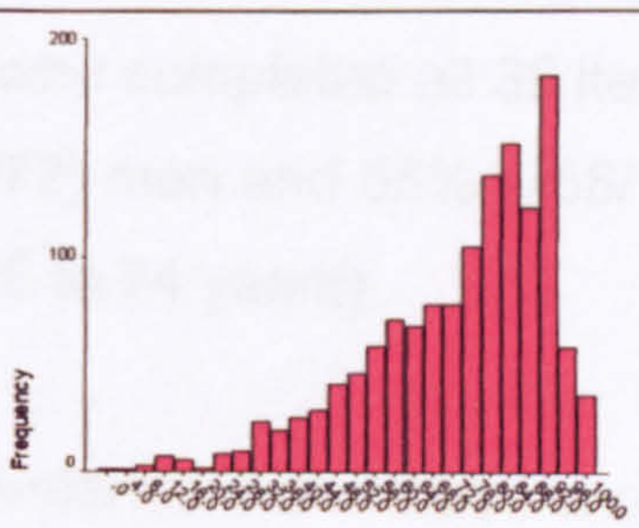
f)



g)



h)



There are advantages in being able to treat HRQoL scales as continuous (e.g. for statistical analysis and economic evaluations) and, therefore, it is important to examine such discrepancies for different instruments and populations. To illustrate the various statistical techniques, throughout this thesis we will use data from a variety of cross-sectional and longitudinal studies, including randomised controlled trials and observational studies that have used the SF-36.

The datasets

There now follows a brief description of the datasets which are used throughout the rest of this thesis. These datasets illustrate the use of HRQoL outcomes across a variety of study designs. There are three types of study: observational (both cross-sectional and with baseline and a single follow-up assessment, two group RCT and longitudinal RCT (with several follow-ups).

SF-36 general population data (Brazier et al 1992)

The aim of this study was to test the acceptability, validity, and reliability of the SF-36 and compare it with the Nottingham Health Profile (NHP). The design was a postal survey using a questionnaire booklet together with a letter from the general practitioner. Non-respondents received two reminders at two-week intervals. The questionnaire booklet was sent to 1980 people aged 16-74 years, randomly selected from two general practice lists in Sheffield. The main outcomes were the scores on the eight dimensions on the SF-36. Completed questionnaires were received from 1582 of the 1980 patients surveyed. Our analysis is based on the 1372 responders who completed all 36 items of the SF-36. This sample consisted of 45% (616/1372) men and 55% (756/1372) women with an average age of 40.3 years (range 16 to 74 years).

We will use this dataset to illustrate the non-standard distributions of HRQoL and to calculate some simple effect sizes. We will also use these data as a pilot data set for the bootstrap sample size simulations.

CPSW Data: Costs & effectiveness of community postnatal support workers (CPSW): randomised controlled trial (Morrell et al 2000)

This randomised controlled trial aimed to establish the relative cost-effectiveness of postnatal support in the community compared to the usual care provided by community midwives. Six hundred and twenty-three postnatal women were allocated at random to **Intervention** (n = 311) or **Control** (n = 312) groups. The intervention consisted of up to 10 home visits in the first postnatal month of up to three hours duration by a community postnatal support worker (CPSW). The main outcomes were HRQoL as measured by the SF-36 at six weeks postnatally. This study is unusual since no baseline HRQoL assessment was made. It was felt that it was inappropriate to assess HRQoL just prior to or immediately after childbirth.

Our analysis is based on the 495 responders to the six-week postnatal questionnaire who completed all 36 items of the SF-36. This sample consisted of 241 women in the Control group and 254 women in the Intervention group. We will use this data set to illustrate various methods for two group cross-sectional comparisons of HRQoL ranging from a simple comparison of mean scores (using conventional and bootstrap hypothesis tests) through to more complex regression analysis including ordinal regression.

OA Knee Data (Brazier et al 1999)

The aim of this longitudinal observational study was to evaluate two condition specific and two generic health status questionnaires for measuring HRQoL in patients with Osteoarthritis (OA) of the Knee, and offer guidance to clinicians and researchers in choosing between them. Patients were recruited from two settings, knee surgery waiting listings and rheumatology clinics. Four self-completion questionnaires including the SF-36 were sent to the subjects on two occasions 6 months apart. Two hundred and thirty patients returned the questionnaire at initial assessment, consisting of 118 patients awaiting total knee replacement (**TKR**) **Surgery** and 112 patients attending **Rheumatology** outpatient **Clinics**. At the six-month follow-up assessment, 211 patients returned

the questionnaire (109 and 102 in the Surgery and Rheumatology groups respectively). The data used here are based on the 211 patients returning both assessments. The mean age of Rheumatology clinic respondents, 64.2 (SD 11.3) years, was significantly younger ($p = 0.001$) than the sample of patients undergoing TKR Surgery (71.1 (SD 8.5)), with more than twice as many women as men. Overall 69.6% (71/102) of the sample were females in the Rheumatology outpatient Clinics group compared to 54.1% (59/109) in TKR Surgery group ($p = 0.03$).

Since there was a difference in the baseline HRQoL and sociodemographic characteristics (age and gender) of the Clinic and Surgery groups, we use this dataset to illustrate multiple regression/analysis of covariance (ANCOVA) methods with follow-up HRQoL as the outcome variable and baseline HRQoL, age, gender and group as covariates. We compare the conventional ordinary least squares (OLS) estimates of standard error (SE) and Confidence Interval (CI) for the group regression coefficient with their bootstrap counterparts.

Leg Ulcer data (Morrell et al 1998)

The aim of this randomised controlled trial with one year of follow-up was to establish the relative cost-effectiveness of community leg ulcer clinics that use four layer compression bandaging versus usual care provided by district nurses. Two hundred and thirty-three patients with venous leg ulcers were allocated at random to intervention (120) or control group (113). The intervention consisted of weekly treatment with four layer bandaging in leg ulcer clinic (**Clinic** group) or usual care at home by the district nursing service (**Home** group). The primary outcome was time to complete ulcer healing over the one-year follow-up. Secondary outcomes included HRQoL as measured by the SF-36 at baseline, three months and 12 months follow-up.

We use these data to illustrate various methods for analysing longitudinal data, including ANCOVA, with follow-up HRQoL as the dependent variable and baseline HRQOL and treatment group as covariates, and summary measures

such as the (AUC). We compare OLS estimates of SE and CI for the group regression coefficient with their bootstrap counterparts.

NAMEIT data (Allard et al 2000)

The NAMEIT trial (NEO-BSL-08) was a 48-week, randomised, double blind study to compare Neoral with methotrexate (**Neoral**) versus placebo plus methotrexate (**Placebo**) in patients with early severe rheumatoid arthritis (RA). The primary efficacy variable in this study was the attainment of American College of Rheumatology (ACR) criteria for improvement of rheumatoid arthritis. Secondary efficacy variables included patient assessment of health related quality of life (HRQoL).

In order to assess the impact of the treatments on patients' health related quality of life, the SF-36 was completed by subjects at seven time-points, Week 0 (baseline), Weeks 8, 16, 24, 32, 40, and Week 48 at the end of the study or at the time of premature withdrawal from the trial.

Three hundred and six subjects at 48 centres were actually entered into the study. One hundred and fifty-two subjects receiving methotrexate were randomised to the Neoral treatment group and 154 subjects receiving methotrexate were randomised to the Placebo group. Of the 306 subjects randomised, 227 completed the study. Seventy-nine randomised subjects discontinued from the study prior to completion.

We use these data to illustrate various methods for analysing longitudinal data, including ANCOVA, summary measures using (e.g. AUC) and generalised estimating equations (GEE) and compare bootstrap SEs and CIs for the parameters with their conventionally estimated counterparts.

Chapter 3: An introduction to the bootstrap

Introduction

Permutation tests and Monte Carlo tests, described in Armitage *et al* 2002, make extensive use of random samples and lack of computational power may have lead to their relative neglect in the past. Now that good computational facilities are widely available, permutation and Monte Carlo tests are taking their place in an increasingly important group of techniques known as *computationally intensive methods*. The *bootstrap*, which will be the main subject of this thesis, is another member of this class of techniques and plays an important role in estimation. The *jackknife* is a rather older technique, which still has its uses and is related to the bootstrap.

According to Everitt's (1995) Dictionary of Statistics in the Medical Sciences:
"The bootstrap is a data-based simulation method for statistical inference, which can be used to study the variability of estimated characteristics of the probability distribution of a set of observations, and provide confidence intervals for parameters in situations where these are difficult or impossible to derive in the usual way."

The term bootstrap derives from the phrase '*to pull oneself up by one's bootstraps*'. Efron and Tibshirani (1993) mention that the phrase is thought to be based on one of the eighteenth century Adventures of Baron Munchausen by Rudolph Erich Raspe. (The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps!)

The basic idea of the bootstrap involves repeated random sampling with replacement from the original data, $\mathbf{x} = (x_1, x_2,\ldots,x_n)$ to produce random samples of the same size n of the original sample, each of which is known as a *bootstrap sample*, $\mathbf{x}^*$ and each provides an estimate $\hat{\theta}^*$ of the parameter of interest, θ. We seek to mimic in an appropriate manner the way the sample is collected from the population in the bootstrap samples from the observed

data. The "with replacement" means that any observation can be sampled more than once in each bootstrap sample. It is important because sampling without replacement would simply give a random permutation of the original data, with many statistics such as the mean being exactly the same (Campbell, 2001).

Repeating the process a larger number of times provides the required information on the variability of the estimator, since the standard error is estimated from the standard deviation of the statistics derived from the bootstrap samples. The point about the bootstrap is that it produces a variety of values, whose variability reflects the standard error that would be obtained if samples were repeatedly taken from the whole population.

Confidence Interval estimation

Suppose we wish to calculate a 95% confidence interval for a mean from a sample $\mathbf{x} = (x_1, x_2,\ldots,x_n)$. We take a random sample $\mathbf{x}^*$, with replacement from data, of the same size as the original sample, and calculate the mean of the data, $\hat{\theta}^*$ in this bootstrap random sample. We do this repeatedly, say B times. So we now have B bootstrap samples $\mathbf{x}^*_1, \mathbf{x}^*_2,\ldots,\mathbf{x}^*_B$, and B estimates of the sample mean, one from each bootstrap sample $\left(\hat{\theta}^*_1,\hat{\theta}^*_2,\ldots,\hat{\theta}^*_B\right)$. If these are ordered in increasing value, $\left(\hat{\theta}^*_{(1)},\hat{\theta}^*_{(2)},\ldots,\hat{\theta}^*_{(B)}\right)$, a bootstrap 95% confidence interval for the mean would be from the $0.025B^{th}$ to the $0.975B^{th}$ largest values. For a 100(1-α)% interval the limits would the $(\alpha/2)B^{th}$ and $(1 - \alpha/2)B^{th}$ largest values. This is known as the *percentile method* and although it is an obvious choice, it is not the best method for bootstrapping confidence intervals, because it can have a bias, which one can estimate and correct for. This leads to methods such as the *bias corrected method* and the *bias corrected and accelerated* (BC_a) method, the latter being the preferred option (Davison and Hinckley, 1997; Efron and Tibshirani, 1993). The paper in *Statistics in Medicine* by Carpenter and Bithell (2000) provides a useful practical guide for medical statisticians on bootstrap confidence intervals. (Further details of how to estimate BC_a Confidence Intervals are given in Appendix 2.)

The bootstrap can be applied to data with a more complex structure than the simple single sample example considered above. For comparing two groups, one with distribution *F* and the other with an independent distribution *G*, then a bootstrap approach would proceed from separate estimates $\hat{F}$ and $\hat{G}$, with bootstrap samples chosen independently from each estimated distribution.

Using the bootstrap method, valid bootstrap confidence intervals can be constructed for all common estimators such as the sample mean, median, proportion, difference in means, and difference in proportions.

Linear regression: Model (residual) and case resampling

Standard errors for regression coefficients can also be obtained using bootstrap methods. However, two different approaches are possible, *case* and *model (residual)* resampling.

For example with the simple linear model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$; with $E(\varepsilon_i) = 0$, if the data are $\mathbf{w} = (x_1, y_1), (x_2, y_2),\ldots,(x_n, y_n)$, then case-based resampling involves drawing a bootstrap sample of size *n*, with replacement from these *n* pairs. The bootstrap data set is of the form

$$\mathbf{w}^* = \left\{\left(y_{i_1}^*, x_{i_1}^*\right), \left(y_{i_2}^*, x_{i_2}^*\right), \ldots, \left(y_{i_n}^*, x_{i_n}^*\right)\right\},$$

where $i_1, i_2,\ldots,i_n$ is a random sample of integers 1 through *n*. Ordinary least squares is then used to estimate the regression coefficients $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$, for this bootstrap sample of paired cases. We do this repeatedly, say *B* times, so we now have *B* bootstrap samples and *B* estimates of the regression coefficients, one from each bootstrap sample $\left\{\left(\hat{\beta}_0^*, \hat{\beta}_1^*\right)_1, \left(\hat{\beta}_0^*, \hat{\beta}_1^*\right)_2, \ldots, \left(\hat{\beta}_0^*, \hat{\beta}_1^*\right)_B\right\}$. The standard error of these estimated coefficients $\mathrm{se}\left(\hat{\beta}_0\right)$ and $\mathrm{se}\left(\hat{\beta}_1\right)$ is simply the standard deviation of these *B* estimates. As before if these estimates are ordered in increasing value, $\left\{\left(\hat{\beta}_1^*\right)_{(1)}, \left(\hat{\beta}_1^*\right)_{(2)}, \ldots, \left(\hat{\beta}_1^*\right)_{(B)}\right\}$, a simple 95% bootstrap percentile confidence interval for the coefficient would be from the $0.025B^{\text{th}}$ to the $0.975B^{\text{th}}$ largest values.

Case-based resampling may be entirely natural for situations where it is plausible that the (*x*, *y*) pairs have been drawn from a bivariate distribution function *F(x, y)* of the pairs. However, case based resampling is less appealing if the *x* values were controlled for in some way, perhaps by the design of the study. In this situation the alternative *model or residual* based procedures could be used.

For model based resampling the simple linear regression model $\{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, with $E(\varepsilon_i) = 0\}$, the conventional fitted values y_i^{fit} and residuals e_i are first obtained from the observed data i.e. $y_i^{fit} = \hat{\beta}_0 + \hat{\beta}_1 x_i$ and $e_i = y_i^{obs} - y_i^{fit} = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$. A bootstrap sample of the residuals is drawn $\mathbf{e}^* = \left(e_{i_1}^*, e_{i_2}^*, ..., e_{i_n}^*\right)$, where $i_1, i_2, \ldots, i_n$ is a random sample of integers 1 through *n*. The bootstrap sample for the regression $\mathbf{z}^* = \left(y_i^*, x_i^*\right)$ comprises the *x* values $\left(x_i^* = x_i\right)$ from the original data and *y* values computed by adding the fitted values to the bootstrap residuals i.e. $y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_{i_n}^*$. (Note $\hat{\beta}_o$ and $\hat{\beta}_1$ are the estimates from the original sample). The bootstrap data set is of the form

$$\mathbf{z}^* = \left\{\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + e_{i_1}^*, x_1\right), \left(\hat{\beta}_0 + \hat{\beta}_1 x_2 + e_{i_2}^*, x_2\right), \ldots, \left(\hat{\beta}_0 + \hat{\beta}_1 x_n + e_{i_n}^*, x_n\right)\right\},$$

where $i_1, i_2, \ldots, i_n$ is a random sample of integers 1 through *n*. Ordinary least squares is then used to estimate the regression coefficients $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$, for this bootstrap sample. As before the process (resampling of the residuals, adding them to the fitted values and estimating the regression coefficients) is repeated *B* times to estimate standard errors and confidence intervals for the *B* regression coefficients from the bootstrap samples.

Thus model based resampling is an example of the "*parametric bootstrap*" when the residuals from a parametric model are bootstrapped to give estimates of the standard error of the parameters.

Hypothesis testing with the bootstrap

Bootstrap methods have also been used for hypothesis testing. These bootstrap tests give similar results to the much older statistical technique of

permutation tests. The bootstrap tests give similar results to permutation tests when both are available. The bootstrap tests are more widely applicable though less accurate (Efron and Tibshirani, 1993). Unlike the permutation test, the *p*-value from bootstrap hypothesis test has no interpretation as an exact probability, and like all bootstrap estimates is only guaranteed to be accurate as the sample size goes to infinity.

The two quantities that we must choose when carrying out a bootstrap hypothesis test are a *test statistic* $t(\mathbf{x})$ and a *null distribution* $\hat{F}_0$ *for the data under the null hypothesis* (H_0). Given these, we generate B bootstrap values of the test statistic $t(\mathbf{x}^*)$ under the null distribution for the data $\hat{F}_0$ and estimate the *achieved significance level* (ASL) by calculating the proportion of the bootstrap values of the B test statistics $t(\mathbf{x}^*)$, which are greater than or equal to the observed value of the test statistic $t(\mathbf{x})$ from the original data. Full details of permutation tests and hypothesis testing with the bootstrap can be found in Efron and Tibshirani (1993, Chapters 15 and 16).

Sample size estimation with the bootstrap

The choice of the test statistic will determine the power of the test. That is the chance that we reject H_0 when it is false. Again bootstrap methods are well suited to answering power and sample size questions, as we will see later on in Chapter 6.

Suppose we have two independent random samples $\mathbf{x} = (x_1, x_2,\ldots,x_m)$ and $\mathbf{y} = (y_1, y_2,\ldots,y_n)$. The **x**'s and **y**'s are random samples from continuous distributions having cumulative distribution functions (cdfs), F_x and F_y respectively. We consider the simple situation where the distributions have the same shape, but the locations may differ. Thus if δ denotes the location difference (i.e. mean (**y**) - mean (**x**) = δ), then $F_y(y) = F_x(y - \delta)$, for every y. We focus on the null hypothesis H_0: $\delta = 0$ against the alternative H_A: $\delta \neq 0$. Then we can test this hypothesis using an appropriate significance test (e.g. Mann-Whitney or *t*-test), and will let π (F, δ, α, n) denote the power function of the test.

The bootstrap strategy (Collings and Hamilton, 1988) is to use pilot data to provide a non-parametric estimate, $\hat{F}$ of *F* and to use a simulation method for finding the power of the test associated with any specified sample size *n* if the data follow the estimated distribution function. If we denote the distribution function estimate by $\hat{G}$, under the alternative hypothesis δ, we can estimate the approximate power, $\hat{\pi}(G, \delta, \alpha, n)$ by the following computer simulation procedure.

(1) Draw a random sample with replacement of size 2*n* from $\hat{F}$. The first *n* observations in the sample form a simulated sample of x's, denoted by $x_1^*,\ldots,x_n^*$, with estimated cdf $\hat{F}^*$. Then δ is added to each of the other *n* observations in the sample to form the simulated sample of y's, denoted by $y_1^*,\ldots,y_n^*$, with estimated cdf $\hat{G}^*$. (The y*'s and x*'s have been generated from the same distribution except that the distribution of the y*'s is shifted δ units to the right.)

(2) The test statistic *t(x, y)*, (Mann-Whitney or *t*-test) is calculated for the x*'s and y*'s, yielding *t(x*, y*)*. If *t(x*, y*)* ≥ $T_{1-\alpha/2}$, (where $T_{1-\alpha/2}$ is the critical value of the test statistic) a success is recorded; otherwise a failure is recorded.

(3) Steps 1 and 2 are repeated *B* times. The estimated power of the test, $\hat{\pi}(G, \delta, \alpha, n)$, is approximated by the proportion of successes among the *B* repetitions.

The jackknife

The jackknife is a technique for the estimation of the bias and standard error of an estimate and is described more comprehensively in Chapter 11 of Efron and Tibshirani (1993). The jackknife pre-dates the bootstrap and bears close similarities to it. Quenouille (1949) first proposed the idea of the jackknife for the estimation of bias. Tukey (1958) recognised the jackknife's potential for estimating standard errors, and gave it its name.

Suppose we have a sample $\mathbf{x} = (x_1, x_2,\ldots,x_n)$ and an estimator $\hat{\theta} = s(x)$. We wish to estimate the bias and standard error of $\hat{\theta}$. The jackknife focuses on the samples that *leave out one observation at a time*:

$$\mathbf{x}_{(i)} = (x_1, x_2,\ldots,x_{i-1}, x_{i+1},\ldots,x_n)$$

For $i = 1, 2,\ldots n$, called *jackknife samples*. The i^{th} jackknife sample consists of the data set with the i^{th} observation removed. Let $\hat{\theta}_{(i)} = s(x_{(i)})$ be the i^{th} jackknife replication of $\hat{\theta}$. The jackknife estimate of bias is defined by

$$\hat{bias}_{jack} = (n-1)(\hat{\theta}_{(.)} - \hat{\theta}) \text{ where } \hat{\theta}_{(.)} = \sum_{i=1}^{n} \hat{\theta}_{(i)} / n .$$

The jackknife estimate of standard error is defined by

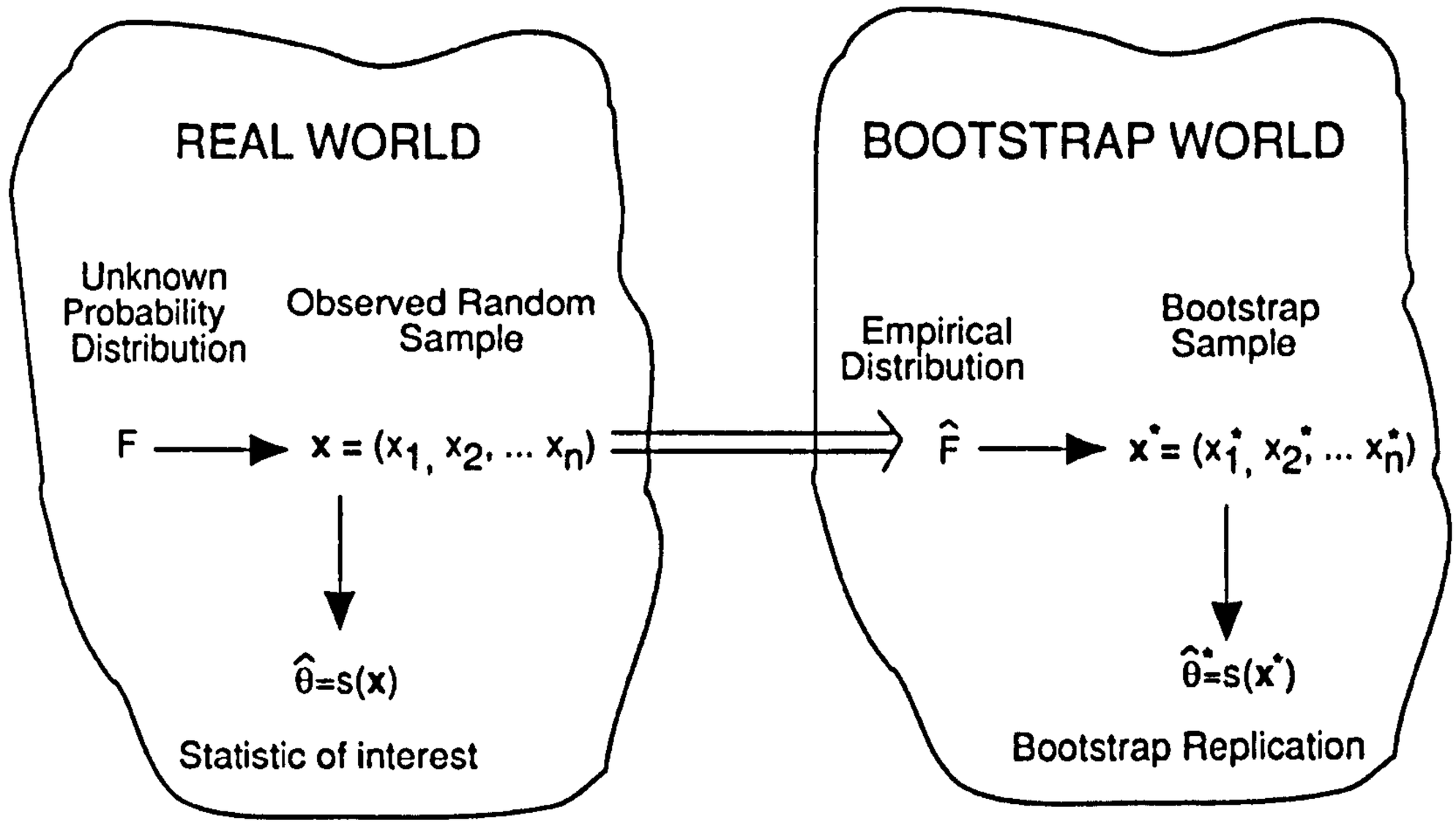
$$\hat{se}_{jack} = \left[\frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{\theta}_{(i)} - \hat{\theta}_{(.)} \right)^2 \right]^{1/2} .$$

Since it requires computation only of $\hat{\theta}$ for n jackknife data sets, the jackknife will be easier to compute than the bootstrap. However by looking only at the n jackknife samples, the jackknife uses only limited information about the statistic $\hat{\theta}$, and thus one might guess that the jackknife is less efficient than the bootstrap. In fact in turns out that the jackknife can be viewed as an approximation to the bootstrap, for the estimation of standard errors and bias, although the jackknife can fail miserably if the statistic $\hat{\theta}$ is not "smooth". A simple example of a non-smooth statistic is the median (Efron and Tibshirani, 1993).

Graphical representation of the bootstrap for general data structures

The bootstrap method can readily be adapted to more complicated data structures. Figure 3.1 (taken from Efron and Tibshirani, 1993) is a simple schematic diagram as it applies to one-sample problems. On the left is the real world, where an unknown distribution F has given the observed data $\mathbf{x} = (x_1, x_2,\ldots,x_n)$ by random sampling. We have calculated a statistic of interest from $\mathbf{x}$, $\hat{\theta} = s(\mathbf{x})$, and wish to know something about $\hat{\theta}$'s statistical behaviour, perhaps its standard error $se_F(\hat{\theta})$.

Figure 3.1: Schematic diagram of the bootstrap as it applies to ones sample problems (taken from Efron and Tibshirani, 1993).



On the right side of the diagram is the bootstrap world. In the bootstrap world, the empirical distribution $\hat{F}$ gives bootstrap samples $\mathbf{x}^* = (x^*_1, x^*_2, \ldots, x^*_n)$ by random sampling with replacement, from which we calculate bootstrap replications of the statistics of interest, $\hat{\theta}^*=s(\mathbf{x})$. The big advantage of the bootstrap world is that we can calculate as many replications of $\hat{\theta}^*$ as we want, or at least as many as we can afford. This allows us to do the probabilistic calculations directly, for example using the observed variability of the $\hat{\theta}^*$'s to estimate the unobservable quantity, $se_F(\hat{\theta})$.

The double arrow "⇒" in Figure 3.1 indicates the calculation of $\hat{F}$ from *F*. Conceptually, this is the crucial step in the bootstrap process, even though it is computationally simple. Every other part of the bootstrap picture is defined by analogy (Efron and Tibshirani 1993). *F* gives **x** by random sampling, so $\hat{F}$ gives $\mathbf{x}^*$ by random sampling; $\hat{\theta}$ is obtained from **x** via the function *s(***x***)*, so $\hat{\theta}^*$ is obtained from $\mathbf{x}^*$ in the same way. Bootstrap calculations for more complex probability mechanisms turn out to be straightforward, once we know how to carry out the double arrow process - estimating the entire probability

mechanism from the data. Fortunately this is easy to do for all of the common data structures.

To facilitate the study of more complicated data structures, we use the notation, $P \rightarrow \mathbf{x}$, to indicate that an unknown probability model P has yielded the observed data set **x**.

Figure 3.2: Schematic diagram of the bootstrap applied to problems with a general data structure $P \rightarrow \mathbf{x}$ (taken from Efron and Tibshirani, 1993).

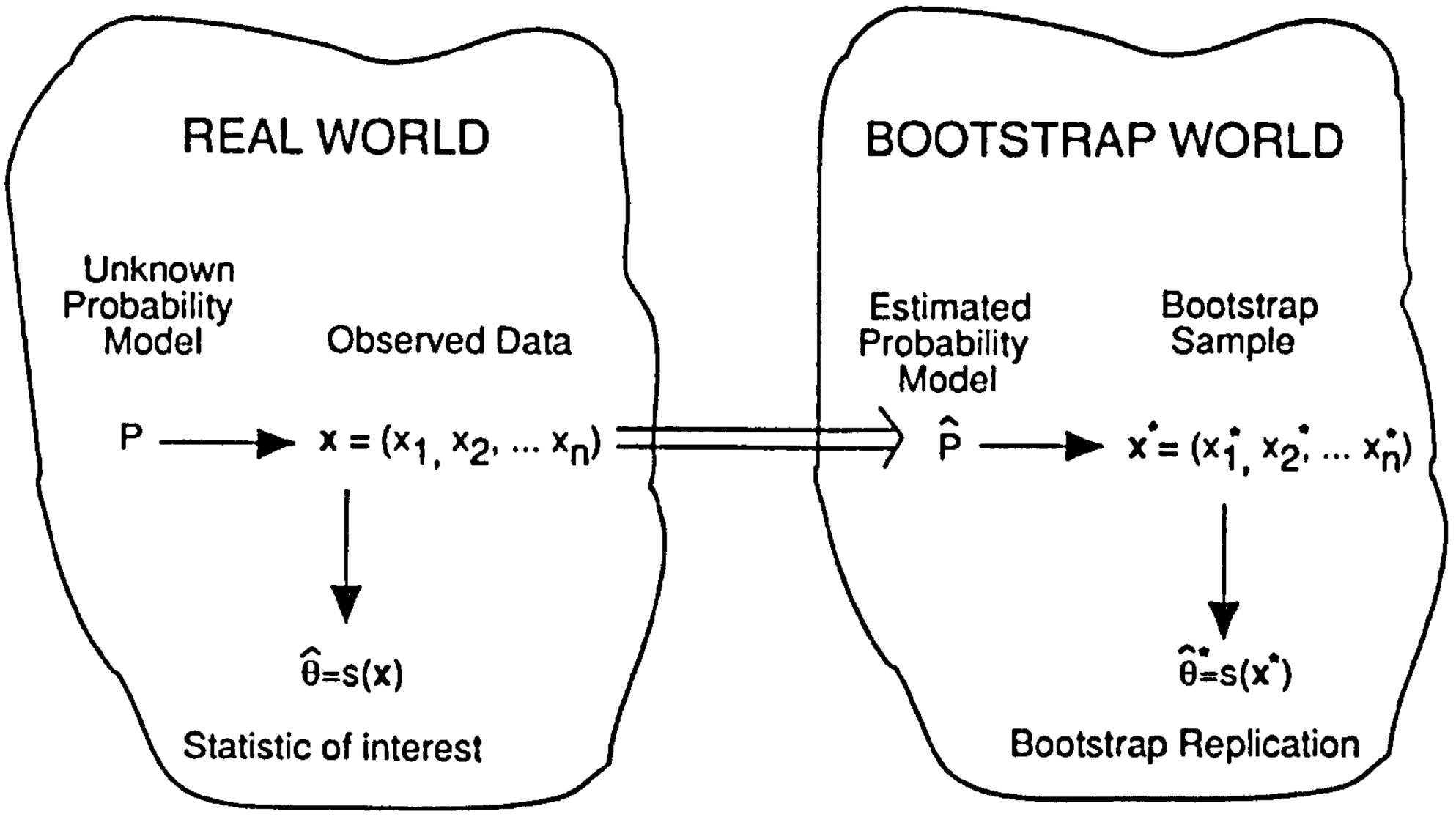


Figure 3.2 (again taken from Efron and Tibshirani, 1993) is a version of Figure 3.1 that applies to general data structures $P \rightarrow \mathbf{x}$. There is not much conceptual difference between the two figures, except for the level of generality involved. In the real world, an unknown probability mechanism P gives an observed data set **x**, according to the rule of construction indicated by the arrow "$\rightarrow$ ". In specific applications we need to define the arrow more carefully, for example if we have two samples. The data set **x** may no longer be a single vector. It has a form dependent on the data structure, for example $\mathbf{x} = (\mathbf{z}, \mathbf{y})$ in the two-sample problem. Having observed **x**, we calculate a statistic of interest $\hat{\theta}$ from **x** according to the function $s(\cdot)$.

The bootstrap side of Figure 3.2 is defined by the analogous quantities in the real world: the arrow in $\hat{P} \rightarrow \mathbf{x}^*$ is defined to mean the same thing as the arrow in $P \rightarrow \mathbf{x}$. The function mapping $\mathbf{x}^*$ to $\hat{\theta}^*$ is the same function $s(\cdot)$ as from $\mathbf{x}$ to $\hat{\theta}$.

Two practical problems arise in actually carrying out a bootstrap analysis based on Figure 3.2.

(1) We need to estimate the entire probability mechanism P from the observed data $\mathbf{x}$. This is the step indicated by the double arrow, $\mathbf{x} \Rightarrow \hat{P}$. It is surprisingly easy to do for most familiar data structures. No general prescription is possible, but quite natural ad hoc solutions are available in each case, for example $\hat{P} = (\hat{F}, \hat{G})$ for the two-sample problem.

(2) We need to simulate the bootstrap data from $\hat{P}$ according to the relevant data structure. This is the step $\hat{P} \rightarrow \mathbf{x}^*$ in Figure 3.2. This step is conceptually straightforward, being the same as $P \rightarrow \mathbf{x}$, but can require some care in the programming if computational efficiency is necessary. Usually the generation of the bootstrap data $\hat{P} \rightarrow \mathbf{x}^*$ requires less time, often much less time, than the calculation of $\hat{\theta}^* = s(\mathbf{x}^*)$.

Notation (following Efron and Tibshirani, 1993)

We have already introduced some notation for the bootstrap in this chapter. This section briefly summarizes some of the notation used in the rest of this thesis. Lower case bold letters such as $\mathbf{x}$ refer to vectors, that is, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Matrices are denoted by upper case bold letters such $\mathbf{X}$, while a plain upper case variable like X refers to a random variable. The transpose of a vector is written as $\mathbf{x}^T$. A superscript "*" indicates a bootstrap random variable: for example $\mathbf{x}^*$ indicates a bootstrap data set generated from data set $\mathbf{x}$. Parameters are denoted by Greek letters such as θ or β, while α is used for error rates in connection with significance tests and confidence sets. A hat on a letter indicates an estimate, such as $\hat{\theta}$. Thus the estimator $t(\mathbf{x})$ has

observed value $t(\boldsymbol{x})_{obs}$, which may be an estimate of the unknown parameter θ.

We use pdf, cdf and edf as shorthand for probability density function, cumulative density function and empirical density function. The letters *F* and *G* are used for cdfs. Notation such as $\#\{\hat{\theta} \geq 1.96\}$ means a count of the number of $\hat{\theta}$s greater than 1.96. The letter *B* is reserved for the number of replicate simulations or bootstrap samples. Simulated quantities of a statistic $t(\boldsymbol{x})$ are denoted by $t(\boldsymbol{x}^*_b)$, $b = 1,\ldots,B$, whose ordered values are $t(\boldsymbol{x}^*_{(1)})$ $t(\boldsymbol{x}^*_{(2)})$ $t(\boldsymbol{x}^*_{(B)})$.

Summary

More details of bootstrap confidence intervals are given in Appendix 2. The bootstrap has the potential for such wide applications in statistics, that the present discussion cannot do it justice. The specialist texts on the bootstrap, by Efron and Tibshirani (1993) and Davison and Hinckley (1997) provide details of many aspects of the technique. The power of the technique makes it one of the most important advances in statistical methodology in recent years (Armitage *et al* 2002). The main advantage of the bootstrap is that it frees the investigator from making inappropriate assumptions about the distribution of an estimator $\hat{\theta}$ in order to make inferences (Campbell, 2001). Therefore, in theory the bootstrap should prove a useful technique for sample size estimation, hypothesis testing, and confidence interval estimation for data with non-standard distributions.

HRQoL outcomes with their discrete, bounded and skewed data distributions should be ideal candidates for the application of bootstrap methods. So it is somewhat surprising that two otherwise comprehensive textbooks on the design and analysis of studies with HRQoL outcomes (Fayers and Machin, 2000; Fairclough, 2002) make no mention of the use of the bootstrap. Therefore subsequent chapters of this thesis will compare and contrast conventional methods of sample estimation and analysis of HRQoL outcomes with bootstrap methods.

Chapter 4: Review of methods of sample size estimation for HRQoL outcomes

Introduction and background

Sample size calculations are now mandatory for many research protocols and are required to justify the size of clinical trials in papers before they will be accepted by journals (Altman *et al* 2000). Thus, when an investigator is designing a study to compare the outcomes of an intervention, an essential step is the calculation of sample sizes that will allow a reasonable chance (*power*) of detecting a predetermined difference (*effect size*) in the outcome variable, at a given level of significance. Sample size is critically dependent on the summary measure, the proposed effect size and the method of calculating the test statistic. For example, for a given power and significance level, the sample size is inversely proportional to the square of the effect size, so halving the effect size will quadruple the sample size.

Whatever type of study design is used the problem of sample size must be faced. Sometimes we may wish to show that a new treatment is clinically equivalent in efficacy to the standard treatment. Machin *et al* (1997) describe statistical methods for calculating the appropriate sample sizes for demonstrating equivalence between two treatments. For simplicity in this chapter we will assume that the primary outcome for the study is a HRQoL measure and that we are interested in comparing the effectiveness (or superiority) of a new treatment compared to a standard control treatment at a single point in time.

Since HRQoL measures are being used more frequently in clinical trials and HSR, as both primary and secondary endpoints, investigators are now asking statisticians for advice on how to plan and analyse studies using HRQoL measures, and this includes questions on the sample size.

However, as we have seen in Chapter 2 the ordinal scaling of HRQoL measures may lead to several problems in determining sample size and analysing the data. To illustrate this, we use some HRQoL data from a RCT

that aimed to compare the difference in health status in a group of women who were offered postnatal support (intervention) from a community midwifery support worker compared with a control group of women who were not offered support (Morrell *et al* 2000). This study is briefly described in Chapter 2.

HRQoL outcome data may not meet the distributional requirements (usually that the data have a Normal distribution) of parametric methods of analysis. Therefore non-parametric methods are most often used to analyse HRQoL data. The main aim of this chapter is to review, describe and compare several methods of sample size estimation (parametric and non-parametric) when using HRQoL measures as outcome in comparative clinical trials.

This chapter is based on two papers published in the *Health Services and Outcomes Research Methodology* journal (Walters *et al* 2001b; Walters and Brazier, 2003b see Appendix 3) and two earlier departmental discussion papers (Walters *et al* 2000; Walters and Brazier 2002).

The remainder of this chapter is structured into the following sections. Section 2 summarises the methods and the sample size formulae. What researchers actually do in practice is discussed in Section 3. The consequences if different sample size formulae are applied are explored in Section 4. Section 5 discusses multiple end-points. The final sections (6 and 7) talk about the choice of sample size method with HRQoL outcomes and conclusions.

Which sample size formulae?

In principle, there are no major differences in planning a study using HRQoL assessment to those using conventional clinical outcomes. Pocock (1983) outlines five key questions regarding sample size:

1. *What is the main purpose of the trial?*
2. *What is the principal measure of patient outcome?*
3. *How will the data be analysed to detect a treatment difference?*
4. *What type of results does one anticipate with standard treatment?*
5. *How small a treatment difference is it important to detect and with what degree of certainty?*

Therefore after deciding on the purpose of the study and the principle outcome measure, the investigator must decide how the data are to be summarised and analysed to detect a treatment difference. Thus, the investigator must choose an appropriate summary measure of this outcome and then calculate a sample size based on the smallest treatment difference in this summary measure that is of such clinical value that it would be very undesirable to fail to detect. Given answers to all of the five questions above, we can then calculate a sample size.

An appropriate summary measure of the outcome data will usually be the sample mean, median, or a rate or proportion. When comparing two groups or a single group over time, appropriate comparative summary measures may include the difference between sample means, difference in medians, difference in rates or proportions, the relative risk (RR) or the odds ratio (OR).

The mean is often chosen as a suitable summary measure, although there are several reasons against using it. One reason would be that the HRQoL outcome measure of interest is an ordinal not a continuous variable, and therefore means are hard to interpret (see points 1 to 7, in Chapter 2). Also the HRQoL outcome may have a skewed distribution and the median or the proportion of the sample in a given category (or less) may be a more useful summary of HRQoL outcome.

For individual patients, the outcome of treatment is usually dichotomous (the treatment either works or the treatment does not work) or ordinal (the effect of treatment worsens the patients' HRQoL, has no effect, or improves HRQoL). In this case, given the probability of a successful outcome (improved HRQoL) from the control treatment (p_C), and the OR that the new treatment is beneficial (compared to the control treatment), then the probability that the new treatment will work for an individual patient (p_T) is:

$$p_T = \frac{ORp_C}{ORp_C + 1 - p_C}. \qquad (4.1)$$

Therefore the OR may be a suitable comparative summary measure for the effect of treatment at an individual level. We can calculate the risk difference ($p_T - p_C$) or absolute risk reduction (ARR) as it is sometimes known, and hence the reciprocal of the risk difference, 1/ARR or $1/(p_T - p_C)$ which is the Number Needed to Treat (NNT),

$$\text{NNT} = \frac{1}{p_T - p_C}. \tag{4.2}$$

The NNT is the number of patients who need to be treated with the new treatment rather than the standard control treatment in order for one additional patient to benefit. Thus, the NNT is a useful summary measure for clinicians to compare two treatments, although it does require the HRQoL outcome to be dichotomised (Laupacis *et al* 1988).

The mean (and mean difference) is a more suitable summary measure for the effect of treatment on average (i.e. at a hospital level). The mean indicates the effect of treatment on average in this group of patients. This summary measure is useful for health care providers (or hospitals) in deciding whether or not to offer a new treatment to its population.

Campbell *et al* (1995) outline the ways of calculating sample sizes in two group studies for binary, ordered categorical and continuous outcomes. Further details, examples and tables are given in the book by Machin *et al* (1997).

Method 1: Continuous Normally distributed HRQoL data – comparing two means

If the HRQoL outcome is assumed to be continuous and plausibly sampled from a Normal distribution, then the best summary statistic for a location parameter is the mean, and the usual hypothesis test for a difference or shift in location parameters between two independent samples is the two-sample *t* test.

For two independent groups with continuous and Normally distributed data, the standardised effect size is the expected mean value of the intervention outcome minus the expected mean value of the control outcome divided by a standard deviation of the outcomes. That is

$$\Delta_{Normal} = \frac{\mu_T - \mu_C}{\sigma}, \qquad (4.3)$$

where Δ_{Normal} is the standardised effect size index, μ_T and μ_C are the expected group means of outcome variable under the null and alternative hypotheses and σ is the standard deviation of outcome variable (assumed to be the same under the null and alternative hypotheses).

In a two-group study comparing mean HRQoL between the two groups, the number of subjects per group *n* for a two-sided significance level α and power $1 - \beta$ is given by equation (4.4),

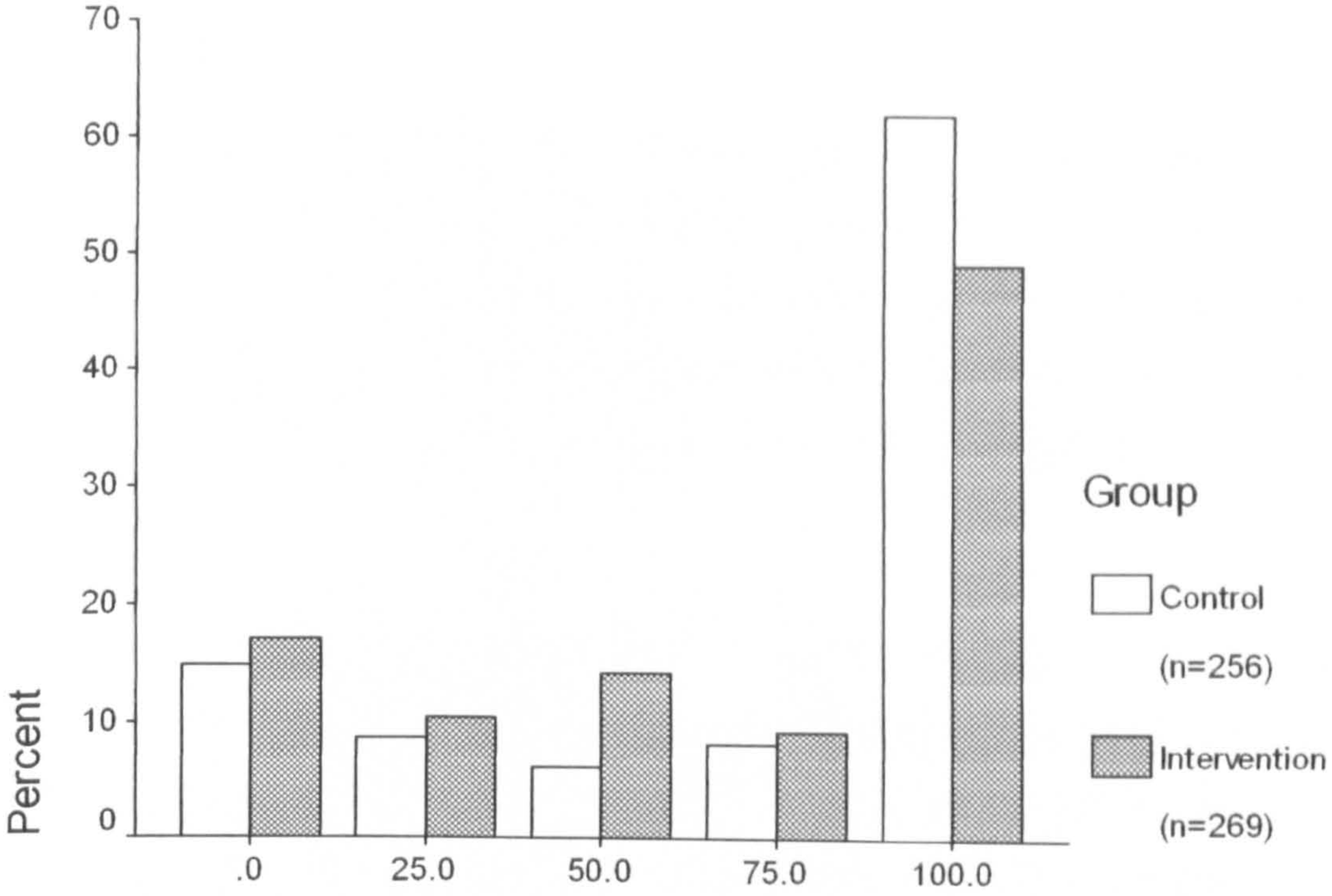
$$n_{Normal} = \frac{2\left[z_{1-\alpha/2} + z_{1-\beta}\right]^2}{\Delta^2_{Normal}}, \qquad (4.4)$$

where, $z_{1-\alpha/2}$ and $z_{1-\beta}$ are the appropriate values from the standard Normal distribution for the 100 $(1 - \alpha/2)$ and 100 $(1 - \beta)$ percentiles respectively.

If the sample size is "sufficiently large", then the Central Limit Theorem (CLT) guarantees that the sample means will be approximately Normally distributed (Hogg and Tanis, 1988). Thus, if the investigator is planning a large study and the sample mean is an appropriate summary measure of the HRQoL outcome, then pragmatically there is no need to worry about the distribution of the HRQoL outcome and we can use equation (4.4) to calculate sample sizes. Although the Normal distribution is strictly only the limiting form of the sampling distribution of the sample mean as the sample size *n* increases to infinity, it provides a remarkably good approximation to the sampling distribution even when *n* is small and the distribution of the data is far from Normal (Armitage *et al* 2002). Generally, if *n* is greater than 30, these approximations will be good. However, if the underlying distribution is symmetric, unimodal and continuous, a value of *n* as small as 4 can yield a very adequate approximation (Hogg and Tanis, 1988). Figure 4.1 clearly

illustrates the bounded, discrete and skewed nature of HRQoL data and shows that a large sample size may be required for the assumption of Normality to be valid.

Figure 4.1: Distribution of SF-36 Role Limitations Physical outcome by group from the CPSW study (Morrell *et al* 2000).



The skewed distribution of the HRQoL (the RP dimension of the SF-36) data from the CPSW study in Figure 4.1 also implies that the sample mean and mean difference may not be suitable summary measures to compare the two groups. The mean score of the Control group was 73.5 (SD 38.4) compared with a mean score of 65.7 (SD 39.2) in the Intervention group at six weeks postnatally, a mean difference of 7.8 (95% CI: 1.2 to 14.2; t = 2.31 on 523 df, p = 0.02). The median scores were 100 and 75 respectively, with over 62% of the control group and 49% of the intervention group scoring 100.

Suppose we are planning a two-group study comparing HRQoL between the groups, using the RP dimension of the SF-36 as the primary outcome. We believe that the mean difference in HRQoL scores between the two groups is

an appropriate comparative summary measure. Assuming a standard deviation σ of 38 and that a mean difference (μ_T - μ_C) of 8 or more points is clinically and practically relevant gives a standardised effect size (from equation 4.3) of 0.21. Using this effect size in equation (4.4) with a two-sided 5% significance level ($z_{1-\alpha/2}$ = 1.96) and 80% power ($z_{1-\beta}$ = 0.84) gives the estimated number of subjects per group as 356.

Transformations

If the HRQoL outcome data is continuous but has a skewed distribution, it may be transformed using a logarithmic transformation. The transformed variable may have a more symmetric distribution that approximates better to the Normal form. The problem is that certain HRQoL measures, such as the SF-36 are scored on 0 to 100 scales and the natural logarithm of zero does not exist.

If we recode the RP dimension so that a score of 0 = 1, 25 = 2, 50 = 3, 75 = 4 and 100 = 5, then the mean RP score in the control group (on a 1 to 5 scale) is now 3.94 (SD 1.54). If we take natural logarithms ($\log_e$) of the recoded RP score, then the mean log-transformed score is 1.24 (SD 0.59). Equation (4.4) can now be applied to the log-transformed scale once the standardised effect size Δ_{Normal} is specified. Unfortunately, there is no simple interpretation for the log-transformed RP scale, and so the inverse transformation is used to obtain scores corresponding to the recoded (1 to 5) RP scale. The mean RP score (on the 1 to 5 scale) using the inverse transformation is now exp (1.24) = 3.46 compared to the original value of 3.94.

If one third of a category or unit change on the recoded RP scale is considered the minimum clinically important difference to detect, this is approximately equivalent to an eight-point difference on the original 0 to 100 scale of the SF-36 RP dimension as a one category change corresponds to 25 points. Then the untransformed effect size from (4.3) is: (4.27 – 3.94)/1.55 = 0.21. Using equation (4.4), this leads to a sample size of 356 patients per group.

Using the log-transformed scale of the RP, a third of a unit increase is approximately from 3.46 to 3.79. This is then expressed as an anticipated effect on the log transformed scale as $\Delta_{Normal} = (\mu_T - \mu_C)/\sigma = [\log_e(3.46 + 0.33) - \log_e(3.46)]/0.59 = 0.15$. Using equation (4.4) with $\Delta_{Normal} = 0.15$ gives $n_{Normal} = 697$ patients per group.

We have used a logarithmic transformation for non-Normal data and made the sample size calculations accordingly. Other possible transformations for this purpose are the reciprocal or square root. A difficulty with the use of transformations is that they distort HRQoL scales and make interpretation of treatment effects difficult (Fayers and Machin, 2000). In fact, only the logarithmic transformation gives results interpretable on the original scale (Bland and Altman, 1996). The logarithmic transformation expresses the effect as a ratio of the geometric mean for patients in the treatment group to the geometric mean for patients in the control group. This is because the difference between two logarithms is the logarithm of the ratio: log (T) – log (C) = log (T/C).

However, this ratio will vary in a way that depends on the geometric mean value of the control treatment (Fayers and Machin, 2000). For example, if the geometric mean for the control treatment C is 2, and treatment T induces a change in RP of 1 unit compared to this level, then this implies an effect size of $\log_e$ (3/2) = 0.41. On the other hand, for geometric mean of 4 for the treatment C, the same numerical change of one unit implies an effect size of $\log_e$ (5/4) = 0.22. Thus, although in this example the effect size is a one unit difference in HRQoL in both cases when expressed on the untransformed RP scale, the logarithmic transformation results in a second effect size which is almost half (0.22/0.41 = 0.54) the first. This makes interpretation difficult.

Method 2: Continuous HRQoL data with no distributional assumptions

If the HRQoL outcome is assumed to be continuous and plausibly not sampled from a Normal distribution then the most popular (not necessarily the most efficient) non-parametric test for comparing two independent samples is

the two-sample *Mann-Whitney U*, also known as the *Wilcoxon rank sum* test (Lehman, 1975).

Suppose we have two independent random samples $X_1, X_2,\ldots,X_m$ and $Y_1, Y_2,\ldots,Y_n$ and we want to test the hypothesis that the two samples have come from the same population against the alternative that the Y observations tend to be larger than the X observations. As a test statistic we can use the Mann-Whitney (*MW*) statistic *U*, i.e., $U = \#(Y_j > X_i)$, $i = 1,..,m$; $j = 1,\ldots,n$, which is a count of the number of times the Y_js are greater than the X_is. The magnitude of *U* has a meaning, because *U/nm is* an estimate of the probability that an observation drawn at random from population Y would exceed an observation drawn at random from population X.

Noether (1987) derived a sample size formula for the *MW* test (see equation 4.6 below), using an effect size $p_{Noether}$, (see equation 4.5) which is probability that an observation drawn at random from population Y would exceed an observation drawn at random from population X,

$$p_{Noether} = \Pr(Y > X), \tag{4.5}$$

that makes no assumptions about the distribution of the data (except that it is continuous), and can be used whenever the sampling distribution of the test statistic *U* can be closely approximated by the Normal distribution, an approximation that is usually quite good except for very small *n* (Collins and Hamilton, 1991).

$$n_{Non-normal} = \frac{\left[z_{1-\alpha/2} + z_{1-\beta}\right]^2}{6\left(p_{Noether} - 0.5\right)^2} \tag{4.6}$$

Thus to determine the sample size, we have to find the 'effect size' $p_{Noether}$. There are several ways of estimating $p_{Noether}$, (Simonoff *et al* 1986) under various assumptions, one possibility is $p_{Noether} = U/nm$ (Lesaffre *et al* 1993). Let μ_X, σ^2_X, μ_Y, and σ^2_Y be the mean and variance of the X and Y variables respectively. Then if $X \sim N(\mu_X, \sigma^2_X)$ and $Y \sim N(\mu_Y, \sigma^2_Y)$ then Simonoff *et al* (1986) show that the maximum likelihood estimator of Pr(Y > X) using the sample estimates of the mean and variance ($\hat{\mu}_X, \hat{\sigma}^2_X, \hat{\mu}_Y, \hat{\sigma}^2_Y$) is:

$$p = \Pr(Y > X) = \Phi\left(\frac{\hat{\mu}_Y - \hat{\mu}_X}{\left(\hat{\sigma}_X^2 + \hat{\sigma}_Y^2\right)^{1/2}}\right), \tag{4.7}$$

where Φ is the Normal cumulative distribution function.

If we assume the SF-36 RP dimension is Normally distributed (unlikely as we have seen) and $\sigma_X = \sigma_Y = \sigma$ then equation (4.7) allows the calculation of two comparable 'effect sizes' $p_{Noether}$ and Δ_{Normal} thus enabling the two methods of sample size estimation (Equations 4.4 and 4.6) to be directly contrasted. If the SF-36 RP is not Normally distributed then we cannot use equation (4.7) to calculate comparable effect sizes and must rely on the empirical estimates calculated post hoc from the data.

Suppose we are planning a two-group study comparing HRQoL (using the RP as the primary outcome) between the groups. We believe the RP outcome to be continuous, but not Normally distributed and are intending to compare RP scores in the two groups with a Mann-Whitney *U* test. Therefore Noether's method will be appropriate. As before if we assume a mean difference of 8 and a standard deviation of 38 for the RP, then using equation (4.7) this leads to an effect size $p_{Noether} = \Pr(Y > X)$ of 0.56. Substituting $p_{Noether} = 0.56$ in equation (4.6) with a two-sided 5% significance level and 80% power gives the estimated number of subjects per group as 363.

The two methods have given similar sample size estimates ($n_{Normal} = 356$ and $n_{non\text{-}Normal} = 363$). The two methods can be regarded as equivalent when the two distributions have the same shape and equal variances. When the two distributions are Normally distributed with equal variances, the *MW* test will require about 5% more observations than the two-sample *t*-test to provide the same power against the same alternative. For non-Normal populations, especially those with long tails, the *MW* test may not require as many observations as the two-sample *t*-test (Elashoff, 1999).

Method 3: Dichotomous categorical HRQoL data – comparing two proportions

If the HRQoL outcomes are measured on a binary or dichotomous categorical scale, for example, "good health" and "poor health", then an appropriate summary measure of the outcome data will usually be the sample rate or proportion in the sample with "good" HRQoL. When comparing two groups or a single group over time, appropriate comparative summary measures may include the difference in rates or proportions, the relative risk or the odds ratio.

The statistical hypothesis test used to compare two independent groups when the outcome is binary is the Pearson chi-squared test for a 2 x 2 contingency table. In this situation, the anticipated effect size is $\delta_{Binary} = (\pi_T - \pi_C)$, where π_T and π_C are the proportions in the two treatment groups with 'good health'. In a two-group study comparing differences in rates or proportions between the groups, the number of subjects per group n_{Binary} for a two-sided significance level α and power $1 - \beta$ is given by equation (4.8).

$$n_{Binary} = \frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \left[\pi_T(1-\pi_T) + \pi_C(1-\pi_C)\right]}{\left(\pi_T - \pi_C\right)^2}. \qquad (4.8)$$

Alternatively, the same difference between treatments may be expressed through the odds ratio (OR), which is defined as:

$$OR_{Binary} = \left(\frac{\pi_T}{1-\pi_T}\right) \Big/ \left(\frac{\pi_C}{1-\pi_C}\right) = \frac{\pi_T(1-\pi_C)}{\pi_C(1-\pi_T)}. \qquad (4.9)$$

This formulation leads to an alternative for equation (4.8) for the sample size. Thus,

$$n_{OR} = \frac{2\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 / \left(\log OR_{Binary}\right)^2}{\bar{\pi}(1-\bar{\pi})}, \qquad (4.10)$$

where $\bar{\pi} = (\pi_T + \pi_C)/2$.

Equations (4.8) and (4.10) are quite dissimilar, but Julious and Campbell (1996) show that for all practical purposes they give very similar sample sizes, with divergent results only occurring for relative large (or small) OR_{Binary}.

Figure 4.1 indicates that approximately 60% of patients in the control group scored 100, i.e. "good health". Suppose it is anticipated that this may improve to 70% having good health with treatment *T*. The anticipated treatment effect is thus, $\delta_{Binary} = (\pi_T - \pi_C) = 0.70 - 0.6 = 0.10$. This equates to a sample size of $n_{Binary} = 353$ per group from equation (4.8).

Alternatively, this anticipated treatment effect can be expressed (using 4.9) as $OR_{Binary} = (0.70/0.30)/(0.60/0.40) = 1.56$. Using this in equation (4.10) with $\bar{\pi} = (0.70 + 0.60)/2 = 0.65$ gives a sample size per group of $n_{OR} = 349$ patients. As we indicated previously, there is usually only a small and inconsequential difference between the calculations from the alternative formulae.

Method 4: Ordered categorical (Ordinal) HRQoL data

If the HRQoL outcomes are measured on an ordinal scale, then the statistical hypothesis test used in this instance (to compare two independent groups) is the Mann-Whitney *U* test with allowance for ties or a Chi-squared test for trend (Altman, 1991).

Whitehead (1993) presents the sample size formula for ordinal data in a key paper. To use Whitehead's formula we need to specify an effect size. For ordinal data Whitehead suggested the odds ratio ($OR_{Ordinal}$), which is the odds of a subject being in a given category or lower in one group compared with the odds in the other group.

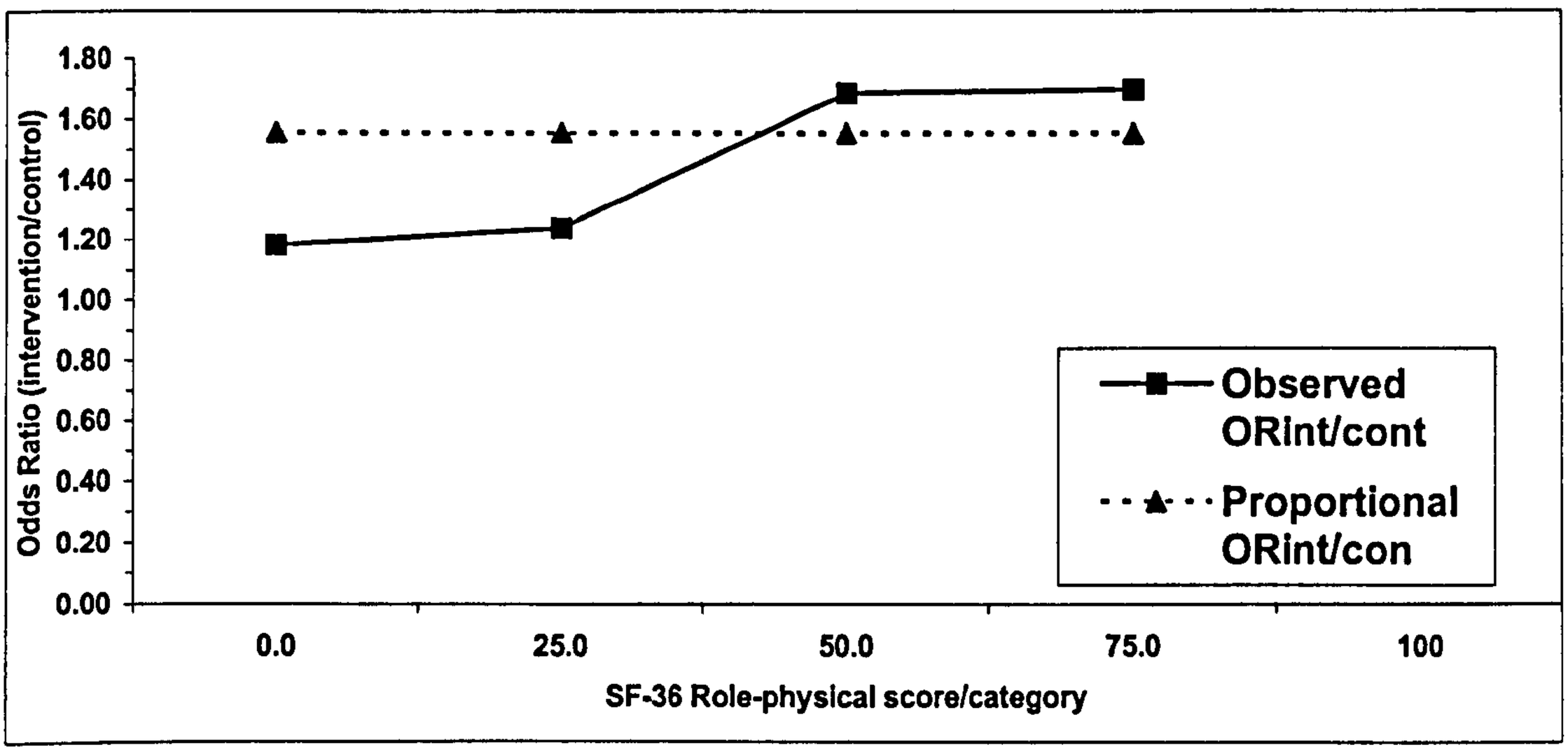
Suppose we have two groups, treatment (*T*) and control (*C*), and the HRQoL outcome measure of interest *Y* has *k* ordered categories y_i denoted by $i = 1,2,\ldots,k$. Let π_{iT} be the probability of being in category *i* in group *T* and γ_{iT} be the expected cumulative probability of being in category *i* or less in group *T* (*i.e.* $\gamma_{iT} = Pr(Y \leq y_i)$, with $\gamma_{kT} = Pr(Y \leq y_k) = 1$). For category *i*, where *i* takes values from 1 to *k-1*, the $OR_{Ordinal}$ is given by

$$\mathrm{OR}_{\mathrm{Ordinali}} = \left(\frac{\gamma_{iT}}{(1-\gamma_{iT})}\right) \Big/ \left(\frac{\gamma_{iC}}{(1-\gamma_{iC})}\right). \qquad (4.11)$$

The assumption of proportional odds specifies that the $OR_{Ordinall}$ will be the same for all categories from *i = 1 to k-1*, and is equal to $OR_{Ordinal}$. This is the proportional odds assumption which underlies the proportional odds model and hence the derivation of the formula

Figure 4.1 illustrates that the HRQoL outcome has five categories, which implies four cut-offs (RLP scores = 0, ≤ 25, ≤ 50 and ≤ 75) and therefore four separate ORs. As the proportional odds model assumes a constant OR for all categories, Figure 4.2 shows the four observed ORs compared to the estimated common OR of 1.56 (95% CI: 1.12 to 2.17) from the proportional odds model. All observed ORs are greater than 1 and seem similar to the model estimate.

Figure 4.2: Odds Ratios for SF-36 Role Limitations Physical categories based on observed data and proportional odds model



A chi-squared score test of proportional odds was $\chi^2 = 6.27$ on 3 df, p = 0.10. This suggests that the proportional odds assumption is plausible. Although in other cases the test may lack sufficient power to detect meaningful departures from proportional odds, (Peterson and Harrell, 1990; Brant, 1990). The model is robust to mild departures from the assumption of proportional odds. A crude test would be to examine the odds ratios and, if they are all greater than unity, or all less than unity, then assume a proportional odds model will suffice

(Walters *et al* 2001a). With increasing numbers of categories it is less likely that proportional odds assumption remains true.

The proportional odds model OR estimate implies that patients in the intervention group have 1.56 times the odds of being in a given category or below (i.e. have worse HRQoL) than patients in the control group.

Whitehead's (1993) method can be regarded as a 'non-parametric' approach as the derivation of the sample size formula and analysis of data is based on the Mann-Whitney *U* test, although it still relies on the assumption of a constant odds ratio for the data. Whitehead's method also assumes a relatively small log odds ratio and a large sample size, which will often be the case in HRQoL studies where dramatic effects are unlikely. Equation (4.12) gives the number of subjects per group *n* for a two-sided significance level α and power 1 - β.

$$n_{Ordinal} = \frac{6\left[\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 / \left(\log \text{ OR }_{\text{Ordinal}}\right)^2\right]}{\left[1 - \sum_{i=1}^{k} \bar{\pi}_i^3\right]} \qquad (4.12)$$

Here $\bar{\pi}_i$ is the average proportion of subjects anticipated in category *i*, that is, $\bar{\pi}_i = \left(\pi_{iT} + \pi_{iC}\right)/2$.

Suppose (as before) we are planning a two-group study to compare HRQoL (using the RP as the primary outcome) between the groups. We believe that the mean difference in HRQoL scores between the two groups is *not* an appropriate comparative summary measure. However, the odds of patient in the intervention group having an HRQoL score in a given category or below compared to the odds for a patient in the control group is felt to be an appropriate comparative summary measure. Approximately 60% of patients in the control group scored 100 i.e. "good health", with 40% scoring less than good health. As before, suppose it is anticipated that this may improve to 70% having good health with treatment *T*, implying that 30% have less than good health. Using equation (4.8),

$$\mathrm{OR}_{\mathrm{Ordinali}} = \left(\frac{0.3}{1-0.3}\right) \Big/ \left(\frac{0.4}{1-0.4}\right) = \frac{0.43}{0.67} = 0.64\,, \qquad (4.13)$$

leads to an $OR_{Ordinal} = 0.64$ which is the reciprocal of OR_{Binary}.

If we assume proportions (π_{iC}) of patients of 0.15, 0.09, 0.06, 0.08 and 0.62 respectively in the five RP categories (0, 25, 50, 75 and 100) in the control group, the cumulative proportions γ_{iC} in category *i* for the control treatment *C* (*i* = 1 to 5) are 0.15, 0.24, 0.30, 0.38 and 1.0. Then, for a given constant $OR_{Ordinal} = 0.64$, the anticipated cumulative proportions (γ_{iT}) for each category of treatment *T* are given by:

$$\gamma_{iT} = \frac{OR_{Ordinal}\gamma_{iC}}{OR_{Ordinal}\gamma_{iC} + (1-\gamma_{iC})} \qquad i = 1 \text{ to } k\text{-}1. \qquad (4.14)$$

After calculating the cumulative proportions (γ_{iT}), the anticipated proportions falling in each treatment category, π_{iT} can be determined from the difference in successive γ_{iT}. Finally, the combined mean ($\bar{\pi}_i$) of the proportions of treatments *C* and *T* for each category is calculated.

Using equation (4.12) with this OR and ($\bar{\pi}_i$) with a two-sided 5% significance level and 80% power gives the estimated number of subjects per group as 340. With a sample size of 340 and proportions of 0.10, 0.06, 0.05, 0.07 and 0.72 scoring 0, 25, 50, 75 and 100 respectively in the treatment group, the mean RP score will be 81.1 in the treatment group compared to a mean RP score of 73.5 in the control group. This is a mean difference of 7.6 points, which is slightly smaller than the eight point mean difference used in equation (4) to calculate n_{Normal}.

If the number of categories is large it is difficult to postulate the proportion of subjects who would fall in a given category. Both Whitehead (1993) and Julious *et al* (1997) point out that there is little increase in power (and hence saving in the number of subjects recruited) to be gained by increasing the number of categories beyond five. Categories that are equally likely to occur lead to the greatest efficiency.

Julious and Campbell (1996) show that, with two categories only, the method given by Whitehead is approximately equivalent to one described by Machin *et al* (1997) for the binary case, even though at first sight the equations are very dissimilar. They state that the practical importance of this is to give the choice of two alternative measures of differences between groups: differences in proportions or odds ratios.

Alternative approaches: exact methods and simulation

Hilton and Mehta (1993) describe methods for sample size determinations based on either exact power or a very precise Monte Carlo estimate of it. This involves the use of an algorithm for computing the exact probabilities of each marginal table for a given fixed sample size. The conditional probability of a particular permutation of the data can be obtained by using the generalised hypergeometric distribution. Even with an efficient algorithm, processing the number of permutations is a considerable computational task and so its use is restricted to studies involving small sample sizes and a small number of ordered categories (Rabbee *et al* 2003). This is likely to be an unrealistic scenario with HRQoL outcomes such as the SF-36 where most dimensions have more than 11 ordered categories (the exceptions being the two Role dimensions, RE, and RE with four and five categories respectively). Large dramatic differences in HRQoL scores between groups (using the SF-36 outcome) are unlikely (see Chapter 5). Therefore larger sample sizes will be required to detect such differences.

Rabbee *et al* (2003) describe a method for computing power and sample size for linear rank tests of differences between two ordered multinomial populations. Again this method, like Whitehead's' (1993) is asymptotic, although it more closely approximates Hilton and Mehta's (1993) exact method. Rabbee's method overcomes some of the computational limitations of the exact methods, but it is still not a very practical way of sample size estimation and so will not be considered any further.

Limitations of Whitehead's (1993) method when model assumptions (constant OR, relatively small log OR and a large sample size) are violated have been

pointed out by Kolassa (1995) and Hilton (1996) respectively. However, as we show in the next chapter dramatic effects are unlikely with HRQoL outcomes, so larger samples sizes are required to detect statistically significant differences. Therefore, the assumptions of a large sample size and relatively small effects are not unreasonable for HRQoL studies (Walters *et al* 2001a).

Julious and Campbell (1998) discuss the problem of calculating the number of subjects required in a matched or paired study in which the outcome variable is ordinal. In the two category (binary) case, the sample size is dependent on the expected number of discordant pairs. They suggest that, as a rule of thumb, the required discordant sample size for the binary/two category case can be used as an approximation to the total sample size when the number of categories is greater than two.

Method 5 Bootstrap methods

We have already mentioned briefly in Chapter 3, that bootstrap methods can be used to answer sample size and power questions. We will expand on this in more detail in Chapter 6. Briefly, the choice of the test statistic (*t*-test, *MW* test, Chi-squared test etc) will determine the power of the test. We can use bootstrap simulation to compare the power of Methods 1 to 4 of sample size estimation above for detecting differences in HRQoL between two groups.

The bootstrap strategy is to use pilot HRQoL data to a provide a non-parametric estimate, $\hat{F}$ of *F* and to use a simulation method for finding the power of the test associated with any specified sample size *n* if the data follow the estimated distribution function. If we denote the distribution function estimate by $\hat{G}$, under the alternative hypothesis δ, we can estimate the approximate power using the algorithm described in Chapter 6 (Collings and Hamilton, 1988; Hamilton and Collings, 1991; Walters and Brazier, 2003b).

Tsodikov *et al* (1998) and Troendle (1999) also use the bootstrap and Collings and Hamilton's (1988) algorithm on bounded outcome data. Both assume that a historical or reliable pilot data set is available to use in estimating the shape of the distribution. Troendle shows that the bootstrap is a reasonably accurate

method of power estimation under the location shift alternative hypothesis. Conversely Tsodikov *et al* suggest some caution in using the bootstrap for power estimation if the pilot data set is small and the anticipated treatment effect is based on the results from the pilot sample.

Lesaffre *et al* (1993) describe a method that involves the use of pilot data to estimate power and sample size for bounded outcome scores. They used a variety of computer simulations including Monte Carlo and bootstrap methods to estimate power for a fixed sample size. Again, the bootstrap methods are based on Collings and Hamilton's (1988) algorithm. The bootstrap methods perform reasonably well, and gave fairly unbiased estimates of power, though for small pilot samples with large variability.

What sample size methods do investigators actually use?

King (1996) mentions the importance of effect sizes in calculating sample sizes for clinical trials and also discusses the alternative parametric and non-parametric approaches, although she does not give a recommendation for either one. King notes that there can be quite marked differences between sample sizes calculated from parametric and non-parametric methods, particularly for HRQoL outcome measures that have a highly skewed distribution.

A few papers (Julious *et al* 1995; Campbell *et al* 1995, Julious *et al* 1997 Fayers and Machin, 2000) appear to have used Whitehead's (1993) non-parametric method for ordinal outcomes. Bolland *et al* (1998) applied Whitehead's method to a three category ordinal outcome (good recovery (GR), moderate disability (MD) severe disability/vegetative state/dead (SD/V/D)) in a randomised trial of patients suffering from severe head injury. They assumed a common odds ratio (proportional odds) of 1.84; proportions of patients of 0.17, 0.30 and 0.53 respectively in the three categories GR, MD and SD/V/D in the control group; and no effect of prognostic factors on outcome. This led to an initial sample size of 400 patients. Due to the uncertainty about these assumptions, the authors planned a blinded sample size calculation after approximately 100 patients were recruited. The review

was performed on the first 93 patients to respond and led to an increase in sample size from 400 to 450. On completion of the study the authors note that the proportional odds assumption, *"whilst not fully valid, was not misleading"*.

Roset *et al* (1999) apply parametric and non-parametric sample size methods to two datasets comprising the EQ-5D. They recommend parametric methods when the outcome variables are thought to be reasonably symmetrical and non-parametric methods when the data are skewed. This conclusion is rather unexciting and follows conventional statistical thinking that non-parametric methods be used for data with non-Normal distributions.

What happens when different sample size formulae are applied?

The sample sizes per group with similar anticipated treatment effects calculated for our example using equations (4.4, 4.6, 4.8, 4.10 and 4.12) respectively were $n_{Normal} = 356$, $n_{Non\text{-}normal} = 363$, $n_{Binary} = 353$, $n_{OR} = 349$ and $n_{Ordinal} = 340$. The binary and ordinal calculations gave lower estimated sample sizes than for the continuous case, which may reflect the skewed nature of the RP outcome data, although for practical purposes the sample size estimates are broadly similar.

Three papers (Julious *et al* 1995; Julious *et al* 1997; Machin and Fayers, 1998) have highlighted the discrepancies between sample sizes for intervention studies using HRQoL outcomes (HADS and SF-36) calculated using conventional parametric techniques and non-parametric approaches. In order to make the alternative hypotheses comparable, the authors used the distribution of the outcome for the control, and shifted it by one category for the intervention. The odds ratios between categories and groups formed by such a shift were calculated, so that the parametric and non-parametric methods are calculated using the same alternative hypothesis.

Using the SF-36, Julious *et al* (1995) show that the results given by the parametric and non-parametric methods are similar in some dimensions of the SF-36 but are very different in dimensions where the scores are highly skewed. For such asymmetric distributions, the parametric methods give the

same sample sizes for effects that are one unit above and one unit below the population mean. This is because the parametric method assumes a symmetric (Normal) distribution, whereas the non-parametric method may give different sample sizes according to the expected direction of the effect. For example they show that the non-parametric estimate of the sample size to detect a change of one category for the GH dimension (which is quite symmetric) is similar for one category up or down, but for the MH dimension (which is asymmetric) the sample size required to detect a change of one category down is three times that to detect one category up.

In all three articles the authors stress that (Julious *et al* 1995; Julious *et al* 1997; Machin and Fayers, 1998). *"In general, statistics such as means and standard deviations are not suitable summary measures for non-Normal distributions, and neither are standardised differences (effect sizes) a suitable basis for the calculation of sample sizes."*

Julious *et al* (1997) recommend that the frequency distributions of HRQoL scores should always be given so that one can assess if non-parametric methods should be used for sample size calculations and analysis. Given the skewed/asymmetric distribution of the majority of HRQoL outcomes in general, they recommend that ordered categorical methods be used for sample size calculations.

Prieto *et al* (1996) in a letter to the editor about Julious *et al's* (1996) paper, strongly disagree with this recommendation. Firstly, Prieto and colleagues argue that between ordinal and continuous scales there are a number of instruments (such as the SF-36) that can be labelled 'summated scales', in which the total score is the sum of a set of ordinal rankings, and so these scales are 'between' ordinal and continuous. They do not claim that equal increments in the observed score along the summated scale represent equal increments in the underlying latent variable being measured, but the mode of construction of the instrument suggests that deviations from interval properties may not be extreme. Secondly, they argue that failure to meet the assumptions required for the use of parametric methods does not appear to

have serious consequences in most instances. Therefore they suggest that parametric techniques should be used for SF-36 sample size calculations. They note, however, that the minimum clinically important difference (MCID) for the SF-36 scales is still unknown and further research is needed to clarify the clinical significance of score changes on the SF-36 scales.

In reply, Campbell *et al* (1996) stated that the parametric method requires one to specify an effect size based on the standard deviation of the outcome. It is the distribution of the population not the estimate, that is important, and the standard deviation for data that are not Normally distributed is uninterpretable in terms of the distribution of the data. Thus one cannot expect 95% of the observations to be within plus or minus two standard deviations of the mean. The problems are exacerbated when there are a limited number of categories; for example, one dimension of the SF-36 (role limitations emotional (RE)) can take only four values (0, 33, 67, 100), and in one study most of the population scored 100 (Brazier *et al* 1992). In practice, an apparent continuous scale is composed of several correlated binary responses and the final response scale is effectively binary (< 100 or = 100). In this case, Julious' methods (1995, 1997) demonstrate that the required sample size approaches the size required for a binary variable.

If the frequency distribution of the HRQoL outcome data is symmetric (or expected to be reasonably symmetric) then the mean and median will tend to coincide and either can be used as a suitable summary measure. If the HRQoL outcome has a discrete, skewed distribution and the proportion of the sample at the upper or lower bounds is large, then the relative frequency or cumulative probabilities may be a more appropriate summary measure for the data. In this case non-parametric methods and the proportional odds model would be more appropriate for analysis of the data.

However, a limitation of the non-parametric and proportional odds model approaches is in their interpretation. The effect sizes $p_{Noether}$ and $OR_{Ordinal}$ are more difficult to interpret than the simple mean difference δ and its standardised effect size Δ_{Normal}. Secondly, it is more difficult to quantify the

minimum important difference (MID) for the non-parametric and proportional odds model methods.

We believe, if the main goal of the analysis is to assess the magnitude of the treatment effect on the HRQoL outcome (i.e. interest lies in comparing location between treatments), then it is more sensible and appealing to assign numerical scores to the ordered categories and to use statistical methods (for sample size estimation and analysis) appropriate for comparing means (e.g. *t*-tests and multiple linear regression). The mean is still a useful summary measure for ordinal scaled HRQoL data if we are prepared to assume an underlying continuous latent variable that quantifies the response of interest and that the actual measured HRQoL outcomes, the ordered categories reflect contiguous intervals along this continuum. If interest lies elsewhere, for example in comparing the relative frequencies or cumulative probabilities in the ordinal categories between treatments, then the proportional odds model would be more appropriate for sample size estimation and analysis.

Multiple end points

We have based the above calculations on the assumption that there is a single identifiable endpoint, or HRQoL outcome, upon which treatment comparisons are based. Sometimes there is more than one endpoint of interest; HRQoL outcomes are multi-dimensional (e.g. the SF-36 has eight dimensions including RP). If one of these dimensions is regarded as more important than the others, it can be named as the primary endpoint and the sample size estimates calculated accordingly. The remainder should be consigned to exploratory analyses or descriptions only.

A problem arises when there are several outcome measures that are all regarded as equally important. One approach is to repeat the sample-size estimates for each outcome measure in turn and then select the largest number as the sample size required to answer all the questions of interest. Here, it is essential to note the relationship between significance tests and power: it is well recognised that *p*-values become distorted if many endpoints are each tested for significance, and adjustments should be made.

To guard against false statistical significance as a consequence of multiple hypothesis testing, it is a sensible precaution to examine the consequences of replacing the significance level α in the various equations by a significance level adjusted using the Bonferroni correction (Bland and Altman, 1995). The Bonferroni correction is:

$$\alpha_{Bonferroni} = \alpha/K, \qquad (4.15)$$

where K is the number of endpoints or hypothesis tests to be performed. Such a correction will clearly lead to larger sample sizes. The Bonferroni approach to adjusting for multiple comparisons tends to be conservative as it assumes all the different endpoints are uncorrelated (Altman *et al* 2000). In the case of HRQoL outcomes there is likely to be a strong correlation between the different dimensions. This conservativeness implies that utilising criterion (4.15) will lead to failure to reject the null hypothesis on too many occasions. Fairclough (2002) gives a more comprehensive discussion of multiple endpoints and suggests several alternative methods to the Bonferroni approach when analysing HRQoL outcomes.

Choice of sample size method with HRQoL outcomes

It is important to make maximum use of the information available from other related studies or extrapolation from other unrelated studies. The more precise the information, the better we can design the trial. We would recommend that researchers planning a study with HRQoL as the primary outcome pay careful attention to any evidence on the validity and frequency distribution of the proposed HRQoL instrument.

The frequency distribution of HRQoL scores from previous studies should be assessed to see if means, rates or proportions are appropriate summary measures for the data, and hence whether parametric or non-parametric methods should be used for sample size calculations and analysis. Given the skewed distribution of the majority HRQoL outcome measures, summary measures of central location such as means and summary measures of variability such as standard deviations may not be appropriate; and so standardised differences (effect sizes) and parametric methods may not be a

suitable basis for calculation of sample size. It is difficult to interpret an effect defined by equation (4.3) when the data are skewed. We would suggest that investigators consider clinically meaningful effect sizes, and do not rely on generic 'small', 'medium' or 'large' ones as suggested by Cohen (1988).

There may be considerable uncertainties in estimates of such quantities as the standard deviation and the treatment effect. Sample size calculations are sometimes based on estimates "pulled out of thin air". If an investigator is uncomfortable with the assumptions, then it is good practice to calculate sample sizes under a variety of scenarios so that the sensitivity to assumptions can be assessed (Julious *et al* 1997). We would recommend that various anticipated benefits be considered, ranging from the optimistic to the more realistic, with sample sizes being calculated for several scenarios within that range. It is a matter of judgement, rather than an exact science, as to which of the options is chosen for the final study size (Fayers and Machin, 2000).

If there is little prior knowledge of the full distribution of scores for the HRQoL outcome, sample size calculation may not be too problematical. Using the ordinal approach to sample size calculation, knowledge of the anticipated distribution within four or five broad categories is usually sufficient to determine the required number of subjects (Whitehead, 1993; Campbell *et al* 1995; Julious *et al* 1997).

The guidance presented here is not meant to imply that other more fundamental design factors such as whether a randomised controlled design can be used are not important or should not be considered. However, to date, the points made about calculating sample sizes for HRQoL measures have not been well recognised. Perhaps the adoption of some of the above recommendations by the developers of HRQoL instruments and in guidelines used by medical journals for refereeing HRQoL studies would help facilitate change.

Conclusions

Given that the end goal of using HRQoL outcomes in research studies is to assess a patient's health and well being, using the right type of HRQoL outcome in the right setting with an appropriate sample size calculation is crucial. Much time and energy is devoted to developing and validating HRQoL measures. Developers and researchers need to complement this effort with clearer descriptions of the distribution of such outcomes and what is an appropriate summary measure, the mean or the proportion with a certain score.

Finally we would stress the importance of a sample size calculation (with all its assumptions), and that any such estimate is better than no sample size calculation at all, particularly in a trial protocol (Williamson *et al* 2000). The mere fact of calculation of a sample size means that a number of fundamental issues have been thought about: what is the main outcome variable, what is a clinically important effect, and how is it measured? The investigator is also likely to have specified the method and frequency of data analysis. Thus protocols that are explicit about sample size are easier to evaluate in terms of scientific quality and the likelihood of achieving objectives.

Chapter 5: Summary statistics and observed effect sizes from the various HRQoL datasets

Introduction

We have already seen from the previous chapter that amongst other factors sample size estimation is dependent on the "effect size" which is related to the smallest treatment difference that it is important to detect.

This chapter calculates several effect sizes indices Δ_{Normal}, $p_{Noether}$, δ_{Binary}, OR_{Binary}, and $OR_{Ordinal}$ (see equations 4.3, 4.5, 4.8, 4.9 and 4.11) for the various data sets (described in Chapter 2). We will use a distribution-based approach, although other anchor-based ways of determining effect sizes are available, as are ways of determining the smallest important difference in a HRQoL measure that it is worthwhile detecting (Walters and Brazier, 2003a).

Interpretation of HRQoL scores

HRQoL outcome measures are being increasingly used in research trials, but less so in routine clinical practice. The interpretation of HRQoL scores raises many issues. The scales and instruments used may be unfamiliar to many clinicians and patients, and they may be uncertain of the meaning of the scale values and summary scores.

Repeated experience and familiarity with a wide variety of physiological measures such as blood pressure or forced expiratory volume, has allowed clinicians to make meaningful interpretation of the results. In contrast, the meaning of a change in score of five points on a HRQoL instrument such as the SF-36 is less intuitively apparent, not only because the scale has unfamiliar units, but also because health professionals seldom use HRQoL measures in routine clinical practice.

In clinical trials, where HRQoL instruments are being increasingly used as primary outcome measures, it is relatively simple to determine the statistical significance of a change in HRQoL (as I shall describe later on in Chapters 7

and 8), but placing the magnitude of these changes in a context that is meaningful for health professionals, patients and other stakeholders (Pharmaceutical and Medical Device Developers, Insurance Payers, Regulators, Governments) has not been so easy. Ascertaining the magnitude of change that corresponds to a minimal important difference would help address this problem (Juniper *et al* 1994). So when an investigator is determining an important change standard the perspective can influence the assessment approach and the way in which an important difference is determined (Frost *et al* 2002). The minimal important difference (*MID*), from the patient perspective, can be defined as *"the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management"*. (Jaeschke *et al* 1989).

Thus standards of individual change are needed to provide meaningful interpretation of HRQoL intervention and treatment effects and to classify patients based on this standard as improved, stable or declined. To date two broad strategies have been used to interpret differences or changes in HRQoL following treatment (Norman *et al* 2001): distribution based approaches - the *effect size* (ES); and anchor-based measures - the *minimum clinically important difference* (MCID).

Distribution based approaches rely on relating the difference between treatment and control groups to some measure of variability. The most popular approach uses Cohen's standardised effect size (Cohen, 1988), the mean change divided by the standard deviation to serve as an "effect size index", that is suitable for sample size estimation. Cohen suggested that standardised effect sizes of 0.2 to 0.5 should be regarded as "small", 0.5 to 0.8 as "moderate" and those above 0.8 as "large". Cohen's effect size may be influenced by the degree of homogeneity or heterogeneity in the sample. Distribution-based methods rely on expressing an effect in terms of the underlying distribution of the results. Investigators may express effects in terms of between-person standard deviation units, within-person standard deviation units, and the standard error of measurement (Guyatt *et al* 2002).

Four statistics commonly used to index responsiveness are (Hays *et al* 1998):

(1) effect size (Kazis *et al* 1989);

(2) *t*-test comparisons (Liang *et al* 1995);

(3) the standardised response mean (Liang *et al* 1990);

(4) the responsiveness statistic (Guyatt *et al* 1987).

The formulae for these statistics are as follows, where D = raw score change on measure; SE = standard error of the difference; SD_{time1} = standard deviation at time 1; $SD_{difference}$ = standard deviation of D; SD_{stable} = standard deviation of D among stable subjects (those whose true status is constant over time):

$$\text{Paired } t\text{-statistics} = D/SE \tag{5.1}$$

$$\text{Effect size (ES) statistic} = D/SD_{time1} \tag{5.2}$$

$$\text{Standardised response mean (SRM)} = D/SD_{difference} \tag{5.3}$$

$$\text{Responsiveness statistic} = D/SD_{stable} \tag{5.4}$$

The paired *t*-statistic is best suited to pre-post assessments of interventions of known efficacy. The effect size statistic relates change over time to the standard deviation of baseline scores. The standardised response mean compares change to the standard deviation of change. The responsiveness statistic looks at HRQoL change relative to variability for clinically stable respondents. The effect size statistic ignores variation in change entirely, the *t*-statistic ignores information about variation in scores for clinically stable respondents, and the responsiveness statistic ignores information about variation in scores for clinically unstable responders. We have already seen that similar effect size statistics to equation (5.2) can be used in the estimation of sample sizes.

Therefore, we shall use the effect size statistic (5.2) which is analogous to equation (4.3) in Chapter 4 to determine Δ_{Normal}. We shall use the pooled standard deviation of the two groups as our estimate of σ rather than the standard deviation of baseline scores. For example, for the CPSW study where we were interested in comparing the HRQoL of new mothers six weeks

postnatal in the intervention and control groups, the effect size will be calculated as the mean HRQoL score in the Intervention group minus the mean HRQoL score in the Control group divided by the pooled standard deviation. The 'pooled' estimate of the standard deviation is given by,

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}, \qquad (5.5)$$

where s_1 and s_2 are the sample standard deviations and n_1 and n_2 the sample sizes for the two groups respectively.

We shall estimate, $p_{Noether}$, which is the probability that an observation drawn at random from population X would exceed an observation drawn a random from population Y, by U/n_1n_2, where U is the value of the Mann-Whitney test statistic and n_1 and n_2 are the sample sizes in the two groups. U_{XY} is the number of pairs for which $x_i > y_i$ and U_{YX} is the number of pairs for which $y_i > x_i$. Any pairs for which $y_i = x_i$, count ½ a unit to both U_{XY} *and* U_{YX}. Either of these statistics can be used for the *MW* test, with exactly equivalent results (Armitage *et al* 2002). We shall use U_{XY} and henceforth refer to it as U without the subscripts. For example, for the CPSW study, $p_{Noether}$, will be estimated from the U value from the *MW* test comparing the equality of the distribution of HRQoL scores in between the (Intervention and Control) groups divided by the product of the samples sizes for the Intervention and Control groups.

If possible, δ_{Binary}, and OR_{Binary} will be estimated from the proportions of patients scoring 100 "good" health vs. "less than good health" (i.e. a HRQoL score < 100) in the two groups.

Finally the $OR_{Ordinal}$ will be estimated by fitting a proportional odds or cumulative logit model to the data (McCullagh and Nelder, 1989) with the HRQoL score as the dependent outcome variable and the treatment group as a covariate or independent or predictor variable. (Full details of the proportional odds model will be given in Chapter 7, Walters *et al* 2001a and Lall *et al* 2002). We shall also carry out a score test of the proportional odds assumption which underlies this model.

The developers of the SF-36 have suggested that using the GH dimension a five-point difference (on the 0-100 scale) is the smallest score change achievable by an individual and considered as "*clinically and socially relevant*" (Ware *et al* 1993). Angst *et al* (2001) found the MCID ranged from 3.3 to 5.3 points on the PF dimension and 7.2 to 7.8 points on the BP dimension in patients with osteoarthritis of the hip or knee. Hays and Morales (2001) also provide information on what a clinically important difference is for the SF-36 scales. They conclude that the MCID for the SF-36 is "*typically in the range of 3-5 points*", although they also recommend caution in interpreting 3-5 points on the SF-36 dimensions as the MCID.

The studies

The data used in this chapter comes from five studies which used the SF-36 including cross-sectional surveys, randomised controlled trials, and observational studies. All of the effect sizes are for simple cross-sectional two group comparisons.

Observed Effect sizes for the CPSW study

As we have already mentioned the effect sizes for the CPSW study will be based on comparison of six-week HRQoL scores between women randomised to the Intervention (extra post-natal support) and Control groups. Table 5.1 shows that patients in the Control group had significantly higher scores (compared to the Intervention group patients) on two dimensions of the SF-36, the RP and SF dimensions, using the *p*-values from the two independent samples *t*-test and a cut off $p \leq 0.05$ for statistical significance,. The observed Normal or Gaussian effect sizes Δ_{Normal} ranged from –0.01 to 0.23 and were all in the 'small' range using Cohen's (1988) classification. The probability that a randomly chosen value (or subject) from the Control group was greater (i.e. had a better HRQoL) than a randomly chosen value (or subject) from the Intervention group, i.e. $p_{Noether}$ ranged from 0.499 to 0.568.

SF-36 Dimension		*Group*	*n*	*mean*	*sd*	*Mean diff δ*	*ES Δ_{Normal}*	*t-test P-value*	*Pr(X > Y) $p_{Noether}$*	*MW test P-value*	*$OR_{Ordinal}$*	*P_{100}*	*δ_{Binary}*	*OR_{Binary}*
Physical Function	*X*	Control	241	89.9	14.5	2.6	0.17	0.060	0.561	0.015	1.49	0.42	0.09	1.48
	Y	Intervention	254	87.3	15.8							0.33		
Role Physical	*X*	Control	241	74.3	38.1	9.1	0.23	0.009	0.568	0.004	1.66	0.63	0.14	1.79
	Y	Intervention	254	65.2	39.5							0.48		
Bodily Pain	*X*	Control	241	75.6	23.7	4.0	0.17	0.065	0.552	0.040	1.39	0.27	0.05	1.28
	Y	Intervention	254	71.6	23.8							0.23		
General Health	*X*	Control	241	77.7	17.7	2.4	0.13	0.139	0.542	0.147	1.26	0.10	0.02	1.22
	Y	Intervention	254	75.3	18.5							0.08		
Vitality	*X*	Control	241	51.1	20.7	1.3	0.06	0.498	0.514	0.596	1.09	0.00	0.00	0.00
	Y	Intervention	254	49.8	21.7							0.00		
Social Function	*X*	Control	241	81.6	22.7	4.7	0.20	0.025	0.561	0.015	1.48	0.43	0.09	1.45
	Y	Intervention	254	76.9	24.2							0.34		
Role Emotional	*X*	Control	241	77.9	36.4	1.1	0.03	0.734	0.515	0.503	1.13	0.68	0.04	1.18
	Y	Intervention	254	76.8	35.5							0.64		
Mental Health	*X*	Control	241	72.9	17.2	-0.2	-0.01	0.902	0.499	0.972	0.99	0.03	0.00	0.89
	Y	Intervention	254	73.1	16.7							0.03		

Key for Tables 5.1 to 5.17.

1. Effect size Δ_{Normal} = mean difference divided by the pooled standard deviation. 2. Effect size $p_{Noether}$ $Pr(X_{Control} > Y_{Intervention})$, based on U/nm, where U = MW test statistic (with allowance for ties). 3. $OR_{Ordinal}$ estimate from the proportional odds model with group as covariate. 4. P_{100} proportion in each group scoring 100 on the SF-36 dimension, δ_{Binary} = $p_{100\_Control} - p_{100\_Intervention}$. 5. $\theta' = Pr(X > Y)/Pr(X \leq Y) = OR$. 6. $\lambda' = Pr(X > Y) - Pr(X \leq Y) = ARR$ (Absolute Risk Reduction). 7. $NNT_{\lambda'}$ (Number needed to treat) = 1/ARR.

The $OR_{Ordinal}$ estimates from the proportional odds model were greater than unity for seven out of the eight dimensions of the SF-36 (the exception being the MH dimension). Odds ratios greater than unity imply that the odds of being in a given category or less (i.e. having better HRQoL) is greater for patients in the Control group compared to the odds of being in a given category or less in the Intervention group. The *p*-values from the Wald test for the significance of the regression coefficient for the group term were almost identical to the *p*-values obtained from the *MW* test and so are not reported. Ordinal regression is equivalent to the *MW* test when there is only one independent 0/1 variable in the regression (Campbell, 2001). Although the advantage of ordinal regression over non-parametric methods is that we get an efficient estimate of a regression coefficient and we can extend the analysis to allow for other confounding variables.

Finally the test of proportional odds which underlies the model suggested that on three dimensions of the SF-36, PF ($p = 0.032$), RP ($p = 0.04$) and GH ($p = 0.001$) this assumption may not be valid.

The effect sizes, δ_{Binary}, and OR_{Binary} also reflect that more patients in the Control group reported "good" health i.e. scoring 100 compared to the Intervention group. Again Odds Ratios greater than one imply that the odds of scoring 100 (i.e. having good HRQoL) is greater for patients in the Control group compared to the odds of scoring 100 being in the Intervention group. For the V dimension the OR_{Binary} statistic could not be calculated as no person in either group scored 100.

Observed Effect sizes for the OA Knee study

For the OA Knee study the effect sizes are based on a comparison of the difference in HRQoL scores between the Rheumatology and Surgical groups at initial assessment.

Table 5.2 shows that patients in the Rheumatology group had significantly different scores (compared to the Surgery patients) on four dimensions

Table 5.2: OA Knee Study comparison baseline Effect Sizes for Rheumatology vs. Surgical Groups

SF-36 Dimension		*Group*	*n*	*mean*	*sd*	*Mean diff δ*	*ES Δ_{Normal}*	*t-test P-value*	*Pr(X > Y) $p_{Noether}$*	*MW test P-value*	*$OR_{Ordinal}$*	P_{100}	*δ_{Binary}*	*OR_{Binary}*
Physical Function	*X*	Rheumatology	97	28.2	22.4	7.0	0.34	0.019	0.595	0.022	1.79	0.01	0.01	∞
	Y	Surgical	95	21.2	18.2							0.00		
Role Physical	*X*	Rheumatology	96	11.5	22.0	-1.4	-0.06	0.684	0.501	0.987	0.99	0.01	-0.02	0.33
	Y	Surgical	99	12.9	26.3							0.03		
Bodily Pain	*X*	Rheumatology	100	32.0	19.5	-4.3	-0.20	0.154	0.454	0.245	0.75	0.00	-0.04	0.00
	Y	Surgical	104	36.3	23.4							0.04		
General Health	*X*	Rheumatology	94	43.9	22.9	-13.4	-0.57	0.001	0.339	0.001	0.37	0.00	-0.04	0.00
	Y	Surgical	96	57.3	23.8							0.04		
Vitality	*X*	Rheumatology	98	36.9	19.0	-5.4	-0.28	0.050	0.419	0.049	0.61	0.01	0.01	∞
	Y	Surgical	99	42.3	19.3							0.00		
Social Function	*X*	Rheumatology	100	53.1	30.6	-0.5	-0.02	0.910	0.494	0.876	0.96	0.12	0.02	1.24
	Y	Surgical	101	53.6	27.6							0.10		
Role Emotional	*X*	Rheumatology	95	41.1	44.2	-3.0	-0.07	0.632	0.482	0.639	0.88	0.31	-0.03	0.88
	Y	Surgical	99	44.1	44.6							0.33		
Mental Health	*X*	Rheumatology	99	62.7	20.9	-5.5	-0.28	0.054	0.420	0.051	0.61	0.03	0.00	1.00
	Y	Surgical	100	68.2	18.8							0.03		

of the SF-36, the PF, GH, V and MH dimensions, using the *p*-values from the two independent samples *t*-test and a cut off *p*-value ≤ 0.05 for statistical significance.

Ignoring the sign the absolute values of the Normal effect sizes Δ_{Normal} ranged from 0.02 to 0.57 and were in the 'small' to 'moderate' range using Cohen's (1988) classification.

$p_{Noether}$ ranged from 0.34 to 0.60, reflecting that for some dimensions e.g. PF, ($p_{Noether}$ = 0.60) patients in the Rheumatology group had a better HRQoL than patients in the Surgical group. That is the probability of a randomly chosen patient from the Rheumatology group having a better HRQoL score than a randomly chosen subject from the Surgical group was greater than a half. For other dimensions e.g. GH ($p_{Noether}$ = 0.34) patients in the Rheumatology group were more likely to have a poorer HRQoL than patients in the Surgical group.

The $OR_{Ordinal}$ estimates from the proportional odds model varied from 0.37 to 1.79. Odds ratios greater than one imply that the odds of being in a given category or less (i.e. having better HRQoL) is greater for patients in the Rheumatology group compared to the odds of being in a given category or less in the Surgical group. Conversely odds ratios less than one imply that the odds of being in a given category or less (i.e. having better HRQoL) is lower for patients in the Rheumatology group compared to the odds of being in a given category or less in the Surgical group (i.e. patients in the Rheumatology group have poorer HRQoL than patients in the Surgical group). Finally the test of proportional odds which underlies the model suggested that on three dimensions of the SF-36, PF, GH and V ($p \leq 0.001$), this assumption may not be valid.

In this study the proportion of patients in both groups scoring 100 on any dimension was low, less than 5% except for the SF and RE dimensions. So the absolute value of the effect size, δ_{Binary}, was less than 0.05 for all eight dimensions of the SF-36.

Observed Effect sizes for the Leg Ulcer data

For the Leg Ulcer data the effect sizes are based on a comparison of the difference in baseline HRQoL scores between those leg ulcer patients who walk freely without an aid and the leg ulcer patients who walked with an aid or were bed bound.

Table 5.3 shows that leg ulcer patients able to walk freely without an aid had significantly higher scores (compared to the Intervention group patients) on five dimensions of the SF-36, (PF, RP, BP, V and SF), using the *p*-values from the two independent samples *t*-test and a cut off $p \leq 0.05$ for statistical significance. The observed Normal effect sizes Δ_{Normal} ranged from 0.08 to 1.45 and were all in the 'small' to 'moderate' range using Cohen's (1988) classification, except for the PF dimension which had a 'large' effect size $\Delta_{Normal} = 1.45$.

The probability that a randomly chosen value (or subject) from the Control group was greater (i.e. had a better HRQoL) than a randomly chosen value (or subject) from the Intervention group, i.e. $p_{Noether}$ ranged from 0.518 to 0.833.

The $OR_{Ordinal}$ estimates from the proportional odds model were greater than unity for all eight dimensions of the SF-36. Odds ratios greater than one imply that the odds of being in a given category or less (i.e. having better HRQoL) is greater for patients in the Walk freely group compared to the odds of being in a given category or less in the Walk with aid group. The *p*-values from the Wald test for the significance of the regression coefficient for the group term were almost identical to the *p*-values obtained from the *MW* test and so are not reported. Finally the test of proportional odds which underlies the model suggested that on three dimensions of the SF-36, GH ($p = 0.001$), V ($p = 0.033$) and SF ($p = 0.005$) the assumption of proportional odds may not be valid.

Table 5.3: Leg Ulcer Study comparison baseline Effect Sizes for Walks freely vs. Walks with aid groups

SF-36 Dimension		*Group*	*n*	*mean*	*Sd*	*Mean diff δ*	*ES* Δ_{Normal}	*t-test P-value*	*Pr(X > Y)* $p_{Noether}$	*MW test P-value*	$OR_{Ordinal}$	P_{100}	δ_{Binary}	OR_{Binary}
Physical	*X*	Walk freely	108	62.4	28.1	35.9	1.45	0.001	0.833	0.001	1.27	0.06	0.06	∞
Function	*Y*	Walk with aid	124	26.5	21.6							0.00		
Role	*X*	Walk freely	108	60.6	41.0	18.9	0.47	0.001	0.623	0.001	2.25	0.43	0.18	2.32
Physical	*Y*	Walk with aid	124	41.7	39.9							0.24		
Bodily	*X*	Walk freely	108	62.9	27.5	12.8	0.46	0.001	0.629	0.002	2.23	0.15	0.04	1.36
Pain	*Y*	Walk with aid	124	50.1	28.5							0.11		
General	*X*	Walk freely	108	66.8	22.1	4.4	0.20	0.139	0.564	0.094	1.44	0.05	0.01	1.41
Health	*Y*	Walk with aid	123	62.4	22.3							0.03		
Vitality	*X*	Walk freely	108	57.2	22.3	7.6	0.36	0.007	0.611	0.004	1.99	0.00	-0.01	0.00
	Y	Walk with aid	123	49.6	20.4							0.01		
Social	*X*	Walk freely	108	73.9	27.1	13.7	0.46	0.001	0.624	0.001	2.18	0.36	0.13	1.85
Function	*Y*	Walk with aid	124	60.2	31.8							0.23		
Role	*X*	Walk freely	108	69.8	40.9	6.6	0.16	0.234	0.544	0.202	1.38	0.60	0.09	1.42
Emotional	*Y*	Walk with aid	124	63.2	42.7							0.52		
Mental	*X*	Walk freely	108	70.5	20.9	1.7	0.08	0.550	0.518	0.630	1.12	0.06	0.03	2.41
Health	*Y*	Walk with aid	124	68.8	21.8							0.02		

Observed Effect Sizes for the NAMEIT data

The observed effect sizes for the NAMEIT data were based on a comparison of the HRQoL scores between the Neoral and Placebo group patients at 48-week follow-up.

Table 5.4 shows that patients in the Neoral group had significantly higher scores (compared to the Placebo group patients) on one dimension of the SF-36, the RP dimension, using the *p*-values from the two independent samples *t*-test and a cut off $p \leq 0.05$ for statistical significance. The effect sizes Δ_{Normal} ranged from 0.2 to 0.32 and were all in the 'small' range using Cohen's (1988) classification.

The probability that a randomly chosen value (or subject) from the Neoral group was greater (i.e. had a better HRQoL) than a randomly chosen value (or subject) from the Placebo group, i.e. $p_{Noether}$ ranged from 0.501 to 0.576.

The $OR_{Ordinal}$ estimates from the proportional odds model were greater than unity for all eight dimensions of the SF-36. Odds ratios greater than one imply that the odds of being in a given category or less (i.e. having better HRQoL) is greater for patients in the Neoral group compared to the odds of being in a given category or less in the Placebo group. The *p*-values from the Wald test for the significance of the regression coefficient for the group term were almost identical to the *p*-values obtained from the *MW* test and so are not reported. Finally the test of proportional odds which underlies the model suggested that on three dimensions of the SF-36, GH, V and SF this assumption may not be valid.

Observed Effect sizes for the SF-36 general population data

For these data we show the results of three separate effect size calculations, with various different groups:

(1) Male vs. female subjects

Table 5.4: NAMEIT Study comparison 48 week Effect Sizes for Neoral vs. Placebo Groups

SF-36 Dimension		*Group*	*n*	*mean*	*sd*	*Mean diff δ*	*ES Δ_{Normal}*	*t-test P-value*	*Pr(X > Y) $p_{Noether}$*	*MW test P-value*	*$OR_{Ordinal}$*	P_{100}	*δ_{Binary}*	*OR_{Binary}*
Physical Function	*X*	Neoral	113	55.3	27.1	5.1	0.20	0.141	0.560	0.119	1.44	0.04	*0.02*	*1.98*
	Y	Placebo	114	50.2	24.3							0.02		
Role Physical	*X*	Neoral	113	45.1	43.9	13.1	0.32	0.017	0.576	0.036	1.67	0.33	*0.16*	*2.42*
	Y	Placebo	114	32.0	37.8							0.17		
Bodily Pain	*X*	Neoral	113	58.1	22.8	4.3	0.19	0.154	0.560	0.113	1.45	0.05	*0.03*	*2.10*
	Y	Placebo	114	53.8	22.5							0.03		
General Health	*X*	Neoral	111	46.2	21.7	3.2	0.15	0.277	0.550	0.193	1.36	0.09	*0.09*	∞
	Y	Placebo	113	43.0	21.7							0.00		
Vitality	*X*	Neoral	113	51.6	19.5	3.4	0.17	0.209	0.550	0.195	1.35	0.01	*-0.01*	*0.50*
	Y	Placebo	114	48.2	21.0							0.02		
Social Function	*X*	Neoral	113	62.9	22.4	5.0	0.23	0.086	0.570	0.066	1.54	0.30	*0.13*	*2.03*
	Y	Placebo	114	57.9	21.6							0.18		
Role Emotional	*X*	Neoral	113	61.2	43.3	0.7	0.02	0.897	0.506	0.874	1.04	0.50	*0.03*	*1.11*
	Y	Placebo	113	60.5	42.2							0.47		
Mental Health	*X*	Neoral	113	63.6	18.2	0.3	0.02	0.916	0.501	0.989	1.00	0.01	*-0.01*	*0.50*
	Y	Placebo	114	63.3	20.0							0.02		

(2) General practitioner consultation in the previous 2 weeks (Yes vs. No)

(3) Outpatient Attendance in the previous 3 months (Yes vs. No)

Tables 5.5, 5.6 and 5.7 show the various effect sizes for the three comparisons. The absolute values of Δ_{Normal} ranged from 0.06 to 0.82 and were mainly in the 'small' to 'moderate' range using Cohen's (1988) classification. The values for $p_{Noether}$ ranged from 0.302 to 0.644.

The $OR_{Ordinal}$ estimates from the proportional odds model ranged from 0.26 to 2.45 for the three different effect size comparisons for the general population data. The assumption of proportional odds appears to be valid for seven out of eight dimensions of the SF-36 for the gender (male vs. female) comparison. The exception being the GH dimension, where there is some reliable statistical evidence that this assumption may not be valid ($p = 0.001$). The assumption of proportional odds appears to be reasonable for eight dimensions for the GP consultation comparison. However for the outpatient attendance in the last three months comparison there is some suggestion that the proportional odds assumption may not be valid for five dimensions (PF, RP, GH, SF and RE) with $p < 0.05$ from the test of proportional odds.

Tables 5.1 to 5.7 suggest that large dramatic differences in HRQoL between groups are unlikely and the observed effect sizes Δ_{Normal}, were mainly in the 'small' to 'moderate' range (0.2 to 0.5) using Cohen's criteria (Cohen, 1988). The results of fitting a proportional odds model also suggest that dramatic effects (differences in HRQoL) are unlikely. The results also suggest that with increasing numbers of categories it is less likely that the proportional odds assumption is true.

Combining effect sizes

The five effect sizes, Δ_{Normal}, $p_{Noether}$, δ_{Binary}, OR_{Binary} and $OR_{Ordinal}$, can all be regarded as different numerical expressions of treatment efficacy.

If *a* and *c* are the number of positive responses from a binary outcome for the Intervention and Control groups respectively for a classical two group randomised control trial (see Table 5.8) then we have already seen that

$$OR_{Binary} = \frac{\left\{\left(\frac{a}{a+b}\right)\Big/\left(\frac{b}{a+b}\right)\right\}}{\left\{\left(\frac{c}{c+d}\right)\Big/\left(\frac{d}{c+d}\right)\right\}} = \frac{ad}{bc}, \tag{5.6}$$

Table 5.8: Binary outcomes from a standard two group randomised controlled trial

		Group 1 (Intervention) X	Group 2 (Control) Y
Outcome	+ ve	*a*	*c*
	- ve	*b*	*d*
		$a + b = n_1$	$c + d = n_2$

and the Absolute Risk Reduction:

$$ARR = \frac{a}{a+b} - \frac{c}{c+d} = \delta_{Binary}, \tag{5.7}$$

and the Number-Needed-to-Treat:

$$NNT_{Binary} = \frac{1}{ARR}. \tag{5.8}$$

The effect size statistics Δ_{Normal}, δ_{Binary} and OR_{Binary} have the weakness of being applicable to only certain data types and therefore cannot be universally applied, making the comparison of effect sizes across different outcomes and studies problematic.

The problem is that the standardised effect size Δ_{Normal}, is only appropriate to continuous outcomes and the δ_{Binary}, OR_{Binary} effect size statistics are only applicable to binary outcomes. Three possible solutions to this problem (Shepstone, 2001) are:

Table 5.5: Sheffield General Population (SGP) Survey Comparison of Effect Sizes for Male vs. Female Groups

SF-36 Dimension		*Group*	*n*	*mean*	*sd*	*Mean diff δ*	*ES Δ_{Normal}*	*t-test P-value*	*Pr(X > Y) $p_{Noether}$*	*MW test P-value*	*$OR_{Ordinal}$*	*P_{100}*	*δ_{Binary}*	*OR_{Binary}*
Physical Function	*X*	Male	616	88.3	20.0	2.0	0.10	0.063	0.548	0.002	1.36	0.43	0.06	1.29
	Y	Female	756	86.3	20.1							0.37		
Role Physical	*X*	Male	616	86.2	29.4	5.7	0.18	0.001	0.544	0.001	1.55	0.78	0.09	1.56
	Y	Female	756	80.5	33.4							0.69		
Bodily Pain	*X*	Male	616	82.1	21.9	4.0	0.18	0.001	0.553	0.001	1.41	0.44	0.10	1.50
	Y	Female	756	78.1	23.6							0.35		
General Health	*X*	Male	616	73.0	20.4	1.3	0.06	0.253	0.518	0.247	1.12	0.06	0.00	1.08
	Y	Female	756	71.7	20.9							0.05		
Vitality	*X*	Male	616	66.2	20.7	8.5	0.41	0.001	0.621	0.001	2.10	0.03	0.02	0.00
	Y	Female	756	57.7	21.1							0.01		
Social Function	*X*	Male	616	91.0	17.7	5.8	0.29	0.001	0.580	0.001	1.88	0.69	0.15	1.88
	Y	Female	756	85.2	22.2							0.54		
Role Emotional	*X*	Male	616	86.1	29.9	7.2	0.22	0.001	0.552	0.001	1.70	0.79	0.10	1.72
	Y	Female	756	78.9	34.9							0.69		
Mental Health	*X*	Male	616	77.7	17.6	8.6	0.46	0.001	0.644	0.001	2.45	0.04	0.02	1.93
	Y	Female	756	69.1	19.3							0.02		

Table 5.6: SGP Survey Comparison of Effect Sizes for GP consultation in last 2 weeks vs. No GP consultation

SF-36 Dimension	*Group*	*n*	*mean*	*sd*	*Mean diff δ*	*ES Δ_{Normal}*	*t-test P-value*	*Pr(X > Y) $p_{Noether}$*	*MW test P-value*	*$OR_{Ordinal}$*	*P_{100}*	*δ_{Binary}*	*OR_{Binary}*
Physical Function	GP consultation	253	81.7	23.0	-6.7	-0.34	0.001	0.410	0.001	0.55	0.31	-0.10	0.64
	No consultation	1112	88.4	19.1							0.41		
Role Physical	GP consultation	253	66.0	40.9	-21.0	-0.68	0.001	0.363	0.001	0.30	0.53	-0.25	0.33
	No consultation	1112	87.0	28.0							0.78		
Bodily Pain	GP consultation	253	68.0	26.5	-14.6	-0.66	0.001	0.333	0.001	0.33	0.21	-0.22	0.35
	No consultation	1112	82.6	21.2							0.43		
General Health	GP consultation	253	64.1	23.6	-10.0	-0.49	0.001	0.375	0.001	0.45	0.02	-0.04	0.36
	No consultation	1112	74.1	19.5							0.06		
Vitality	GP consultation	253	52.5	22.4	-11.0	-0.53	0.001	0.354	0.001	0.40	0.01	-0.01	0.00
	No consultation	1112	63.5	20.5							0.02		
Social Function	GP consultation	253	77.1	26.3	-13.1	-0.66	0.001	0.345	0.001	0.32	0.40	-0.26	0.35
	No consultation	1112	90.2	18.1							0.65		
Role Emotional	GP consultation	253	73.0	38.1	-11.1	-0.34	0.001	0.423	0.001	0.50	0.61	-0.15	0.50
	No consultation	1112	84.1	31.4							0.76		
Mental Health	GP consultation	253	65.6	20.9	-9.0	-0.48	0.001	0.370	0.001	0.44	0.02	-0.01	0.54
	No consultation	1112	74.6	18.3							0.03		

Table 5.7: SGP Survey Comparison of Effect Sizes for Outpatient Attendance in last 3 months vs. No Outpatient Attendance

SF-36 Dimension	*Group*	*n*	*mean*	*sd*	*Mean diff δ*	*ES Δ_{Normal}*	*t-test P-value*	*Pr(X > Y) $p_{Noether}$*	*MW test P-value*	*$OR_{Ordinal}$*	*P_{100}*	*δ_{Binary}*	*OR_{Binary}*
Physical Function	Outpatient visit	192	75.7	28.4	-13.4	-0.69	0.001	0.355	0.001	0.37	0.25	-0.17	0.47
	No visit	1167	89.1	17.6							0.42		
Role Physical	Outpatient visit	192	64.7	42.0	-21.5	-0.70	0.001	0.365	0.001	0.30	0.53	-0.24	0.35
	No visit	1167	86.2	28.5							0.76		
Bodily Pain	Outpatient visit	192	64.3	27.2	-18.1	-0.82	0.001	0.302	0.001	0.26	0.18	-0.24	0.30
	No visit	1167	82.4	21.0							0.42		
General Health	Outpatient visit	192	60.8	23.8	-13.5	-0.67	0.001	0.327	0.001	0.33	0.03	-0.04	0.40
	No visit	1167	74.3	19.4							0.06		
Vitality	Outpatient visit	192	54.5	23.2	-8.0	-0.38	0.001	0.399	0.001	0.53	0.01	-0.01	0.00
	No visit	1167	62.5	20.8							0.02		
Social Function	Outpatient visit	192	76.2	28.9	-13.6	-0.68	0.001	0.363	0.001	0.35	0.43	-0.21	0.43
	No visit	1167	89.8	18.0							0.63		
Role Emotional	Outpatient visit	192	74.3	39.8	-9.1	-0.28	0.001	0.451	0.005	0.62	0.67	-0.08	0.69
	No visit	1167	83.4	31.5							0.75		
Mental Health	Outpatient visit	192	68.0	22.2	-5.7	-0.30	0.001	0.432	0.002	0.65	0.02	-0.01	0.77
	No visit	1167	73.7	18.4							0.03		

(1) Dichotomise the continuous outcome;

(2) Categorise the continuous outcome into more than 2 categories;

(3) Use an approximation.

Shepstone (2001) suggests option (3) and that a useful statistic for the quantification of the treatment effect in a two group (Control and Intervention) context is the *A* statistic. If *X* and *Y* are the values of an outcome (higher values more preferable) for randomly selected individuals from the Intervention and Control groups respectively, then $A_{XY} = \Pr(X > Y)$ i.e. the probability that the Intervention patient has an outcome preferable to that of the Control patient, this is equivalent to the effect size statistic (4.5) $p_{Noether}$.

If we let $A_{YX} = \Pr(Y > X)$ i.e. the probability that a random individual from group 2 (Control) has a better outcome than a random individual from group 1 (Intervention) and

$$\lambda = A_{XY} - A_{YX} = \Pr(X > Y) - \Pr(Y > X) \tag{5.9}$$

and

$$\theta = \frac{A_{XY}}{A_{YX}} = \frac{\Pr(X > Y)}{\Pr(Y > X)}. \tag{5.10}$$

For a 2 x 2 table (Table 5.8 above), Pr(*X* > *Y*) i.e. the probability that an individual in group 1 (Intervention) has a positive outcome (a/n_1) and an individual in group 2 (Control) has a negative outcome (d/n_2) is $\frac{a}{n_1} \cdot \frac{d}{n_2}$ (since the two events are independent). Therefore $A_{XY} = \Pr(X > Y) = \frac{ad}{n_1 n_2}$ and $A_{YX} = \Pr(Y > X) = \frac{bc}{n_1 n_2}$. Hence,

$$\lambda = A_{XY} - A_{YX} = \frac{ad - bc}{n_1 n_2} = \frac{a(c+d) - c(a+b)}{n_1 n_2} = \frac{an_2 - cn_1}{n_1 n_2} = \frac{a}{n_1} - \frac{c}{n_2} = ARR, \tag{5.11}$$

and

$$\theta = \frac{A_{XY}}{A_{YX}} = \frac{\frac{ad}{n_1 n_2}}{\frac{bc}{n_1 n_2}} = \frac{ad}{n_1 n_2} \cdot \frac{n_1 n_2}{bc} = \frac{ad}{bc} = OR. \qquad (5.12)$$

Therefore for a binary outcome $\lambda = Pr(X > Y) - Pr(Y > X)$ is equivalent to the Absolute Risk Difference (ARR), which is the inverse of the NNT and θ = Pr(X > Y)/Pr(Y > X) is the equivalent of the OR_{Binary}.

If we consider the case of two treatment groups with an ordered categorical outcome (higher scores are more desirable) with three levels 0, 1, and 2 (Table 5.9).

Table 5.9: Ordinal outcomes from a standard two group randomised controlled trial (higher scores are more desirable)

			Outcome			
	Group		**0**	**1**	**2**	
Group 1	*Intervention*	*X*	*a*	*b*	*c*	n_1
Group 2	*Control*	*Y*	*d*	*e*	*f*	n_2
			m_0	m_1	m_2	

Then

$$A_{XY} = \Pr(X > Y) = \frac{b}{n_1} \cdot \frac{d}{n_2} + \frac{c}{n_1} \cdot \frac{d}{n_2} + \frac{c}{n_1} \cdot \frac{e}{n_2} \qquad (5.13)$$

and

$$\Pr(X = Y) = \frac{a}{n_1} \cdot \frac{d}{n_2} + \frac{b}{n_1} \cdot \frac{e}{n_2} + \frac{c}{n_1} \cdot \frac{f}{n_2} \qquad (5.14)$$

and

$$A_{YX} = \Pr(Y > X) = \frac{e}{n_2} \cdot \frac{a}{n_1} + \frac{f}{n_2} \cdot \frac{a}{n_1} + \frac{f}{n_2} \cdot \frac{b}{n_1}. \qquad (5.15)$$

In general, $A_{XY} = \frac{\sum_{i>j} f_{1i} f_{2j}}{n_1 n_2}$, $A_{YX} = \frac{\sum_{j>i} f_{1i} f_{2j}}{n_1 n_2}$ and $\Pr(X = Y) = \frac{\sum_{i=j} f_{1i} f_{2j}}{n_1 n_2}$, where f_{1i} is the cell count for Group 1, row 1 and column *i* of the general 2 by *k* contingency table of outcomes (*i* = 1 to *k* where *k* is the number of ordinal

categories). A_{XY} and A_{YX} can be estimated by $A_{XY} = \frac{U_{XY}}{n_1 n_2}$ and $A_{YX} = \frac{U_{YX}}{n_1 n_2}$ where U_{XY} and U_{YX} are the values of the Mann-Whitney U statistics.

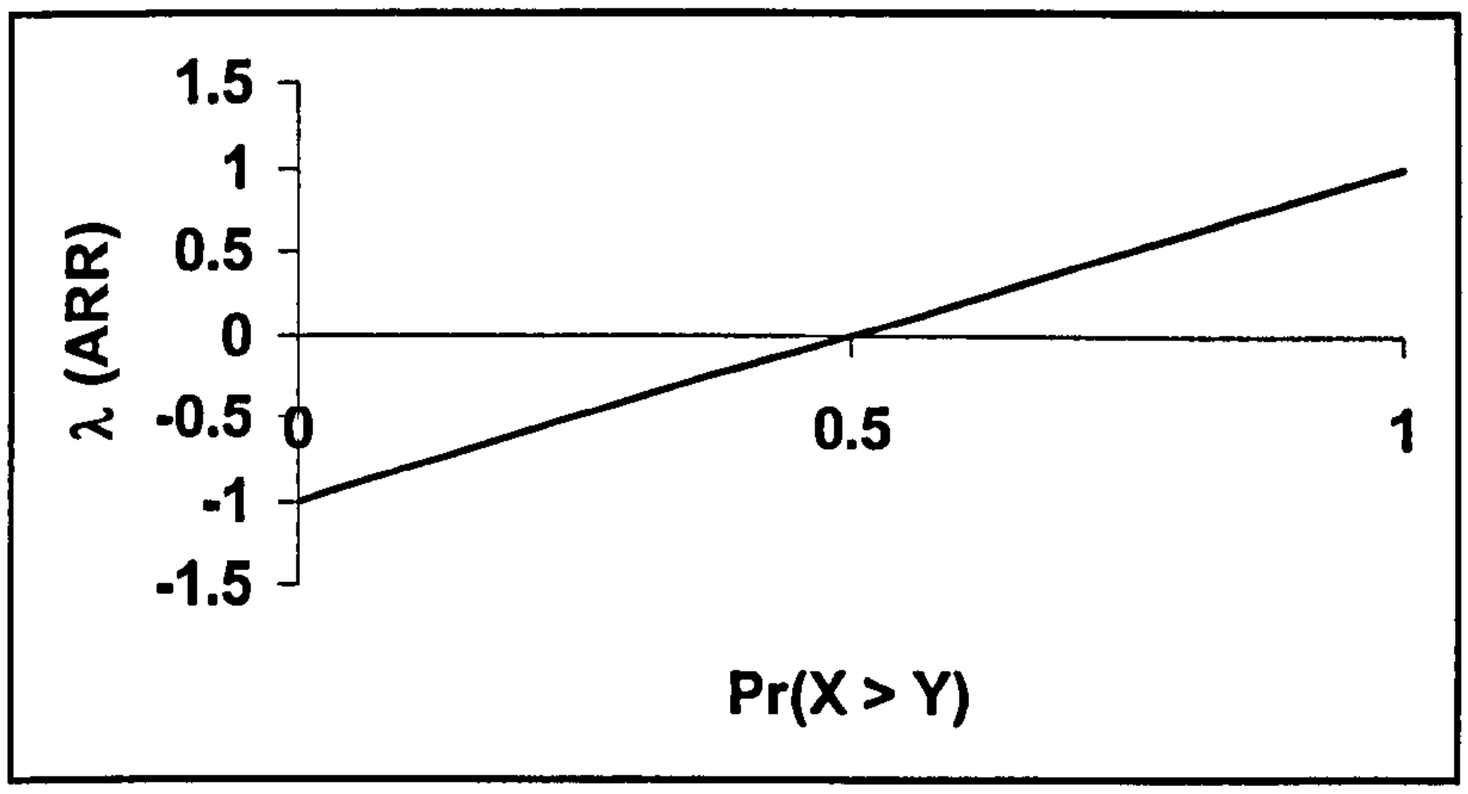
If the outcomes are continuous and/or can be fully ranked and there are no ties in the data then $Pr(X = Y) = 0$ and $\lambda = A_{XY} - A_{YX} = Pr(X > Y) - Pr(Y > X)$ and $\theta = A_{XY}/A_{YX} = Pr(X > Y)/Pr(Y > X)$ can be estimated exactly. Conversely, if there are a large number of ties in the data, i.e. $x_i = y_i$, (which is likely for HRQoL outcomes, with their discrete response categories) then $Pr(X = Y) > 0$. In this case any pairs for which $x_i = y_i$, contribute ½ a unit to both U_{XY} and U_{YX}. Hence the two A statistics A_{XY} and A_{YX} can only be estimated approximately and thus the approximate estimates of θ and λ in the case of ties will be denoted by θ' and λ' respectively.

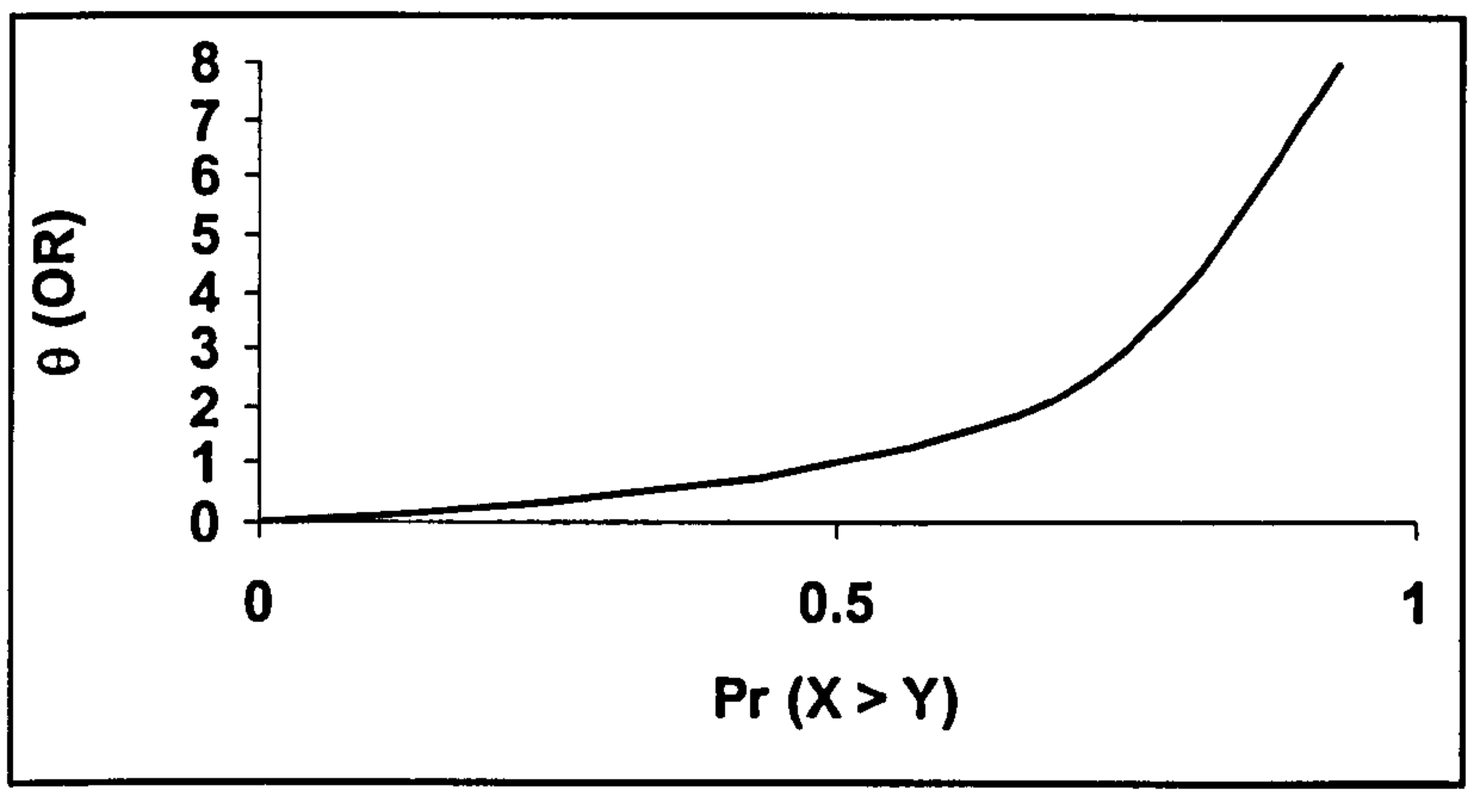
Shepstone (2001) shows that by re-conceptualising the ARR and the OR in terms of A_{XY} and A_{YX} a generalisation can be made to non-binary outcomes. We have already seen (Chapter 4) that A_{XY} and A_{YX}, or their equivalent statistics $Pr(X > Y)$ and $Pr(Y > X)$ can be calculated by either a parametric approach for continuous outcomes (equation 4.7) via a theoretical distribution (e.g. Normal) or a non-parametric approach without any distributional assumptions via the Mann-Whitney U statistic.

Thus, these statistics A_{XY}, A_{YX}, λ and θ can be generalised to ordinal and continuous outcomes with no distributional assumptions. Therefore, the NNT and OR statistics can be generalised to all data types with analogous interpretations. Figure 5.1 shows the relationship between $Pr(X > Y)$ or its equivalent A_{XY} and λ (the ARR) and θ (the OR).

In the case of a Normal or Gaussian outcome, Shepstone (2001) shows that the generalised NNT_λ is a function of the effect size Δ_{Normal} (see Table 5.10). Therefore by conceptualising the ARR and the OR in this fashion a simple and universal approach to expressing group differences and effect sizes is obtained.

Figure 5.1: The relationship between Pr(X >Y) and λ (ARR) and θ (OR)





Tables 5.11 to 5.17 show that Odds Ratio effect size estimates θ and $OR_{Ordinal}$ are of similar magnitude, but are not identical. The $OR_{Ordinal}$ estimate consistently tends to be larger than the corresponding estimate θ. The $OR_{Ordinal}$ is parametric estimate as it assumes that the proportional odds assumption for the cumulative logit model is valid. On the other hand the estimate of θ we have used (which is derived from the Mann-Whitney U test) is a non-parametric estimate and makes no distributional assumptions. Also the cumulative logit model is estimating $\log\left[\frac{\gamma_j}{1-\gamma_j}\right]$, where $\gamma_j = Pr(Y <= y_j)$ i.e. the probability of being in category *j or less*, not $Pr(Y < y_j)$. If there are a large number of ties in the data, i.e. $x_i = y_i$, (which is likely for HRQoL outcomes, with their discrete response categories) then $Pr(X = Y) > 0$. In this case any pairs for which $x_i = y_i$, contribute ½ a unit to both U_{XY} and U_{YX}. Hence the two odds ratio effect size estimates θ and $OR_{Ordinal}$ are likely to diverge further.

Table 5.10: The relationship between the Gaussian effect size Δ_{Normal}, λ, θ, Cohen's criteria and the NNT_λ (adapted from Shepstone 2001)

Effect Size Δ_{Normal}	(ARR) A_{XY}-A_{YX} λ	(OR) A_{XY}/A_{YX} θ	NNT_λ[1]	Interpretation[2]
0	0	1.0	∞	*No effect*
0.2	0.11	1.25	8.9	*Small effect*
0.5	0.28	1.76	3.6	*Moderate effect*
0.8	0.43	2.50	2.3	*Large Effect*
1.0	0.52	3.17	1.9	
2.0	0.84	11.66	1.2	

1. The inverse of λ can be interpreted as the NNT.

2. Based on Cohen's (1988) interpretation of the effect size, Δ_{Normal}.

Again using Cohen's (1988) interpretation (from Table 5.10) the θ, λ and NNT statistics are mainly in the 'no effect' to 'moderate effect' range. Thus large dramatic differences in HRQoL between groups are inconsistent with the observed data.

Summary

In this chapter we have calculated several effect size statistics for the eight SF-36 dimensions across a variety of studies with different populations. We have also shown how the different numerical expressions of effect size can be unified via the λ (ARR), θ (OR) and NNT statistics. The results suggest that large differences in HRQoL (as measured by the SF-36) between groups are unlikely, particularly from the randomised controlled trial comparisons and the observed effect sizes were mainly in the 'small' to 'moderate' range (0.2 to 0.5) using Cohen's (1988) criteria.

We shall use these estimates of effect size δ, Δ_{Normal} and *OR* as our MID in the calculation of sample sizes in the next chapter. In this chapter we will compare the power of various methods of sample size estimation described in the

preceding chapter for simple two group cross-sectional comparisons via bootstrap simulation.

Table 5.11: CPSW Study Generalised Effect Sizes and NNTs for Control vs. Intervention Groups

SF-36 Dimension	*Mean diff* δ	*ES* Δ_{Normal}	*Pr(X > Y)* $p_{Noether}$	$OR_{Ordinal}$	ARR λ'	OR θ'	$NNT_{\lambda'}$
Physical Function	2.6	0.17	0.561	1.49	0.123	1.28	8.1
Role Physical	9.1	0.23	0.568	1.66	0.137	1.32	7.3
Bodily Pain	4.0	0.17	0.552	1.39	0.105	1.23	9.5
General Health	2.4	0.13	0.542	1.26	0.084	1.18	11.9
Vitality	1.3	0.06	0.514	1.09	0.028	1.06	36.4
Social Function	4.7	0.20	0.561	1.48	0.122	1.28	8.2
Role Emotional	1.1	0.03	0.515	1.13	0.029	1.06	34.1
Mental Health	-0.2	-0.01	0.499	0.99	-0.002	1.00	541.7

Table 5.12: OA Knee Study Generalised Effect Sizes and NNTs for Rheumatology vs Surgical Groups

SF-36 Dimension	*Mean diff* δ	*ES* Δ_{Normal}	*Pr(X > Y)* $p_{Noether}$	$OR_{Ordinal}$	ARR λ'	OR θ'	$NNT_{\lambda'}$
Physical Function	7.0	0.34	0.595	1.79	0.190	1.47	5.3
Role Physical	-1.4	-0.06	0.501	0.99	0.001	1.00	950.4
Bodily Pain	-4.3	-0.20	0.454	0.75	-0.092	0.83	-10.8
General Health	-13.4	-0.57	0.339	0.37	-0.321	0.51	-3.1
Vitality	-5.4	-0.28	0.419	0.61	-0.162	0.72	-6.2
Social Function	-0.5	-0.02	0.494	0.96	-0.013	0.97	-78.9
Role Emotional	-3.0	-0.07	0.482	0.88	-0.036	0.93	-27.5
Mental Health	-5.5	-0.28	0.420	0.61	-0.160	0.72	-6.3

Table 5.13: Leg Ulcer Study Generalised Effect Sizes and NNTs for Walks freely vs. Walks with aid groups

SF-36 Dimension	*Mean diff* δ	*ES* Δ_{Normal}	*Pr(X > Y)* $p_{Noether}$	$OR_{Ordinal}$	ARR λ'	OR θ'	$NNT_{\lambda'}$
Physical Function	35.9	1.45	0.833	1.27	0.666	4.98	1.5
Role Physical	18.9	0.47	0.623	2.25	0.246	1.65	4.1
Bodily Pain	12.8	0.46	0.629	2.23	0.258	1.69	3.9
General Health	4.4	0.20	0.564	1.44	0.127	1.29	7.9
Vitality	7.6	0.36	0.611	1.99	0.221	1.57	4.5
Social Function	13.7	0.46	0.624	2.18	0.247	1.66	4.0
Role Emotional	6.6	0.16	0.544	1.38	0.088	1.19	11.4
Mental Health	1.7	0.08	0.518	1.12	0.037	1.08	27.3

Table 5.14: NAMEIT Study Generalised Effect Sizes and NNTs for Neoral vs. Placebo Groups

SF-36 Dimension	*Mean diff* δ	*ES* Δ_{Normal}	*Pr(X > Y)* $p_{Noether}$	$OR_{Ordinal}$	ARR λ'	OR θ'	$NNT_{\lambda'}$
Physical Function	5.1	0.20	0.560	1.44	0.119	1.27	8.4
Role Physical	13.1	0.32	0.576	1.67	0.153	1.36	6.5
Bodily Pain	4.3	0.19	0.560	1.45	0.121	1.27	8.3
General Health	3.2	0.15	0.550	1.36	0.101	1.22	9.9
Vitality	3.4	0.17	0.550	1.35	0.099	1.22	10.1
Social Function	5.0	0.23	0.570	1.54	0.139	1.32	7.2
Role Emotional	0.7	0.02	0.506	1.04	0.011	1.02	88.1
Mental Health	0.3	0.02	0.501	1.00	0.001	1.00	990.9

Table 5.15: Sheffield General Population Survey Generalised Effect Sizes and NNTs for Male vs. Female Groups

SF-36 Dimension	*Mean diff* δ	*ES* Δ_{Normal}	*Pr(X > Y)* $p_{Noether}$	$OR_{Ordinal}$	ARR λ'	OR θ'	$NNT_{\lambda'}$
Physical Function	2.0	0.10	0.548	1.36	0.096	1.21	10.4
Role Physical	5.7	0.18	0.544	1.55	0.087	1.19	11.4
Bodily Pain	4.0	0.18	0.553	1.41	0.106	1.24	9.4
General Health	1.3	0.06	0.518	1.12	0.036	1.08	27.6
Vitality	8.5	0.41	0.621	2.10	0.241	1.64	4.1
Social Function	5.8	0.29	0.580	1.88	0.160	1.38	6.3
Role Emotional	7.2	0.22	0.552	1.70	0.104	1.23	9.6
Mental Health	8.6	0.46	0.644	2.45	0.288	1.81	3.5

Table 5.16: Sheffield General Population Survey Generalised Effect Sizes and NNTs for GP consultation in last 2 weeks vs No GP consultation

SF-36 Dimension	*Mean diff* δ	*ES* Δ_{Normal}	*Pr(X > Y)* $p_{Noether}$	$OR_{Ordinal}$	ARR λ'	OR θ'	$NNT_{\lambda'}$
Physical Function	-6.7	-0.34	0.410	0.55	-0.180	0.70	-5.6
Role Physical	-21.0	-0.68	0.363	0.30	-0.274	0.57	-3.7
Bodily Pain	-14.6	-0.66	0.333	0.33	-0.335	0.50	-3.0
General Health	-10.0	-0.49	0.375	0.45	-0.251	0.60	-4.0
Vitality	-11.0	-0.53	0.354	0.40	-0.291	0.55	-3.4
Social Function	-13.1	-0.66	0.345	0.32	-0.310	0.53	-3.2
Role Emotional	-11.1	-0.34	0.423	0.50	-0.154	0.73	-6.5
Mental Health	-9.0	-0.48	0.370	0.44	-0.261	0.59	-3.8

Table 5.17: Sheffield General Population Survey Generalised Effect Sizes and NNTs for Outpatient Attendence in last 3 months vs. No Attendance

SF-36 Dimension	*Mean diff* δ	*ES* Δ_{Normal}	*Pr(X > Y)* $p_{Noether}$	$OR_{Ordinal}$	ARR λ'	OR θ'	$NNT_{\lambda'}$
Physical Function	-13.4	-0.69	0.355	0.37	-0.291	0.55	-3.4
Role Physical	-21.5	-0.70	0.365	0.30	-0.270	0.57	-3.7
Bodily Pain	-18.1	-0.82	0.302	0.26	-0.396	0.43	-2.5
General Health	-13.5	-0.67	0.327	0.33	-0.347	0.48	-2.9
Vitality	-8.0	-0.38	0.399	0.53	-0.203	0.66	-4.9
Social Function	-13.6	-0.68	0.363	0.35	-0.274	0.57	-3.7
Role Emotional	-9.1	-0.28	0.451	0.62	-0.098	0.82	-10.2
Mental Health	-5.7	-0.30	0.432	0.65	-0.136	0.76	-7.3

Chapter 6: Comparing the power of various methods of sample size estimation via bootstrap simulation for simple two group cross-sectional designs

Introduction

We have already seen that HRQoL outcome measures may not meet the distributional requirements (usually that the data have a Normal distribution) of parametric methods of sample size estimation and analysis. Therefore non-parametric methods are most often used to analyse HRQoL data. In this chapter, we describe a non-parametric bootstrap method for estimating sample size and power, when the primary outcome of the study is HRQoL measure (which usually have bounded and non-standard distributions). The aim of this chapter is to compare the power of various methods of sample size estimation (described in Chapter 4) when using HRQoL measures as outcomes using bootstrap simulation and to provide pragmatic guidance to researchers on what method to use.

For simplicity in this paper we will assume the SF-36 is being used as the primary HRQoL endpoint in a two group comparative clinical study, at a single time point, to assess the superiority (not equivalence) of a new treatment over a control treatment.

To illustrate this, we use some HRQoL data from a randomised controlled trial, the Community Postnatal Support Worker (CPSW) Study, which aimed to compare the difference in health status in a group of women who were offered postnatal support (Intervention) from a community midwifery support worker compared with a control group of women who were not offered support. The primary outcome (used to estimate sample size for this study) was the GH dimension of the SF-36 at six weeks postnatally.

In Chapter 4, we described five methods of sample-size estimation for the SF-36 for a simple two group cross-sectional comparison of HRQoL.

Method (1) Equation (4.4) assumes the various individual dimensions of the SF-36 are continuous and Normally distributed. The effect size is Δ_{Normal} (4.3).

Method (2) Equation (4.6) assumes the SF-36 dimensions are continuous but with no other distributional assumptions. The effect size is $p_{Noether}$ (4.5).

Method (3) Equations (4.8 or 4.10) assumes the HRQoL outcomes are measured on binary scale or that the HRQoL outcomes can be dichotomised into a binary scale. The effect size is OR_{Binary} (4.9).

Method (4) Equation (4.12) assumes the SF-36 is an ordered categorical outcome. The effect size is $OR_{Ordinal}$ (4.11).

Method (5) Bootstrap approach (random sampling with replacement).

Figure 6.1g shows the overall distribution of the SF-36 GH dimension in a general population sample of women aged 16 to 45 years. The GH dimension does not appear to be Normally distributed and appears to be negatively skewed, with more people reporting better health in this general population sample.

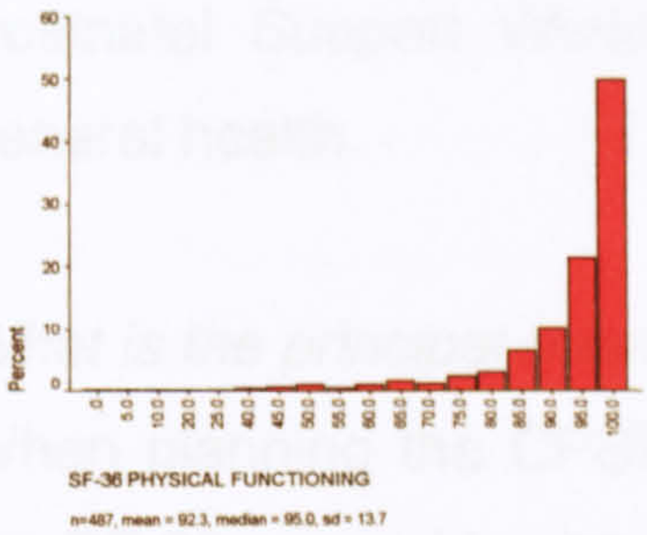
Method 1: Continuous Normally distributed HRQoL data

Suppose we have two independent random samples $\mathbf{x} = (x_1, x_2,\ldots,x_m)$ and $\mathbf{y} = (y_1, y_2,\ldots,y_n)$ of size m and n respectively. The **x**'s and **y**'s are random samples from continuous distributions having cdfs, F_x and F_y respectively. We will consider situations where the distributions have the same shape, but the locations may differ. Thus if δ denotes the location difference (i.e. mean (**y**) - mean (**x**) = δ), then $F_y(y) = F_x(y - \delta)$, for every y. We shall focus on the null hypothesis H_0: $\delta = 0$ against the alternative H_A: $\delta \neq 0$. We can test these hypotheses using an appropriate significance test (e.g. *MW* or *t*-test). With a Normal distribution under the location shift assumption and with $n = m$, the necessary sample size to achieve a certain power is given by (4.4).

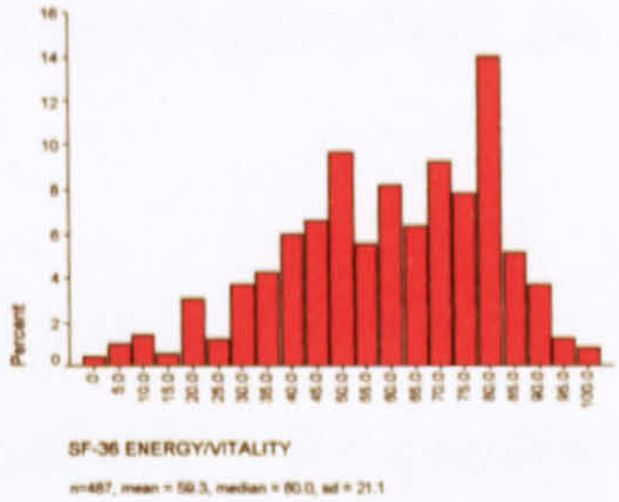
When planning the CPSW study we went through Pocock's (1983) five key questions regarding sample size.

Figure 6.1: Distribution of the eight SF-36 dimensions in the Sheffield population, females aged 16-45 (n = 487)

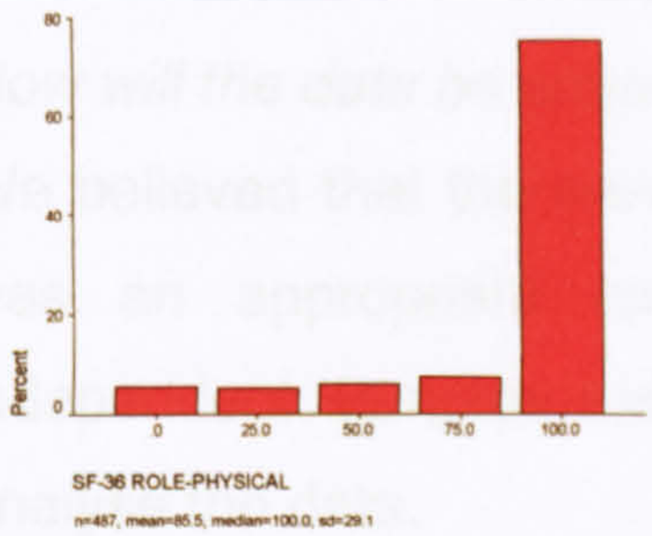
a)



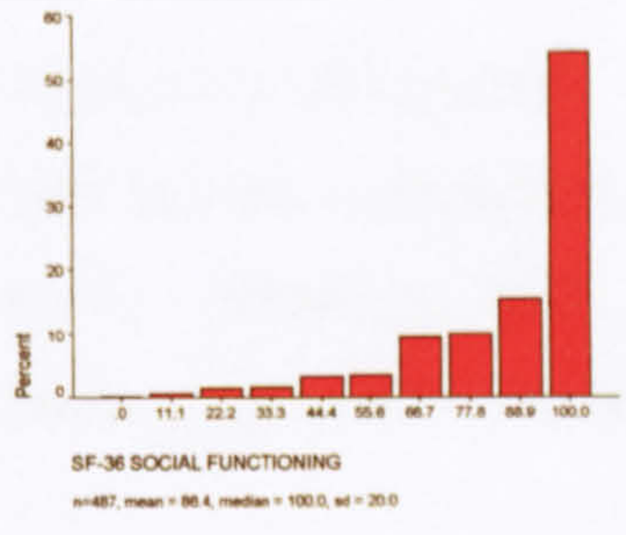
b)



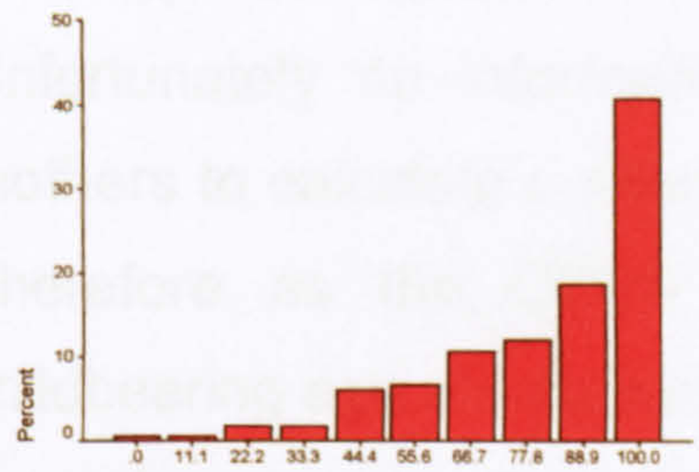
c)



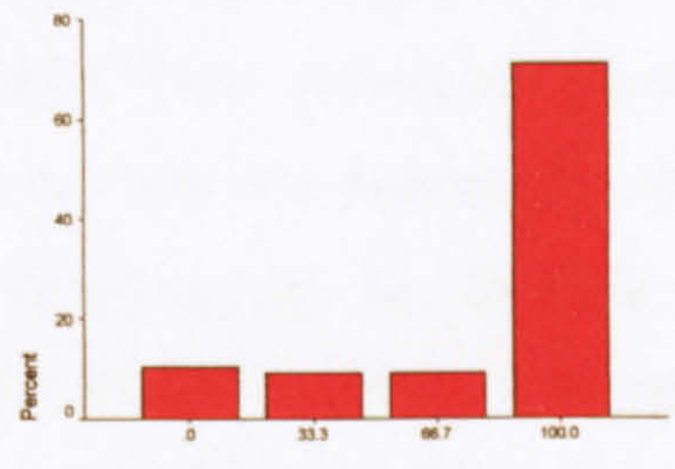
d)



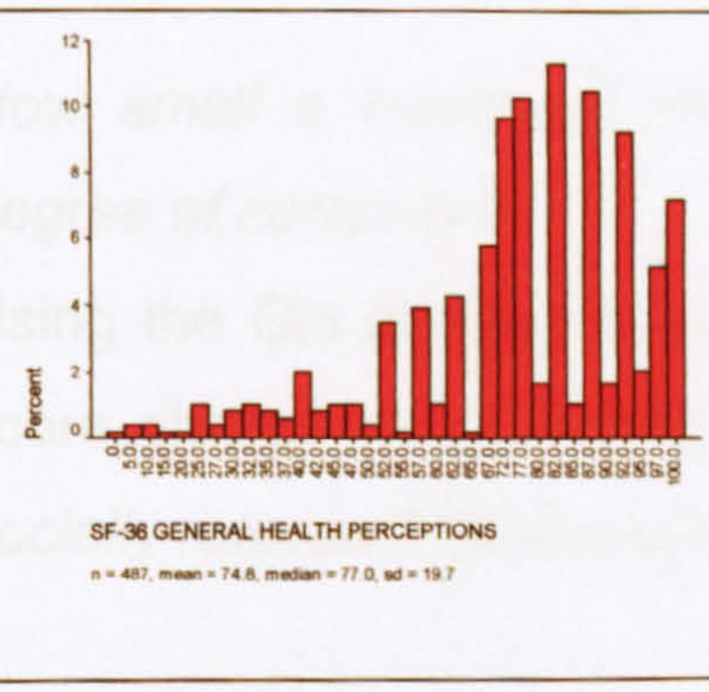
e)



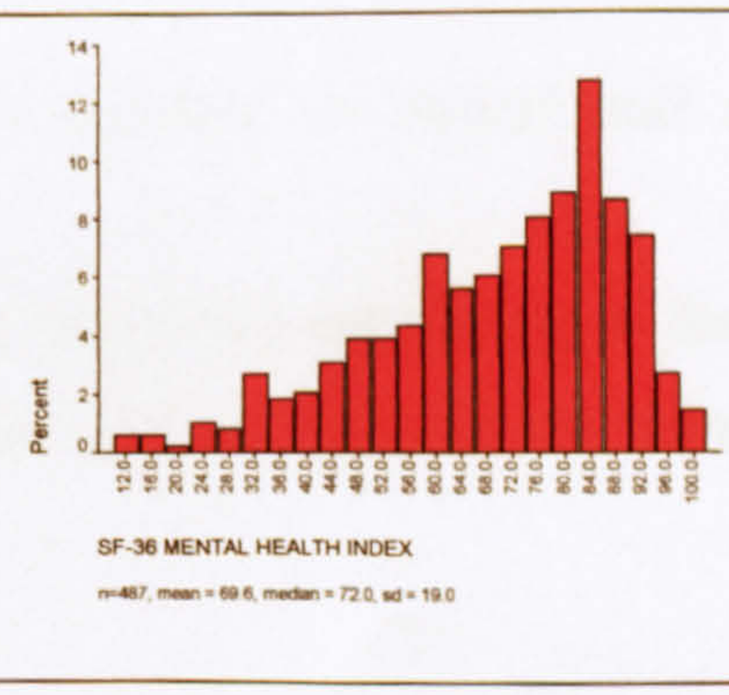
f)



g)



h)



What is the main purpose of the trial?

To assess whether additional postnatal support by trained Community Postnatal Support Workers could have a positive effect on the mother's general health.

What is the principal measure of patient outcome?

When planning the CPSW study we decided that the primary outcome was the SF-36 general health perception (GH) dimension at six weeks postnatally.

How will the data be analysed to detect a treatment difference?

We believed that the mean difference in GH scores between the two groups was an appropriate comparative summary measure and that a two-independent samples *t*-test or multiple linear regression would be used to analyse the data.

What type of results does one anticipate with standard treatment?

Unfortunately no information was available from previous studies of new mothers to calculate a sample size based on the GH dimension of the SF-36. Therefore as the CPSW study was only going to involve women of childbearing age we estimated the standard deviation of the GH outcome from a previous survey of the Sheffield general population using females aged 16 to 45 (Figure 6.1g). This gave us an estimated SD of 20.

How small a treatment difference is it important to detect and with what degree of certainty?

Using the GH dimension of the SF-36, a five-point difference is the smallest score change achievable by an individual and considered as "*clinically and socially relevant*" (Ware *et al* 1993).

Therefore, using Method 1, assuming a standard deviation σ of 20 and a location shift or mean difference ($\mu_T - \mu_C$) of 5 or more points (i.e. $\delta = 5$) between the two groups is clinically and practically relevant gives a

standardised effect size, Δ_{Normal} (from 4.3), of 0.25. This is equivalent to a 'small' effect using Cohen's (1988) interpretation and is of similar magnitude to the effect sizes calculated from the various datasets described in Chapter 5. Using this standardised effect size, equation (4.4) with a two-sided 5% significance level and 80% power gives the estimated number of subjects per group, n_{Normal}, as 244.

Transformations

If the SF-36 outcome data were continuous but had a skewed distribution they may be transformed using a logarithmic transformation. The transformed variable may have a more symmetric distribution that is better approximated by the Normal form. One problem with transforming data is that some HRQoL measures (such as the SF-36 dimensions) are scored on 0 to 100 scales and the natural logarithm of zero does not exist. Another difficulty with the use of transformations is that they distort HRQoL scales and make interpretation of treatment effects difficult (Fayers and Machin, 2000; Walters *et al* 2001b). Unfortunately log-transforming the GH data in Figure 6.1g did not make the distribution of the data more symmetric.

Method 2: Continuous HRQoL data with no distributional assumptions

If the GH dimension outcome of the SF-36 is assumed to be continuous and plausibly not sampled from a Normal distribution then the most popular (not necessarily the most efficient) non-parametric test for comparing two independent samples is the two-sample *Mann-Whitney U* test.

Noether derived a sample size formula for the *MW* test equation (4.6), using an effect size $p_{Noether}$, (4.5) which is an estimate of the probability that an observation drawn at random from population X would exceed an observation drawn at random from population Y i.e. (Pr(X > Y)), that makes no assumptions about the distribution of the data (except that it is continuous), and can be used whenever the sampling distribution of the test statistic *U* can be closely approximated by the Normal distribution, an approximation that is usually quite good except for very small *n*.

Thus to determine the sample size, we have to find the 'effect size' $p_{Noether}$. There are several ways of estimating $p_{Noether}$, under various assumptions, one non-parametric possibility is $p_{Noether} = U/nm$. Unfortunately, this can only be estimated after we have collected the data and calculated the *U* statistic or by computer simulation (as we shall see later). If we assume that $X \sim N(\mu_X, \sigma^2_X)$ and $Y \sim N(\mu_Y, \sigma^2_Y)$ then a *parametric* estimate of Pr(X > Y) using the sample estimates of the mean and variance $(\hat{\mu}_X, \hat{\sigma}^2_X, \hat{\mu}_Y, \hat{\sigma}^2_Y)$ is given by (4.7).

If we assume the SF-36 is Normally distributed then equation 4.7 allows the calculation of two comparable 'effect sizes' $p_{Noether}$ and Δ_{Normal} thus enabling the two methods of sample size estimation (4.4 and 4.6) to be directly contrasted. If this SF-36 is not Normally distributed then we cannot use (4.7) to calculate comparable effect sizes and must rely on the empirical estimates of $p_{Noether} = U/nm$ calculated post hoc from the data. Alternatively, under the location shift assumption, we can use bootstrap methods to estimate $p_{Noether}$.

As before if we assume a mean difference of 5 (i.e. $\delta = \hat{\mu}_X - \hat{\mu}_Y = 5$) and a common standard deviation of 20 (i.e. $\hat{\sigma} = \hat{\sigma}_X = \hat{\sigma}_Y = 20$) for the GH dimension of the SF-36, then using (4.7) this leads to a parametric estimate of the effect size $p_{Noether}$ = Pr(X > Y) of 0.57. Which leads to an estimated ARR, λ' = 0.57 - 0.43 = 0.14 and an estimated OR θ = 0.57/0.43 = 1.33. Substituting $p_{Noether}$ = 0.57 in (4.6) with a two-sided 5% significance level and 80% power gives the estimated number of subjects per group, $n_{Non\text{-}normal}$ as 258.

Method 1 (n_{Normal} = 244) gave a slightly smaller sample size estimate than Method 2 ($n_{Non\text{-}normal}$ = 258). The two methods can be regarded as equivalent when the two distributions have the same shape and equal variances. When the two distributions are Normally distributed with equal variances, the *MW* test will require about 5% more observations than the two-sample *t*-test to provide the same power against the same alternative (Elashoff, 1999). For non-Normal populations, especially those with long tails, the *MW* test may not require as many observations as the two-sample *t*-test.

Table 6.1 shows the estimated samples sizes for the other seven dimensions of the SF-36, besides GH, using both Method 1 and Method 2 and with a parametric estimate (4.7) of $p_{Noether}$. The $n_{Non\text{-}normal}$ estimate using Method 2 of the required sample size is about 5% greater than the n_{Normal} estimate (Method 1) for all eight dimensions of the SF-36.

Empirically, calculating a parametric estimate of Pr(X > Y) from the observed effect size data (using the observed sample means and standard deviations) in Tables 5.1 to 5.7 (of Chapter 5), leads to values very similar to the non-parametric estimate (data and calculations not shown). For example for the GH dimension in the CPSW data in Table 5.1, the observed non-parametric (N-P) estimate of Pr(X > Y) was 0.542 compared to a parametric (P) estimate of 0.537. An absolute difference (N-P estimate – P estimate) of 0.005 and a relative difference or ratio (P: N-P) of 0.991.

However, Noether's method is very sensitive to variations in the effect size $p_{Noether}$. From the bootstrap resampling (assuming bounded outcomes and a location shift of five points, see below for more details) the mean observed effect size estimate from the bootstrap replications was $p_{Noether}$ = 0.58, compared to 0.57 assuming a Normal distribution and equation (4.7). Although only a difference of 0.01, this leads to estimated sample size of 205 subjects per group when $p_{Noether}$ = 0.58 in equation (4.6) with a two-sided 5% significance level and 80% compared to 258 when $p_{Noether}$ = 0.57.

Method 3: Dichotomous categorical HRQoL data – comparing two proportions

From equations (5.9) and (5.10) once we have an estimate of Pr(X > Y) we can calculate two other statistics λ and θ. For a binary outcome, from (5.9) λ = $Pr(X > Y) - Pr(Y > X)$ is equivalent to the Absolute Risk Difference (ARR), which is the inverse of the NNT and from (5.10), θ = Pr(X > Y)/Pr(Y > X) is the equivalent of the OR_{Binary}.

Table 6.1 shows the estimates of θ (the OR) based on parametric estimates of Pr(X > Y) and Pr(Y > X) for the eight dimensions of the SF-36. For the GH dimension a parametric estimate of the OR θ is 1.33.

If we assume the HRQoL outcomes are measured on a binary categorical scale, for example, "good health" and "poor health", or can be dichotomised into a binary scale then an appropriate summary measure of the outcome data will usually be the sample rate or proportion in the sample with "good" HRQoL. When comparing two groups or a single group over time, appropriate comparative summary measures may include the difference in rates or proportions, the relative risk or the odds ratio.

We dichotomise the SF-36 dimensions into two categories i.e. those with a score of 100 classified as "good health" and those with a score less than 100 classified as having "less than good health" or "poor health". Table 6.1 shows the estimated proportions scoring 100 (P_{100}) or good health in the Control or reference group. For the GH dimension the proportion in the control scoring 100 or reporting "good health" was 7%.

If we assume a proportion of patients of 0.07 scoring 100 in the control group (i.e. $\pi_C = 0.07$, Then, for a given OR = OR_{Binary} = 1.33, the anticipated proportion (π_T) of patients scoring 100 or good health in the treatment T group is given by:

$$\pi_T = \frac{OR_{Binary}\pi_C}{OR_{Binary}\pi_C + (1-\pi_C)}, \tag{6.1}$$

$$\pi_T = \frac{1.33 \times 0.07}{(1.33 \times 0.07) + (1-0.07)} = 0.09\ . \tag{6.2}$$

This is equivalent to a δ_{Binary} = (π_T - π_C) = 0.09 - 0.07 = 0.02. Using OR_{Binary} = 1.33 and $\bar{\pi} = (0.09 + 0.07)/2 = 0.08$ in equation (4.10) gives a sample size per group of n_{OR_Binary} = 2509 patients.

SF-36 Dimension	Reference (control) population mean	sd	P_{100} [1]	Mean diff δ	ES Δ_{Normal} [2]	Parametric Pr(X > Y) $p_{Noether}$ [3]	Parametric OR θ' [4]	δ_{Binary}	Estimated sample size per group 5% (two-sided significance) and 80% power n_{Normal}	$n_{Non\text{-}Normal}$	$n_{Ordinal}$	n_{OR}
Physical Function	92.3	13.7	0.50	5.0	0.36	0.602	1.51	0.10	118	127	336	372
Role Physical	85.5	29.1	0.75	5.0	0.17	0.548	1.21	0.14	532	559	2393	2334
Bodily Pain	81.4	22.0	0.41	5.0	0.23	0.564	1.29	0.05	304	321	393	967
General Health	74.8	19.7	0.07	5.0	0.25	0.571	1.33	0.02	244	258	584	2509
Vitality	59.3	21.1	0.01	5.0	0.24	0.567	1.31	0.00	280	296	651	23959
Social Function	86.4	20.0	0.54	5.0	0.25	0.570	1.33	0.09	251	266	353	805
Role Emotional	80.2	34.3	0.71	5.0	0.15	0.541	1.18	0.04	738	776	2793	2901
Mental Health	69.6	19.0	0.01	5.0	0.26	0.574	1.35	0.00	227	240	527	11004

1. P_{100} proportion in each group scoring 100 on the SF-36 dimension, $\delta_{Binary} = P_{100\_Intervention} - P_{100\_Control}$.
2. Effect size Δ_{Normal} = mean difference (δ) divided by the pooled standard deviation (sd), see equation (4.3).
3. Parametric estimate of effect size $p_{Noether} = Pr(X_{Intervention} > Y_{Control})$ based on equation (4.7).
4. Parametric estimate of OR, θ' based on Pr(X > Y)/Pr(X < Y) where Pr(X > Y) is based on equation (4.7).

Table 6.1 shows that the n_{OR} estimates of sample size are far greater than the n_{Normal} and $n_{Non\text{-}normal}$ estimates for all eight dimensions of the SF-36.

Method 4: Ordered categorical (Ordinal) HRQoL data

If the HRQoL outcomes are measured on an ordinal scale, then the most popular (but not necessarily the most efficient) statistical hypothesis test used in this instance (to compare two independent groups) is the Mann-Whitney *U* test with allowance for ties or Chi-squared test for trend.

For sample size estimation for ordinal outcomes, Whitehead (1993) suggested an effect size of the log odds ratio θ_R ($OR_{Ordinal}$), which is the odds of a subject being in a given category or lower in one group compared with the odds in the other group. To estimate the required sample size using Method 4 (equation 4.12) we must also know or specify the proportion of subjects (π_j) in each scale category for one of the groups. Whitehead's method relies on the assumption of a constant odds ratio for the data and also assumes a relatively small log odds ratio and a large sample size, which will often be the case in HRQoL studies where dramatic effects are unlikely (as we have seen in Chapter 5).

If we assume $OR_{Ordinal} = OR_{Binary} = OR = 1.33$ and proportional odds. The assumption of proportional odds specifies that the $OR_{Ordinali}$ will be the same for all 34 categories of the GH dimension. If we also assume the proportion of subjects in each category in the control group is the same as in Figure 6.1g. Then (4.12) gives the number of subjects per group *n* for a two-sided significance level α and power 1 - β.

Under the assumption of proportional odds $OR_{Ordinal} = 1.33$, the anticipated cumulative proportions (γ_{iT}) for each category of treatment *T* are given by:

$$\gamma_{iT} = \frac{OR_{Ordinal}\gamma_{iC}}{OR_{Ordinal}\gamma_{iC} + (1 - \gamma_{iC})} \qquad i = 1 \text{ to } k\text{-}1. \qquad (6.3)$$

After calculating the cumulative proportions (γ_{iT}), the anticipated proportions falling into each treatment category, π_{iT} can be determined from the difference in successive γ_{iT}. Finally, the combined mean ($\bar{\pi}_i$) of the proportions of treatments *C* and *T* for each category is calculated.

Substituting $OR_{Ordinal} = 1.33$ and $\sum_{i=1}^{k} \bar{\pi}_i^3 = 0.0067$ in (4.12) with a two-sided 5% significance level and 80% power gives the estimated number of subjects per group as 584.

With a sample size of 584, proportions in each category in the control group as shown in Figure 6.1g and an $OR_{Ordinal}$ of 1.33, then the estimated mean GH score will be 77.6 in the treatment group compared to an estimated mean GH score of 75.0 in the control group. This is an estimated mean difference of 2.6 points, which is smaller than the five-point mean difference used in (4.4) to calculate n_{Normal}.

Assuming proportions in each category in the control group as shown in Figure 6.1g and proportional odds shift. Then an $OR_{Ordinal}$ of 1.63 and $\sum_{i=1}^{k} \bar{\pi}_i^3 = 0.007$ with a two-sided 5% significance level and 80% power leads to a sample size estimate of 199 subjects per group. With 199 subjects per group and proportional odds of 1.63 this leads to estimated sample mean GH scores of 74.8 and 79.8 in the control and intervention groups respectively i.e. a mean difference of approximately five-points between the groups.

If the number of categories is large it is difficult to postulate the proportion of subjects who would fall in a given category. Whitehead (1993), Campbell *et al* (1985) and Julious *et al* (1997) point out that there is little increase in power (and hence saving in the number of subjects recruited) to be gained by increasing the number of categories beyond five. Categories that are equally likely to occur lead to the greatest efficiency.

Table 6.1 shows the estimated sample sizes for the eight dimensions of the SF-36 using Methods 1 to 4. The smallest estimates of the required sample come from Method 1 (n_{Normal}). All of the effect size estimates for the eight dimensions are in the 'small' to 'moderate' range using Cohen's (1988) classification and are consistent with the empirical estimates observed in Chapter 5 (Tables 5.1 to 5,17).

Figure 6.2 shows the estimated power curves for the SF-36 GH dimension for the four methods for a range of sample sizes. The curves for Method 1 and Method 2 are broadly similar. On the scale shown the curve for the Method 3 does not even reach the 80% power level and so large sample sizes are required (over 2500 subjects per group (see Table 6.1). The GH dimension of the SF-36 has over 30 discrete categories (most of which are occupied in the control population (Figure 6.1g) so it would seem sensible to regard this scale as either ordinal or continuous. Therefore for the bootstrap simulation for the GH dimension we will ignore Method 3.

Method 5:– computer simulation – Bootstrap methods

Figure 6.1g shows the skewed distribution of the GH dimension and that the underlying assumption of Normality of the distribution required for Method 1 may not be appropriate. Furthermore the, GH dimension is bounded by 0 and 100. Thus, if a new mother already has a GH score of 100 in the control group, then under the intervention no extra improvement can be seen, at least not by the GH dimension of the SF-36. Seven percent of women (35/487) in the Sheffield data had a GH score of 100 and 14.2% (70/487) had a score of 95 or more.

Individual improvements of five points on the GH dimension should correspond to a lower difference on average. A mean difference of 5 implies an average improvement ascribable to the intervention of only 4.5 on the GH scale.

Figure 6.2: Power curves for Methods 1 to 4 for estimating sample sizes for CPSW study with the SF-36 general health dimension as the primary outcome

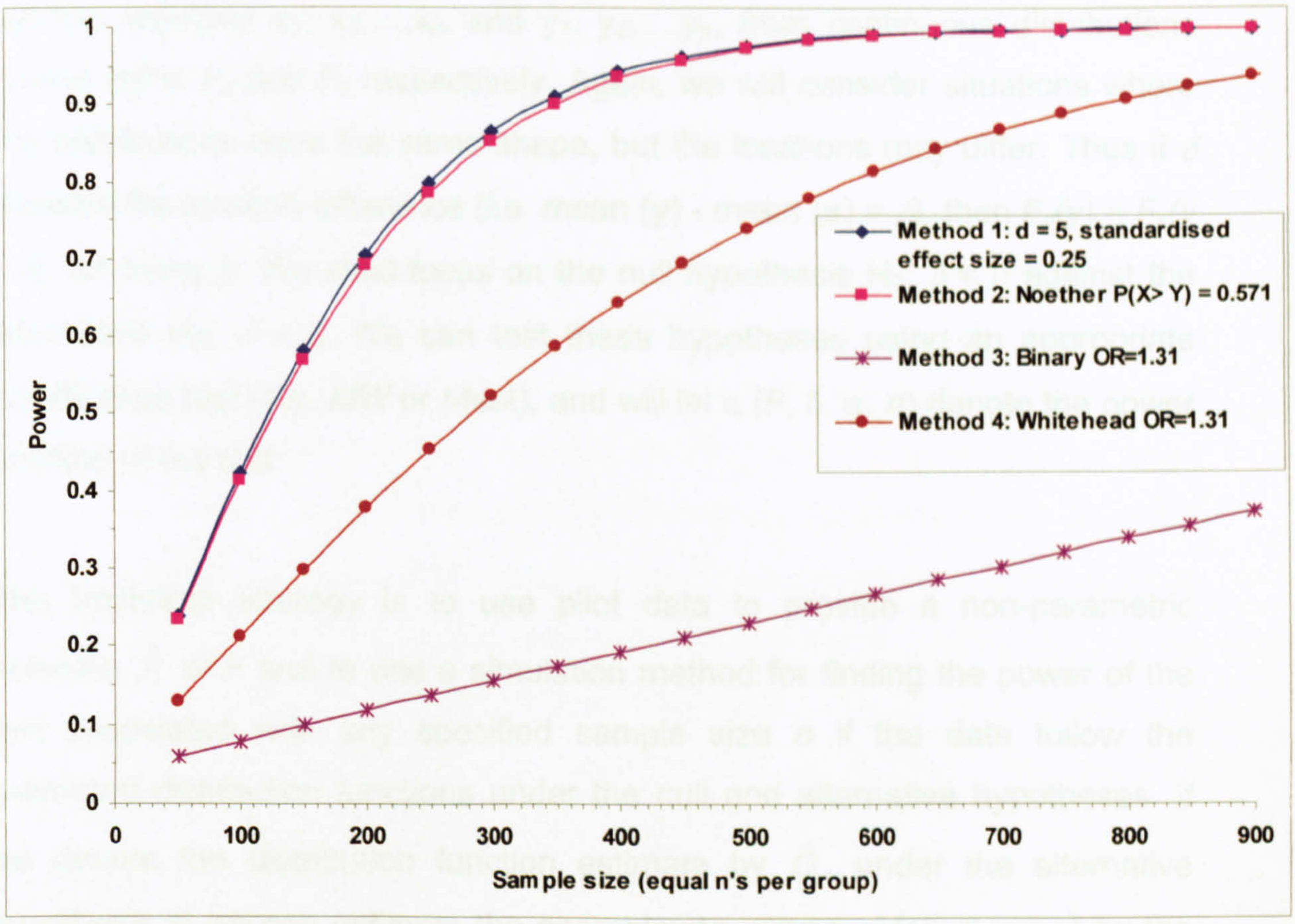


Table 6.1 shows that for other dimensions of the SF-36 the proportion scoring 100 (P_{100}) in the reference/control group is greater than 50% for four out of the eight dimensions (PF, RP, SF and RE).

Methods 1 and 2 assume the outcome is continuous and the simple location shift model is appropriate. Here this would imply that, on a certain scale, the difference in effect of the intervention compared to the control is constant or, at least that the intervention shifts the distribution of the GH scores under the control to the right (or to the left if the intervention is harmful) thereby keeping its shape. However, the boundedness of the SF-36 outcomes renders this simple location shift assumption questionable, especially if the proportion of cases at each bound is high.

We used bootstrap methods to compare the power of the *t*-test and *MW* test with allowance for ties for detecting a shift in location using the SF-36 GH dimension as an outcome. Suppose (as before) we have two independent random samples $x_1, x_2,\ldots,x_m$ and $y_1, y_2,\ldots,y_n$, from continuous distributions having cdf's, F_x and F_y respectively. Again, we will consider situations where the distributions have the same shape, but the locations may differ. Thus if δ denotes the location difference (i.e. mean (**y**) - mean (**x**) = δ), then $F_y(y) = F_x(y - \delta)$, for every *y*. We shall focus on the null hypothesis H_0: $\delta = 0$ against the alternative H_A: $\delta \neq 0$. We can test these hypotheses using an appropriate significance test (e.g. *MW* or *t*-test), and will let π (F, δ, α, *n*) denote the power function of the test.

The bootstrap strategy is to use pilot data to provide a non-parametric estimate $\hat{F}$ of *F* and to use a simulation method for finding the power of the test associated with any specified sample size *n* if the data follow the estimated distribution functions under the null and alternative hypotheses. If we denote the distribution function estimate by $\hat{G}$, under the alternative hypothesis δ, we can estimate the approximate power, $\hat{\pi}(G, \delta, \alpha, n)$ by the following computer simulation procedure.

Algorithm 6.1

Power and sample size estimation using the bootstrap

(1) Draw a random sample with replacement of size 2*n* from $\hat{F}$. The first *n* observations in the sample form a simulated sample of **x**'s, denoted by $x_1^*,\ldots,x_n^*$, with estimated cdf $\hat{F}^*$. Then δ is added to each of the other *n* observations in the sample to form the simulated sample of **y**'s, denoted by $y_1^*,\ldots,y_n^*$, with estimated cdf $\hat{G}^*$. (The y*'s and x*'s have been generated from the same distribution except that the distribution of the y*'s is shifted δ units to the right.)

(2) The test statistic *t(**x***, ***y***)*, (Mann-Whitney or *t*-test) is calculated for the **x***'s and **y***'s, yielding *t*(**x***, **y***). If *t*(**x***, **y***) $\geq T_{1-\alpha/2}$, (where $T_{1-\alpha/2}$ is the

critical value of the test statistic) a success is recorded; otherwise a failure is recorded.

(3) Steps 1 and 2 are repeated *B* times. The estimated power of the test, $\hat{\pi}(G,\delta,\alpha,n)$, is approximated by the proportion of successes among the *B* repetitions. (In all cases discussed in this paper, *B* = 10,000).

The above bootstrap procedure assumes a simple constant location shift model. For bounded outcome scores the procedure is in principle the same but more imagination is needed to specify the effect of the new treatment in comparison with the control treatment. Under the simple location shift model, individual improvement of δ is assumed and this is equal to the improvement of the population scale measured by a location parameter. For bounded outcome scores we have to assume an effect δ(x) such that x + δ(x) remains in the interval determined by the lower and upper boundary. In the case of the SF-36 GH dimension between 0 and 100. One function is to assume a constant treatment effect whenever possible i.e. $\delta_1(x)$. For $\delta_1(x)$, we assumed a constant additional effect of 5 points, until a GH score of 95. Patients with a GH score of 95 or more were truncated at 100. This is shown in Figure 6.3.

Draw a random sample with replacement of size 2*n* from $\hat{F}$. The first *n* observations in the sample form a simulated sample of x's, denoted by $x_1^*,\ldots,x_n^*$. Then $\delta_1(x)$, is added to each of the other *n* observations in the sample to form the simulated sample of y's, denoted by $y_1^*,\ldots,y_n^*$. The y*'s and x*'s have been generated from the same distribution except that the distribution of the y*'s is shifted $\delta_1(x)$, units to the right.

The software Resampling Stats was used for the bootstrapping (Simon, 2000). Two examples of the Resampling Stats programs for carrying out the bootstrapping are listed in Appendix 4. The bootstrap computer simulation procedure involved using SF-36 data from a general population survey based on 487 women aged 16-74 as the pilot dataset. Figure 6.1g shows the non-symmetric distribution of the GH dimension.

Figure 6.3: Effect of treatment with bounded outcomes. d1(x) represents a constant treatment effect of 5, until 95.

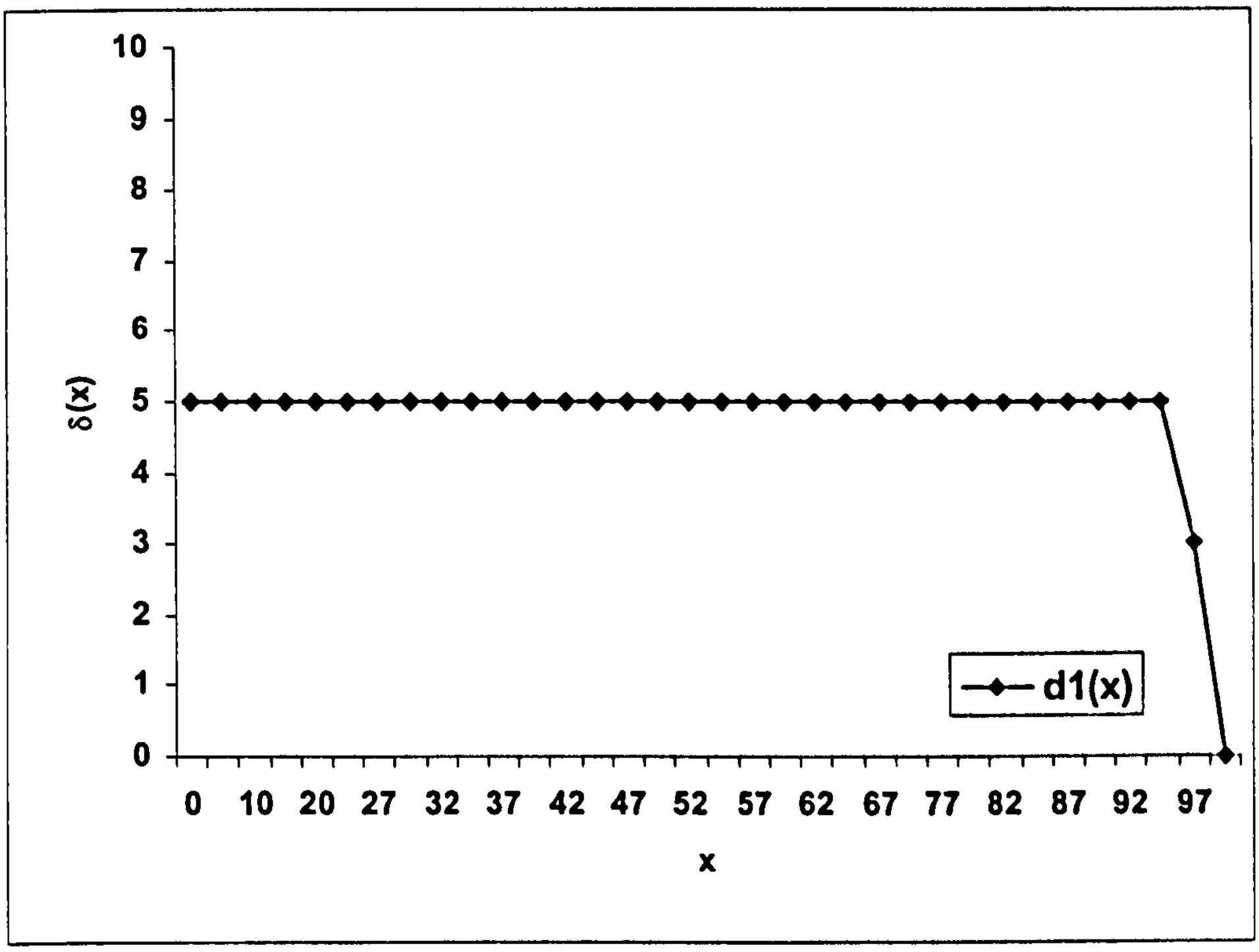


Figure 6.4 shows the estimated power curves for Methods 1, 2, 4 and 5 (the two bootstrap methods, *t* and *MW* tests) at the 5% two-sided significance level for detecting a location shift (mean difference) $\delta = 5$ in the SF-36 GH dimension using the data from the general population as our pilot sample, for sample sizes per group varying from 50 to 600. For these general population data a location shift of $\delta = 5$ is equivalent to a standardised effect size $\Delta_{Normal} = 0.25$ and a parametric estimate of $p_{Noether} = Pr(X > Y) = 0.57$. The bootstrap methods taking into account the bounded and non-normal distribution of the data suggest an observed mean difference *d* of approximately 4.5 and an observed $p_{Noether} = Pr(X > Y) = 0.58$. (These values are calculated from the average observed effect sizes across the bootstrap replications.)

The power curve for Method 4 is based on an $OR_{Ordinal} = 1.63$ effect size, which with 199 subjects per group (the estimated sample size to have an 80% chance of detecting this $OR_{Ordinal}$ effect size as statistically significant at the 5% two-sided level) and proportions in the control group as in Figure 6.1g leads to estimated group means of 75 and 80 points in the control and

intervention groups respectively. This is equivalent to an observed mean difference between groups *d* of approximately five points on the GH scale.

The GH dimension (Figure 6.1g) of the SF-36 has a large number (> 30) of discrete values or categories, most of which are occupied, and the proportion of scoring 0 or 100 is low. So the power curves shown in Figure 6.4, do not diverge too greatly and thus, the simple location shift hypothesis is a useful working model.

However, Figure 6.5 shows the estimated power curves for the RP dimension of the SF-36, which can only take one of five discrete values (Figure 6.1c), for detecting a location shift (mean difference) $\delta = 5$. For these data a location shift of $\delta = 5$ is equivalent to a standardised effect size $\Delta_{Normal} = 0.17$ and $p_{Noether} = Pr(X > Y) = 0.55$. Since three-quarters of the pilot sample scored 100, the bootstrap methods under the location shift model, taking into account the bounded and non-Normal distribution of the data suggest an observed mean difference *d* of 1.2 and p = Pr(X > Y) = 0.51. So the power curves shown in Figure 6.5 diverge greatly and the simple location shift model alternative hypothesis may be inappropriate for this outcome.

Figures 6.6 to 6.11 show the power curves for the other six dimensions of the SF-36. Again the curves diverge for four out of six dimensions (BP, PF, SF and RE), where the proportions at the upper boundary scoring 100 is more than 40%, but are reasonably close for the two other dimensions (V and MH), where the proportions scoring 100 are small (less than 1%)

Figure 6.4: Estimated power curves for the SF-36 General Health dimension using general population data (females aged 16-45)

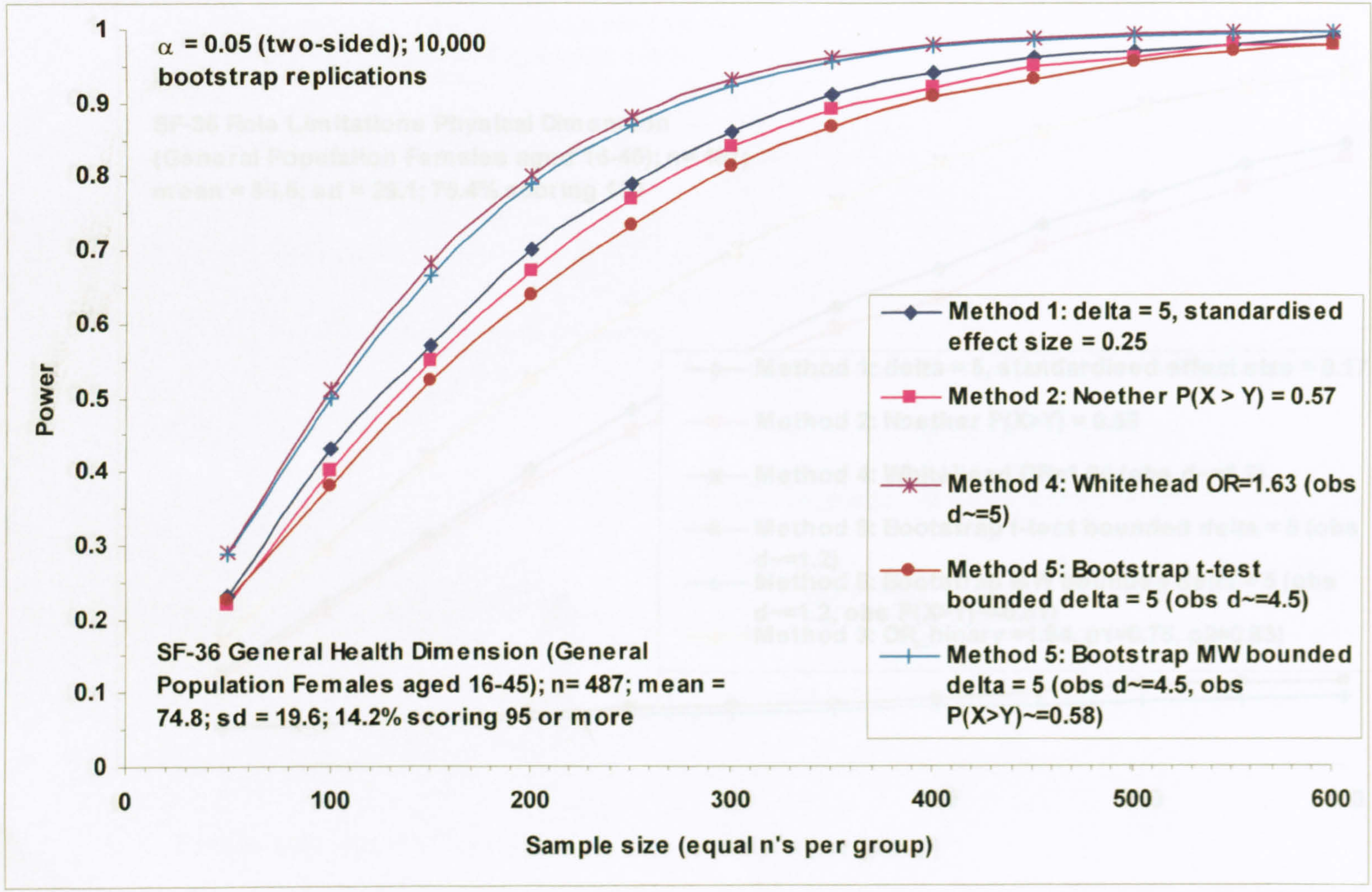


Figure 6.5: Estimated power curves for the SF-36 Role Physical dimension using general population data (females aged 16-45)

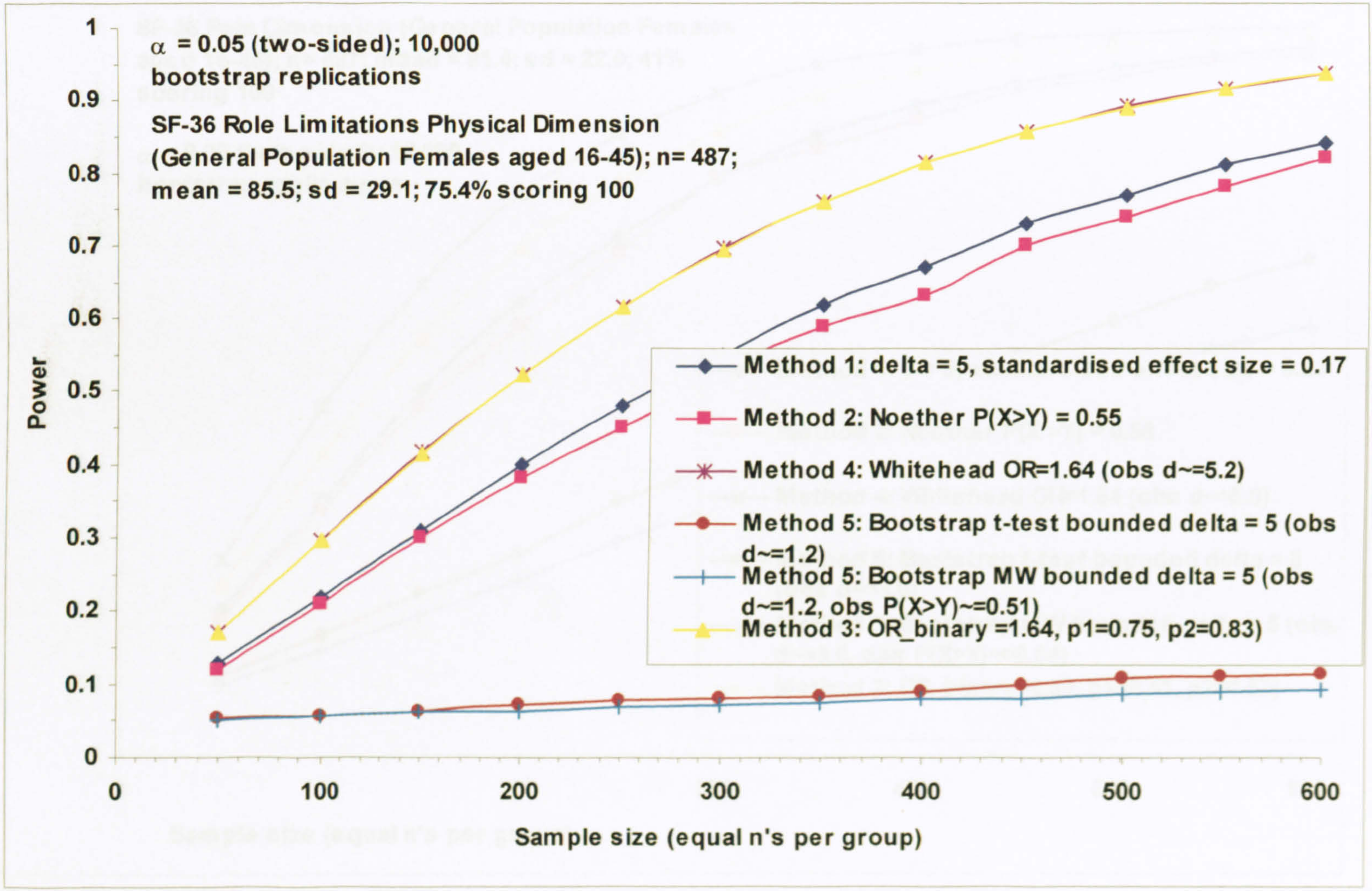


Figure 6.6: Estimated power curves for the SF-36 Bodily Pain dimension using general population data (females aged 16-45)

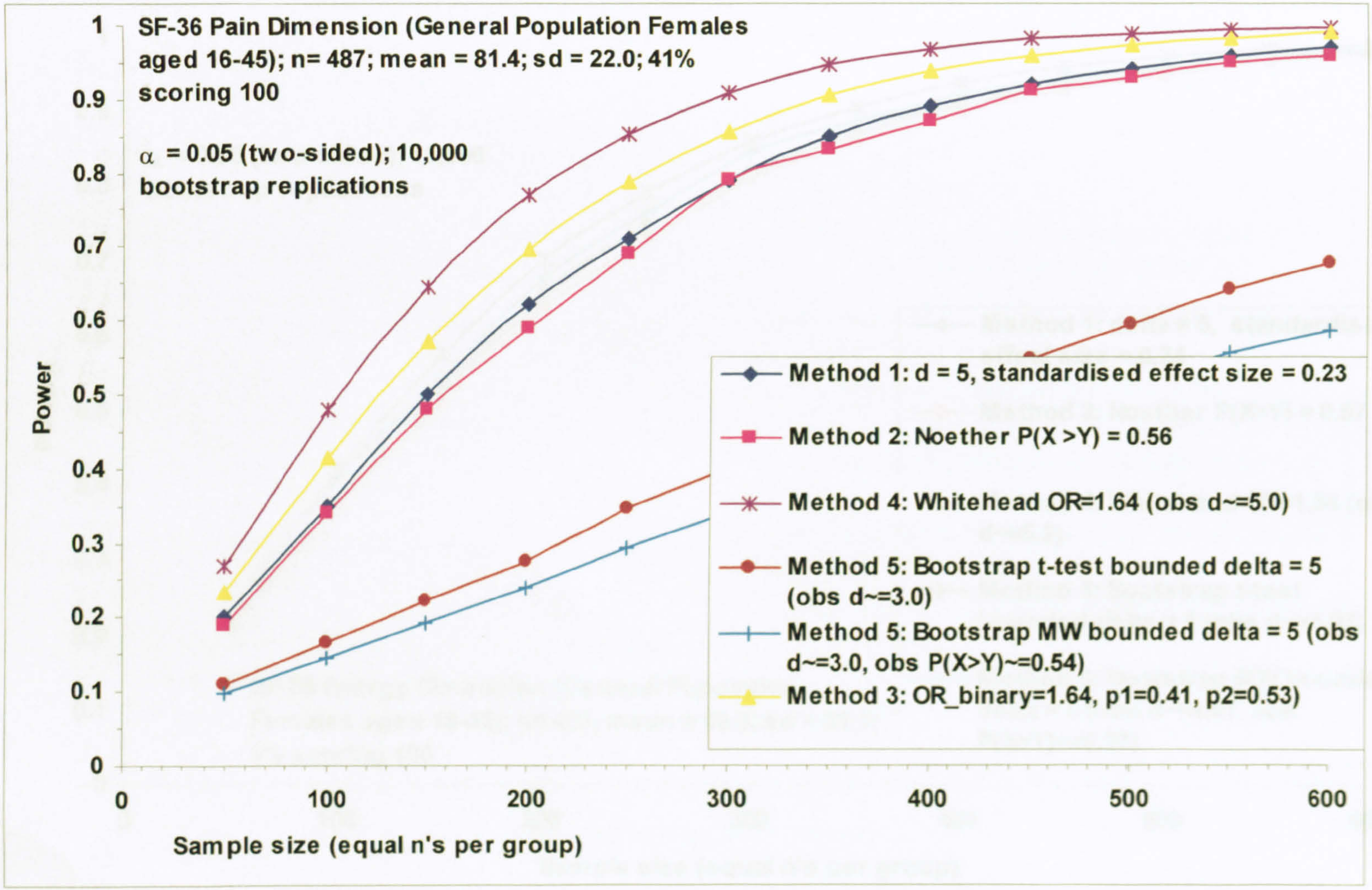


Figure 6.7: Estimated power curves for the SF-36 Vitality dimension using general population data (females aged 16-45)

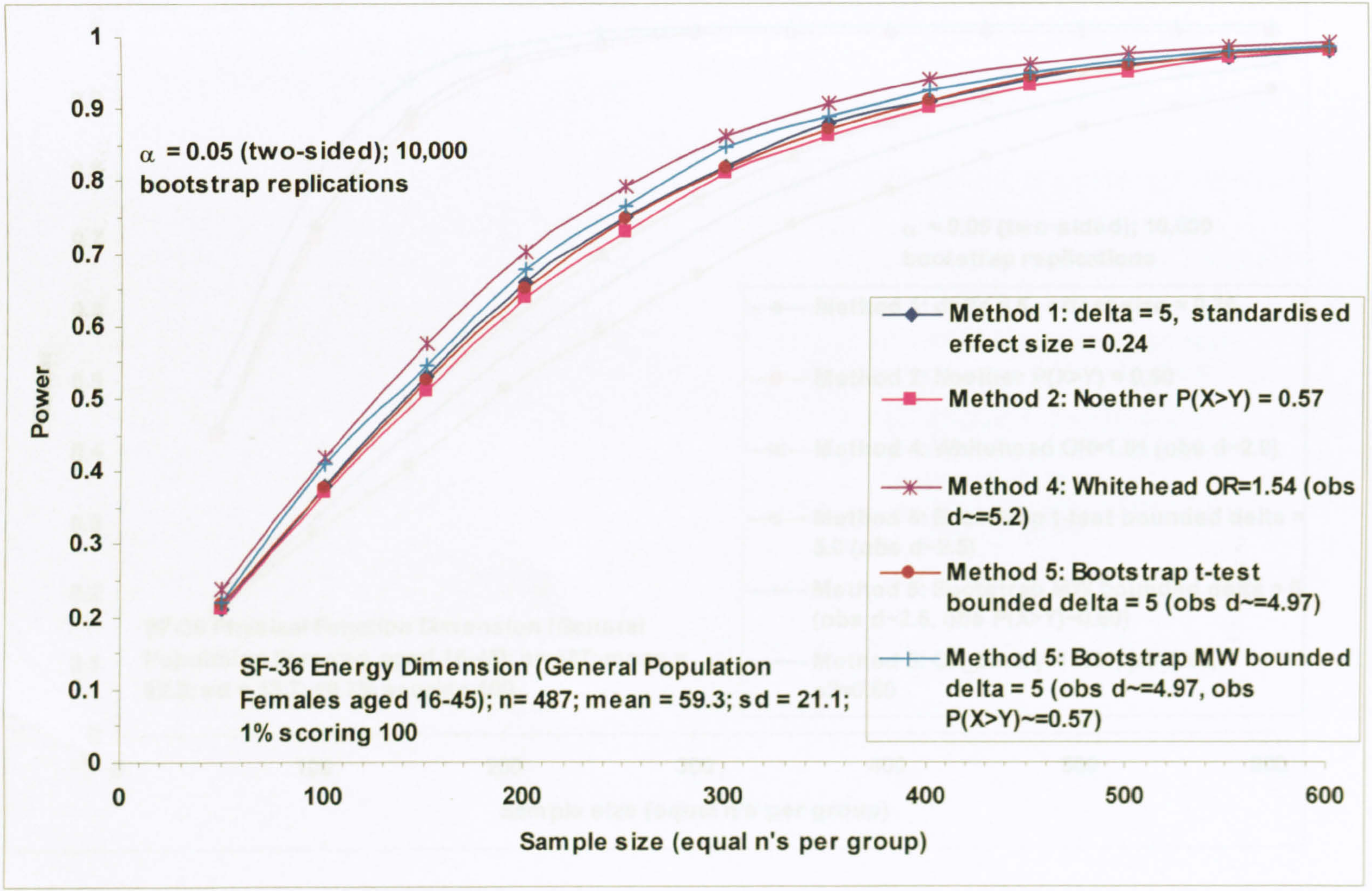


Figure 6.8: Estimated power curves for the SF-36 Physical Function dimension using general population data (females aged 16-45)

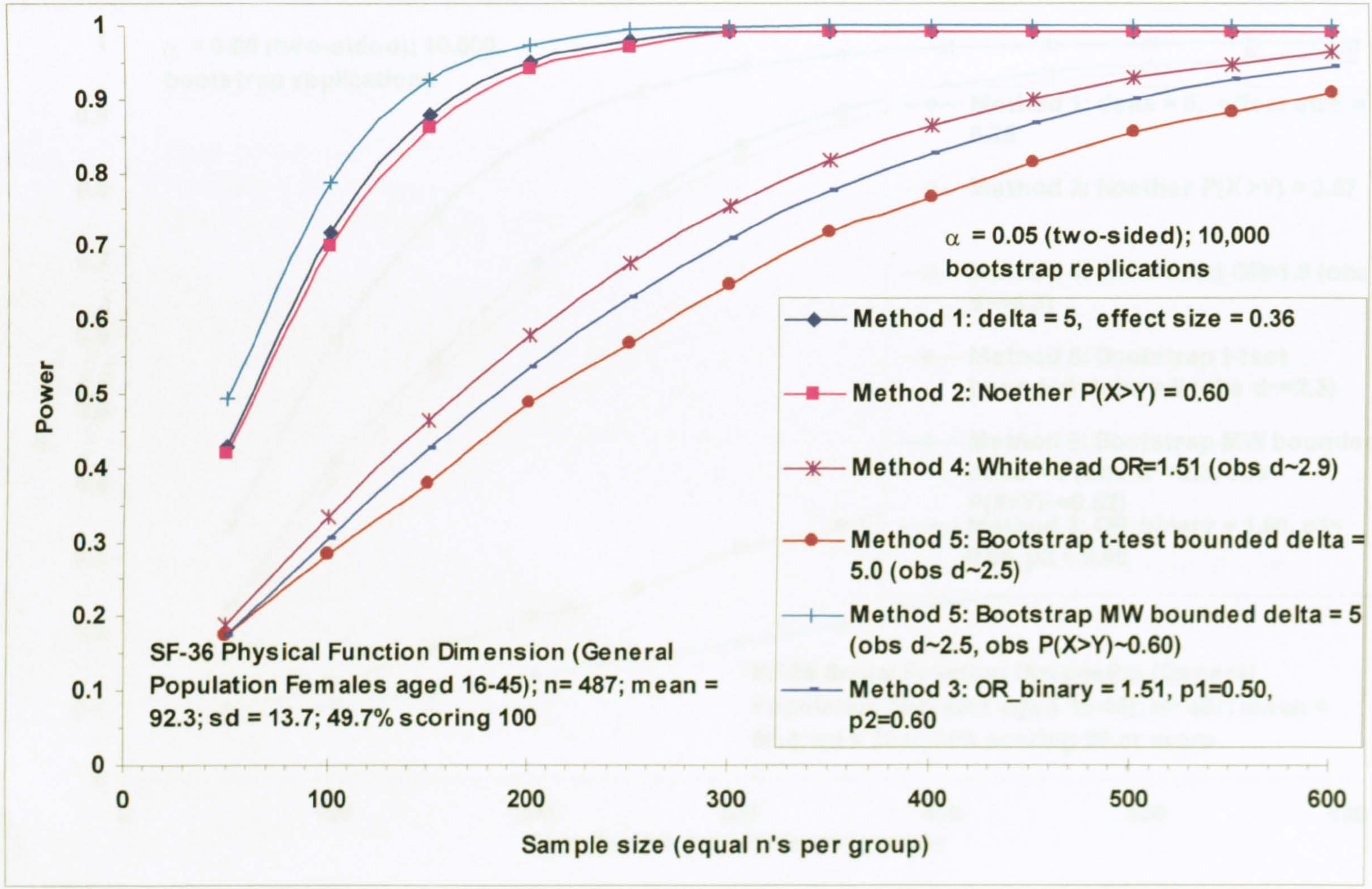


Figure 6.9: Estimated power curves for the SF-36 Social Function dimension using general population data (females aged 16-45)

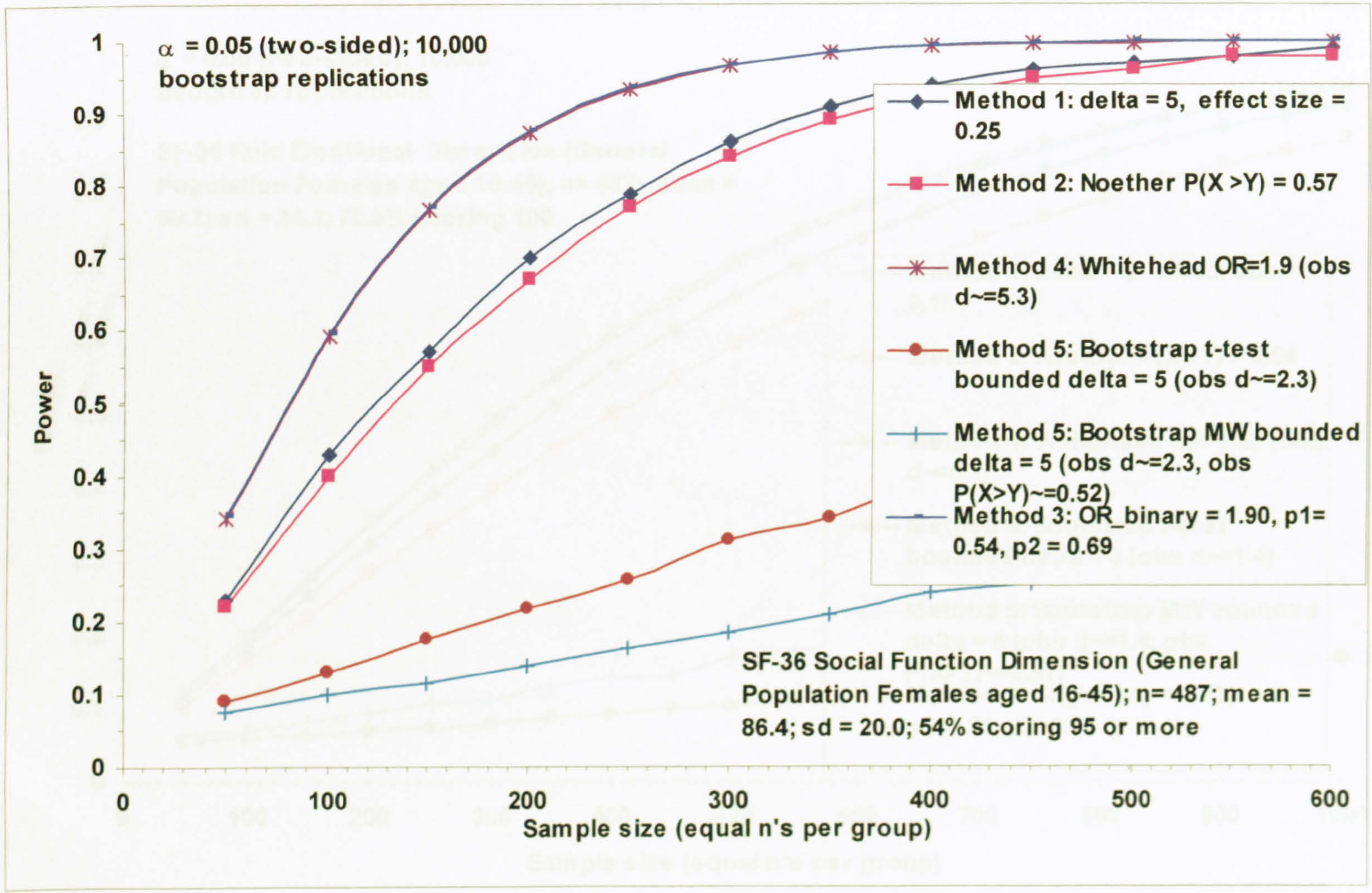


Figure 6.10: Estimated power curves for the SF-36 Role Emotional dimension using general population data (females aged 16-45)

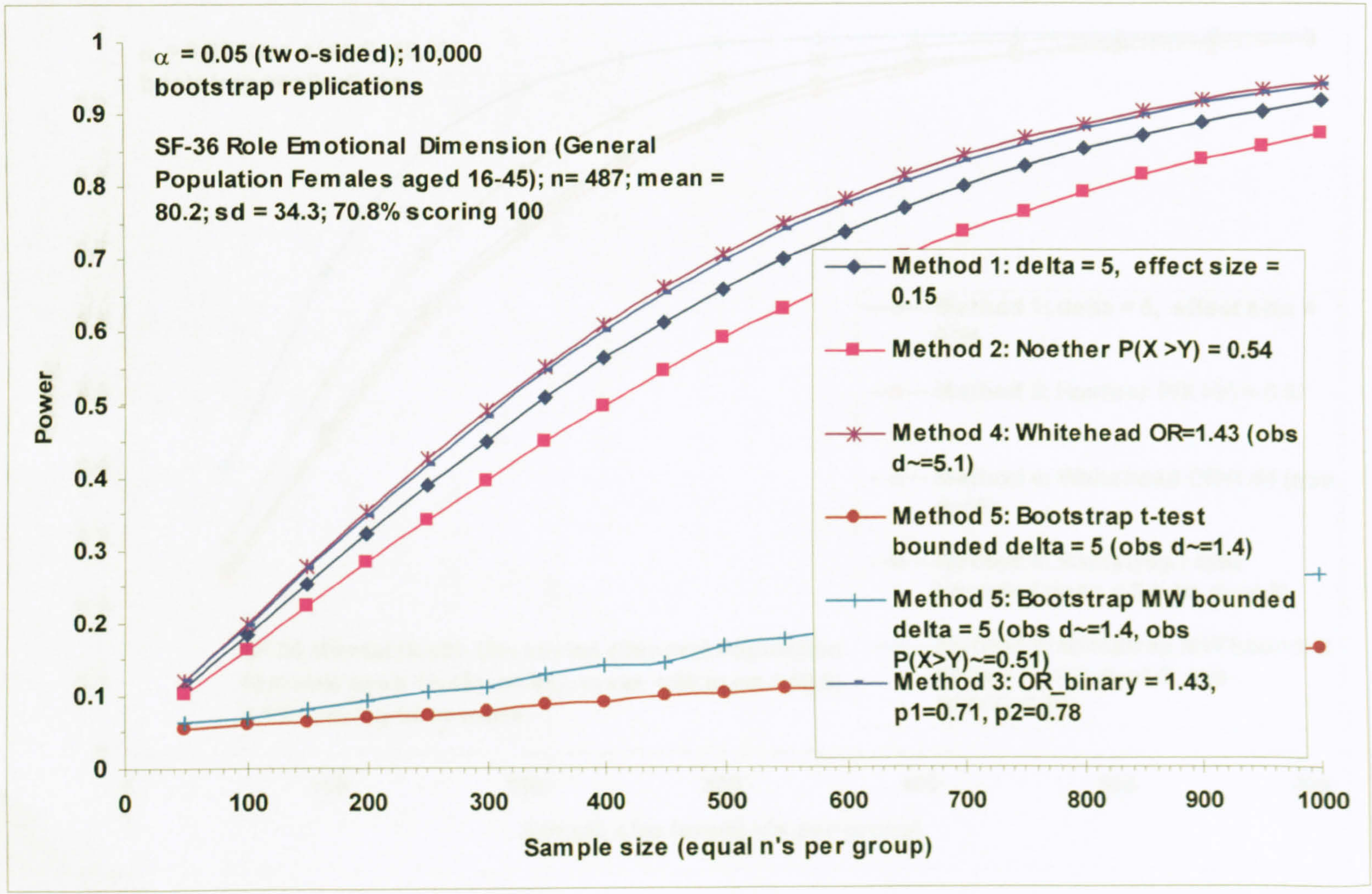
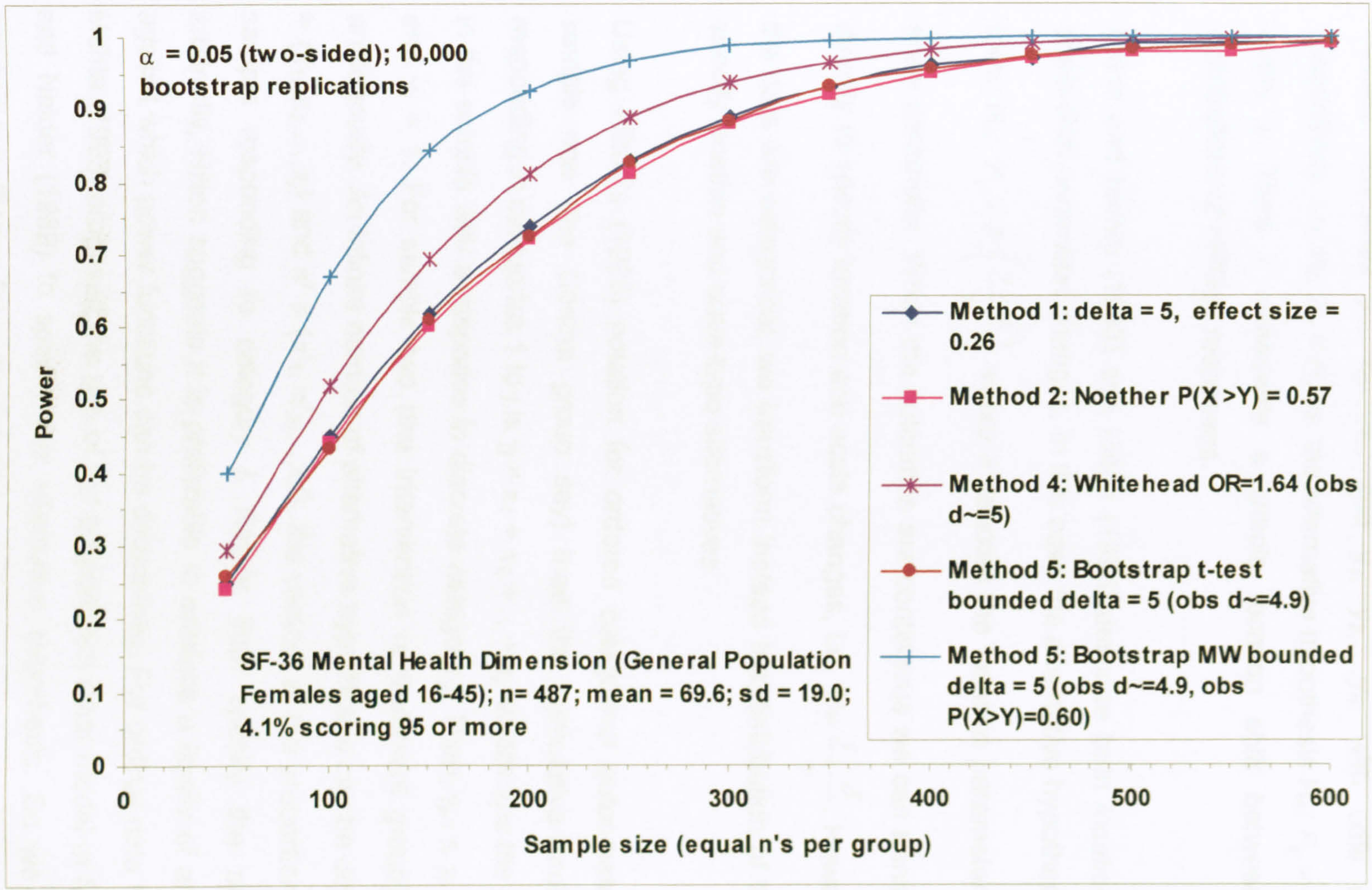


Figure 6.11: Estimated power curves for the SF-36 Mental Health dimension using general population data (females aged 16-45)



Location and scale changes

We have considered a simple *location* shift alternative hypothesis for estimating sample sizes for comparing the distributions of two groups of HRQoL outcomes $x_1, x_2,\ldots,x_m$ and $y_1, y_2,\ldots,y_n$, with cdfs F_x and F_y respectively. I.e. H_0: $F_x = F_y$ vs. the alternative hypothesis H_A: $F_y = F_x(y - \delta)$ for every y. Here δ represents a simple location shift between the two distributions of HRQoL responses.

Hilton and Mehta (1993) and Hilton (1996) describe both *location* shift and *scale* shift parameter changes. In this case the alternative hypothesis is of the form, H_A: $F_y = F_x\left(\frac{y-\delta}{\tau}\right)$ where δ denotes the location parameter and τ the scale parameter. When the outcomes are continuous we can transform them directly to specify location and scale changes, i.e. $y' = \frac{y-\delta}{\tau}$. However, when the data are categorical, we transform instead the distribution of the data to specify location and scale-type alternatives

Using Hilton's (1996) notation for ordered categorical outcomes, then for sample one (the Control group say) then the cumulative probability of responding in categories 1 to j is $\gamma_j = \pi_1 + \pi_2 + \ldots + \pi_j$, where π_j is the proportion in the sample with a response in discrete category j. Then, $\gamma_{j-1} \leq \gamma_j$, $j = 1,\ldots,k$ and $\gamma_k = 1$. For sample two (the Intervention or treatment group) define γ' analogously. An infinite number of alternative hypotheses can be defined by $\pi = (\pi_1, \pi_2,\ldots,\pi_j)$ and $\pi' = (\pi'_1, \pi'_2,\ldots,\pi_j')$, the vectors of the proportions in each sample responding to category j. Rather than specify the parameters arbitrarily, Hilton suggests it is preferable to produce a family of alternatives against which power functions can be determined. For ordinal data Hilton and Mehta (1993) proposed the use of the proportional odds model of McCullagh and Nelder (1989) to specify this alternative hypothesis. So we can now express hypotheses for ordinal data in terms of $\gamma = (\gamma_1, \gamma_2,\ldots,\gamma_{k-1})$ and of γ' =

(γ'_1, γ'_2,..., γ'_{k-1}). The null hypothesis is H_0: $\gamma'_j = \gamma_j$ for all *j* and an alternative hypothesis is H_A: $\gamma'_j \geq \gamma_j$ or $\gamma'_j \leq \gamma_j$, for all *j*, with inequality for at least one *j*.

The proportional odds model is:

$$\text{logit}(\gamma'_j) = \text{logit}(\gamma_j) - \Delta, \qquad j = 1,\ldots,k\text{-}1, \qquad (6.4)$$

where $\delta \in (-\infty, \infty)$ with $\Delta = 0$ representing the null case. It expresses the difference between the two distributions (γ and γ') in terms of a single 'location-type' parameter Δ. Here Δ is the log odds ratio that compares the two samples' odds of responses in categories 1 to *j* (versus *j+1* to *k*). (It should not be confused with the effect size Δ_{Normal}). Hilton (1996) extends this model to also allow differences in scale:

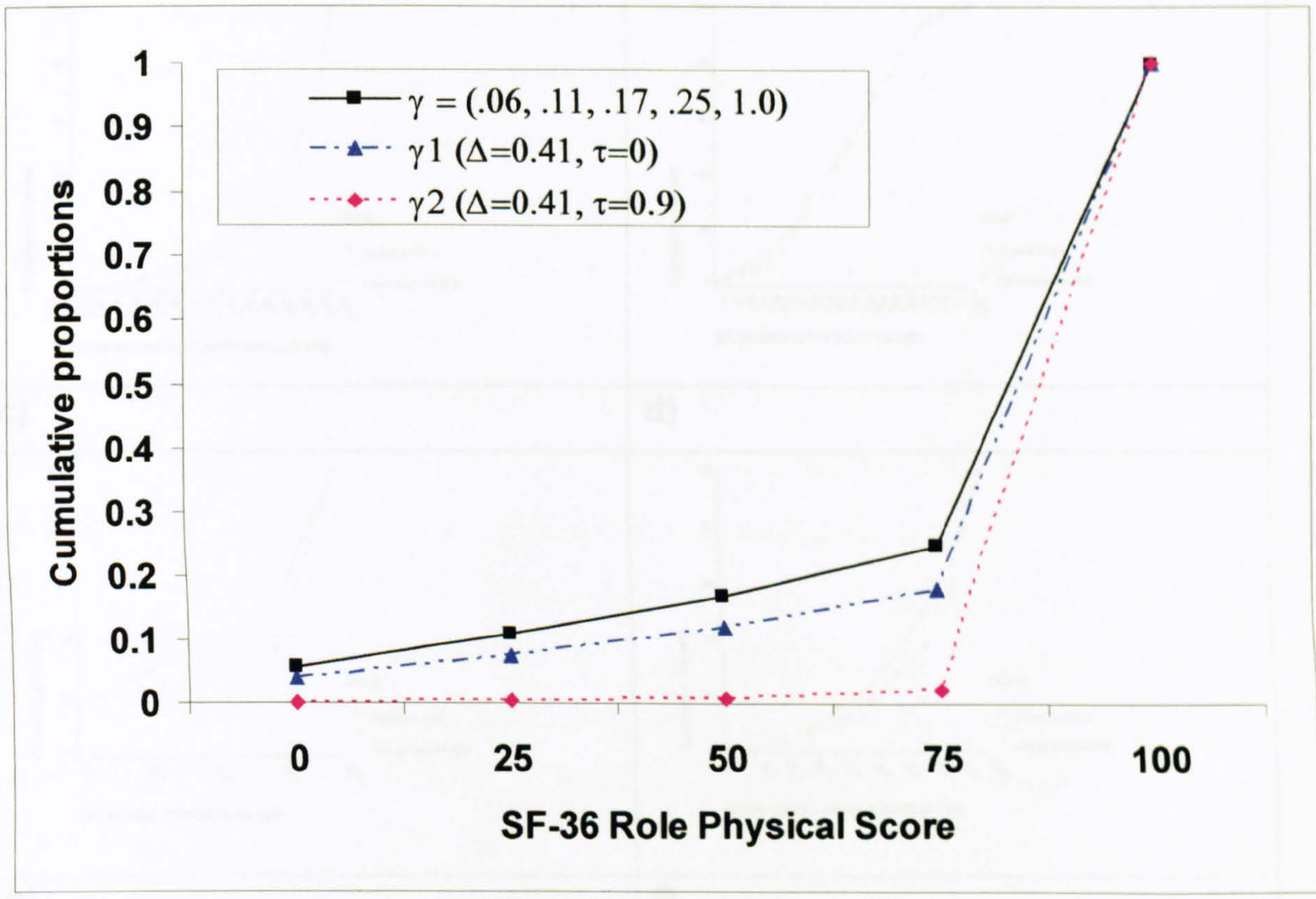
$$\text{logit}(\gamma'_j) = \frac{\text{logit}(\gamma_j) - \Delta}{\exp(-\tau)} \qquad j = 1,\ldots, k\text{-}1, \qquad (6.5)$$

where $\tau \in (-\infty, \infty)$ with $\Delta = 0$ and $\tau = 0$ under H_0.

Hilton (1996) shows how the effects of the scale and location parameters τ and Δ can be illustrated graphically by plotting the cumulative distribution functions of γ and γ' respectively. Figure 6.12 shows three cumulative probability distributions for the five category SF-36 RP dimension, the reference population is the cumulative distribution for females aged 16-45 from the Sheffield population with π = (0.06, 0.05, 0.06, 0.08, 0.75) and λ = (0.6, 0.11, 0.17, 0.25, 1.0). The two other curves represent the effect of a location shift alone (γ1), with an odds ratio of 1.5 (or log odds ratio Δ = 0.41) and both a scale and location shift (γ2) with scale and location parameters of Δ = 0.41 and τ = 0.9 respectively.

In the case of scale changes, τ tends to make γ' cross γ, so that neither $\gamma'_j \geq \gamma_j$ for all *j* nor $\gamma'_j \leq \gamma_j$ for all *j* holds (Hilton, 1996). So in the presence of both scale and location shift changes the cumulative probability distribution function curves will tend to cross. Figure 6.12 illustrates that this is not always the case and that it is difficult to distinguish location changes from scale and location effects combined.

Figure 6.12: Three cumulative distribution functions for the SF-36 Role Physical dimension under reference (γ), and location (γ1) and scale and location changes (γ2)

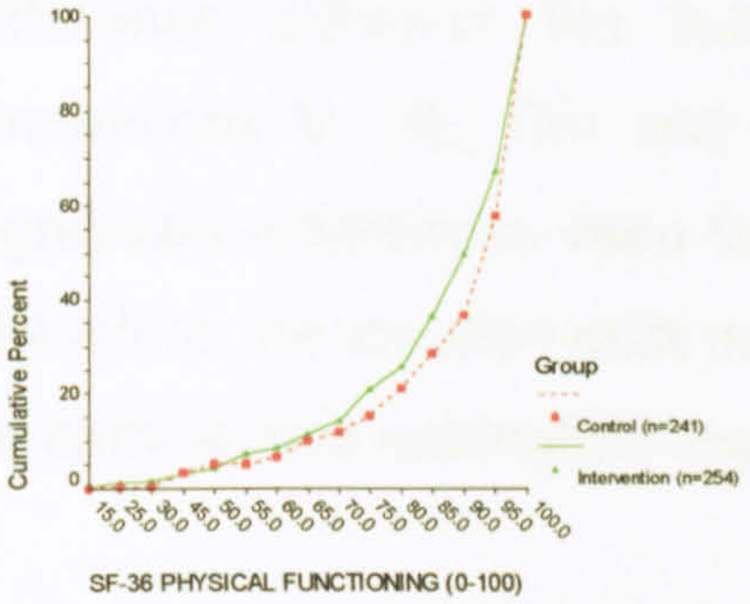


Hilton (1996) also compared the power of three tests, the Mann-Whitney statistic, O'Brien's generalised Wilcoxon statistic (O'Brien, 1988) and the omnibus Smirnov statistic (Smirnov, 1939) for two sample studies with ordinal responses. Hilton found that when location shifts alone were present, the *MW* test had the greatest power of the three tests. As the influence of the scale parameter grew relative to the location parameter, so did the power of the Smirnov and O'Brien test until they exceeded the power of the *MW* test.

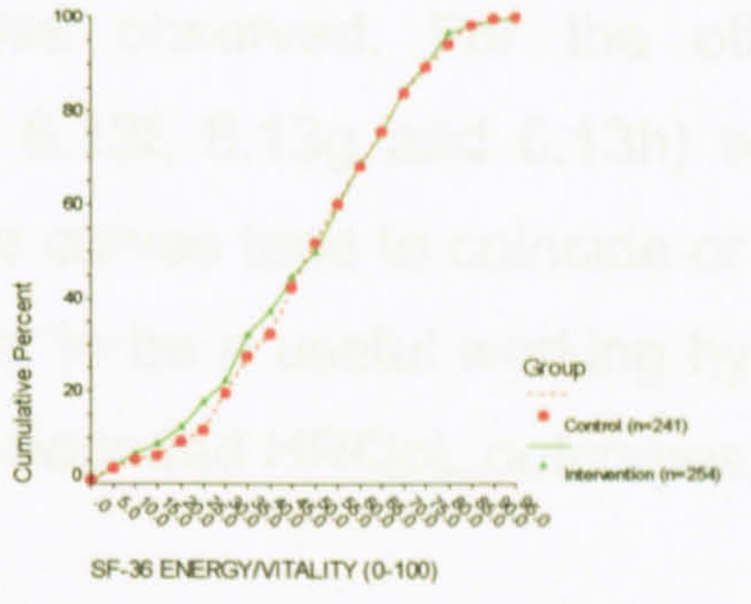
The difference in distributions between groups with HRQoL outcomes may result from a location shift alone or to a combination of location and scale changes. Figure 6.13 which compares the outcomes for the Intervention and Control group mothers from the CPSW study suggests that the difference in HRQoL is predominately a shift in location.

Figure 6.13: Cumulative distribution function graphs for the eight dimensions of the SF-36 for the Control and Intervention Groups

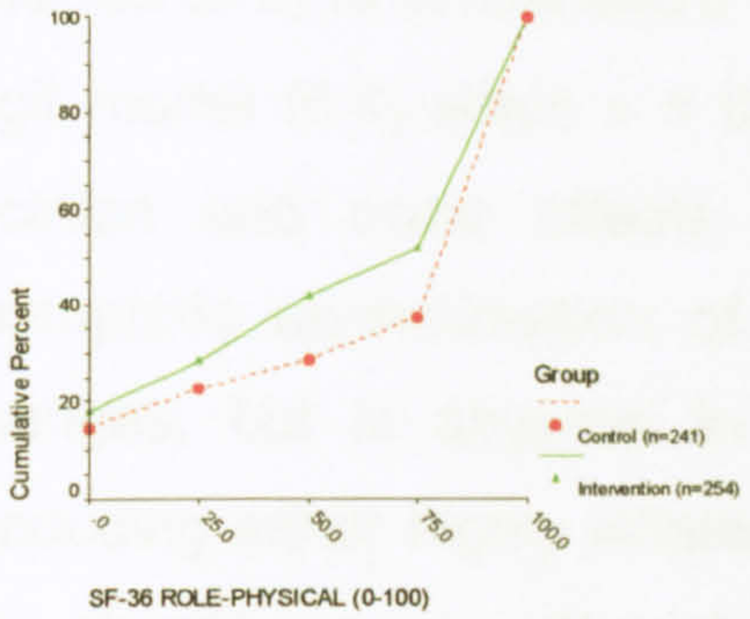
a)



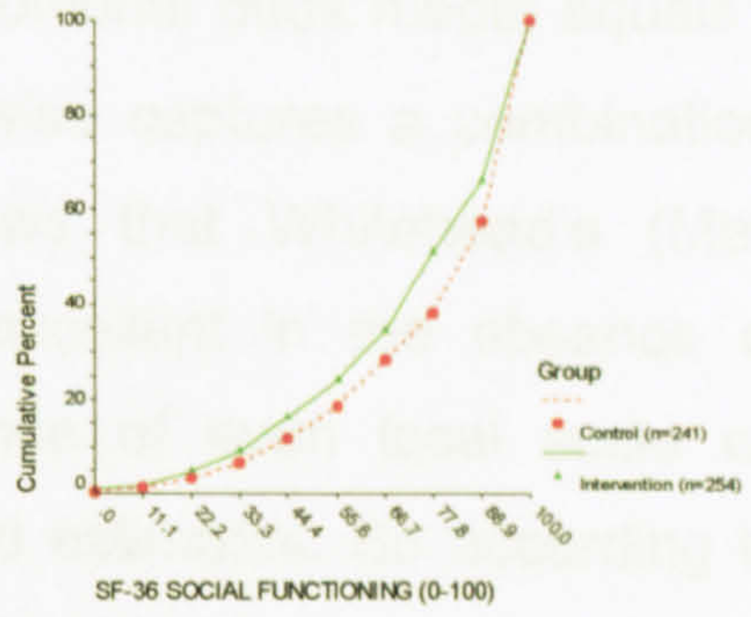
b)



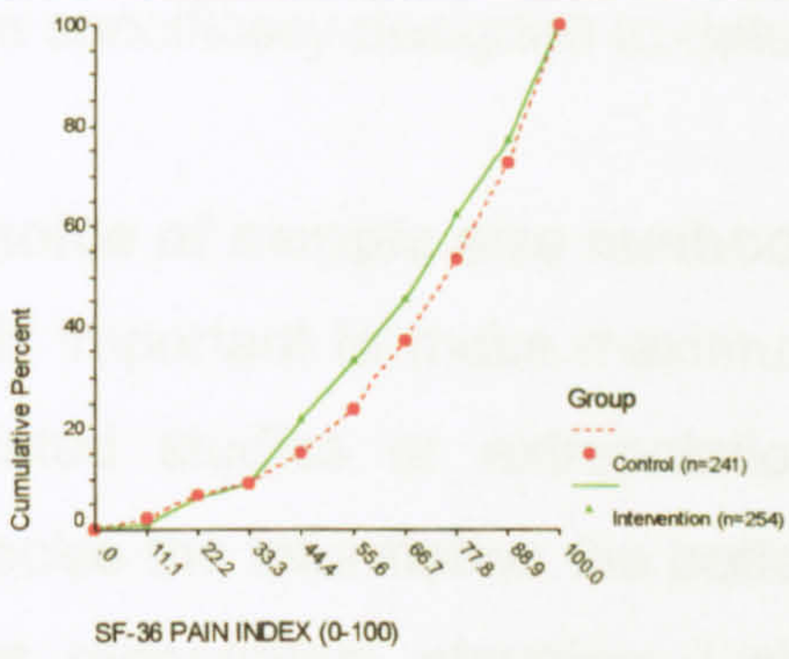
c)



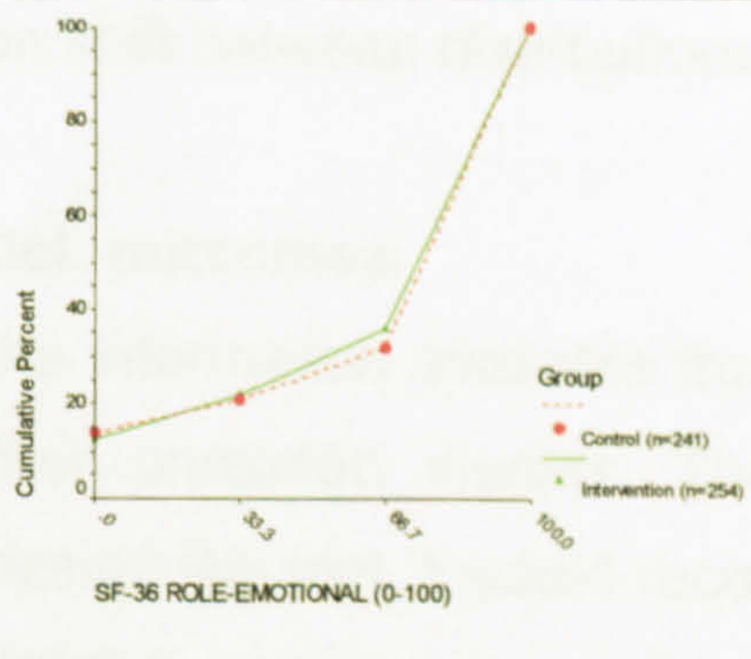
d)



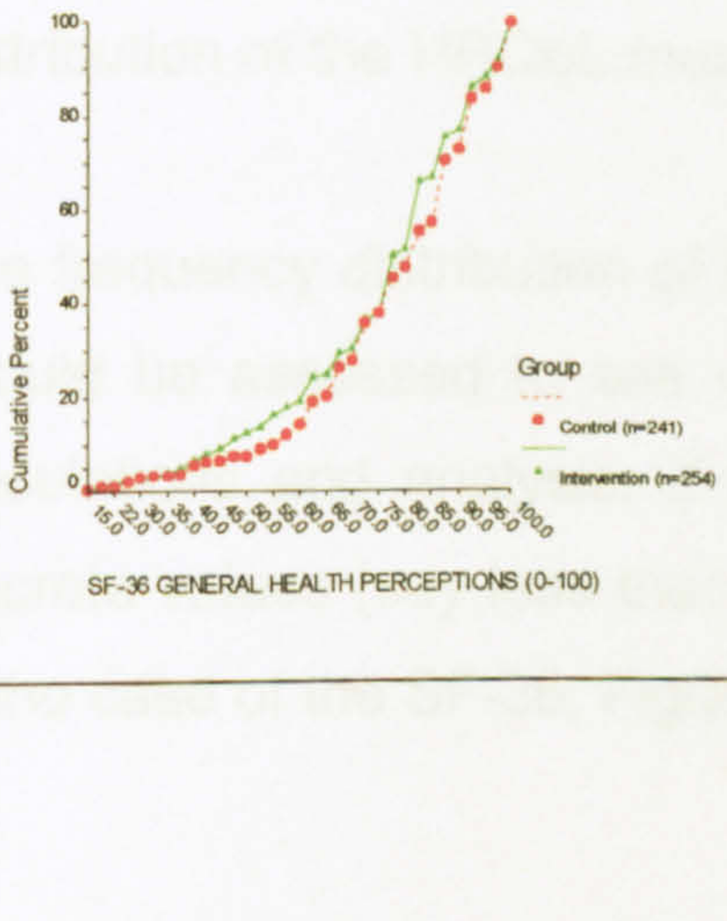
e)



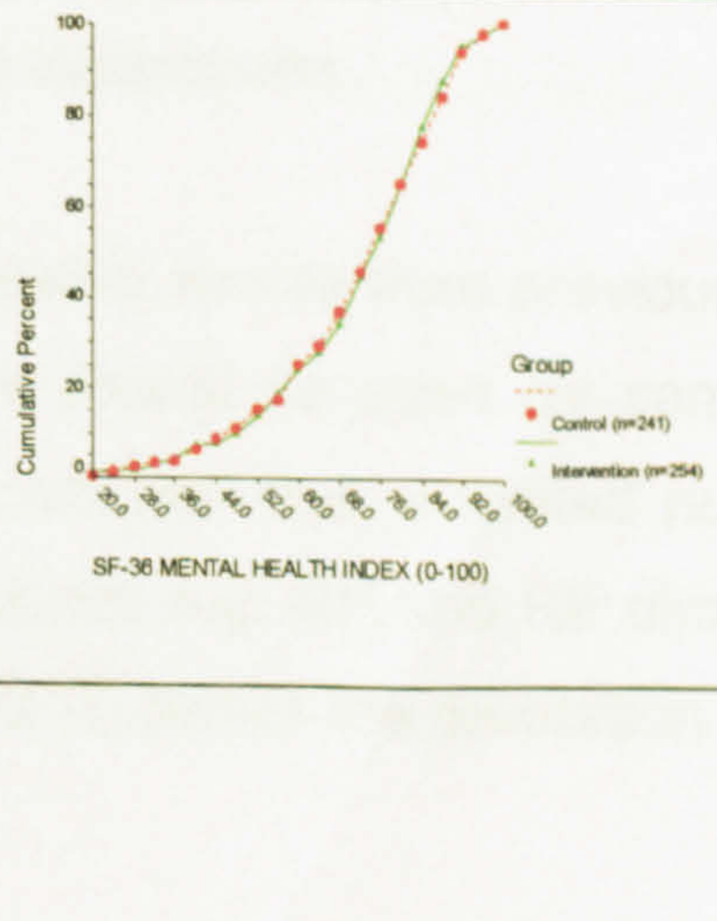
f)



g)



h)



at the upper bound (i.e. scoring 100) is high (e.g. PF, SF and BP dimensions in our general population sample example dataset, Figures 6.1a, 6.1d and 6.1e), then we would recommend using Whitehead's Method 4 to estimate the required sample size. In this case the alternative hypothesis of a simple location shift model is questionable and the proportional odds model may provide a suitable alternative with such bounded discrete outcomes (Figures 6.4, 6.5 and 6.6), although with larger numbers of ordinal categories it is less likely for the proportional odds assumption to hold.

If the HRQoL outcome has a larger number of discrete values (greater than or equal to seven categories), most of which are occupied and the proportion of cases at the upper or bounds (i.e. scoring 0 or 100, in the case of the SF-36) is low (e.g. MH, VT and GH dimensions in our general population sample example dataset, Figures 6.1e, 6.1f and 6.1h), then the simple location shift model appears to be a useful working hypothesis. We would therefore recommend using Methods (1) or (2) to estimate the required sample size.

Computer simulation (Figure 6.7) has suggested that if the distributions of the HRQoL dimensions are reasonably symmetrical, and the proportion of patients at each bound is low, then under the simple location shift alternative hypothesis, the power curves for the *t*-test and *MW* test tend to coincide and the differences in power for a given sample size *n* are small. Therefore if the distribution of the HRQoL outcomes is symmetrical or expected to be reasonably symmetric and the proportion of patients at the upper or lower bounds is low then pragmatically Method 1 could be used for sample size calculations and analysis. The use of parametric methods for analysis (i.e. *t*-test) also enables the relatively easy estimation of confidence intervals, which is regarded as good statistical practice.

If the distribution of the HRQoL outcome is expected to be skewed then the *MW* test appears to be more powerful at detecting a location shift (difference in means) than the *t*-test. So in these circumstances the *MW* test is preferable to the *t*-test and Method 2 could be used for sample size calculations and

analysis. However, using Method 2 for sample size estimation requires the effect size to be defined in terms of Pr(X > Y), which is difficult to quantify and interpret.

If the HRQoL data have a symmetric distribution the mean and median will tend to coincide so either measure is a suitable summary measure of location. If the HRQoL data have an asymmetric distribution, then conventional statistical advice would suggest that the median is the preferred summary statistic (Altman, 2000). However, a case when the mean and mean difference might be preferred (even for skewed outcome data) as a summary measure is when heath care providers are deciding whether to offer a new treatment or not to its population. The mean (along with the sample size) provides information about the total benefit (and total cost) from treating all patients, which is needed as the basis for health care policy decisions (Thomson and Barber, 2000). We cannot estimate the total benefit (or cost) from the sample median.

If the sample size is "sufficiently large" then the CLT guarantees that the sample means will be approximately Normally distributed (Hogg and Tanis, 1988). So, if the investigator is planning a large study (which is likely to be the case with HRQoL outcomes, where large effects are unlikely) and the sample mean is an appropriate summary measure of the HRQoL outcome, pragmatically there is no need to worry about the distribution of the HRQoL outcome and we can use equation (4.4) to calculate sample sizes. Strictly speaking, the Normal distribution is only the limiting form of the sampling distribution of the sample mean as the sample size *n* increases to infinity, it provides a remarkably good approximation to the sampling distribution even when *n* is small and the distribution of the data is far from Normal. Generally, if *n* is greater than 25, these approximations will be good.

If a reliable pilot or historical dataset of HRQoL data is readily available (to estimate the shape of the distribution) then bootstrap simulation (Method 5) will provide a more accurate and reliable sample size estimate than Methods

1 to 4, as it allows us to check the sensitivity of various assumptions including the treatment effect.

Use of the bootstrap and other issues

Bootstrap simulation has illustrated that under the simple location shift assumption with bounded outcome scores (like the SF-36) then conventional methods of sample size determination (Methods 1 and 2) can underestimate the required sample size (or power). This is because the observed mean difference d between the groups is likely to be less than the hypothesised mean difference δ, as is the observed $p_{Noether}$, due to the bounded nature of the outcome variables.

Thus with bounded HRQoL outcome scores the simple location shift alternative hypothesis becomes increasingly questionable, particularly when the proportion of patients at the upper (or lower) bound of the HRQoL outcome is large. A useful alternative to the location shift model, especially for HRQoL outcomes which tend to have a limited number of discrete values is the proportional odds model.

The bootstrap power and sample size method can easily be extended to apply to other models and statistical tests. The extension is accomplished by using the desired test statistic and critical value in place of T and T_α in step 2 of the Algorithm 6.1 described under Method 5. The bootstrap method has a straightforward extension to studies with three or more groups, in which for example, it could be used to estimate the sample size for a one-way ANOVA or Kruskal-Wallis test.

One can also use bootstrapping to create an adaptive non-parametric test. That is to produce an algorithm for selecting a powerful test procedure and associated sample size to use for the main experiment. For example, one could estimate the power curve for several alternative test statistics and base the sample size calculations on the test statistic having the largest estimated power.

The bootstrap sample size estimator of this chapter was applied in the situation in which the alternative to the null hypothesis was a simple location shift. The bootstrap method, can however, be adapted to alternative hypotheses other than a location shift (such as an odds ratio transformation). For example, consider the alternative $F_y = F^v_x$ for some $v \geq 1$. Then the power and the required sample size are functions of v instead of δ. In this situation, one could alter step 1 of the simulation procedure in Algorithm 6.1 as follows:

Draw a random sample with replacement of n x*'s, from G_x; then draw a sample of n y*'s, from G^v_x.

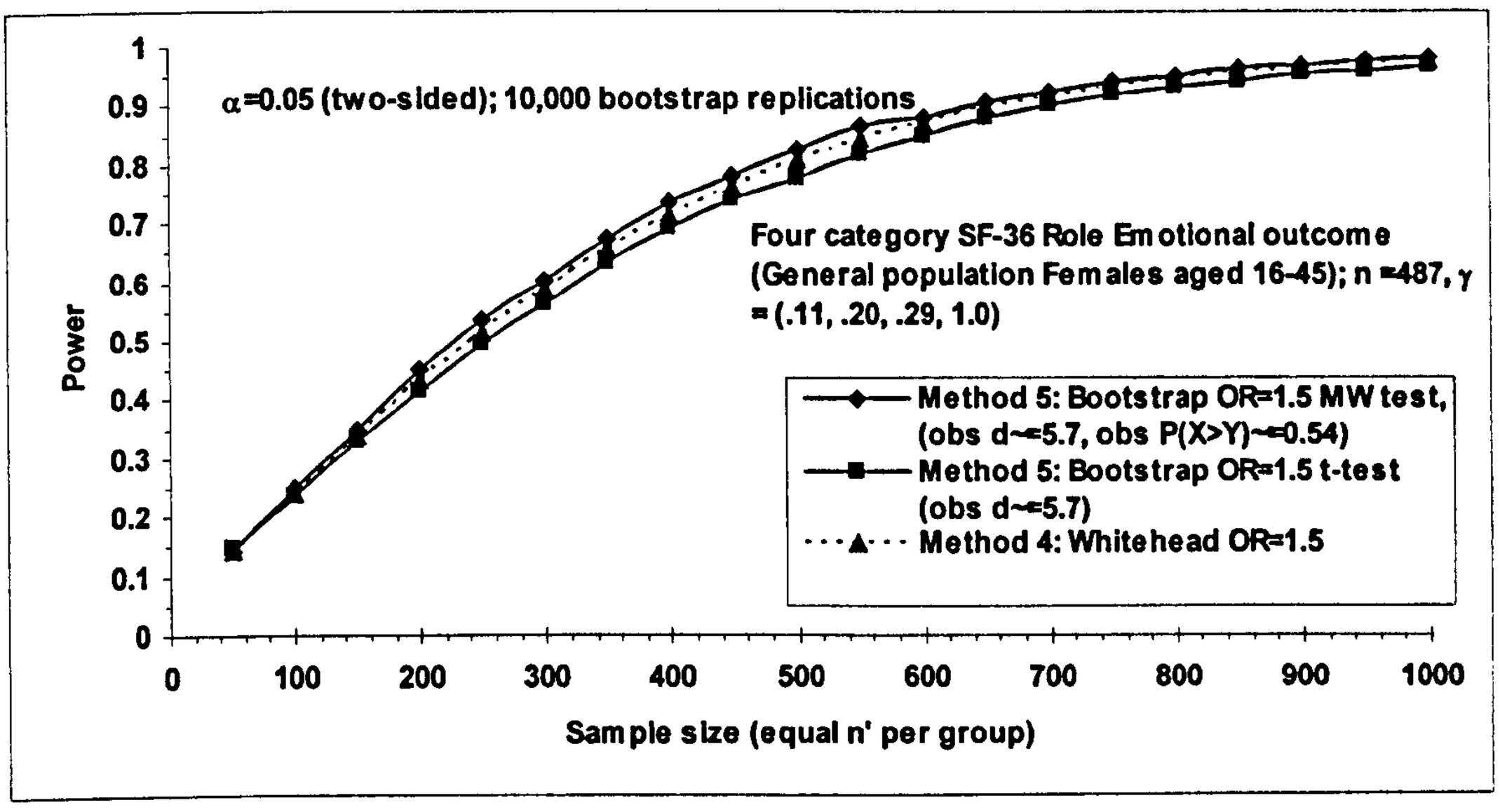
When using the proportional odds model to estimate sample size, Whitehead (1993) and Julious *et al* (1997) have pointed out that there is little increase in power (and hence saving in the number of subjects recruited) to be gained by increasing the number of categories beyond five. Although the model is robust to mild departures from the assumption of proportional odds, with increasing numbers of categories it is less likely that the proportional odds assumption remains true. Therefore we shall use the four and five discrete category outcomes respectively of the RE and RP dimensions of the SF-36 to illustrate the effect of the bootstrap sample size estimator when the alternative to the null hypothesis is an odds ratio transformation.

Figures 6.14 and 6.15 show the power curves for *t*-test and *MW* test for the RP and RE dimensions of the SF-36 assuming the alternative hypothesis is proportional odds shift in HRQoL of $OR_{Ordinal} = 1.50$. Ordinal regression is equivalent to the *MW* test when there is only on independent 0/1 variable in the regression (Campbell, 2001). Although the advantage of ordinal regression over non-parametric methods is that we get an efficient estimate of a regression coefficient and we can extend the analysis to allow for other confounding variables.

As one would expect the bootstrap power curves in Figures 6.14 and 6.15 show that the *MW* test or the equivalent proportional odds model is more

powerful than the *t*-test when the alternative hypothesis is an odds ratio shift, although the differences in power for a given sample size are small.

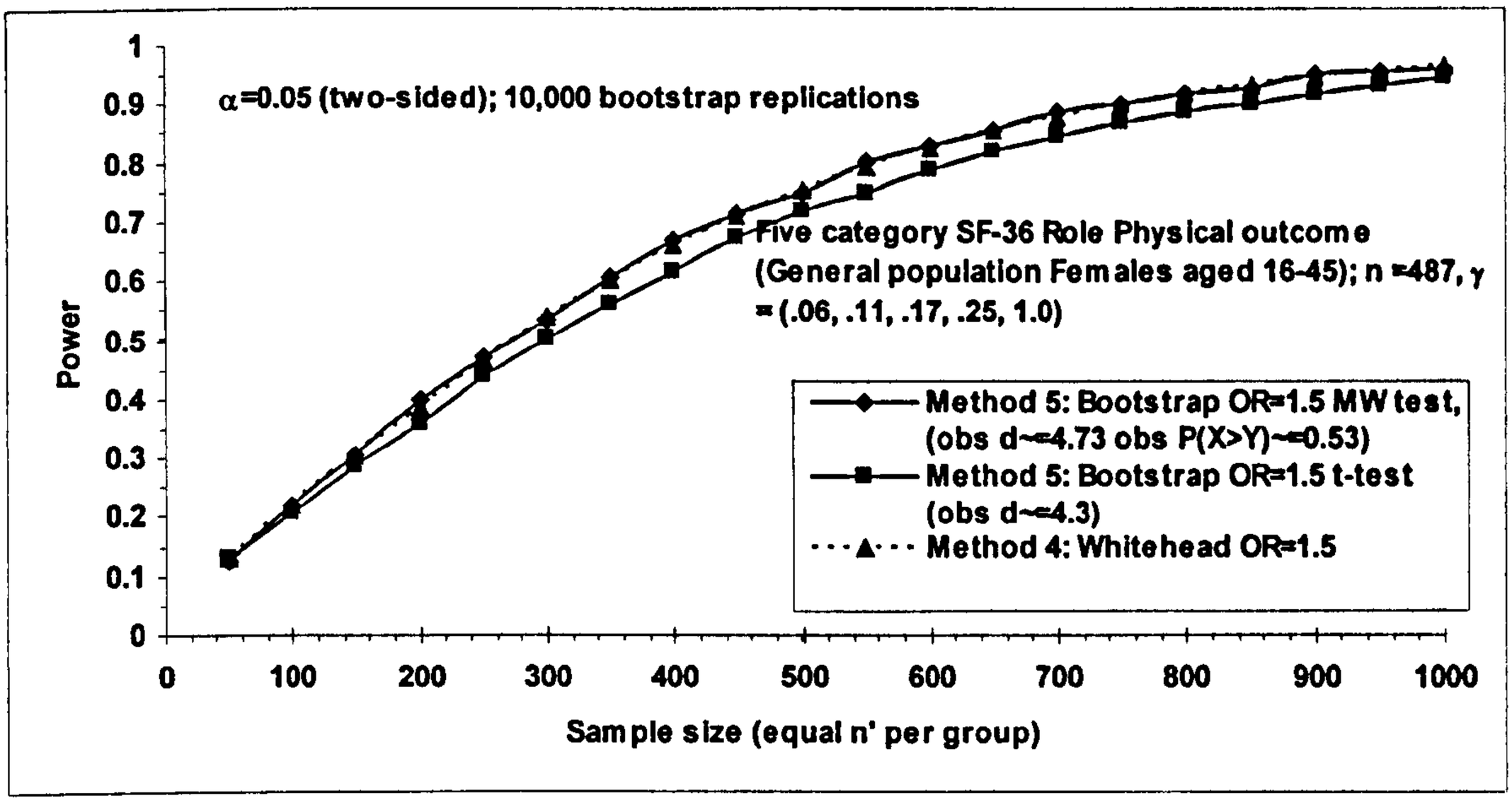
Figure 6.14 Estimated power curves for the SF-36 Role Emotional dimension to detect an Odds Ratio shift using general population data (females aged 16-45)



Sample sizes of over 450 patients per group are required to have an 80% chance of detecting this 'small to moderate' odds ratio (OR = 1.5) effect as statistically significant at the 5% two-sided level. With such 'large' sample sizes then the CLT guarantees that the sample means will be approximately Normally distributed. Thus, ensuring the relatively good performance of the *t*-test in detecting an OR style location shift. The robustness of the two independent samples *t*-test when applied to three-, four- and five-point ordinal scaled data has previously been demonstrated by Heeren and D'Agostino (1987) for far smaller sample sizes than this (as small as 20 per group).

The simulation results show that sample choice with HRQoL outcomes is far from an exact science. In practice one could use the bootstrap determined sample size as a suggestion and then taking account of the uncertainty, adjust the bootstrap sample size up or down as dictated by the circumstances of the experiment at hand.

Figure 6.15: Estimated power curves for the SF-36 Role Physical dimension to detect an Odds Ratio shift using general population data (females aged 16-45)



Longitudinal comparisons

In this chapter and Chapter 4 we have described sample size calculations for simple cross-sectional study designs. Many studies are longitudinal and as with cross-sectional studies, investigators need to know in advance the number of subjects approximately required to achieve a specified power. Diggle *et al* (2002) give sample size formulae for longitudinal studies for comparing two groups for continuous and binary responses. Frison and Pocock (1992) also provide sample size formulae for comparing two groups with continuous outcomes in clinical trials with pre-treatment (before) and post-treatment (after) repeated measurements. A recent paper by Strickland and Lu (2003) describes power and sample size calculations for two-sample ordinal outcomes under before and after study designs.

For estimating sample sizes for longitudinal studies, compared with cross-sectional study designs the following additional quantities are needed, the number of repeated observations per person and the correlation among the repeated observations, ρ. As with the between subject measurement

variance, σ^2, the correlation ρ can sometimes be estimated from either pilot studies or similar studies previously reported in the literature. When this is not possible (as is frequently the case), the statistician must make a reasonable guess at its value.

Pragmatically, with longitudinal designs one could choose one time point as the most important and estimate a sample size using formulas for Methods 1 to 4 for cross sectional designs. Thus, we could base the sample size calculation on a simple cross-sectional comparison (of means say), but use statistical methods for analysing longitudinal data such as relatively simple ANCOVA methods (Frison and Pocock, 1992) or more complex models for longitudinal data (Hand and Crowder, 1996; Diggle *et al* 2002). When the sample size for a longitudinal study is based on a simple cross-sectional comparison, considerably more patients than necessary may be recruited (Phillips and Campbell, 1997), subsequently, possibly increasing the cost and completion time of the study.

However, this over estimation of the required sample size for longitudinal studies is less likely to be a problem with HRQoL outcomes. This is because large or dramatic differences in HRQoL between groups are unlikely (see the small to moderate effect sizes shown in Chapter 5). Therefore, larger sample sizes will be required to detect clinically meaningful differences in HRQoL (as we have seen earlier on in this chapter). Secondly, missing HRQoL assessment data may also be a problem (Fayers and Machin, 2000). So the required sample size may need to be inflated anyway to take into account patients who drop out completely or who do not complete the HRQoL assessments.

Clinically meaningful effect sizes

More work is required on what is a clinically meaningful effect size for the SF-36 and other HRQoL outcomes. To illustrate the various methods of sample size calculation we assumed a mean difference δ of 5.0 in SF-36 scores was the MCID worth detecting.

There maybe considerable uncertainties in estimates of such quantities as the standard deviation and the treatment effect. Sample size calculations are sometimes based on estimates "pulled out of thin air". If an investigator is uncomfortable with the assumptions then it is good practice to calculate sample sizes under a variety of scenarios so that the sensitivity to assumptions can be assessed. We would recommend that various anticipated benefits be considered, ranging from the optimistic to the more realistic, with sample sizes being calculated for several scenarios within that range. It is a matter of judgement, rather than an exact science, as to which of the options is chosen for the final study size (Fayers and Machin, 2000; Walters *et al* 2001b).

In this chapter we have concentrated on the issue that HRQoL outcome data (such as the SF-36) may not meet the distributional requirements of parametric methods of sample size estimation and statistical analysis. There are other equally important problems with HRQoL measures such as ordinal scaling, linearity of the scale, floor/ceiling effects, non-constant variance and missing data which are discussed more fully in Walters *et al* 2001a; 2001b.

Conclusions

Given that the end goal of using HRQoL outcomes in research studies is to assess a patient's health and well being, using the right type of HRQoL outcome in the right setting with an appropriate sample size calculation is crucial.

If the HRQoL outcome has a limited number of discrete ordered values (less than seven categories) and/or the proportion of cases at either of the bounds is high, then we would recommend using Whitehead's (1993) Method 4 to estimate the required sample size. In this case the alternative hypothesis of a simple location shift model is questionable and the proportional odds model may provide a suitable alternative with such bounded discrete outcomes. Method 4 is particularly appealing if interest lies in comparing the relative

frequencies or cumulative probabilities in the ordered categories between treatment groups.

If the number of categories is large it is difficult to postulate the proportion of subjects who would fall into a given category. Even if there is little prior knowledge of the full distribution of scores for the HRQoL outcome, sample size calculation may not be too problematical. Using the ordinal approach to sample size calculation, knowledge of the anticipated distribution within four or five broad categories is usually sufficient to determine the required number of subjects (Whitehead, 1993; Julious *et al* 1997).

If the HRQoL outcome has a larger number of discrete values (greater than or equal to seven categories say), most of which are occupied and the proportion of cases at the upper or bounds are low, the simple location shift model is a useful working alternative hypothesis. This implies our interest lies in the comparison of means of the outcome variable between the two treatments. I would therefore recommend using Methods (1) or (2) to estimate the required sample size. If the distribution of the HRQoL outcome is expected to be reasonably symmetric then Method 1 is more appropriate for sample size calculations. If the distribution of the HRQoL outcome is expected to be skewed then the *MW* test appears to be more powerful at detecting a location shift (difference in means) than the *t*-test. So in these circumstances the *MW* test is preferable to the *t*-test and Method 2 could be used for sample size calculations and analysis, although pragmatically we would recommend Method 1 as the effect size Δ_{Normal} is rather easier to quantify and interpret than the effect size $p_{Noether}$ required for sample size estimation using Method 2. If a reliable pilot or historical dataset is readily available (to estimate the shape of the distribution) then bootstrap simulation (Method 5) may provide a more accurate and reliable sample size estimate than Methods 1 to 4.

We had a reliable historical data set of over 400 subjects so we had a large sample to estimate the cdfs F_x and F_y under the null and alternative hypotheses using Method 5. Lesaffre *et al* (1993) show that bootstrap can

give fairly unbiased estimates of power, though for small pilot samples with large variability. In the absence of a reliable pilot set, bootstrapping is not appropriate and conventional methods of sample size estimation or simulation models will need to be used. Fortunately with the increasing use of HRQoL outcomes in research, historical datasets are becoming more readily available. White and Thompson (2003) suggest the estimation of $\hat{F}$ (and hence $\hat{G}$) should be derived from a pilot dataset, and that the use of baseline data or related data sets (which we have used) is somewhat less satisfactory. They suggest a third possibility for estimating $\hat{F}$ is to use follow-up data viewed in a blinded manner, although only when the blinding can demonstrably be preserved.

Strictly speaking our results and conclusions only apply to the SF-36 HRQoL outcome measure. Further empirical work is required to see whether or not these results hold true for other HRQoL outcomes. However, the SF-36 has many features in common with other HRQoL outcomes, such as the NHP and QLQ-C30, i.e. multi-dimensional, ordinal or discrete response categories with upper and lower bounds, and skewed distributions. Therefore we see no theoretical reason why these results and conclusions with the SF-36 may not be appropriate for other HRQoL measures.

Chapter 7: Analysing HRQoL data (one outcome measurement or one outcome and a baseline measurement) using the bootstrap

Introduction

In Chapters 4 to 6 we discussed the estimation of sample sizes for studies with HRQoL outcomes. Once we have carried out the study and collected the data we will then need to analyse the HRQoL outcomes. So how should investigators analyse HRQoL data? Conventional methods of analysis of HRQoL outcomes are extensively discussed in Fayers and Machin (2000) and Fairclough (2002). However, neither of these texts used the bootstrap to analyse HRQoL outcomes. Therefore in this chapter and the next we will concentrate on comparing and contrasting the bootstrap with standard methods of analysing HRQoL outcomes as described in Fayers and Machin (2000) and Fairclough (2002).

In this chapter we will concentrate on the analysis of HRQoL data collected from simple cross-sectional designs or studies with two HRQoL assessments (baseline and follow-up). The subsequent chapter will look at more complex longitudinal studies where HRQoL outcomes are collected at three or more time points.

We will apply conventional and bootstrap methods to the data from the CPSW study and the OA Knee study. The CPSW study was a two group RCT, but it was unusual since no baseline HRQoL assessment was made. It was felt that it was inappropriate to assess HRQoL just prior to or immediately after childbirth. We will use this data set to illustrate various methods for two group cross-sectional comparisons of HRQoL ranging from a simple comparison of mean scores (using conventional and bootstrap hypothesis tests) through to more complex regression analyses including ordinal regression.

The OA Knee study involved the collection of HRQoL data at two time-points (baseline and follow-up) six months apart. Since there was a difference in the

baseline HRQoL and socio-demographic characteristics (age and gender) of the Rheumatology clinic and TKR surgery groups, we use this dataset to illustrate multiple regression/ANCOVA methods with follow-up HRQoL as the outcome variable and baseline HRQoL, age, gender and group as covariates. We will compare conventional OLS estimates of SE and CI for the group regression coefficient with their bootstrap counterparts.

Analysis of CPSW data

Figures 7.1 and 7.2 show the histograms of the SF-36 dimension scores at six weeks post-natally for Intervention and Control groups. The graphs clearly show the skewed and discrete nature of the outcome data for the SF-36 from this study.

Suppose the HRQoL measure has a large number of ordered categories, most of which should be occupied if the underlying scale really is continuous, but the scale is measured imperfectly by an instrument with a limited number of discrete values. It is often worth treating this discrete scale as if it were continuous. An informal rule of thumb (Walters *et al* 2001a) is that this discrete scale should be treated as continuous if it has seven or more categories and as ordinal otherwise. So for example with the eight dimensions of the SF-36 six out of eight have more than seven discrete categories and only two i.e. the RP and RE scales have less than seven discrete categories.

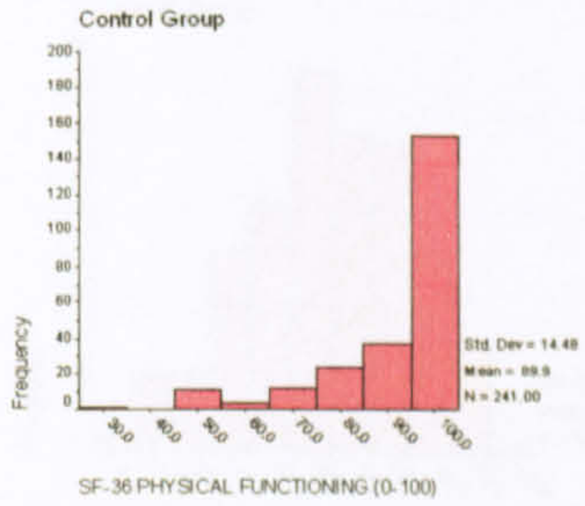
Walters' *et al* (2001a) base this informal rule of thumb on Whitehead's (1993) sample size formula for ordinal data. Whitehead illustrates the dependence of sample size *n* on the number of categories *c*. It is assumed that all categories are equally probable ($\bar{\pi}_1 = \bar{\pi}_2 = ...\bar{\pi}_k$ for all *c*). It follows from (7.1) that the sample size (denoted by *n(c)*) required when there are *c* equally probably categories, keeping *OR*, α and β constant, is,

$$n(c) = \frac{0.75}{1 - 1/c^2} n(2). \qquad (7.1)$$

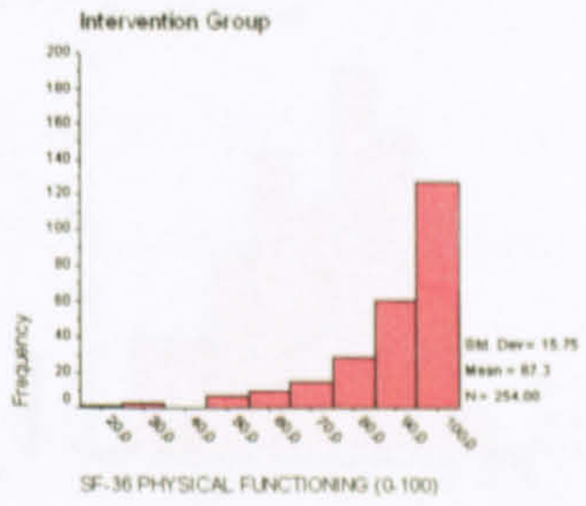
In the limit as $c \to \infty$, $n(c) \to 0.75n$.

Figure 7.1: Distribution SF-36 dimensions from CPSW data by group

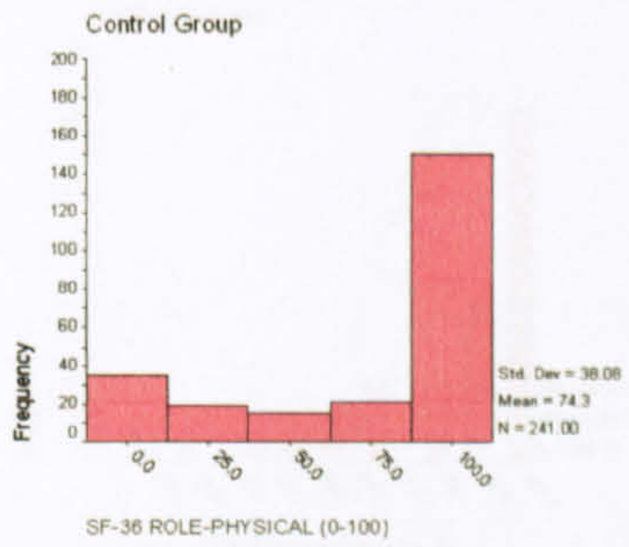
a)



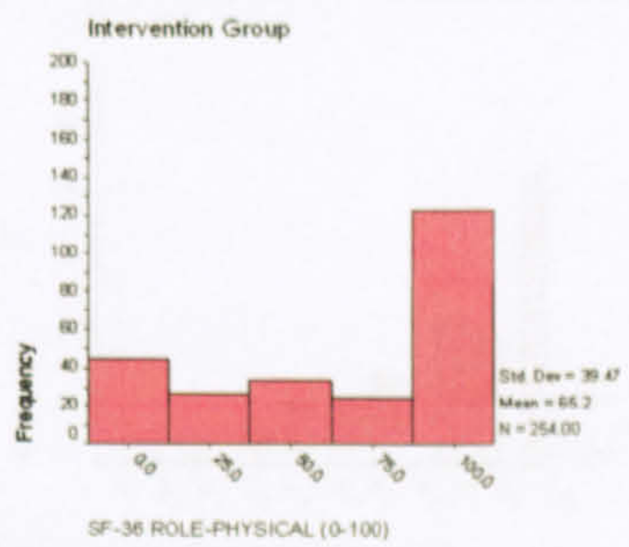
b)



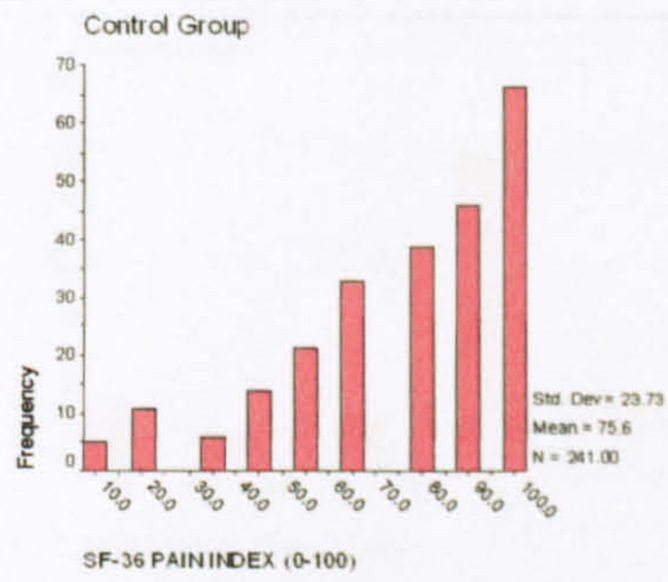
c)



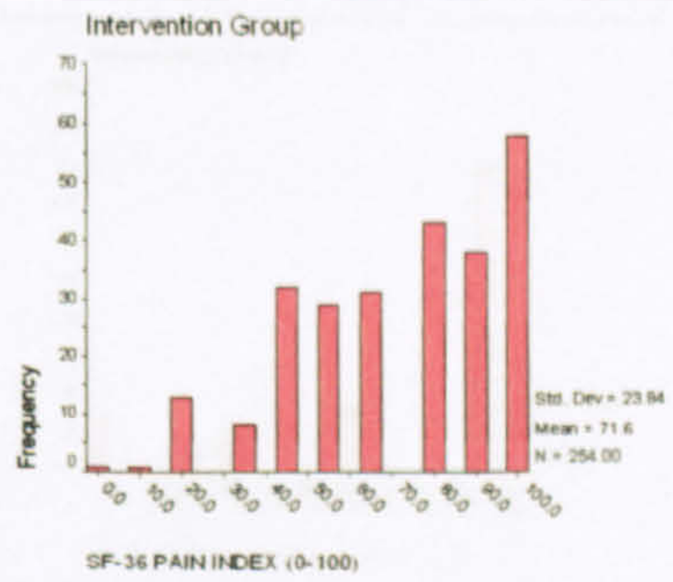
d)



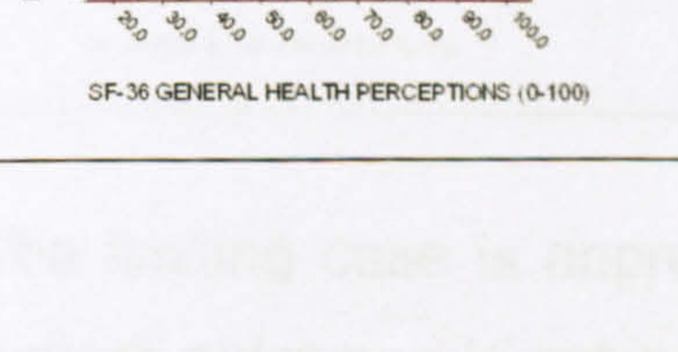
e)



f)



g)



h)

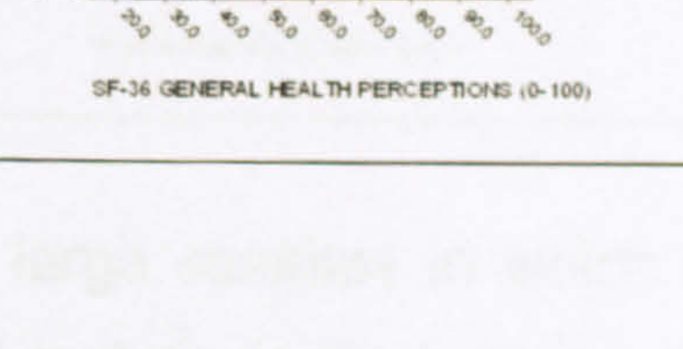
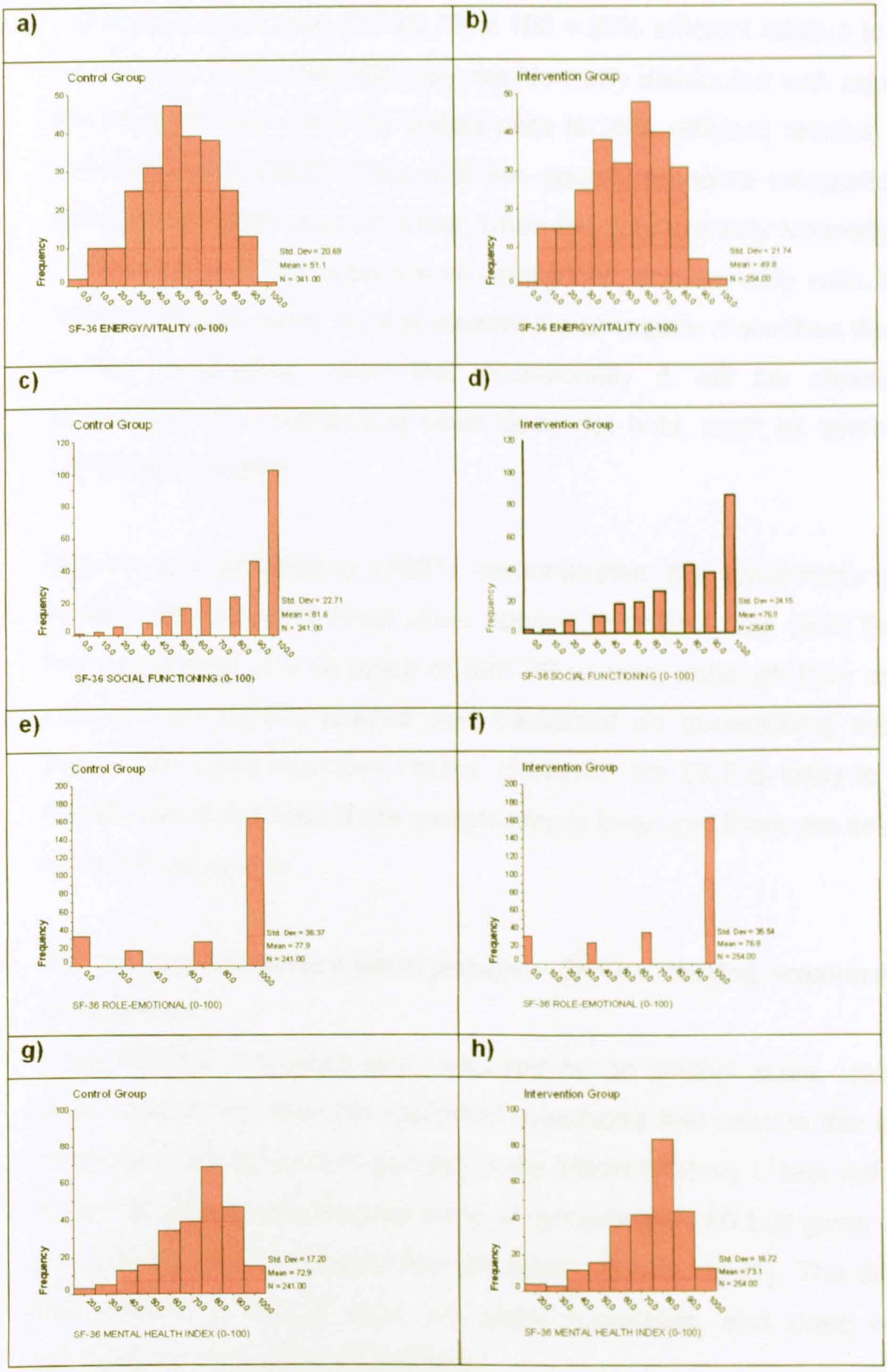


Figure 7.2: Distribution SF-36 dimensions from CPSW data by group



The limiting case is approached in large samples in which a full ranking of patient outcomes is achieved. A full ranking is equivalent to a categorisation with one patient in each category. So for continuous data the sample size using equation (4.12) tends to be 75% of the binary outcome case. Equation

(7.1) says that little is gained by using more than five categories, as the hypothesis test will be (0.75/0.78) x 100 = 96% efficient relative to the use of a full ranking. When the data truly are Normally distributed with equal variances the Mann-Whitney test for untied data is 96% efficient relative to the *t*-test (Armitage *et al* 2002). Thus with five equally probable categories the test is 92% efficient relative to the *t*-test, when the data are truly Normally distributed. These relative efficiencies are all asymptotic, and are only valid for moderate to large sample sizes, so it is reasonable to require more than five categories in the assumption. Note that occasionally it will be obvious that the assumption of a continuous scale does not hold, such as when one of the categories is death.

Heeren and D'Agostino (1987) demonstrated the robustness of the two-independent samples *t*-test when applied to ordinal data (with three, four or five categories) and samples of size 20 or less, although they assumed the scales were equally spaced and cautioned on generalising the results to scales with more than five values. However, the CLT is likely to ensure the robustness of the *t*-test if the sample size is large and there are seven or more occupied categories.

Comparing two independent groups - Ordinal HRQoL measures (with < 7 categories)

If the HRQoL outcomes are measured on an ordinal scale, with less than seven categories, then the statistical hypothesis test used in this instance (to compare two independent groups) is the Mann-Whitney *U* test with allowance for ties or Chi-squared test for trend. In general the *MW* test gives very similar *p*-values to the Chi-squared test for trend (Altman, 1991). The difficulty with this method is that it does not allow covariates, and does not provide estimates of population parameters.

The simplest approach to analysing ordinal data is to dichotomise the data and use logistic regression. However this method ignores useful information in the data, may not be very powerful (Armstrong and Sloan, 1989) and introduces the problem of where to choose the cut point. If one were to keep

the ordinal structure then there are number of models possible (Ananth and Kleinbaum, 1997; Manor *et al* 2000; Walters *et al* 2001a; Lall *et al* 2002). These include proportional odds, continuation ratio, polytomous and stereotype. We will illustrate these various models using the CPSW data and the five category RP dimension of the SF-36. Table 7.1 shows the frequency distribution of the RP dimension at six weeks postnatally for new mothers in the Control and Intervention groups respectively. Table 7.1 also summarises three potential confounding variables or covariates, age at delivery, parity and type of delivery.

Proportional odds or cumulative logit model

The proportional odds or cumulative logit model is based on the cumulative response probabilities rather than the category probabilities (McCullagh and Nelder, 1989).

For example consider an HRQoL outcome variable Y with c categorical outcomes y_i denoted *by* $i = 1,\ldots,c$ and let p be a set of covariates $(x_1, x_2,\ldots,x_p)$. The cumulative logit or proportional odds model is

$$\gamma_i = \Pr\left(Y \le y_i \middle| x_1, x_2, \ldots, x_p\right) = \frac{\exp\left(\alpha_i + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\right)}{1 + \exp\left(\alpha_i + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\right)} \quad i = 1,\ldots,c\text{-}1. \quad (7.2)$$

This can be expressed equivalently in logit (γ_i) form as

$$\text{logit}\left(\gamma_i\right) = \log\left[\frac{\gamma_i}{1-\gamma_i}\right] = \log\left[\frac{\Pr\left(Y \le y_i \middle| x_1, \ldots, x_p\right)}{\Pr\left(Y > y_i \middle| x_1, \ldots, x_p\right)}\right] = \alpha_i + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

$$i = 1,\ldots,c\text{-}1, \quad (7.3)$$

where $\gamma_i = \Pr(Y \le y_i \mid x_1, x_2,\ldots,x_p)$ is the cumulative probability of being in category i or lower given the set of covariates (note that for $i = c$; $\Pr(Y \le y_i \mid x_1, x_2,\ldots,x_p) = 1$). The α_i $(i = 1,\ldots,c\text{-}1)$ and $\{\beta_1, \beta_2, \ldots, \beta_p\}$ parameters are treated as unknown and the intercept parameters α_i must satisfy the condition $\alpha_1 \le \alpha_2 \le \ldots \le \alpha_{c-1}$ (McCullagh and Nelder, 1989). The regression coefficient β_p for a binary explanatory variable x_p (e.g. Control or Intervention group) is the log-odds ratios for the Y by x_p association controlling for the other covariates in the model. That is the treatment effect on HRQoL after adjusting for prognostic factors such as age, parity and type of delivery.

Table 7.1: Descriptive statistics for SF-36 Role Physical Dimension at 6 weeks for Control and Intervention group mothers from the CPSW trial

	Control			*Intervention*			
	N=241			*n=254*			*Odds Ratio*
	N	*(%)*	*Cumulative %*	*n*	*(%)*	*Cumulative %*	Int/Con
RP Score							
0.0	35	(14.5)	*14.5*	45	(17.7)	*17.7*	0.79
25.0	19	(7.9)	*22.4*	27	(10.6)	*28.3*	0.73
50.0	15	(6.2)	*28.6*	34	(13.4)	*41.7*	0.56
75.0	21	(8.7)	*37.3*	25	(9.8)	*51.6*	0.56
100.0	151	(62.7)	*100.0*	123	(48.4)	*100.0*	
Total	241	*100.0*		254	*100.0*		
Mean	74.3			65.2			
SD	38.1			39.5			
Median	100.0			75.0			
25th percentile	50.0			25.0			
75th percentile	100.0			100.0			
Age (years)							
Mean	28.1			28.0			
SD	5.6			5.7			
Parity	N	(%)		n	(%)		
First child	119	(49.4)		133	(52.4)		
Second or more	122	(50.6)		121	(47.6)		
Normal Delivery							
No	76	(31.5)		81	(31.9)		
Yes	165	(68.5)		173	(68.1)		

The $\{\beta_1, \beta_2, ..., \beta_p\}$ regression parameters do not depend on the category *i*, so that the model (7.3) assumes that the relationship between each of the covariates and *Y* (HRQoL) is independent of *i* (the response category). This assumption of identical log-odds ratios across the *c* categories is the proportional odds assumption.

The proportional odds model is useful when one believes HRQoL is a continuum, which is measured imperfectly by an instrument with a limited number of values. The proportional odds model is invariant when the codes for the response *Y* are reversed (i.e. y_1 recoded as y_c, y_2 recoded as y_{c-1} and

so on). Also the proportional odds model is invariant under the collapsibility of adjacent categories of the ordinal response implying that when y_1 and y_2 are combined, the estimate of the odds ratio remains essentially the same as the odds ratios obtained for the individual categories (Greenland, 1994).

Continuation ratio model.

An alternative method to the proportional odds model is the *Continuation ratio model*. This may be relevant when an ordinal HRQoL scale may be thought of as a progression through various stages, so that people start with 'excellent' and deteriorate to 'poor' and are unlikely to reverse this trend. The cumulative probabilities $\gamma_i = \Pr(Y \leq y_i \mid x_1, x_2, \ldots, x_p)$ of being in category *i* or lower in the cumulative logit model (7.3) are replaced by the probability of being in category *i* [i.e. $\pi_i = \Pr(Y = y_i)$] divided by the probability of being in a category higher than *i* [i.e. $\Pr(Y > y_i)$] for the continuation ratio model.

$$\text{logit}\left(\frac{\pi_i}{1-\gamma_i}\right) = \log\left[\frac{\left(\frac{\pi_i}{1-\gamma_i}\right)}{\left(1-\frac{\pi_i}{1-\gamma_i}\right)}\right] = \log\left[\frac{\Pr\left(Y = y_i \middle| x_1,\ldots,x_p\right)}{\Pr\left(Y > y_i \middle| x_1,\ldots,x_p\right)}\right] = \alpha_i + \left(\beta_1 x_1 + \beta_2 x_2 + .. + \beta_p x\right.$$

$$i = 1,\ldots,c\text{-}1. \qquad (7.4)$$

When the 'logit' expansion is replaced by the 'complementary log-log' link function in model (7.4), the resulting model (7.5) is

$$\log\left[-\log\left(\frac{\pi_i}{1-\gamma_i}\right)\right] = \alpha_i + (\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p) \qquad i = 1,\ldots,c\text{-}1, \quad (7.5)$$

which is the Cox proportional-hazards model for survival data in discrete time.

The continuation ratio model is not invariant under the collapsing or reversal of categories. The continuation ratio model is best suited to circumstances where the individual categories of the HRQoL scale are of interest and a monotonic progression through the individual categories is expected. Armstrong and Sloan (1989) have given a useful comparison of the proportional odds and continuation ratio models.

Chi-squared (χ^2) score tests are available for tests of the proportional odds assumption but these lack power (Brant, 1990; Peterson and Harrell, 1990). Also the model is robust to mild departures from the assumption of proportional odds. A crude test would be to examine the odds ratios and if they are all greater than unity, or all less than unity, then a proportional odds model will suffice (Walters *et al* 2001a), although with increasing numbers of categories it is less likely that proportional odds assumption remains true.

The ordinal regression method also allows us to adjust the treatment effect for other prognostic factors and covariates (such as centre, sex and age). The regression coefficients and their standard errors also enable confidence intervals to be calculated. The statistical packages SPSS, SAS and STATA have procedures for fitting proportional odds or continuation ratio models.

Stereotype logistic model

For the cumulative logit model (7.3), the HRQoL outcome variable *Y* is assumed to have an unobserved underlying variable (say *Z*), which takes on a continuous form. For example, '*Age*' may be represented by ordered categories, which take on the form '*young*', *'middle-aged'*, '*old*' and '*very old*'. In this case, there is an underlying variable, calendar age.

HRQoL scales are sometimes constructed in such a way that there is no underlying variable that directly links to the ordered *y*-response categories. (Although as mentioned in the Introduction this is one of my key assumptions, that there *actually* is an underlying latent continuous HRQoL variable.) For instance when assessing 'pain' one may use a rating scale of the form 'none', 'mild', 'moderate' and 'severe'. Here pain is rated depending on other factors such as its severity and type of pain. Although the rating scale is in principle ordered, there may be no underlying variable (continuous or otherwise) that directly relates the factors and links these up with the categories on the scale. Anderson (1984) recognised these types of ordered categories as being truly discrete and referred to the response as a *judged* or *assessed* variable. As the cumulative logit model would be inappropriate for analysing such variables, Anderson introduced another model known as the *stereotype*

model. One of the main advantages of the stereotype model over other regression models is that it does not assume a priori ordering of the *y*-response categories.

The stereotype model is based on the polytomous regression model (Anderson, 1984), which does not impose any restrictions on the ordering of the categories. The ordinality is in-built into it by imposing a structure on the regression coefficients. Consider an HRQoL outcome variable *Y* with *c* ordered categorical outcomes y_i denoted by $i = 1, 2,\ldots,c$, and let $x_1, x_2,\ldots,x_p$ denote a set of *p* covariates. The ordinary polytomous regression model can be written as

$$\Pr(Y = y_i \mid x_1, x_2, \ldots, x_p) = \frac{\exp\left(\alpha_i + \beta_{i1}x_1 + \beta_{i2}x_2 + \ldots + \beta_{ip}x_p\right)}{\sum_{i=1}^{c} \exp\left(\alpha_i + \beta_{i1}x_1 + \beta_{i2}x_2 + \ldots + \beta_{ip}x_p\right)}, \quad (7.6)$$

where $\alpha_c = 0$ and $\beta_{ck} = 0$ $(k = 1,\ldots,p)$ to assure identifiability. The log-probability ratios are formed for model (7.6) by comparing each response category (y_i) with a reference category (y_c). The choice of the reference category is arbitrary but we shall use the first category. Thus, the log-probability ratio can be represented by a linear model of the form

$$\log\left[\frac{\Pr\left(Y = y_i \middle| x_1, x_2, \ldots, x_p\right)}{\Pr\left(Y = y_1 \middle| x_1, x_2, \ldots, x_p\right)}\right] = \alpha_i + \beta_{i1}x_1 + \beta_{i2}x_2 + \ldots + \beta_{ip}x_p \qquad i = 2 \text{ to } c. \quad (7.7).$$

The regression coefficient β_{ip} for the p^{th} covariate x_p corresponds to the log-probability ratio comparing $(Y = y_i)$ versus $(Y = y_1)$ for a unit increase in x_p.

From model (7.7) it is clear that the ordinal nature is not accounted for in any way. The ordinality can be built into this model by imposing a structure on the regression coefficients β_{ik} $(k = 1,\ldots,p)$. Anderson (1984) proposed modelling the regression coefficients, β_{ik}, by imposing the relationship

$$\beta_{ik} = \phi_i \beta_k \qquad i = 2,\ldots,c;\ k = 1,\ldots,p, \quad (7.8)$$

where β_k is a list of new parameters and the ϕ_i's can be thought of as the scores attached to the response y_i. Note that since $\beta_{1k} = 0$, we have $\phi_1 = 0$, and a further constraint, $\phi_c = 1$ (in order to uniquely identify the parameters

when using estimated scores). Substituting (7.8) into (7.7) yields the stereotype model

$$\log\left[\frac{\Pr(Y=y_i|x_1,x_2,...,x_p)}{\Pr(Y=y_1|x_1,x_2,...,x_p)}\right]=\alpha_i+\phi_i(\beta_1x_1+\beta_2x_2+...+\beta_px_p)\ \ i=2\ to\ c.\quad(7.9)$$

Thus, it can be seen that the stereotype model determines a set of parameters {ϕ_i's} for the dependent variable and a single parameter β_k for each covariate. The ϕ_i's are decided upon for the response variable and are directly tied up with the effect of the explanatory variables. Thus, with a positive β_k, when the log probability ratios {$\phi_i\beta_k$'s} form a decreasing trend, the ϕ_i parameters also become ordered such that:

$$\phi_c\beta_k \geq,.....,\phi_2\beta_k \geq \phi_1\beta_k = 0 \text{ then } 1 = \phi_c \geq,...,\phi_2 \geq \phi_1 = 0. \quad (7.10)$$

Here we can say that the effect of the covariates upon the first log probability ratio is greater than their effect upon the second and so on, (or the effect is the same upon the consecutive log probability ratios), and that provided (7.10) holds model (7.9) is an ordered regression model.

The model fitted does not necessarily require the ϕ_i's to be ordered; whether there is ordering or not is purely determined by the empirical evidence provided by the data. Two categories denoted by c_1 and c_2 are indistinguishable with respect to the covariates if $\phi_{c1} = \phi_{c2}$, that is the effect of the covariates is the same in the two categories. The product $\phi_i\beta_p$ for the p^{th} covariate x_p corresponds to the log-probability ratio comparing ($Y = y_i$) versus ($Y = y_1$) for a unit increase in x_p.

Greenland (1984) argues strongly in favour of the stereotype model when there is no underlying continuum that is directly related to the response categories, but where each state is assessed. The statistical package STATA has a procedure for performing stereotype regression via constrained polytomous logistic regression (Hendrickx, 2000). The stereotype model can also be fitted in SAS using PROC CATMOD. Further details of the stereotype model when applied to HRQoL outcomes are given in Lall *et al* 2002 (See Appendix 3 for more details).

Table 7.2 shows the results of fitting a binary logistic model, cumulative logit model, continuation ratio model, polytomous model and stereotype model to the RP data with group, delivery and parity as factors and age as a covariate. For the logistic model the outcome variable is dichotomised into a score of 100 "good health" and less than 100 ("less than good health"). The OR of 0.54 (95% CI: 0.37 to 0.78) for the binary logistic model implies that the odds of new mothers in the Intervention group having 'good' health is 0.54 times that of mothers in the Control group after allowing for age, parity and delivery. That is new mothers in the Intervention group are significantly less likely to report good health than Control group mothers.

Similarly the proportional odds model implies that having been randomised to the Intervention group carried with it an odds ratio of 0.56 (95% CI: 0.40 to 0.80) compared to that of women randomised to the Control group for being in a given category or below (i.e. better HRQoL) after allowing for age, parity and delivery. As the proportional odds model assumes a constant OR for all categories, Table 7.1 and Figure 7.3 show how the four (*c - 1*) observed ORs compare with the estimated common OR of 0.56 from the model. All observed ORs are less than 1 and seem similar to the model estimate. However, a chi-squared score test of proportional odds was $\chi^2 = 112.1$ on 12 df, $p = 0.0001$. Thus, there is strong statistical evidence to reject the assumption of proportional odds. So the proportional odds model may not be appropriate, although the model is robust to mild departures from the assumptions.

Similarly, the continuation ratio model OR estimate implies that being randomised to the Intervention group carried with it an odds ratio of 0.60 (95% CI: 0.44 to 0.80) compared to that of new mothers randomised to the Control group for better HRQoL after allowing for age, parity and delivery.

The polytomous logistic model implies that the probability of having a RP score of 25 compared to 0 is 1.04 times the probability for women in the Intervention group compared to women in the Control group; the probability of having a RP score of 50 as opposed to 0 is 1.64 times the probability for

Table 7.2: Results of fitting various binary and ordinal models to the SF-36 Role Physical dimension score at six weeks postnatally for the Intervention and Control group mothers from the CPSW study.

Model[a]	$\hat{\beta}$	$SE(\hat{\beta})$	*P-value*	*OR*[f]	*95% CI for OR*		χ^2 *Goodness of fit*[g]
Logistic regression (binary)[b]	-0.623	0.19	0.001	0.54	0.37	0.78	40.6 on 4 df, p = 0.0001
Proportional Odds[c]	-0.576	0.18	0.001	0.56	0.40	0.80	59.7 on 4df , p = 0.0001
Continuation ratio	-0.518	0.15	0.001	0.60	0.44	0.80	52.0 on 4 df, p=0.0001
Polytomous Model[d]							
25 vs 0	0.039	0.387	0.919	1.04	0.49	2.22	90.9 on 16 df, p = 0.0001
50 vs 0	0.5	0.4	0.213	1.64	0.75	3.60	
75 vs 0	-0.136	0.383	0.723	0.87	0.41	1.85	
100 vs 0	-0.528	0.2805	0.06	0.59	0.34	1.02	
Stereotypical[e]							
$\hat{\beta}$	-0.469	0.28	0.093	0.63	0.36	1.10	
25 vs 0 $(\phi_2\hat{\beta})$	-0.361			0.70			
50 vs 0 $(\phi_3\hat{\beta})$	-0.424			0.65			
75 vs 0 $(\phi_4\hat{\beta})$	-0.342			0.71			
100 vs 0 $(\phi_5\hat{\beta})$	-0.469			0.63			

a. All models include age, parity and delivery as covariates.
b. The response variable was dichotomised into (0 or 25 or 50 or 75) vs. 100.
c. Response variable is RP score I.e. 0, 25, 50, 75,100.
d. The response variable is the RP score with 0 as the reference category.
e. With ϕ_1= 0, ϕ_2 = 0.77, ϕ_3 = 0.90, ϕ_4 = 0.73, ϕ_5 = 1.00
f. Odds ratios except for the polytomous and stereotype models where it is the ratio of probability ratios.
g. Likelihood-ratio statistics for testing null model (no covariates) against the extended model (with covariates).

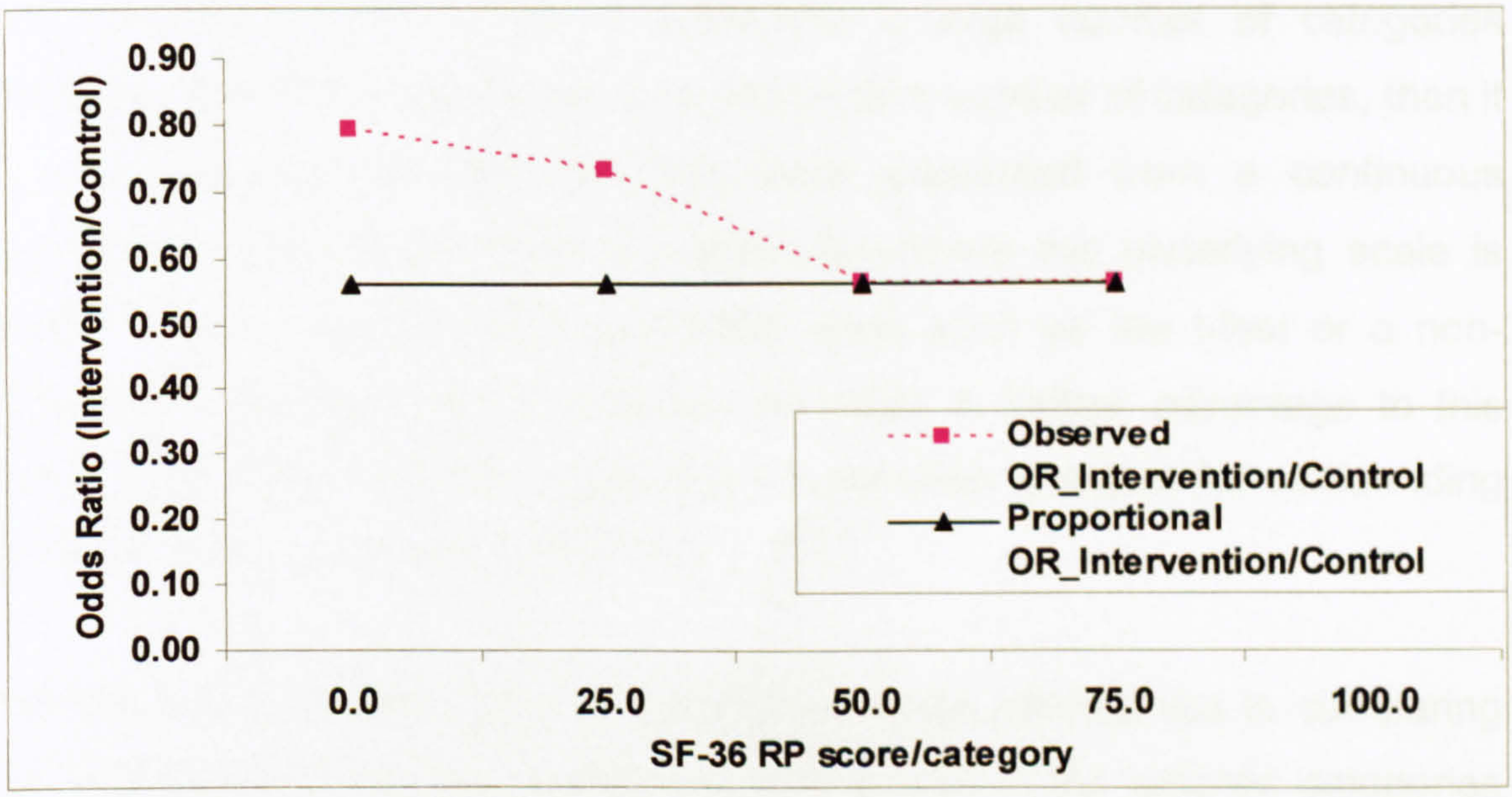
Table 7.2: Results of fitting various binary and ordinal models to the SF-36 Role Physical dimension score at six weeks postnatally for the Intervention and Control group mothers from the CPSW study.

Model[a]	$\hat{\beta}$	$SE(\hat{\beta})$	*P-value*	*OR*[f]	*95% CI for OR*		χ^2 *Goodness of fit*[g]
Logistic regression (binary)[b]	-0.623	0.19	0.001	0.54	0.37	0.78	40.6 on 4 df, p = 0.0001
Proportional Odds[c]	-0.576	0.18	0.001	0.56	0.40	0.80	59.7 on 4df , p = 0.0001
Continuation ratio	-0.518	0.15	0.001	0.60	0.44	0.80	52.0 on 4 df, p=0.0001
Polytomous Model[d]							
25 vs 0	0.039	0.387	0.919	1.04	0.49	2.22	90.9 on 16 df, p = 0.0001
50 vs 0	0.5	0.4	0.213	1.64	0.75	3.60	
75 vs 0	-0.136	0.383	0.723	0.87	0.41	1.85	
100 vs 0	-0.528	0.2805	0.06	0.59	0.34	1.02	
Stereotypical[e]							
$\hat{\beta}$	-0.469	0.28	0.093	0.63	0.36	1.10	
25 vs 0 $(\phi_2\hat{\beta})$	-0.361			0.70			
50 vs 0 $(\phi_3\hat{\beta})$	-0.424			0.65			
75 vs 0 $(\phi_4\hat{\beta})$	-0.342			0.71			
100 vs 0 $(\phi_5\hat{\beta})$	-0.469			0.63			

a. All models include age, parity and delivery as covariates.
b. The response variable was dichotomised into (0 or 25 or 50 or 75) vs. 100.
c. Response variable is RP score I.e. 0, 25, 50, 75,100.
d. The response variable is the RP score with 0 as the reference category.
e. With ϕ_1= 0, ϕ_2 = 0.77, ϕ_3 = 0.90, ϕ_4 = 0.73, ϕ_5 = 1.00
f. Odds ratios except for the polytomous and stereotype models where it is the ratio of probability ratios.
g. Likelihood-ratio statistics for testing null model (no covariates) against the extended model (with covariates).

women in the Intervention group compared to women in the Control group; the probability of having a RP score of 75 as opposed to 0 is 0.87 times the probability for women in the Intervention group compared to women in the Control group and the probability of having a RP score of 100 (as opposed to 0) is 0.59 times the probability for new mothers in the Intervention group compared to new mothers in the Control group. For the last three categories the trend in the probability ratios is monotonic as the health status goes from the 'poor' stage (i.e. score of 50) through to the better stages (i.e. RP scores of 75 or 100).

Figure 7.3: Observed and proportional odds ratio estimates from the SF-36 Role Physical Dimension at six weeks postnatally for Control and Intervention group mothers from the CPSW trial



Attaching a set of scores to the beta parameters in the polytomous model, leads to the formation of the stereotype model (Table 7.2). The ϕ_i's are decided upon for the response variable and are directly tied up with the effect of the explanatory variables (age, delivery, parity and group). The model does not necessarily require the ϕ_i's to be ordered and indeed the empirical evidence provided by the data, suggests the ϕ_i's are not ordered.

In this model the probability of having an RP score of 25 compared to a score of 0 is 0.70 times the probability for subjects in the Intervention group

compared to subjects in the Control group; the probability of having an RP score of 50 as opposed to 0 is 0.65 times the probability for subjects in the Intervention group compared to subjects in the Control group; the probability of having an RP score of 75 as opposed to 0 is 0.71 times the probability for subjects in the Intervention group compared to subjects in the Control group and the probability of having an RP score of 100 (as opposed to 0) is 0.63 times the probability for subjects in the Intervention group compared to subjects in the Control group.

Comparing two independent groups - HRQoL scales with more than seven categories

When the HRQoL scale has more than seven categories, it is important to check in any particular situation that most of the categories are occupied, to rule out having only a few of potentially a large number of categories occupied. Where the distribution is spread over a number of categories, then it is useful to assume that the data were generated from a continuous distribution, especially if there is reason to believe the underlying scale is linear. In this case the usual parametric tests such as the *t*-test or a non-parametric test such as the *MW* can be used. A further advantage to this assumption is that multiple regression can be used to adjust for confounding variables such as baseline covariates.

The proportional odds model is appropriate when interest lies in comparing the relative frequencies or cumulative probabilities in the ordered categories between groups. As such, a limitation of the non-parametric and proportional odds modelling approaches is in their interpretation. If the goal of the analysis is to assess the magnitude of the treatment effect on the ordered HRQoL outcome, then our interest lies in comparing location between two treatments. Therefore a more appealing approach is to assign numeric scores to the ordered categories and to use a more familiar linear regression model for analysis. It is common for HRQoL outcomes to assign numeric scores to the *c* ordered categories (for example, 0, 1, 2, (*c* - 1) (or 0, 25, 50, 75, and 100 in the case of the five category SF-36 RP dimension) and to compare means between the groups. Furthermore, Heeren and D'Agostino (1987) have

demonstrated the robustness of the two-independent samples *t*-test applied to three-, four- and five-point ordinal scaled data when using assigned scores, even for sample sizes as small as 20 per group.

Many HRQoL instruments such as the SF-36, NHP and EORTC QLQ-C30 are multi-dimensional with individual dimensions measured on a variety of discrete scales with both less than seven categories and seven or more categories. Therefore it may be sensible and indeed preferable to use one method of analysis such as linear regression for all dimensions of the HRQoL instrument, rather than use ordinal regression for some dimensions (with less than seven categories) and linear regression for other dimensions (with ≥ 7 categories).

For the rest of this thesis we are going to assume the underlying latent HRQoL variable is continuous but quantification of the outcome is limited to ordered categories and that our interest lies in comparing location between two treatments or groups. So we will concentrate on statistical methods for comparing means such as the *t*-test and linear regression.

If we assume the HRQoL data are continuous and approximately Normally distributed then Fayers and Machin (2000) suggest an appropriate summary statistic is the mean. They go on to suggest that assuming the HRQoL data have an approximately Normal form then two treatment groups can be compared by calculating the difference between the two respective means by a two independent samples *t*-test.

Suppose we wish to test the null hypothesis (H_0) that the means from two populations (μ_z and μ_y), estimated from two independent samples, are equal i.e. ($\mu_z = \mu_y$), against the alternative (H_A) that they are different i.e. ($\mu_z \neq \mu_y$).

If sample 1 has n subjects, with sample mean $\bar{z}$, and sample standard deviation s_z. Similarly, sample 2 has m subjects, mean $\bar{y}$, standard deviation

s_y. Then to test the null hypothesis $\mu_z = \mu_y$, when σ_z and σ_y are unknown but equal ($\sigma_z = \sigma_y = \sigma$) we take:

$$t = \frac{\bar{z} - \bar{y}}{s_p \sqrt{\left\{\frac{1}{n} + \frac{1}{m}\right\}}}, \tag{7.11}$$

where s_p is a pooled estimate (5.5) of the common standard deviation σ.

Under the null hypothesis, (7.11) is distributed as Student's *t*-distribution with $n + m - 2$ degrees of freedom. If we assume the two variances are equal then the two sample *t*-test is not only a comparison of means, but also a comparison of populations or distributions i.e. a test of $F_z = G_y$.

The three assumptions for carrying out a two-sample *t*-test (Campbell and Machin, 1999) are

(1) The data are plausibly Normally distributed.

(2) The data (or samples) are independent.

(3) Standard deviations from the two populations are equal.

As we have seen in Chapter 2, HRQoL data, with their discrete and bounded scores are unlikely to be Normally distributed and may not have constant variance. Although the importance of the Normality assumption should not be overstated, since the method is valid because the sample size is sufficient to take care of non-Normality through operation of the CLT. The two-sample *t*-test is exactly correct if the two samples are from Normal distributions with equal variances, and otherwise is approximately valid, where the approximation improves with increasing *n* and *m* and the closer the distributions are to Normality (Armitage *et al* 2002). With larger *n* and *m* then the more non-Normality can be tolerated.

If we are still worried about the assumptions of Normality and equality of variances we can either try and transform the HRQoL data to Normality and constant variance or use a non-parametric test. If we use a non-parametric test then the most popular (but not necessarily the most efficient) for

comparing two groups is the Mann-Whitney *U* test. The *MW* test is a test of the null hypothesis that the two populations are identical (i.e. $F = G$). What if we wanted to test only whether their means are equal? (Again more on this later).

Analysis of CPSW data - comparison of two means

Table 5.1 in Chapter 5 shows the two sample *t*-test (with equal variances) and *MW* comparisons of the eight SF-36 dimension scores. If we assume a cut-off of $p \leq 0.05$ for statistical significance, then the *t*-test suggests significant differences on two dimensions of the SF-36, RP and SF. On two other dimensions PF ($p = 0.060$) and BP ($p = 0.065$) the *p*-values are close to the arbitrary cut-off of 0.05, suggesting some differences although these may not be statistically reliable. The results of the *MW* tests suggest significant differences on four dimensions (PF, RP, BP and SF) of the SF-36. The only major contrast between the interpretation of the results of the *MW* and *t* tests is on the BP and PF dimensions, where the former test suggests a difference and later not. Although Altman (1991) suggests that the *p*-value cut-off of 0.05 for statistical significance and rejecting the null hypothesis is arbitrary and it is ridiculous to interpret the results of a study differently according to whether the *p*-value obtained was, say 0.055 or 0.045.

Hypothesis testing with the bootstrap

As we mentioned in Chapter 3, bootstrap methods can also be used for hypothesis testing. The two quantities that we must choose when carrying out a bootstrap hypothesis test are a *test statistic* $t(\mathbf{x})$ and a *null distribution* $\hat{F}_0$ *for the data under the null hypothesis* (H_0). Given these, we generate *B* bootstrap values of the test statistic $t(\mathbf{x}^*)$ under the null distribution for the data $\hat{F}_0$ and estimate the *achieved significance level* (ASL) by calculating the proportion of the bootstrap values of the *B* test statistics $t(\mathbf{x}^*)$, which are greater than or equal to the observed value of the test statistic $t(\mathbf{x})_{obs}$ from the original data.

I.e.
$$ASL = \text{Prob}_{H_0}\left\{t\left(x^*\right) \geq t(x)_{obs}\right\}. \qquad (7.12)$$

Suppose we have samples $\mathbf{z} = (z_1, z_2,\ldots,z_n)$ and $\mathbf{y} = (y_1, y_2,\ldots,y_m)$ from possibly different probability distributions F and G respectively, and we wish to test the null hypothesis H_0: $F = G$. A bootstrap hypothesis test, like conventional hypothesis tests is based on a test statistic. To emphasize that a test statistic need not be an estimate of a parameter such as θ, we follow Efron and Tibshirani (1993), and denote the test statistic by $t(\mathbf{x})$. The quantity $t(\mathbf{x})$ is fixed at its observed value, t_{obs} and the random variable $\mathbf{x}$ has a distribution F_0 specified by the null hypothesis H_0.

If we let the combined sample of $\mathbf{z}$ and $\mathbf{y}$ be denoted by $\mathbf{x}$ and let the empirical distribution of this combined sample be $\hat{F}_0$, putting probability $1/(n + m)$ on each member of $\mathbf{x}$. Under H_0, $\hat{F}_0$ provides a non-parametric estimate of the common population that gave rise to both $\mathbf{z}$ and $\mathbf{y}$. Algorithm 7.1 derived from Efron and Tibshirani (1993) shows how the bootstrap test statistic and ASL is computed for testing $F = G$.

Algorithm 7.1

Computation of the bootstrap test statistic for testing $F = G$

1. Draw B samples of size $n + m$ with replacement from $\mathbf{x}$. Call the first n observations $\mathbf{z}^*$ and the remaining m observations $\mathbf{y}^*$.
2. Evaluate $t(.)$ on each sample,

$$t\left(x_b^*\right) = \bar{z}^* - \bar{y}^*, \; b = 1, 2,\ldots,B. \tag{7.13}$$

3. Approximate ASL_{boot} by

$$\hat{A}SL_{boot} = \frac{\#\left\{t\left(x_b^*\right) \geq t_{obs}\right\}}{B}, \tag{7.14}$$

where $t_{obs} = t(\mathbf{x})$ is the observed value of the statistic from the original data.

More accurate testing can be obtained through the use of a studentized statistic (Efron and Tibshirani, 1993). In the above test, instead of $t(x) = \bar{z} - \bar{y}$ we could use,

$$t(x)=\frac{\bar{z}-\bar{y}}{\bar{\sigma}\sqrt{\frac{1}{n}+\frac{1}{m}}}, \text{ where } \bar{\sigma}=\left\{\frac{\left[\sum_{i=1}^{n}(z_i-\bar{z})^2+\sum_{j=1}^{m}(y_j-\bar{y})^2\right]}{[n+m-2]}\right\}^{1/2}. \quad (7.15)$$

This is equivalent to the standard two-sample *t* test statistic (7.11) if we let $\bar{\sigma}$ = s_p. We shall denote the *ASL* from (7.15) by $ASL_{Student7.1}$.

The previous algorithm (7.1) tests the null hypothesis that the two populations are identical, that is *F* = *G*. Suppose we want to test only whether their means are equal. One approach would be to use the two sample *t* statistic. Under the null hypothesis and assuming Normal populations with equal variances, this has a Student's *t* distribution with *n* + *m* - 2 degrees of freedom. It uses a pooled estimate of the standard error $\bar{\sigma}$. If we are not willing to assume that the variances in the two populations are equal, we could base the test on

$$t(x)=\frac{\bar{z}-\bar{y}}{\sqrt{\frac{s_z^2}{n}+\frac{s_y^2}{m}}}, \quad (7.16)$$

where $s_z^2=\frac{\sum_{i=1}^{n}(z_i-\bar{z})^2}{n-1}$ and $s_y^2=\frac{\sum_{i=1}^{m}(y_i-\bar{y})^2}{m-1}$ are the estimated sample variances. With Normal populations, the quantity (7.16) no longer has a Student's *t* distribution and a number of approximate solutions have therefore been proposed (Satterthwaite, 1946; Welch, 1947). Using Satterthwaite's formula $t(x)$ in (7.16) is distributed Student's *t* with ν degrees of freedom, where ν is given by:

$$\nu=\frac{\left[\frac{s_z^2}{n}+\frac{s_y^2}{m}\right]^2}{\left[\frac{(s_z^2/n)^2}{n-1}+\frac{(s_y^2/m)^2}{m-1}\right]}. \quad (7.17).$$

The equal variances assumption is attractive for the *t*-test because it simplifies the form of the resulting distribution. In considering a bootstrap hypothesis for comparing the two means, there is no compelling reason to assume equal variances and so we do not make this assumption. To proceed we need estimates of *F* and *G* that use only the assumption of a common mean.

Letting $\bar{x}$ be the mean of the combined sample, we can translate both samples so that they have mean $\bar{x}$, and then resample each population separately. The procedure is shown in detail in Algorithm 7.2 taken from Efron and Tibshirani (1993).

Algorithm 7.2

Computation of the bootstrap test statistic for testing equality of means

1. Let F put equal probability on the points $\tilde{z}_i = z_i - \bar{z} + \bar{x}$, $i = 1, 2,\ldots,n$, and $\hat{G}$ put equal probability on the points $\tilde{y}_i = y_i - \bar{y} + \bar{x}$, $i = 1, 2,\ldots,m$, where $\bar{z}$ and $\bar{y}$ are the group means and $\bar{x}$ is the mean of the combined sample.
2. Form B bootstrap data sets (**z***, **y***) where **z*** is sampled with replacement from $\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_n$ and y* is sampled with replacement from $\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_m$.
3. Evaluate $t(.)$ defined by (7.16) on each data set,

$$t\left(x_b^*\right) = \frac{\bar{z}^* - \bar{y}^*}{\sqrt{\frac{s_z^{2*}}{n} + \frac{s_y^{2*}}{m}}} \qquad b = 1, 2,\ldots,B. \tag{7.18}$$

4. Approximate ASL_{boot} by

$$\hat{A}SL_{boot} = \frac{\#\left\{t\left(x_b^*\right) \geq t_{obs}\right\}}{B}, \tag{7.19}$$

where $t_{obs} = t(\mathbf{x})$ is the observed value of the statistic from the original sample.

Efron and Tibshirani (1993) state that $\hat{A}SL_{boot}$ has no interpretation as an exact probability, but like all bootstrap estimates is only guaranteed to be accurate as the sample size goes to infinity. On the other hand, the bootstrap hypothesis test does not require the special symmetry that is needed for a permutation test, and so can be applied much more generally. For instance in the two-sample problem a permutation test can only test the null hypothesis $F = G$, whilst the bootstrap can test equal means and equal variances or equal means with possibly unequal variances.

Table 7.3 shows the results of applying Algorithms 7.1 ($ASL_{mean7.1}$ and $ASL_{Student7.1}$) and 7.2 ($ASL_{boot7.2}$) to the CPSW data. Algorithm 7.1 was applied using SPSS v11 (SPSS, 2001) and Algorithm 7.2 in `STATA v8` (StataCorp, 2003), see Appendix 4 for more details and examples of the programs. It compares and contrasts the results of the *p*-values from three bootstrap hypothesis tests ($ASL_{mean7.1}$, $ASL_{Student7.1}$ and $ASL_{boot7.2}$) with the *p*-values from the standard two sample *t*-test with equal variances, the *MW* test and two sample *t*-test with unequal variances. The *p*-values from $ASL_{boot7.2}$ tests are very similar to the unequal variances *t*-test, which is to be expected as the two methods use the same test statistic (7.16).

The other hypothesis tests ($ASL_{mean7.1}$ and $ASL_{Student7.1}$) are effectively tests of the equality of distributions (i.e. $F = G$). Although they report quantitatively different *p*-values, the magnitudes are similar, and if we use a cut-off of $p < 0.05$ for statistical significance then the qualitative interpretation of the tests is the same. So in this example dataset there appears to be little advantage in using the bootstrap hypothesis tests compared to conventional hypothesis tests, such as the *MW* test, for testing equality of distributions.

Estimation and confidence intervals (CI)

A major limitation of non-parametric methods such as the *MW* and the bootstrap hypothesis tests (Algorithms 7.1 and 7.2) is that they do not allow for the estimation of confidence intervals for parameters or allow for the adjustment of confounding variables such as baseline covariates. Journals such as the *British Medical Journal* and *Lancet* now expect scientific papers to contain confidence intervals when appropriate. Indeed several statisticians have argued strongly for a change of emphasis in statistical analysis from hypothesis testing to estimation (Altman *et al* 2000).

One way to estimate non-parametric CIs is via the bootstrap method. Table 7.4 compares and contrasts the Normal *t*-test (equal variances) based confidence intervals with the bootstrap BC_a ones.

Table 7.3: CPSW Study Simple cross-sectional comparison of 6 week HRQoL for Control vs. Intervention Groups

SF-36 Dimension	*Group*	*n*	*mean*	*sd*	*Mean diff*	*t-test equal σ's P-value*	*MW test P-value*	*Unequal σ's t-test P-value*	$ASL_{boot7.2t}$	$ASL_{mean7.1}$	$ASL_{Student7.1}$
Physical Function	Control	241	89.9	14.5	2.6	0.060	0.015	0.059	0.057	0.030	0.028
	Intervention	254	87.3	15.8							
Role Physical	Control	241	74.3	38.1	9.1	0.009	0.004	0.009	0.010	0.005	0.006
	Intervention	254	65.2	39.5							
Bodily Pain	Control	241	75.6	23.7	4	0.065	0.040	0.065	0.062	0.028	0.029
	Intervention	254	71.6	23.8							
General Health	Control	241	77.7	17.7	2.4	0.139	0.147	0.139	0.131	0.071	0.073
	Intervention	254	75.3	18.5							
Vitality	Control	241	51.1	20.7	1.3	0.498	0.596	0.497	0.494	0.248	0.249
	Intervention	254	49.8	21.7							
Social Function	Control	241	81.6	22.7	4.7	0.025	0.015	0.025	0.024	0.014	0.014
	Intervention	254	76.9	24.2							
Role Emotional	Control	241	77.9	36.4	1.1	0.734	0.503	0.734	0.737	0.369	0.369
	Intervention	254	76.8	35.5							
Mental Health	Control	241	72.9	17.2	-0.2	0.902	0.972	0.902	0.904	0.554	0.553
	Intervention	254	73.1	16.7							

$ASL_{boot7.2}$, $ASL_{mean7.1}$, and $ASL_{Student7.1}$ based on 5000 bootstrap replications.

Both sets of confidence intervals were estimated using the `bootstrap` procedure in `STATA v8` (StataCorp, 2003). According to Efron and Tibshirani (1993) each interval $(\hat{\theta}_{lo}, \hat{\theta}_{up})$ can be described by its *length* and *shape*,

$$length = \hat{\theta}_{up} - \hat{\theta}_{lo}, \text{ and } shape = \frac{\hat{\theta}_{up} - \hat{\theta}}{\hat{\theta} - \hat{\theta}_{lo}}. \qquad (7.20).$$

'Shape' measures the symmetry of the interval about the point estimate $\hat{\theta}$. Shape > 1.00 indicates greater distance from $\hat{\theta}_{up}$ to $\hat{\theta}$ than from $\hat{\theta}$ to $\hat{\theta}_{lo}$. Conversely, Shape < 1.00 indicates a greater distance from $\hat{\theta}$ to $\hat{\theta}_{lo}$ than from $\hat{\theta}_{up}$ to $\hat{\theta}$. The standard Normal (or *t*-distribution based) intervals are symmetrical about $\hat{\theta}$, having shape = 1.00 by definition. Therefore shape is a measure of skewness of the CI about the point estimate. A shape > 1.00, implies the CI is 'positively' skewed, with a long 'tail' to the right. Whereas shape, < 1.00 implies the CI is 'negatively' skewed.

The estimates and lengths of the CIs are almost identical. Table 7.4 also shows that the shape of the BC_a CIs is almost symmetric about the point estimate of the mean difference except for the RE dimension, where there is some evidence of asymmetry. So again in this example dataset there appears little advantage in using the bootstrap BC_a confidence intervals compared to conventional methods of confidence interval estimation.

The bootstrap (and Normal) confidence intervals are calculated for a characteristic of the distributions (for example mean difference). The groups may have differences in distributions but similar characteristics e.g. means. (Morrell *et al* 2000). For example the *MW* tests suggests a significant difference (in distributions) for the PF, RP, BP and SF dimensions, but the bootstrap and Normal confidence limits for two out of four of these dimensions (PF and BP) includes zero suggesting no differences in the mean HRQoL between the groups.

Table 7.4: Comparisons of parametric and bootstrap estimates of confidence intervals for the eight dimensions of the SF-36 from the CPSW Study for Control vs. Intervention Groups

SF-36 Dimension		*Mean Difference* $\hat{\theta}$	*CIs Lower* $\hat{\theta}_{lo}$	*Upper* $\hat{\theta}_{up}$	*Interval Length*	*Shape*
Physical	Normal (*t*-test)	-2.6	-5.2	0.1	5.4	1.00
Function	Bootstrap BC_A		-5.2	0.0	5.2	0.98
Role	Normal (*t*-test)	-9.1	-16.0	-2.3	13.7	1.00
Physical	Bootstrap BC_A		-15.8	-2.3	13.5	1.02
Bodily	Normal (*t*-test)	-4.0	-8.2	0.2	8.4	1.00
Pain	Bootstrap BC_A		-8.1	0.3	8.4	1.03
General	Normal (*t*-test)	-2.4	-5.6	0.8	6.4	1.00
Health	Bootstrap BC_A		-5.6	0.8	6.4	0.99
Vitality	Normal (*t*-test)	-1.3	-5.0	2.5	7.5	1.00
	Bootstrap BC_A		-5.1	2.4	7.5	0.98
Social	Normal (*t*-test)	-4.7	-8.9	-0.6	8.3	1.00
Function	Bootstrap BC_A		-8.7	-0.6	8.1	1.03
Role	Normal (*t*-test)	-1.1	-7.5	5.3	12.7	1.00
Emotional	Bootstrap BC_A		-7.1	5.6	12.7	1.11
Mental	Normal (*t*-test)	0.2	-2.8	3.2	6.0	1.00
Health	Bootstrap BC_A		-2.8	3.2	6.0	0.98

Mean difference $\hat{\theta}$ *= Intervention mean - Control mean*

BCa confidence intervals based 5000 bootstrap replications.

When a hypothesis is tested using the bootstrap, the resampling is carried out assuming the null hypothesis H_0 is true. Whereas when confidence intervals for mean differences between two groups are estimated the resampling is carried out separately for each group. A useful analogy is with the comparison of proportions in two independent groups. Here the standard error for the hypothesis test is different to the standard error of the difference between the observed proportions used for estimating a confidence interval (Altman, 1991).

Adjusting for other variables

If we suspect that the observed differences (imbalance) between the groups at the start of the trial may have affected the outcome we can take account of the imbalance in the analysis. Adjusting for other variables requires the use of *analysis of covariance* (ANCOVA) or some form of *multiple regression* analysis. For the rest of this chapter we will be concerned with *linear* regression in which the *mean* of the HRQoL response *Y* observed at a value or vector **x** of *p* explanatory variables or covariates ($x_1, x_2,\ldots,x_p$) is:

$$E(Y \mid x_1, x_2, \ldots, x_p) = \mu(x) = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p. \qquad (7.21)$$

The model is completed by specifying the nature of the random variation, which for independent responses amounts to specifying the form of the variance var($Y \mid \boldsymbol{x}$). The multiple linear regression model is:

$$Y_i = \beta_0 x_{0i} + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon_i \qquad i = 1,\ldots,n, \qquad (7.22)$$

where for models with an intercept $x_{0j} = 1$.

The multiple regression model assumes:

(1) The relationship between the outcome variable (*Y*) and the predictor variables ($x_1, x_2,\ldots,x_p$) is linear.

(2) The variability of *Y*, as assessed by the variance or standard deviation σ, corresponding to a particular set of values $x_1, x_2,\ldots,x_p$ is the same, regardless of $x_1, x_2,\ldots,x_p$.

(3) The values of the outcome variable *Y* should have a Normal distribution for each set of values of the predictor variables $x_1, x_2,\ldots,x_p$

The regression coefficients are estimated using the ordinary least squares (OLS) method (McCullagh and Nelder, 1989; Dobson, 1990).

Multiple regression does not involve any assumptions about the distribution of the *x* values (Armitage *et al* 2002). If the above three assumptions hold then the residuals ε_is should be uncorrelated, Normally distributed with zero mean and have the same variance σ^2 throughout the range of fitted values (i.e. $\varepsilon_i \sim N(0, \sigma^2)$).

These assumptions can be checked graphically using histograms, Normal plots and scatter diagrams of the residuals from the model. Assumption (1) can be checked by plotting the residuals against each of the predictor variables in turn. We expect to see no association if the true relationship is linear. A plot of the residuals against the fitted or predicted values can assess assumption (2). No pattern should be discernable. In particular, the variability of the residuals should be constant across the range of the fitted values. For assumption (3) we can produce a Normal plot of the residuals, which should fall on a straight line if the residuals are Normally distributed.

In general, lack of Normality of the residuals is unlikely to affect seriously the estimates of a regression equation although Campbell and Machin, (1999), have pointed out that it may effect the standard errors (and hence confidence interval estimates) and the size of the *p*-value. Similarly, a lack of constant variance of the residuals is unlikely to seriously affect the estimates, but again will have some influence on the final *p*-value and confidence intervals. In either case Campbell and Machin's advice is to proceed with caution, particularly if the *p*-value is close to some critical value such as 0.05. The lack of linearity is more serious, and would suggest either a transformation of the data before fitting the regression, or a model involving quadratic (squared) or higher terms using multiple regression.

For linear regression with Normal random errors having constant variance, the least squares theory of regression estimation and inference provides clean exact methods for analysis (Davison and Hinckley, 1997). But for generalisations to non-Normal errors and non-constant variance, exact methods rarely exist, and we are faced with approximate methods based on linear approximations to estimators and CLTs. So, just as in the simpler context of hypothesis testing, resampling methods have the potential to provide more accurate analysis.

Bootstrap regression analysis.

Standard errors and confidence intervals for regression coefficients can also be obtained using bootstrap methods. However, as we mentioned in Chapter

3, two different approaches are possible, *case* and *model (residual)* resampling.

Algorithm 7.3

Case-based resampling in linear regression

With the simple linear model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$; with $E(\varepsilon_i) = 0$, if the data are **w** = $(x_1, y_1), (x_2, y_2),\ldots,(x_n, y_n)$.

1. Case-based resampling involves drawing a bootstrap sample of size *n*, with replacement from these *n* pairs. The bootstrap data set is of the form:

$$\mathbf{w}^* = \{(y_{i_1}^*, x_{i_1}^*), (y_{i_2}^*, x_{i_2}^*), \ldots, (y_{i_n}^*, x_{i_n}^*)\},$$

where $i_1, i_2, \ldots, i_n$ is a random sample of integers 1 through *n*.

2. Ordinary least squares is then used to estimate the regression coefficients $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$, for this bootstrap sample of paired cases.
3. We do this repeatedly, say *B* times, so we now have *B* bootstrap samples and *B* estimates of the regression coefficients, one from each bootstrap sample, $\{(\hat{\beta}_0^*, \hat{\beta}_1^*)_1, (\hat{\beta}_0^*, \hat{\beta}_1^*)_2, \ldots, (\hat{\beta}_0^*, \hat{\beta}_1^*)_B\}$.
4. The standard error of these estimated coefficients $\mathrm{se}(\hat{\beta}_0)$ and $\mathrm{se}(\hat{\beta}_1)$ is simply the standard deviation of these *B* estimates. As before if these estimates are ordered in increasing value, $\{(\hat{\beta}_1^*)_{(1)}, (\hat{\beta}_1^*)_{(2)}, \ldots, (\hat{\beta}_1^*)_{(B)}\}$, a simple 95% bootstrap percentile confidence interval for the coefficient would be from the $0.025B^{th}$ to the $0.975B^{th}$ largest values.

For multiple regression a *case* or "pair" in step 1 of the above algorithm consists of a vector $(y_i, x_{1i}, x_{2i},\ldots,x_{pi})$ of the response variable y_i and *p* covariates for the i^{th} case.

Case-based resampling may be entirely natural for situations where is it plausible that the (*x*, *y*) pairs have been drawn from a bivariate distribution function *F (x, y)* of the pairs. However, case based resampling is less appealing if the *x* values were controlled for in some way, perhaps by the design of the study. In this situation the alternative *model or residual* based procedures could be used.

Algorithm 7.4

Model or residual based resampling in linear regression

For model based resampling using the simple linear regression model $\{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, with $E(\varepsilon_i) = 0\}$.

1. Conventional fitted values y_i^{fit} and residuals e_i are first obtained from the observed data using ordinary least squares estimation i.e. $y_i^{fit} = \hat{\beta}_0 + \hat{\beta}_1 x_i$ and $e_i = y_i^{obs} - y_i^{fit} = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$.
2. A bootstrap sample of the residuals is drawn $\mathbf{e}^* = \left(e_{i_1}^*, e_{i_2}^*, ..., e_{i_n}^*\right)$, where $i_1, i_2, \ldots, i_n$ is a random sample of integers 1 through n.
3. The bootstrap sample for the regression $\mathbf{z}^* = \left(y_i^*, x_i^*\right)$ comprises the x values $\left(x_i^* = x_i\right)$ from the original data and y values computed by adding the fitted values to the bootstrap residuals i.e. $y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_{i_n}^*$. Note that $\hat{\beta}_0$ and $\hat{\beta}_1$ are the coefficient estimates from the original sample. The bootstrap data set is of the form
$\mathbf{z}^* = \left\{\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + e_{i_1}^*, x_1\right), \left(\hat{\beta}_0 + \hat{\beta}_1 x_2 + e_{i_2}^*, x_2\right), ..., \left(\hat{\beta}_0 + \hat{\beta}_1 x_n + e_{i_n}^*, x_n\right)\right\}$,
where $i_1, i_2, \ldots, i_n$ is a random sample of integers 1 through n.
4. Ordinary least squares is then used to estimate the regression coefficients $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$, for this bootstrap sample.
5. As before the process is repeated B times to estimate standard errors and confidence intervals for the B regression coefficients from the bootstrap samples.

For the multiple regression, the bootstrap sample in step 3 of the above algorithm consists of a vector $\mathbf{z}^* = \left(y_i^*, x_{1i}^*, x_{2i}^*, ..., x_{pi}^*\right)$ and comprises the x values $\left(x_{1i}^* = x_{1i}, ..., x_{pi}^* = x_{pi}\right)$ from the original data and y values computed by adding the fitted values to the bootstrap residuals.

The model based resampling is an example of the "*parametric bootstrap*" when the residuals from a parametric model are bootstrapped to give estimates of the standard error of the parameters. There is considerable

debate about which form of resampling is more appropriate (see Table 7.5a below).

Both Algorithms (7.3 and 7.4) can be easily implemented in S-PLUS 2000 (MathSoft, 1999) using S-PLUS functions from the bootstrap library developed by Davison and Hinckley (1997). Appendix 4 provides examples of the S-PLUS programs for case and residual resampling and estimation of BC_a confidence intervals.

Adjusted Analysis of CPSW data

For the CPSW study we thought that age, parity and delivery might affect HRQoL. The analysis involved using OLS to fit the linear regression model below,

$$Y_i = \beta_0 + \beta_{Age} x_{Age\_i} + \beta_{Parity} x_{Parity\_i} + \beta_{Delivery} x_{Delivery\_i} + \beta_{Group} x_{Group\_i} + \varepsilon_i . \qquad (7.23)$$

Equation (7.23) has the six week postnatal HRQoL score as the dependent variable, Y_i and x_{Age_i} (age in years); x_{Parity_i} (parity, coded 0 for first child and 1 for second or subsequent child); $x_{Delivery\_i}$ (type of delivery, coded 0 = normal and 1 = abnormal) and treatment group x_{Group_i} (coded 0 = Control, 1 = Intervention) as covariates. The term β_0 is the intercept or constant and ε_i is a $N(0, \sigma^2)$ random error term.

The regression coefficient estimate, $\hat{\beta}_{Group}$, represents the difference in six week HRQoL between the Intervention and Control groups after adjustment for the mother's age, parity and type of delivery. A positive value for the regression coefficient indicates the Intervention group has a better mean HRQoL at six weeks postnatally than the Control group after adjustment for the other covariates.

Table 7.5a: Linear Regression: Model (residual) vs. Case resampling

	Model (residual)	Case
Efron and Tibshirani 1993	Which bootstrap resampling method is better? The answer depends on how far we trust the linear model $\{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, with $E(\varepsilon_i) = 0\}$. This model says that the error between y_i and its mean $\mu_i = \beta_0 + \beta x_i$ does not depend on x_i; it has the same distribution "F" no matter what x_i may be. This is a strong assumption, which can fail even if the model for the expectation $E(y_i) = \mu_i = \beta_0 + \beta x_i$ is correct. The reverse argument also holds! Model $\{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$; with $E(\varepsilon_i) = 0\}$ does not have to hold perfectly in order for bootstrapping residuals to give reasonable results. When we bootstrap residuals, the bootstrap data sets $z^* = \{(x_1, y_1^*), (x_2, y_2^*), \ldots, (x_n, y_n^*)\}$ have covariate vectors $x_1, x_2, \ldots, x_n$ exactly the same as those for the actual data set X. Even when covariates are generated randomly, there are reasons to do the analysis as if they are fixed. Regression coefficients have larger Standard Errors (SE) when the covariates have smaller standard deviation. By treating the covariates as fixed constants we obtain a SE that reflects the precision associated with the sample of covariates actually observed. However as E & T show the difference between x_i fixed and x_i random usually does not affect the SE estimate very much.	Bootstrapping cases is less sensitive to assumptions than bootstrapping residuals. The Standard Error estimated by bootstrapping cases, gives reasonable answers even if the linear model is completely wrong. The only assumption behind case-based resampling is that the original "pairs" $v_i = (x_i, y_i)$ were randomly sampled from some distribution F, where F is a distribution on $(p+1)$-dimensional vectors (X, y). Even if the linear model $\{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$; with $E(\varepsilon_i) = 0\}$ is correct, it is no disaster to bootstrap cases; it can be shown that the answer given by case-based resampling approaches that given by model based resampling as the number of "pairs" *n* grows large.
Davison and Hinckley 1997		Two important differences between case and model. First, with case resampling we make no assumption about variance homogeneity. Indeed we do not even assume that the conditional mean of Y given X = x is linear. This offers the advantage of potential robustness to heteroscedasticity, and the disadvantage of inefficiency if the constant variance model is correct. Secondly, the simulated samples have different designs, because the values of $x_1^*, \ldots, x_n^*$ are randomly sampled. The design fixes the information content of a sample, and in principle our inference should be specific to the information in our data. The variation in $x_1^*, \ldots, x_n^*$ will cause some variation in information, but fortunately this is often not important in moderately large datasets.

Figures 7.4 and 7.5 show the residual plots for the eight dimensions of the SF-36 after using OLS to regress SF-36 dimension score on age, parity, delivery and treatment group. The right hand column of the figures shows a Normal probability plot of standardised residuals. Only on the Vitality dimension is the Normal plot reasonably straight. The Normal plots for the other seven dimensions indicate the distribution of the residuals departs somewhat from Normality particularly for the two Role dimensions.

The plots of residuals against the fitted response (left hand column of Figures 7.4 and 7.5) show the variability of the residuals may not be constant for some dimensions of the SF-36. For example, for the PF, BP and SF dimensions there is some suggestion that the variability of the residuals is decreasing as the fitted values increase. Therefore, if the distribution of errors is very far from Normal or heteroscedastic (unequal variances), then standard OLS results may not be reliable and resampling methods may offer a genuine improvement, particularly case resampling which is robust to the failure of the model assumptions (Davison and Hinckley, 1997).

Table 7.5b compares the OLS and bootstrap standard errors and confidence interval estimates for the treatment group coefficient from the CPSW data. All models include age, delivery and parity as covariates in the regression. For the bootstrap methods the standard errors are the standard deviations of the coefficients from the 5000 bootstrap re-samples. For ease of interpretation and comparison only the estimates for the treatment group coefficient are shown. As can be seen from Table 7.5b the standard error estimates are almost identical for the three methods. Similarly the length of the confidence intervals is virtually the same for all three methods. The bootstrap intervals have a tendency to be non-symmetric around the point estimate of the regression coefficient. Qualitatively all of the intervals from the three methods either include or exclude zero so the interpretation of the treatment group regression coefficient is the same. Therefore in this example dataset there appears to be little advantage in using bootstrap case or model based re-sampling to estimate standard errors and confidence intervals compared to

conventional methods of confidence interval estimation from the OLS multiple regression model.

Figure 7.4: Residual plots from CPSW data (left column Residuals against predicted response; right column Normal plot of residuals)

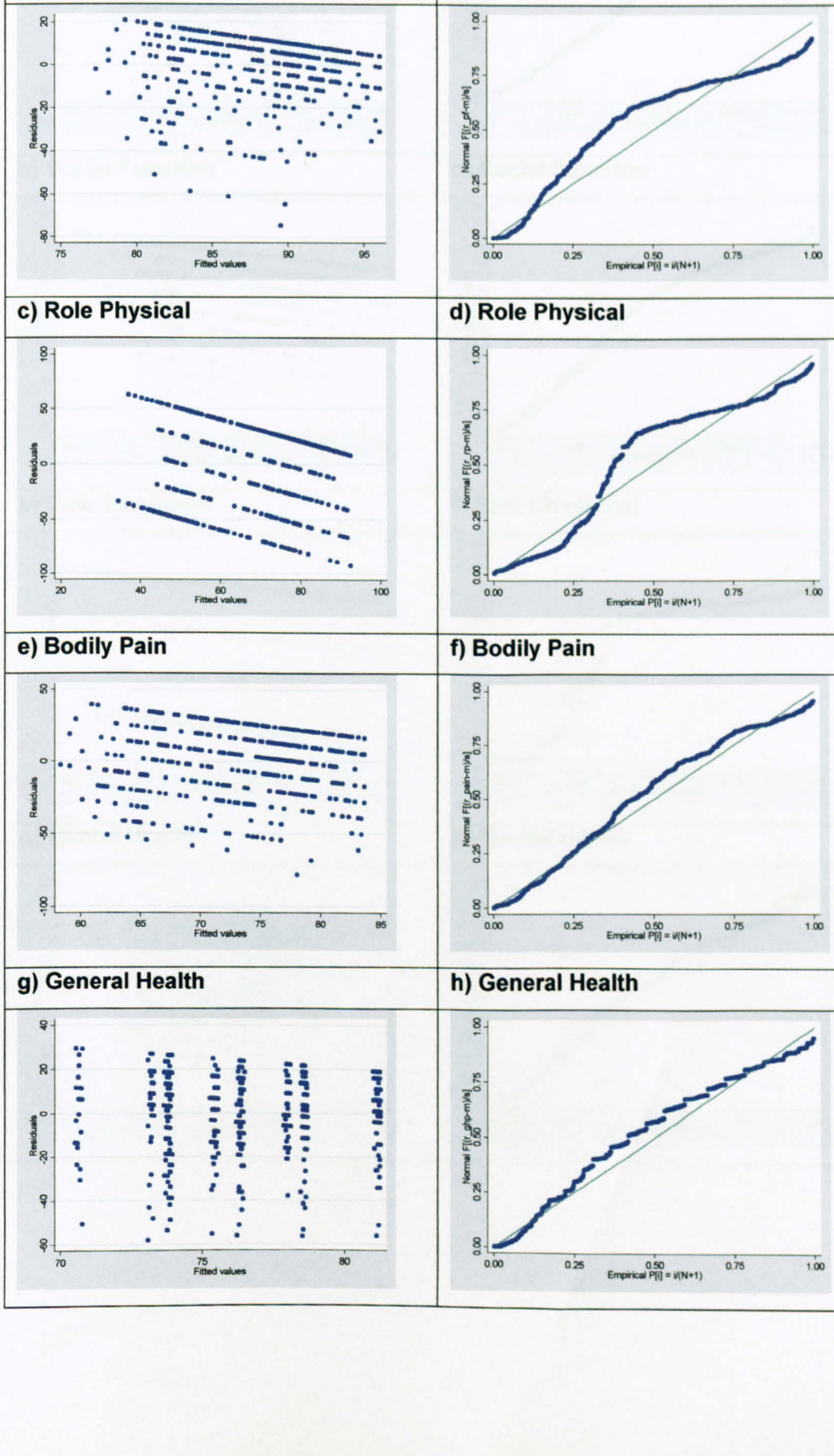


Figure 7.5: Residual plots from CPSW data (left column Residuals against predicted response; right column Normal plot of residuals)

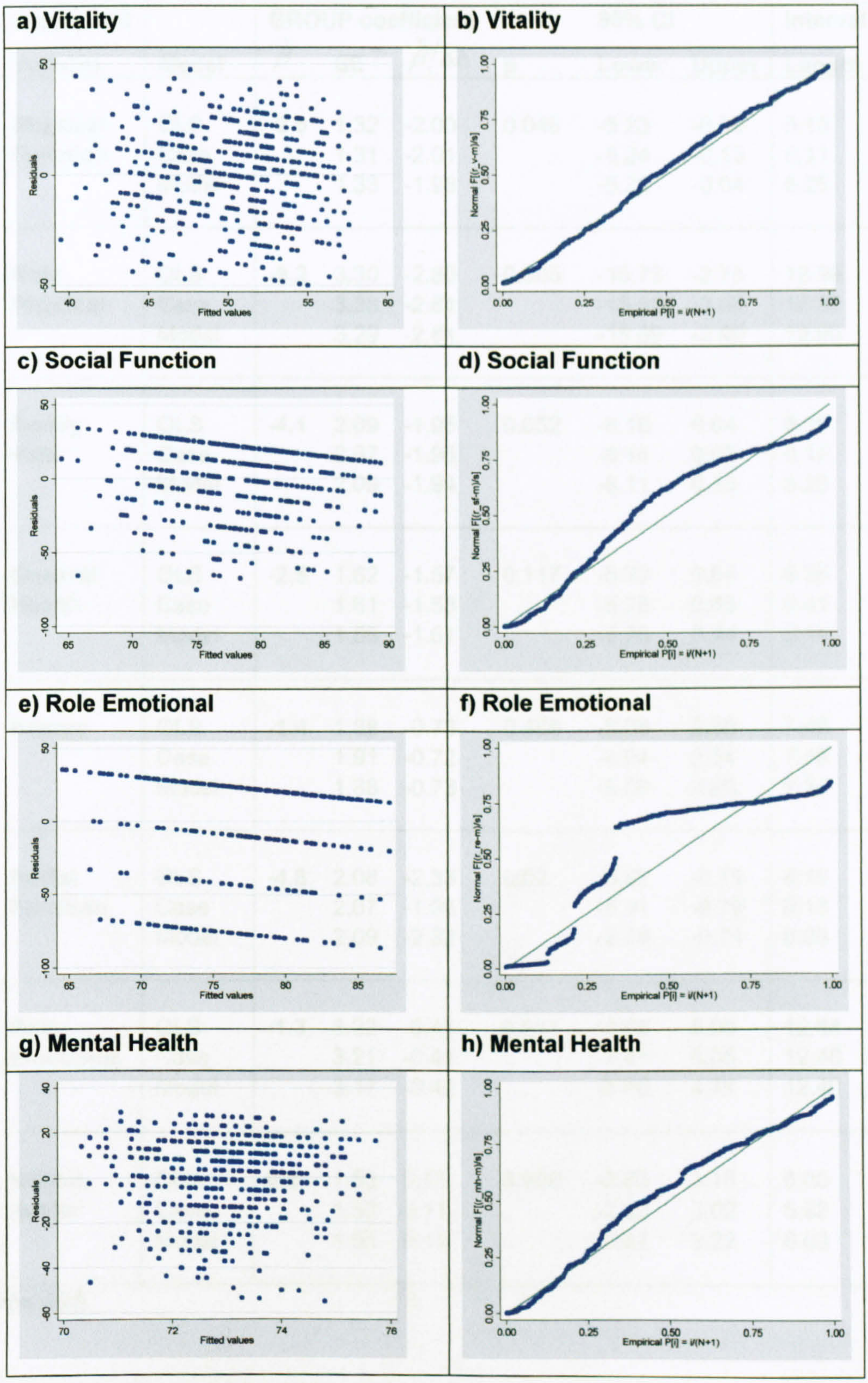


Table 7.5b: Multiple regression, case and model based resampling SE's and Confidence Intervals estimates from the CPSW data

Dependent Variable	Model	GROUP coefficient $\hat{\beta}$	SE	$\hat{\beta}/SE$	p	95% CI Lower	95% CI Upper	Interval Length	Interval Shape
Physical Function	OLS	**-2.6**	1.32	-2.00	0.046	-5.23	-0.05	5.19	1.00
	Case		1.31	-2.01		-5.24	-0.13	5.11	0.96
	Model		1.33	-1.98		-5.29	-0.04	5.25	0.98
Role Physical	OLS	**-9.2**	3.30	-2.80	0.005	-15.73	-2.75	12.98	1.00
	Case		3.28	-2.81		-15.86	-3.06	12.80	0.93
	Model		3.29	-2.81		-15.89	-2.90	12.99	0.95
Bodily Pain	OLS	**-4.1**	2.09	-1.95	0.052	-8.16	0.04	8.19	1.00
	Case		2.07	-1.96		-8.14	0.03	8.17	1.00
	Model		2.09	-1.94		-8.11	0.15	8.26	1.04
General Health	OLS	**-2.5**	1.62	-1.57	0.117	-5.72	0.64	6.36	1.00
	Case		1.61	-1.58		-5.76	0.65	6.41	0.99
	Model		1.58	-1.61		-5.70	0.44	6.14	0.95
Energy	OLS	**-1.4**	1.88	-0.73	0.465	-5.08	2.33	7.40	1.00
	Case		1.91	-0.72		-4.94	2.54	7.48	1.10
	Model		1.88	-0.73		-5.08	2.26	7.34	0.98
Social Function	OLS	**-4.8**	2.08	-2.33	0.02	-8.92	-0.75	8.16	1.00
	Case		2.07	-1.96		-8.91	-0.78	8.13	0.99
	Model		2.09	-2.32		-8.79	-0.74	8.05	1.03
Role Emotional	OLS	**-1.3**	3.22	-0.41	0.681	-7.64	5.00	12.64	1.00
	Case		3.21	-0.41		-7.41	5.05	12.46	1.05
	Model		3.17	-0.42		-7.66	4.74	12.40	0.96
Mental Health	OLS	**0.2**	1.53	0.11	0.909	-2.83	3.18	6.00	1.00
	Case		1.53	0.11		-2.90	3.02	5.92	0.92
	Model		1.53	0.11		-2.81	3.22	6.03	1.02

N= 495.

Changes from baseline

Although it is common to use the patients' status at the end of the study period as the outcome of interest, sometimes it is more appropriate to take the

change (or difference) from the pre-treatment, or baseline, measurement as the prime outcome measure.

When changes from baseline are analysed it is misleading to perform separate analyses (either hypothesis tests or CIs) within each treatment group. A better approach is to calculate each patient's change from baseline, and then compare directly the changes in the different groups.

Analysis of changes from baseline

Given the distribution of the changes or differences in outcome measures are more likely to be symmetric and Normally distributed, parametric tests can be used to compare differences in changes between groups, especially if we assume that for a seven-point HRQoL scale going from 2 to 1 is the same as a change from 7 to 6 say. Parametric CIs for means and their differences can then be calculated. Multiple regression can be used (with change from baseline as the dependent *Y* variable) to compare groups and adjust for other covariates and factors (such as baseline HRQoL, age, sex and treatment centre).

Bajorski and Petkau (1999) describe a non-parametric method of comparing changes from baseline on ordinal responses for two independent groups. (This method is based on performing several *MW* tests on the follow-up up HRQoL scores stratified by baseline HRQoL score. These separate *MW* test statistics for each strata are then weighted and summed together to produce an overall test statistic). Unfortunately, this method is purely a hypothesis test and does not allow estimation of CIs and so does not appear particularly useful for the analysis of HRQoL data. Furthermore, Sullivan and D'Agostino (2003) have demonstrated the robustness and power of the two sample *t*-test on change scores (and ANCOVA) when applied to ordinal scales of three, four, or five points in a clinical trials setting comparing two treatments, with sample sizes as small as 20 per group. This suggests that a simple *t*-test on change scores may be more suitable than a non-parametric test.

Frison and Pocock (1992) demonstrate that ANCOVA is the method of choice for analysis of pre-treatment (baseline) and post-treatment (follow-up) means. They show that ANCOVA is superior to both analysis of post-treatment means and analysis of mean changes. Therefore we will concentrate on such methods for HRQoL outcomes from the OA Knee data, which measured HRQoL at baseline and 6 months follow-up in two groups of patients (Surgical and Rheumatology).

Analysis of the OA Knee Data

Table 7.6 shows the baseline socio-demographic and HRQoL characteristics of the two groups of OA patients those awaiting total knee replacement surgery (Surgical) and those having pharmacological treatment (Rheumatology). The group of patients awaiting surgery is significantly older and has significantly more men than the Rheumatology group. The Surgical group has significantly lower levels of PF prior to total knee replacement surgery than the Rheumatology group. Conversely the Surgical group has significantly higher levels of GH, V and MH compared to the Rheumatology clinic patients. For the other four dimensions of the SF-36 (RP, BP, SF and RE) there was no evidence of any difference in HRQoL between the two groups.

We were interested in seeing whether or not there was a difference in HRQoL in OA patients after TKR surgery compared with pharmacologically treated patients. From previous studies using the SF-36 (Brazier *et al* 1992; Walters *et al* 2001c) we know that HRQoL varies with age and gender. Since there was a difference in the baseline HRQoL and socio-demographic characteristics (age and gender) of the Rheumatology clinic and TKR surgery groups, we use this dataset to illustrate multiple regression/ANCOVA methods with follow-up HRQoL as the outcome variable and baseline HRQoL, age, gender and group (TKR surgery or Rheumatology clinic) as covariates.

Therefore the analysis involved using OLS to fit the multiple regression model below,

Table 7.6: Baseline characteristics of the TKR Surgery and Rheumatology Clinic patients from the OA Knee study.

	Rheumatology			Surgical			Mean	95% CI		P-value
	N	Mean	SD	N	Mean	SD	Diff	Lower	Upper	
Age (years)	102	64.2	(11.3)	109	71.1	(8.5)	-6.9	-9.6	-4.2	0.001
SF-36 Dimensions										
Physical Function	97	28.2	(22.4)	95	21.2	(18.2)	7.0	1.2	12.8	0.019
Role Physical	96	11.5	(22.0)	99	12.9	(26.3)	-1.4	-8.3	5.4	0.684
Bodily Pain	100	32.0	(19.5)	104	36.3	(23.4)	-4.3	-10.3	1.6	0.154
General Health	94	43.9	(22.9)	96	57.3	(23.8)	-13.3	-20.0	-6.6	0.001
Vitality	98	36.9	(19.0)	99	42.3	(19.3)	-5.4	-10.8	0.0	0.050
Social Function	100	53.1	(30.6)	101	53.6	(27.6)	-0.5	-8.6	7.6	0.910
Role Emotional	95	41.1	(44.2)	99	44.1	(44.6)	-3.1	-15.6	9.5	0.632
Mental Health	99	62.7	(20.9)	100	68.2	(18.8)	-5.5	-11.0	0.1	0.054
Gender										
Female	71		(69.6%)	59		(54.1%)	(15.5%)	(2.4%)	(27.8%)	0.021
Male	31		(30.4%)	50		(45.9%)				
Total	102		(100%)	109		(100%)				

$$Y_i = \beta_0 + \beta_{Age} x_{Age\_i} + \beta_{Sex} x_{Sex\_i} + \beta_{Base} x_{Base\_i} + \beta_{Group} x_{Group\_i} + \varepsilon_i, \qquad (7.24)$$

where, Y_i is the six month follow-up HRQoL for subject *i*; x_{Age_i} is the age in years at baseline; x_{Sex_i} is the gender of the patient (coded 0 for males and 1 for females); x_{Base_i} is the baseline HRQoL and x_{Group_i} is the treatment group variable (coded 0 = Clinic, 1 = Surgery). The term β_0 is a constant and ε_i is a $N(0, \sigma^2)$ random error term. Again the regression coefficient estimate, $\hat{\beta}_{Group}$, represents the difference in six-month follow-up HRQoL between the Rheumatology Clinic and TKR Surgery groups after adjustment for the

patient's age, gender and baseline HRQoL. A positive value for the regression coefficient indicates the Surgery group has a better mean HRQoL at six months follow-up than the Clinic group after adjustment for the other covariates.

Figures 7.6 and 7.7 show the residual plots for the eight dimensions of the SF-36 after using OLS to regress SF-36 dimension score at 6 months follow-up on age, gender, baseline HRQoL and group. The right hand column of the figures shows a Normal probability plot of standardised residuals. Only six out of eight of the dimensions the Normal plots are reasonably straight. The Normal plots for the two Role dimensions (RP and RE) indicate the distribution of the residuals departs somewhat from Normality.

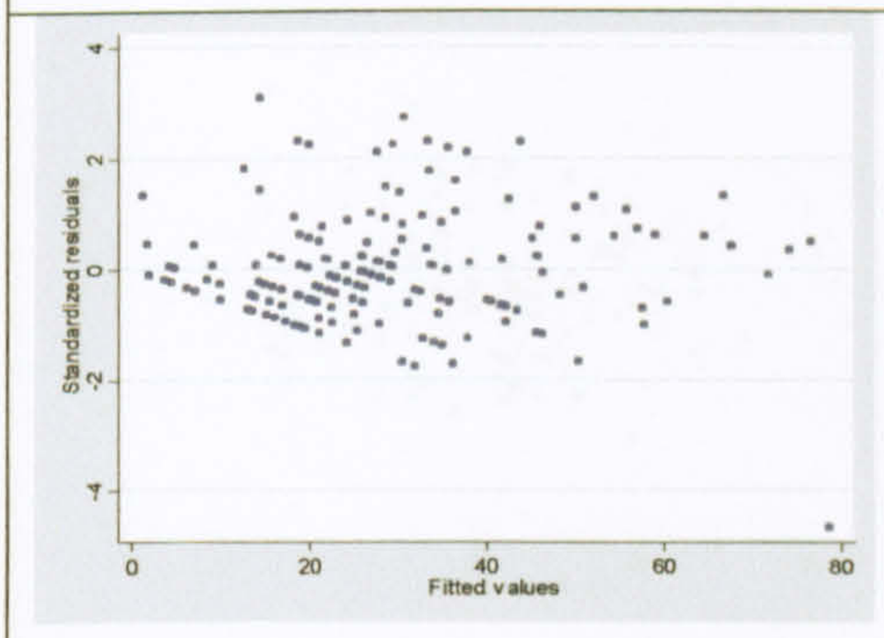
The plots of residuals against the fitted response (left hand column of Figures 7.6 and 7.7) show the variability of the residuals may not be constant for the two Role dimensions of the SF-36. For the other six dimensions, no pattern in the residuals against the fitted values is discernable.

Table 7.7 compares the OLS and bootstrap standard errors and confidence interval estimates for the group coefficient from the OA Knee data. All models include age, baseline HRQoL and gender as covariates in the regression. For the bootstrap methods the standard errors are the standard deviations of the coefficients from the 5000 bootstrap re-samples. For ease of interpretation and comparison only the estimates for the group coefficient are shown.

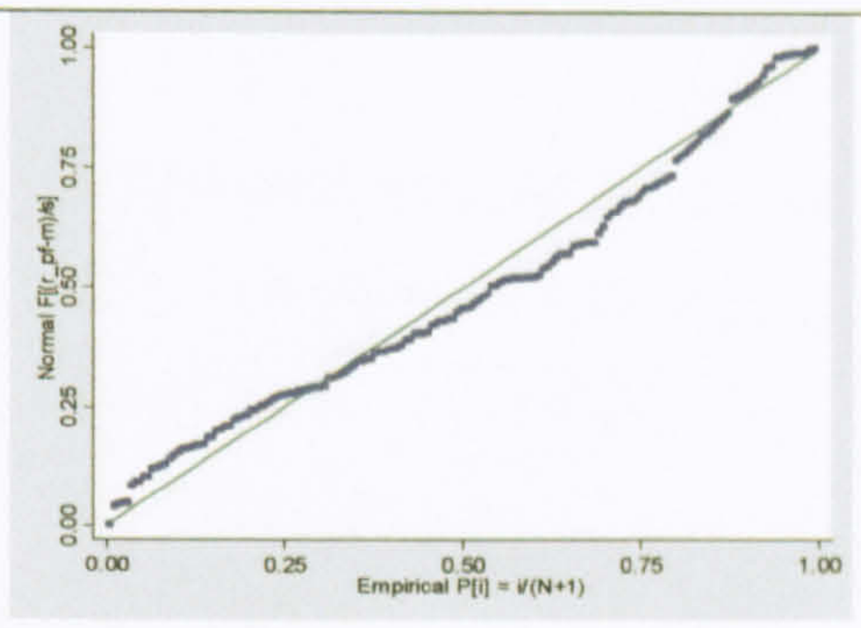
The regression analysis suggests that at six month follow-up TKR surgical patients have significantly better HRQoL than Rheumatology treated clinic patients on five dimensions of the SF-36 (PF, BP, GH, V and SF) after adjustment for age, gender and baseline HRQoL. As can be seen from Table 7.7 the standard error estimates are almost identical for the three methods. Similarly the length of the confidence intervals is virtually the same for all three methods. Although the bootstrap CIs tend to be asymmetric about the point-estimate of the regression coefficient.

Figure 7.6: Residual plots from OA Knee data (left column Residuals against predicted response; right column Normal probability plot)

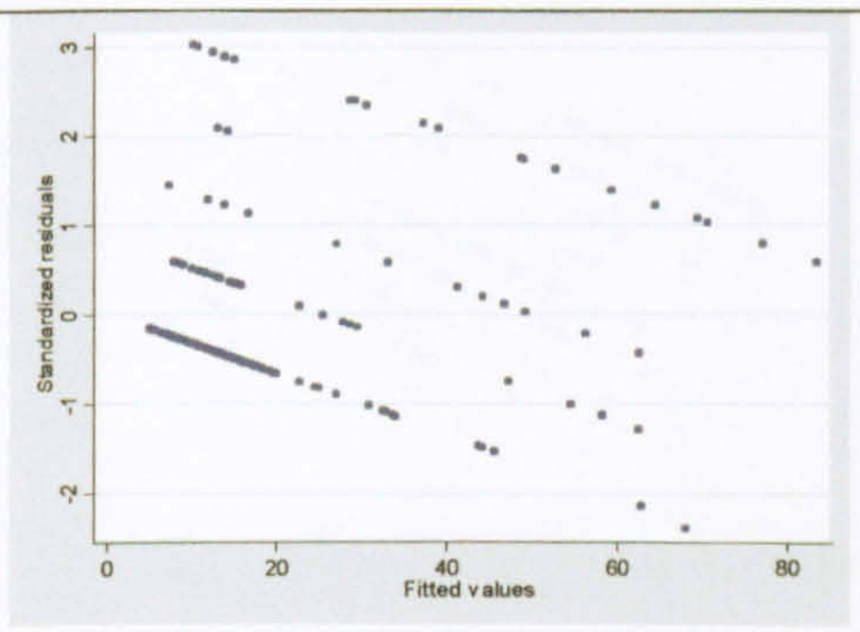
a) Physical Function



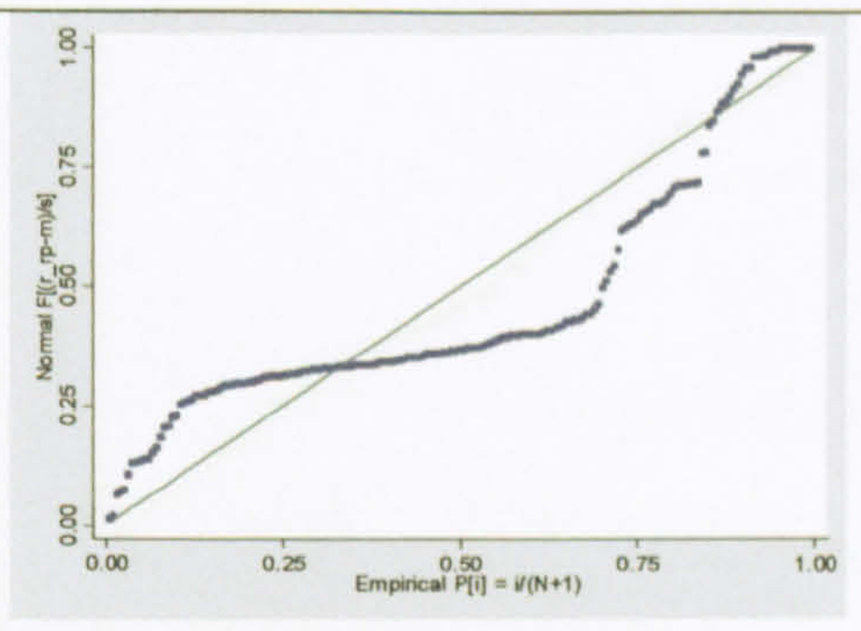
b) Physical Function



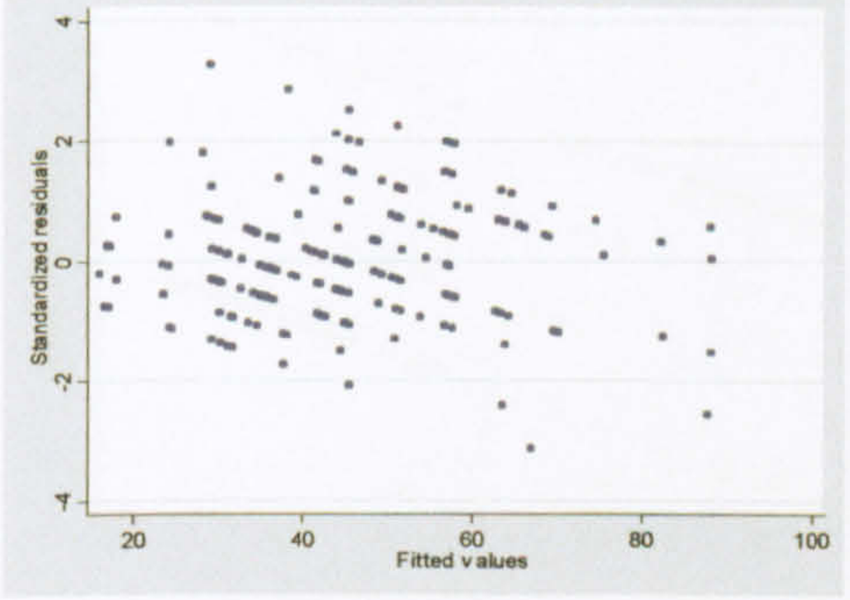
c) Role Physical



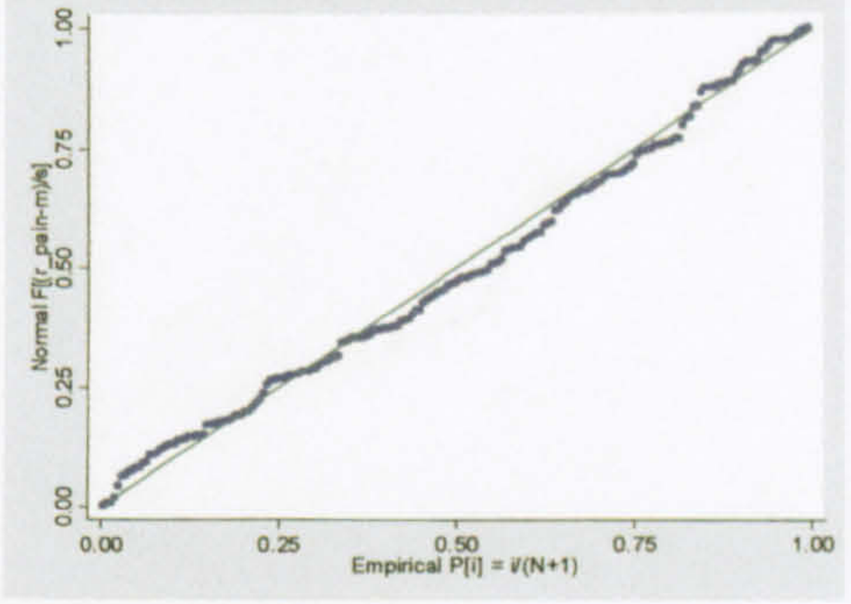
d) Role Physical



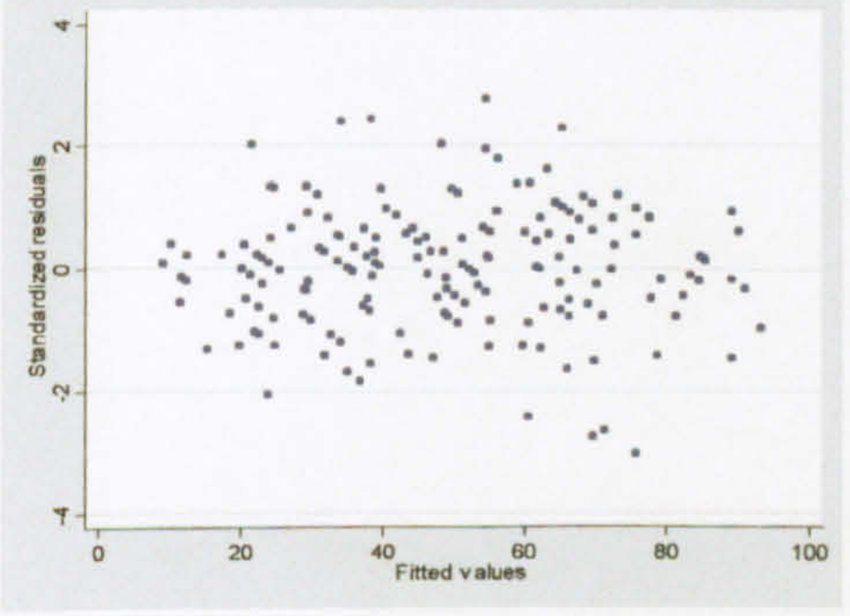
e) Bodily Pain



f) Bodily Pain



g) General Health



h) General Health

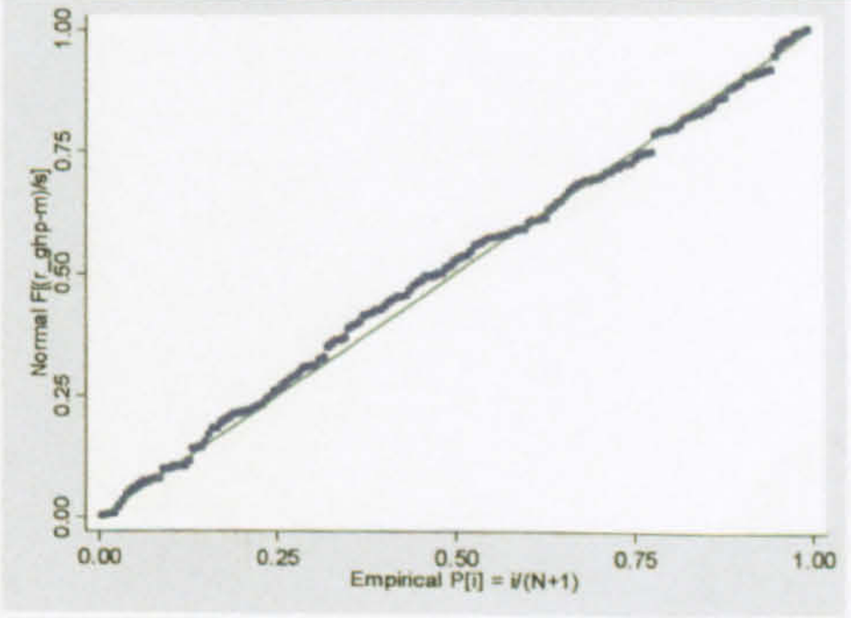
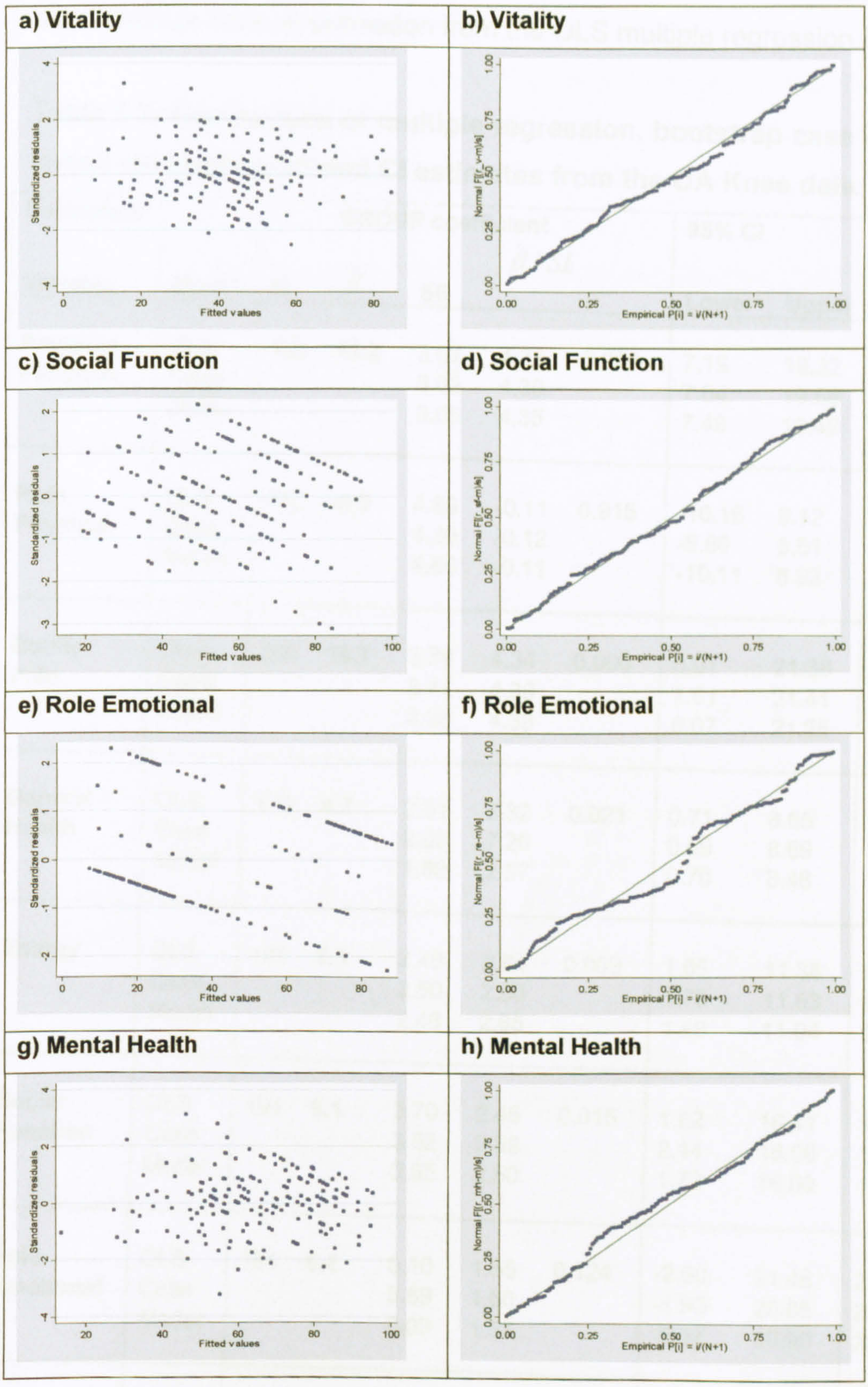


Figure 7.7: Residual plots from OA Knee data (left column Residuals against predicted response; right column Normal probability plot)



Qualitatively all of the intervals from the three methods either include or exclude zero so the interpretation of the group regression coefficient is the same. Therefore again in this example dataset there appears to be little

advantage in using bootstrap case or model based re-sampling to estimate standard errors and confidence intervals compared to conventional methods of confidence interval estimation from the OLS multiple regression model.

Table 7.7: Comparison of multiple regression, bootstrap case and model based resampling SE and CI estimates from the OA Knee data

Dependent Variable	Model	GROUP coefficient N	$\hat{\beta}$	SE	$\hat{\beta} / SE$	p	95% CI Lower	Upper	Interval Length	Shape
Physical Function	OLS	165	**13.3**	3.07	4.31	0.001	7.19	19.32	12.14	1.00
	Case			3.02	4.39		7.64	19.69	12.05	1.15
	Model			3.05	4.35		7.49	19.49	12.00	1.08
Role Physical	OLS	177	**-0.5**	4.89	-0.11	0.915	-10.16	9.12	19.29	1.00
	Case			4.39	-0.12		-8.60	8.51	17.11	1.12
	Model			4.86	-0.11		-10.11	8.93	19.04	0.99
Bodily Pain	OLS	200	**14.7**	3.39	4.34	0.000	8.01	21.38	13.37	1.00
	Case			3.41	4.30		7.81	21.41	13.60	0.98
	Model			3.36	4.38		8.07	21.25	13.18	0.99
General Health	OLS	173	**4.7**	2.01	2.32	0.021	0.71	8.65	7.95	1.00
	Case			2.03	7.26		0.69	8.69	8.00	1.01
	Model			1.98	2.37		0.70	8.48	7.78	0.95
Energy	OLS	185	**6.5**	2.46	2.64	0.009	1.65	11.36	9.72	1.00
	Case			2.50	2.60		1.75	11.63	9.88	1.08
	Model			2.46	2.65		1.49	11.04	9.54	0.90
Social Function	OLS	194	**9.1**	3.70	2.46	0.015	1.82	16.41	14.59	1.00
	Case			3.52	2.59		2.14	16.06	13.92	1.00
	Model			3.65	2.50		1.72	16.09	14.37	0.94
Role Emotional	OLS	184	**9.4**	6.10	1.55	0.124	-2.60	21.48	24.08	1.00
	Case			5.89	1.60		-1.90	20.85	22.75	1.01
	Model			6.03	1.57		-2.37	20.90	23.27	0.97
Mental Health	OLS	191	**1.1**	2.15	0.51	0.613	-3.15	5.33	8.48	1.00
	Case			2.32	0.47		-3.54	5.42	8.96	0.94
	Model			2.14	0.51		-3.17	5.12	8.28	0.95

It should be noted that in the OA knee dataset there is a considerable amount of missing data. Two hundred and two patients were assessed at the six-month follow-up, but for example only 165/202 (82%) completed all ten items of the SF-36 PF dimension. Missing data is a common problem with HRQoL assessment data, particularly in longitudinal studies. The imputation of missing HRQoL scores and the analysis of HRQoL data with missing values is extensively discussed in several papers (Curran *et al* 1998; Fayers *et al* 1998; Troxel *et a*l 1998) and book chapters (Fayers and Machin, 2000 Chapter 11; Fairclough, 2002 Chapters 4 to 8). As the subject of this thesis is the use of computer intensive methods in analysing HRQoL data and not missing value imputation we will not discuss the issue of missing values further.

Summary and conclusions

In this chapter we have shown how the bootstrap can be used for simple hypothesis testing (Algorithms 7.1 and 7.2) and more complex multiple linear regression analysis of cross-sectional HRQoL data and HRQoL data with a baseline and follow-up assessment (Algorithms 7.3 and 7.4). In the dataset studied, hypothesis testing with the bootstrap appears to offer no advantage over conventional significance tests such as the *t*-test and *MW* test. Similarly, in the two datasets studied, both the case and model based bootstrap re-sampling methods for estimating SEs and CIs for linear regression models gave estimates almost identical to the conventional values estimated by OLS. In the next chapter we will look at methods of analysing HRQoL data collected at three or more time points including the simple analysis of summary measures and the more complex modelling of longitudinal HRQoL data. Again we will compare conventional estimates of SE and CI for parameters with their bootstrap counterparts.

Chapter 8: Modelling Longitudinal HRQoL data and summary measures (three or more time points) using the bootstrap

Introduction: Analysis of Longitudinal HRQoL data

With one HRQoL observation on each experimental unit (e.g. the CPSW study), we are confined to modelling the population average of *Y*, called the *marginal* mean response; there is no other choice. With repeated HRQoL measurements, there are several different approaches that can be adopted. I will split these approaches into three broad classifications (Everitt, 2002):

(1) Time by time analysis;

(2) Response feature analysis – the use of summary measures;

(3) Modelling of longitudinal data.

The modelling of longitudinal data takes into account the fact that successive HRQoL assessments by a particular subject are likely to be correlated. The models are an extension of the linear regression model described in Chapter 7. The three alternative modelling approaches I am going to discuss are repeated measures ANOVA, marginal (General Estimating Equations) models and random-effect (multi-level) models. All three models require the specification of the *auto-* or *serial correlation*, which is the strength of the association between successive longitudinal measurements of a single HRQoL variable on the same patient.

In both the marginal and random-effect approaches we model both the dependence of the HRQoL response on the explanatory variables and the autocorrelation among the responses. With cross-sectional data, (as in Chapter 7), only the dependence of *Y* on *x* need be specified; there is no correlation as the responses (*Y*'s) are independent. If we choose to ignore the correlation that exists in longitudinal data then Diggle *et al* (2002) mention three important consequences.

(1) Incorrect inferences about regression coefficients, β;

(2) Estimates of β which are inefficient, that is less precise than possible;

(3) Sub-optimal protection against biases caused by missing data.

Notation

We follow Diggle *et al* (2002)'s notation, and let Y_{ij} represent the HRQoL response variable and $\boldsymbol{x}_{ij}$ a row vector of length p of explanatory variables observed at time t_{ij}, for observation $j = 1,\ldots,n_i$ on subject $i = 1,\ldots,m$. The mean and variance of the HRQoL responses Y_{ij} are represented by $E(Y_{ij}) = \mu_{ij}$ *and* $Var(Y_{ij}) = \upsilon_{ij}$. The set of repeated HRQoL outcomes for subject i are collected into a n_i-vector, $\boldsymbol{Y}_i = \left(Y_{i1}, Y_{i2},\ldots, Y_{in_i}\right)$, with mean $E(\boldsymbol{Y}_i) = \mu_i$ and n_i x n_i covariance matrix $Var(\boldsymbol{Y}_i) = V_i$, where the jk^{th} element of V_i is the covariance between Y_{ij} and Y_{ik}, denoted by $Cov(Y_{ij}, Y_{ik}) = \upsilon_{ijk}$. We shall use R_i for the n_i x n_i (*auto*-) correlation matrix of $\boldsymbol{Y}_i$. The responses for all patients are referred to as $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2,\ldots, \boldsymbol{Y}_m)$, which is an N-vector with length $N = \sum_{i=1}^{m} n_i$.

The longitudinal analyses we will consider will be based on an extension of the linear regression model (7.21),

$$Y_{ij} = \beta_1 x_{ij1} + \beta_2 x_{ij2} + \ldots. + \beta_p x_{ijp} + \varepsilon_{ij}, \qquad (8.1)$$

$$= \boldsymbol{x}_{ij}^{T}\beta + \varepsilon_{ij}$$

where $\beta^T = (\beta_1, \beta_2,\ldots,\beta_p)$ is a (p x 1) vector of unknown regression coefficients and ε_{ij} is a zero-mean random variable representing the deviation from the model prediction, $\boldsymbol{x}_{ij}^{T}\beta$. ($\boldsymbol{x}^T$ denotes the transpose of the matrix). Typically, x_{ij1} = 1 for all i and j, and β_1 is then the intercept term in the linear model.

In matrix notation, the regression equation for the i^{th} subject takes the form

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \varepsilon_i, \qquad (8.2)$$

where $\boldsymbol{X}_i$ is a n_i x p matrix with $\boldsymbol{x}_{ij}$ in the j^{th} row and $\varepsilon_i = \left(\varepsilon_{i1}, \varepsilon_{i2},\ldots, \varepsilon_{in_i}\right)$.

Autocorrelation

If Y_{i1} and Y_{i2} represent the values of two successive HRQoL assessments by the same (i^{th}) patient and m represents the total number of patients completing both assessments in the sample. Then equation (8.3) measures the strength of association or *auto-correlation* between successive longitudinal measurements of HRQoL on the same patient,

$$r_T(1,2) = \frac{\sum_{i=1}^{m}(Y_{i1} - \bar{Y}_1)(Y_{i2} - \bar{Y}_2)}{\sqrt{\sum_{i=1}^{m}(Y_{i1} - \bar{Y}_1)^2 \sum_{i=1}^{m}(Y_{i2} - \bar{Y}_2)^2}}, \qquad (8.3)$$

where $\bar{Y}_1$ and $\bar{Y}_2$ are the sample mean HRQoL scores at times t_1 and t_2 respectively. (This is equivalent to Pearson's product moment correlation coefficient.)

Suppose HRQoL is assessed on numerous occasions and the measurements at different times are Y_{i0}, Y_{i1},..., Y_{iT} for patient i at time T in the study. Then equation (8.3) can be utilised, one pair at a time, with the respective Y_{ij} replacing the Y values. If there are assessments at T + 1 time-points, there will be $(T + 1)T/2$ pairs of assessments leading to separate auto-correlation coefficients. For example, for T = 6 there are (7 x 6)/2 = 21 auto-correlation coefficients that may be calculated.

In the NAMEIT study HRQoL assessment was carried out at 0, 8, 16, 24, 32, 40 and 48 weeks (i.e. 6 + 1 time-points). Table 8.1 summarises the resulting 21 auto-correlation pairs for the assessments until week 48. The pattern of the observed auto-correlation matrix R, gives a guide to the so-called error structure associated with the successive HRQoL measurements. Table 8.1 shows that the autocorrelation coefficients range between 0.19 and 0.85. For three dimensions of the SF-36, PF, GH and MH, the autocorrelation coefficients are moderately large (between 0.5 and 0.85). The pattern of values suggests decreasing correlation as the observations become further apart.

Several underlying patterns of the auto-correlation matrix R are used in the modelling of HRQoL data. The error structure is *independent* (sometimes termed *random*) if the off diagonal terms of the auto-correlation matrix R are zero. The repeated HRQoL observations on the same subject are then independent of each other, and can be regarded as though they were observations from different individuals.

Table 8.1 Auto-correlation matrices for the eight dimensions of the SF-36 from RA patients in the NAMEIT study assessed at seven time points

a) Physical Function (n = 218)

Week	0	8	16	24	32	40	48
0	1.00						
8	0.61	1.00					
16	0.63	0.74	1.00				
24	0.57	0.69	0.75	1.00			
32	0.56	0.68	0.80	0.79	1.00		
40	0.55	0.67	0.77	0.81	0.86	1.00	
48	0.53	0.64	0.74	0.81	0.81	0.85	1.00

b) Role Physical (n = 212)

Week	0	8	16	24	32	40	48
0	1.00						
8	0.40	1.00					
16	0.35	0.53	1.00				
24	0.29	0.39	0.57	1.00			
32	0.19	0.30	0.56	0.67	1.00		
40	0.34	0.42	0.52	0.60	0.61	1.00	
48	0.27	0.40	0.59	0.67	0.64	0.71	1.00

c) Bodily Pain (n = 219)

Week	0	8	16	24	32	40	48
0	1.00						
8	0.43	1.00					
16	0.45	0.55	1.00				
24	0.44	0.47	0.61	1.00			
32	0.37	0.46	0.51	0.68	1.00		
40	0.40	0.42	0.57	0.60	0.69	1.00	
48	0.42	0.46	0.59	0.63	0.68	0.76	1.00

d) General Health (n = 209)

Week	0	8	16	24	32	40	48
0	1.00						
8	0.55	1.00					
16	0.56	0.68	1.00				
24	0.60	0.67	0.80	1.00			
32	0.58	0.67	0.77	0.83	1.00		
40	0.59	0.65	0.72	0.79	0.84	1.00	
48	0.58	0.65	0.75	0.84	0.82	0.85	1.00

e) Vitality (n = 216)

Week	0	8	16	24	32	40	48
0	1.00						
8	0.55	1.00					
16	0.48	0.58	1.00				
24	0.47	0.54	0.71	1.00			
32	0.50	0.59	0.68	0.71	1.00		
40	0.42	0.49	0.67	0.68	0.77	1.00	
48	0.47	0.53	0.66	0.72	0.72	0.76	1.00

f) Social Function (n = 219)

Week	0	8	16	24	32	40	48
0	1.00						
8	0.44	1.00					
16	0.43	0.53	1.00				
24	0.39	0.55	0.63	1.00			
32	0.36	0.46	0.63	0.70	1.00		
40	0.38	0.51	0.58	0.64	0.71	1.00	
48	0.34	0.45	0.58	0.64	0.71	0.71	1.00

g) Role Emotional (n = 206)

Week	0	8	16	24	32	40	48
0	1.00						
8	0.46	1.00					
16	0.35	0.47	1.00				
24	0.34	0.40	0.59	1.00			
32	0.31	0.32	0.56	0.62	1.00		
40	0.34	0.46	0.53	0.56	0.54	1.00	
48	0.31	0.37	0.49	0.58	0.54	0.69	1.00

h) Mental Health (n = 218)

Week	0	8	16	24	32	40	48
0	1.00						
8	0.57	1.00					
16	0.57	0.62	1.00				
24	0.55	0.59	0.72	1.00			
32	0.52	0.55	0.65	0.69	1.00		
40	0.50	0.54	0.70	0.70	0.74	1.00	
48	0.56	0.55	0.68	0.72	0.73	0.77	1.00

On the other hand, if all the correlations are approximately equal or *uniform* then the matrix of correlation coefficients is termed *exchangeable*, or *compound symmetric.* This means that we can re-order (exchange) the successive observations in any way we choose in our data file without affecting the pattern in the correlation matrix.

Frequently, as the time or lag between successive observations increases, the auto-correlation between the observations decreases. Thus, we would expect a higher auto-correlation between HRQoL assessments made only two days apart than between two HRQoL assessments made one month apart. This is in contrast to the uniform correlation model above. In such a situation one may postulate that the relationship between the size of the correlation and the "lag", that is the time t_j and t_k may be of the form

$$\rho_T(t_j, t_k) = \rho^{\phi|t_j - t_k|}. \qquad (8.4)$$

The $|t_j - t_k|$ implies that if the difference between t_j and t_k is negative the sign should be ignored and ϕ takes a constant value that is usually less than one. A correlation matrix of this form is said to have an *autoregressive structure* (sometimes called *multiplicative* or *time series*). Diggle (*et al* 2002) refers to this as the *exponential correlation model*.

The auto-correlation pattern affects the way in which the computer packages estimate the regression coefficients in the corresponding statistical model, and so it should be chosen with care.

We will concentrate on longitudinal data analysis problems where the regression of Y on x is the scientific focus and the number of experimental units (m) or patients in our case is much greater than the number of observations per unit (n). However, before we get into a more detailed discussion of the three longitudinal models (repeated measures ANOVA, marginal, and random-effect), I will briefly describe two simpler methods of analysing longitudinal HRQoL data as outlined in the introduction: the time-by-time analysis; and response feature analysis (the use of summary measures).

1. Time by time analysis

A series of two independent samples *t*-tests (or the non-parametric equivalent) are used to test for differences between the two groups at each time point. (In examples with more than two groups, a series of one-way ANOVAs might be used). The procedure is straightforward but has a number

of serious flaws and weaknesses (Everitt, 2001). Consequently, it will not be pursued further here.

2. Response feature analysis – the use of summary measures

Here the repeated measures for each participant are transformed into a single number considered to capture some important aspect of the participant's response. A simple and often effective strategy (Diggle *et al* 2002) is to:

(1) Reduce the repeated values into one or two summaries.

(2) Analyse each summary as a function of covariates (x_i).

Diggle *et al* (2002) call this strategy a *two-stage* or *derived variable* analysis, and mention that it works well when $x_{ij} = x_i$ for all i and j (i.e. the important explanatory variables do not change over time), since the summary value which results from stage (1) can only be regressed on x_i in stage (2).

Table 8.2: Response features suggested in Matthews *et al* (1990).

Type of data	**Property to be compared between groups**	**Summary measure**
Peaked	Overall value of response	Mean or Area Under the Curve
Peaked	Value of most extreme response	Maximum (minimum)
Peaked	Delay in response	Time to maximum or minimum
Growth	Rate of change of response	Linear regression coefficient
Growth	Final level of response	Final value or (relative) difference between first and last
Growth	Delay in response	Time to reach a particular value

Examples of summary measures include the Area Under the Curve (AUC) or the overall mean of post-randomisation measures. Other possible summary measures are listed in Matthews *et al* (1990) and are shown in Table 8.2. Having identified a suitable summary measure, a simple *t*-test (or ANOVA) can be applied to assess between group differences. If the data for each patient can effectively be summarised by a pre-treatment mean and a post-

treatment mean, then the ANCOVA is the preferred method of choice (Frison and Pocock, 1992). It is superior to both analysis of post treatment means or analysis of mean changes. Diggle *et al* (2002) suggested that provided the data are complete, then the method of derived variables or summary measures can give a simple and easily interpretable analysis with a strong focus on particular aspects of the mean response.

In lieu of reducing the repeated HRQoL responses to summary statistics, we can model the individual Y_{ij} in terms of $\boldsymbol{x}_{ij}$. The next section will discuss three distinct strategies in analysing the repeated HRQoL responses.

3. Modelling of longitudinal data

(i) Repeated measures Analysis of Variance (ANOVA)

In some situations HRQoL assessments may be made over a limited period rather than over an extended time span. In this case in may be reasonable to assume that all the subjects complete all the assessments. Thus instead of having a fragmented data file with the number of observations for each subject varying from subject to subject, the file has a regular or rectangular shape. This enables the repeated measures ANOVA approach to be considered (Fayers and Machin, 2000).

Diggle *et al* (2002) say that ANOVA has limitations that prevent its recommendation as a general approach for longitudinal data. The first is that it fails to exploit the potential gains in efficiency from modelling the covariance among repeated observations. A second limitation is that, ANOVA methods usually require a complete balanced array of data. As Fayers and Machin (2000) point out this is the main difficulty with repeated measures ANOVA, in HRQoL research, since there are seldom equal numbers of QoL assessments recorded per patient. It is therefore better to use a regression modelling approach rather than repeated measures ANOVA for analysing longitudinal HRQoL data.

Diggle *et al* (2002) point out that the use of repeated measures ANOVA implies an exchangeable auto-correlation between any two observations on

the same patient. This may not always be appropriate for HRQoL assessments.

(ii) Marginal Models - Generalised Estimating Equations

The second strategy is to model the marginal mean as in a cross-sectional study. Since repeated values are not likely to be independent, this marginal analysis must also include assumptions about the form of the correlation. For example, in the linear model we can assume that $E(\boldsymbol{Y}_i) = \boldsymbol{X}_i\beta$, and $Var(\boldsymbol{Y}_i) = V_i(\alpha)$ where β and α must be estimated. The marginal model approach has the advantage of separately modelling the mean and covariance. Valid inferences about β can sometimes be made even when an incorrect form for the $V(\alpha)$ is assumed.

Marginal models are appropriate when inferences about the population average are the focus. For example, in a clinical trial the average difference between control and treatment is most important, not the difference for any one individual. In a marginal model, the regression of the response on explanatory variables is modelled separately from the within-person correlation. In regression, we model the marginal expectation, $E(Y_{ij})$, as a function of explanatory variables. By marginal expectation, we mean the average response over the sub-population that shares a common value of x. The marginal expectation is what we modelled in the analysis of the cross-sectional studies in the preceding chapter.

A marginal model has the following assumptions (Diggle *et al* 2002):

(1) the marginal expectation of the response, $E(Y_{ij}) = \mu_{ij}$, depends on explanatory variables $\boldsymbol{x}_{ij}$, by $h(\mu_{ij}) = \boldsymbol{x}_{ij}^T\beta$ where h is a known *link function* such as the logit for binary responses or log for counts;

(2) the marginal variance depends on the marginal mean according to $Var(Y_{ij}) = v(\mu_{ij})\phi$ where v is a known variance function and ϕ is a scale parameter which may need to be estimated;

(3) the correlation between Y_{ij} and Y_{ik} is a function of the marginal means and perhaps of additional parameters α, that is Corr(Y_{ij}, Y_{ik}) = $\rho(\mu_{ij}, \mu_{ik}; \alpha)$ where $\rho(.)$ is a known function.

The marginal regression coefficients, β, have the same interpretation as the coefficients from a cross-sectional analysis. Consider a simple linear regression model for HRQoL over time for a group of hospital patients following surgery say,

$$Y_{ij} = \beta_1 + \beta_2 t_{ij} + \varepsilon_{ij}, \tag{8.5}$$

where t_{ij} is the time of the QoL assessment, say in months post surgery, of patient i at visit j, Y_{ij} is the HRQoL at time t_{ij} post surgery and ε_{ij} is a mean-zero residual. Since patients' HRQoL will not all improve (or deteriorate) at the same rate, the residuals $\varepsilon_{i1}, \varepsilon_{i2}, ..., \varepsilon_{in_i}$, for patient i are likely to be correlated with one another. The marginal modelling approach is to assume

(1) $E(Y_{ij}) = \beta_1 + \beta_2 t_{ij}$;

(2) $Corr(\varepsilon_{ij}, \varepsilon_{ik}) = \rho(t_{ij}, t_{ik}; \alpha)$.

Assumption (1) is that the average HRQoL for all patients in the population at any time t is $\beta_1 + \beta_2 t_{ij}$. The parameter β_2 is therefore the change per month in the population-average HRQoL. Assumption (2) specifies the nature of the autocorrelation; a specific simple example might be that

$$Corr(\varepsilon_{ij}, \varepsilon_{ik}) = \rho(t_{ij}, t_{ik}; \alpha) = \alpha_0, \tag{8.6}$$

I.e. a constant, so that the auto-correlation matrix is exchangeable or compound symmetric. In the marginal approach, we separate the modelling of the regression and the correlation either can be changed without necessarily changing the other (Diggle *et al* 2002).

A more complex example would be to consider the same linear regression but this time let the errors follow an autoregressive structure introduced in (8.4). Again the mean response is $E(Y_{ij}) = \beta_1 + \beta_2 t_{ij}$. The covariance structure is now

given by $Cov(Y_{ij}, Y_{ik}) = \sigma^2 \exp(-\phi|t_j - t_k|)$ and the variance is assumed to be independent of the mean.

The marginal modelling approach uses *Generalized Estimating Equations* (GEEs) to estimate the regression coefficients (Liang and Zeger, 1986). In the GEE approach any required covariance structure and any link function may be assumed and the parameters estimated without specifying the joint distribution of the repeated observations. Estimation is via a multivariate analogue of a *quasi-likelihood* approach (Wedderburn, 1974). We briefly outline this approach in Appendix 5. In the marginal modelling approach, we only need to specify the first two moments of the responses for each person (i.e. the mean and variance). With Normal data, the first two moments fully determine the likelihood, but this is not the case for other members of the *Generalized Linear Model* (GLM) family (Diggle *et al* 2002).

Since the parameters specifying the structure of the correlation matrix are rarely of great practical interest (they are what is known as *nuisance parameters*), simple structures (e.g. exchangeable or 1st order autoregressive) are used for the within subject correlations giving rise to the so called *working correlation matrix*. Liang and Zeger (1986) show that the estimates of the parameters of most interest, i.e. those that determine the mean profiles over time, are still valid even when the correlation structure is incorrectly specified.

Diggle *et al* (2002) conclude that the GEE method of estimation enjoys two useful and important properties:

(1) $\hat{\beta}$ is nearly efficient relative to the maximum likelihood estimates of β in many practical situations provided that Var($\mathbf{Y}_i$) has been reasonably approximated.

(2) $\hat{\beta}$ is consistent as $m \to \infty$, even if the covariance of $\mathbf{Y}_i$ is incorrectly specified.

When the regression coefficients are the scientific focus, as in the examples in this chapter, one should invest the lion's share of time in modelling the mean

structure, while using a reasonable approximation to the covariance. The robustness of the inferences about β can be checked by fitting a final model using different covariance assumptions and comparing the two sets of estimates and their robust standard errors. If they differ substantially, a more careful treatment of the covariance model may be necessary (Diggle *et al* 2002).

The process of fitting marginal models using GEE begins by assuming the simple independence form for the autocorrelation matrix *R*, and fitting the model as if each assessment were from a different patient. Once this model is obtained the corresponding residuals are calculated and these are then used to estimate the autocorrelation matrix assuming it is of the exchangeable (or autoregressive) type. This matrix is then used to fit the model again, the residuals are once more calculated, and the autocorrelation matrix obtained. The iteration process is repeated until the corresponding regression coefficients that are obtained in the successive models converge or differ little on successive occasions (Fayers and Machin, 2000).

(iii) Random effects models

A third approach, the random effects model, assumes that the correlation arises among repeated responses because the regression coefficients vary across individuals. Here we model the conditional expectation of Y_{ij} given the person-specific coefficients, β_i, by $E(Y_{ij}|\beta_i) = \boldsymbol{x}^T_{ij}\beta_i$.

Since there are too little data on a single person to estimate β_i from $(\boldsymbol{Y}_i, \boldsymbol{X}_i)$ alone, we further assume that the β_i's are independent realisations from some distribution with mean β. If we write $\beta_i' = \beta + \boldsymbol{U}_i$ where β is fixed and $\boldsymbol{U}_i$ is a zero-mean random variable, then the basic heterogeneity assumption can be restated in terms of the latent variables, $\boldsymbol{U}_i$. That is, there are unobserved factors represented by the $\boldsymbol{U}_i$'s that are common to all responses for a given person but vary across people, thus indicating the correlation. Random effects models are particularly useful when inferences are to be made about individuals, rather than the population average.

The advantage of the random effects model is that there are fewer regression parameters to estimate. It is based on the assumption that the subjects in the study are chosen at random from some wider patient population. This will seldom be true, at least in the context of a clinical trial for which trial patients are screened for eligibility and entered only after giving informed consent. However it is usually reasonable to assume that trial patients have been chosen at random from a large number of potentially eligible patients, and that they represent a random selection from this artificial population. Thus, a random effects model is frequently applied whenever a study includes a large numbers of patients (Fayers and Machin, 2000).

If we assume a random effects model is appropriate, then models can be fitted using multi-level modelling statistical software which is implemented in MLwiN for example (Goldstein *et al* 1998). Use of multilevel modelling as opposed to marginal modelling allows examination of the "error" parts of the model in more detail.

The random effects model is most useful when the objective is to make inference about individuals rather than the population average. Thus a random effects approach will allow us to estimate the HRQoL status of an individual patient. The regression coefficients, β, represent the effect of the explanatory variables on an individual patient's HRQoL. This is in contrast to the marginal model coefficients, which describe the effect of the explanatory variables on the population average.

Random-effects vs. Marginal Modelling

The two approaches of random effect and marginal modelling, lead to different interpretations of between subject effects. In random effects models, a between subject effect represents the difference between subjects conditional on having the same random effect, whereas the parameters of marginal models represent the average difference between subjects.

Diggle *et al* (2002) demonstrate, that in the linear case, it is possible to formulate the two regression approaches to have coefficients with the same interpretation. That is coefficients from linear random effects models can have marginal interpretations as well. With non-linear link functions, such as the logit this is not the case. So in practice, this distinction is important only if link functions other than the identity link are used (Rabe-Hesketh and Everitt, 2000).

However, we tend to agree with Diggle *et al* (2002) that the choice of model should depend on the scientific question being asked. We will concentrate on marginal models since they are appropriate when inferences about the population average are the focus. In a clinical trial (such as the Leg Ulcer and NAMEIT studies) the average difference in HRQoL between the control and intervention groups is most important, not the difference in HRQoL for any one individual. However, first of all we will demonstrate the use of simpler summary measures or response features to analyse the Leg Ulcer dataset.

Analysis of Leg Ulcer data

The aim of this randomised controlled trial with one year of follow-up was to establish the relative cost-effectiveness of community leg ulcer clinics that use four-layer compression bandaging versus usual care provided by district nurses. Patients with venous leg ulcers were allocated at random to intervention (**Clinic**) or control (**Home**) groups. The intervention consisted of weekly treatment with four layer bandaging in leg ulcer clinic (Clinic group) or usual care at home by the district nursing service (Home group). The primary outcome was time to complete ulcer healing over the one-year follow-up. Secondary outcomes included HRQoL as measured by the SF-36 at baseline, three months and 12 months follow-up. Of the 233 patients randomised 155/233 (66.5%) completed the 12-month HRQoL assessment (77 in the Home group and 78 in the Clinic group). We are interested in comparing the HRQoL over the one-year follow-up between the two groups.

We will base our analysis on these 155 patients, but again it should be noted that missing HRQoL assessments may be a serious problem with this dataset and the reasons for the missing data should be thoroughly investigated. We will assume that the data are Missing Completely at Random, (MCAR) and that this reduced dataset represents a randomly drawn sub-sample of the full dataset and the inferences drawn can be considered reasonable (Fayers and Machin, 2000). There is extensive discussion of the occurrence of dropouts and missing data in a special edition of *Statistics in Medicine* (Volume 17, 1998) and both Fayers and Machin (2000) and Fairclough (2002) devote several chapters to this topic.

We use the Leg Ulcer data to illustrate various simple methods for analysing longitudinal data, including ANCOVA, with mean follow-up HRQoL (i.e. the average of the 3- and 12-month responses) as the dependent variable and baseline HRQoL and treatment group as covariates, and summary measures such as the AUC. We will compare and contrast conventional methods of standard error and confidence interval estimation with the corresponding estimates from the use of various bootstrap resampling methods.

Table 8.3 shows the baseline HRQoL and socio-demographic characteristics of the 155 patients in the Leg Ulcer study. The two groups were well matched at baseline for age, gender and HRQoL, except for the RE dimension of the SF-36, where there was some reliable statistical evidence of a difference ($p = 0.052$).

Table 8.3: Baseline characteristics of the Home and Clinic patients from the Leg Ulcer study

	Home			Clinic			Mean	95% CI		
	N	Mean	SD	N	Mean	SD	Difference	Lower	Upper	P-value[a]
Age (years)	77	73.5	(11.5)	78	73.1	(11.1)	0.4	-3.1	4.0	0.808
SF-36 Dimensions										
Physical Function	77	45.9	(33.0)	77	46.2	(31.1)	-0.3	-10.5	9.9	0.960
Role Physical	77	52.6	(40.9)	77	50.3	(42.2)	2.3	-11.0	15.5	0.735
Bodily Pain	77	55.7	(28.7)	77	59.6	(29.6)	-3.9	-13.2	5.4	0.408
General Health	77	62.9	(24.2)	77	67.5	(19.8)	-4.6	-11.6	2.5	0.201
Vitality	77	51.0	(24.4)	77	56.0	(18.9)	-5.0	-11.9	1.9	0.157
Social Function	77	67.1	(31.8)	77	69.4	(27.0)	-2.3	-11.7	7.1	0.628
Role Emotional	77	72.3	(40.2)	77	59.3	(42.1)	13.0	-0.1	26.1	0.052
Mental Health	77	67.5	(23.7)	77	71.1	(20.4)	-3.6	-10.7	3.4	0.310
Gender *Female*	21		(27%)	28		(36%)	(9%)	(-23%)	(6%)	0.248[b]
Male	56		(73%)	50		(64%)				
Total	77		(100.0%)	78		(100.0%)				

a. P-values and CIs for age and HRQoL estimated from unequal variances two independent samples t-test.
b. P-values for gender estimated from Chi-squared test and CIs via Wilson's method.

Leg Ulcer AUC analysis

The overall HRQoL of the leg ulcer patients over the 12-month study period (and three HRQoL assessments) can be summarised by the AUC. AUC were calculated using the trapezium rule as described in section 1 of Appendix 5. If we set the time units for the AUC calculation as a fraction of a year, then an AUC value of 100 implies the leg ulcer patient has been in "good health" for the entire 12-month follow-up period. Conversely an AUC value of 0 implies the leg ulcer patient has been in "poor health" for the entire 12-month follow-up period. Figures 8.1 and 8.2 show the histograms of the distribution of the AUC summary measure for the eight dimensions of the SF-36 separately for the Home and Clinic groups. Although the distributions are not symmetric, the histograms are not as skewed as the raw data at each time point.

Table 8.4 gives the results of simple comparisons of differences in mean AUC between the groups using the two independent samples *t*-test, with unequal variances (7.16 and 7.17) and the bootstrap hypothesis test (using Algorithm 7.2). We also show the results of the *MW* test. All analyses were carried out in `STATA v8` (StataCorp, 2003).

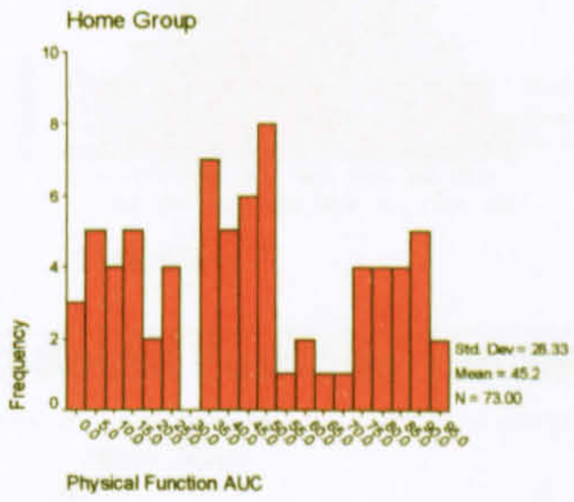
The *p*-values from the *t*-test and the ASL from the bootstrap hypothesis tests are very similar. None of the *p*-values for the eight SF-36 dimensions are less than 0.05. Therefore there is no reliable statistical evidence to suggest a difference in mean AUC between the Clinic and Home treated leg-ulcer patients. Only the results of the *MW* test on the RE dimension of the SF-36 provide ($p = 0.071$) any evidence of a difference in AUC distributions between the groups, although even this *p*-value is not statistically significant using the conventional cut-off of 0.05.

The table also contrasts the Normal theory based CI estimates from the *t*-test with the bootstrap BC_a limits. The lengths of the intervals are very similar, although the bootstrap BC_a intervals tend to have a non-symmetric shape. All the estimated CIs include zero, again suggesting no evidence of a difference

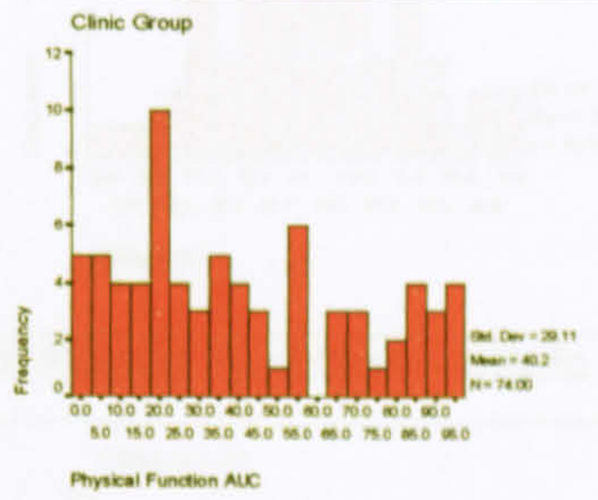
in mean AUC (HRQoL) between the Clinic and Home group patients in the Leg Ulcer study.

Figure 8.1: Histograms of AUC summary from Leg ulcer data by group

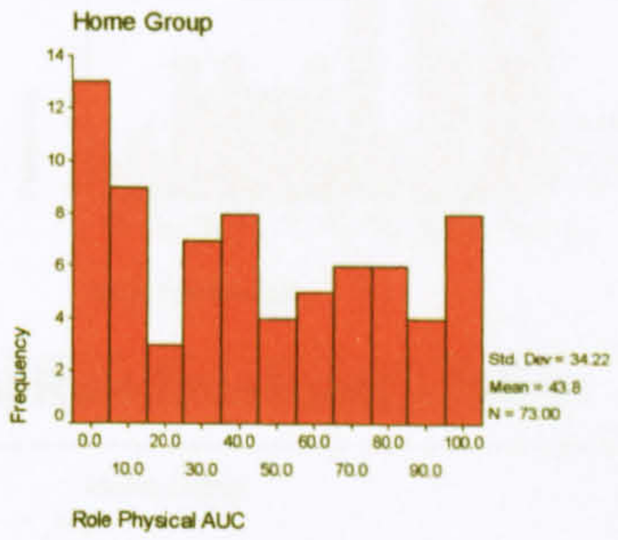
a) Physical Function – Home



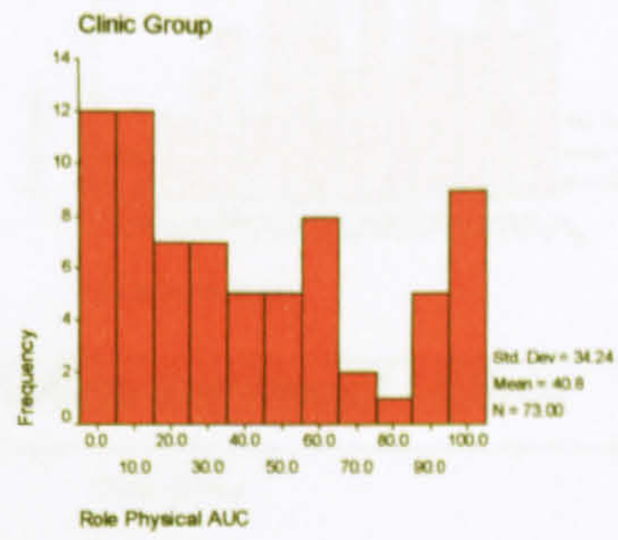
b) Physical Function – Clinic



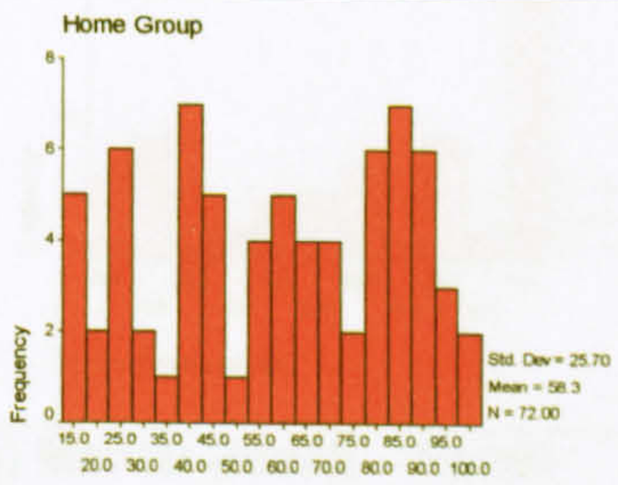
c) Role Physical – Home



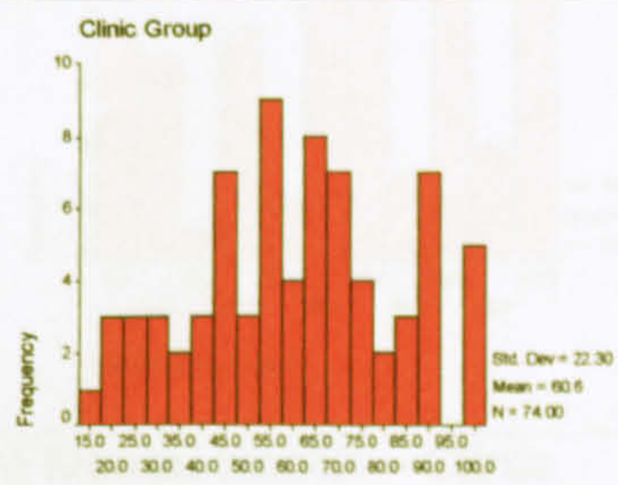
d) Role Physical – Clinic



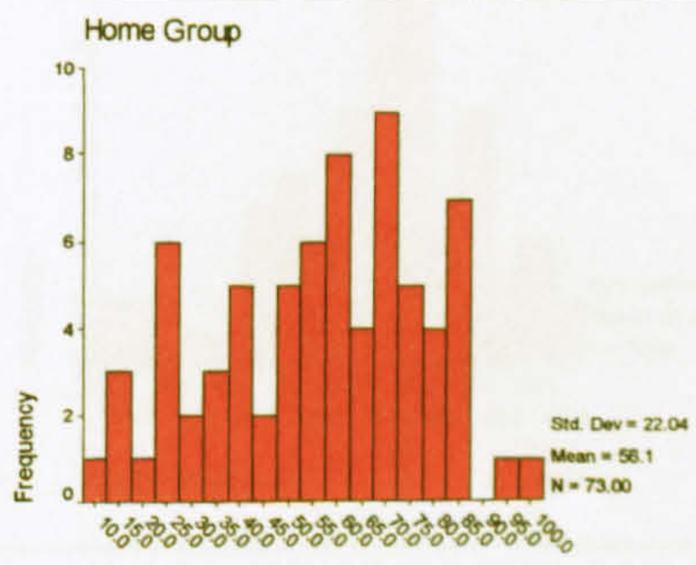
e) Bodily Pain – Home



f) Bodily Pain – Clinic



g) General Health – Home



h) General Health – Clinic

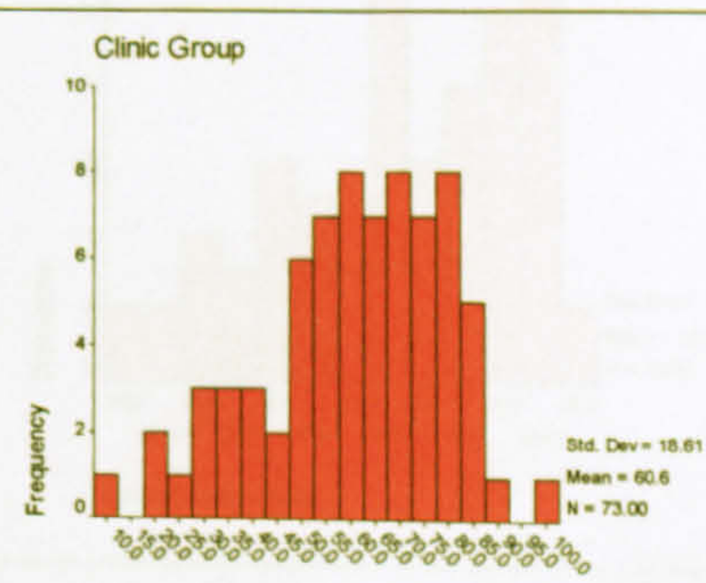
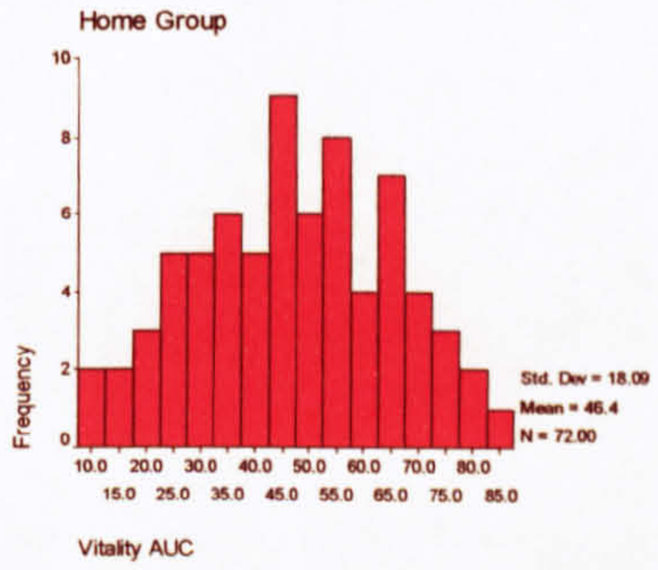
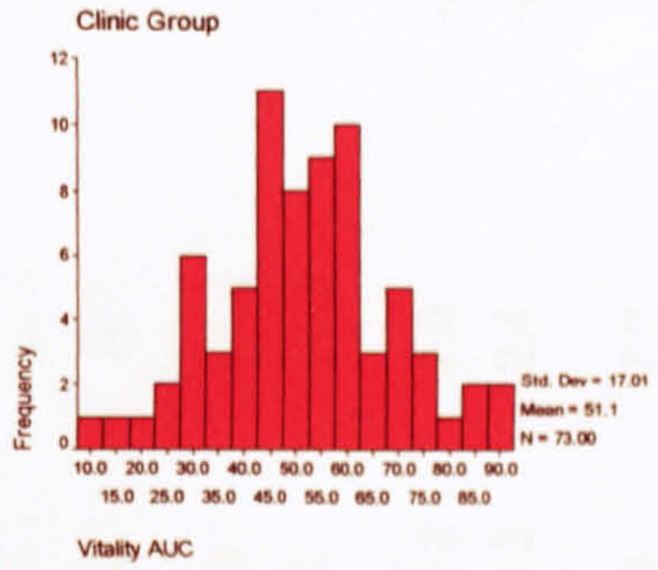


Figure 8.2: Histograms of AUC summary from Leg ulcer data by group

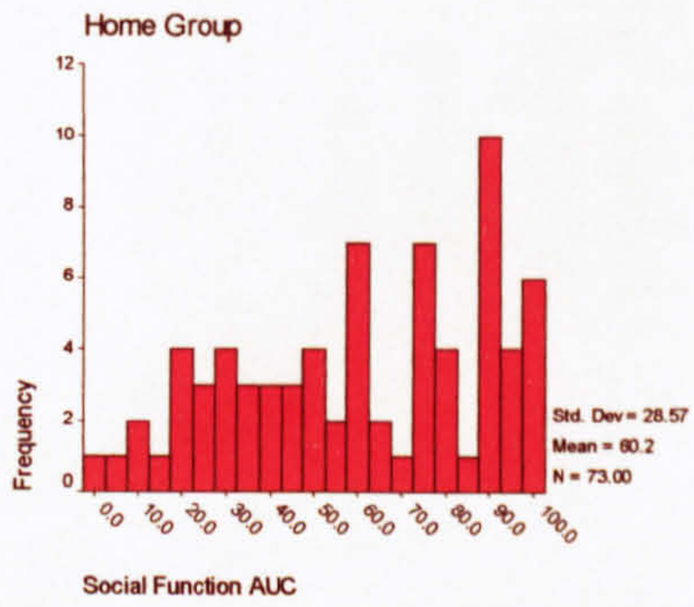
a) Vitality – Home



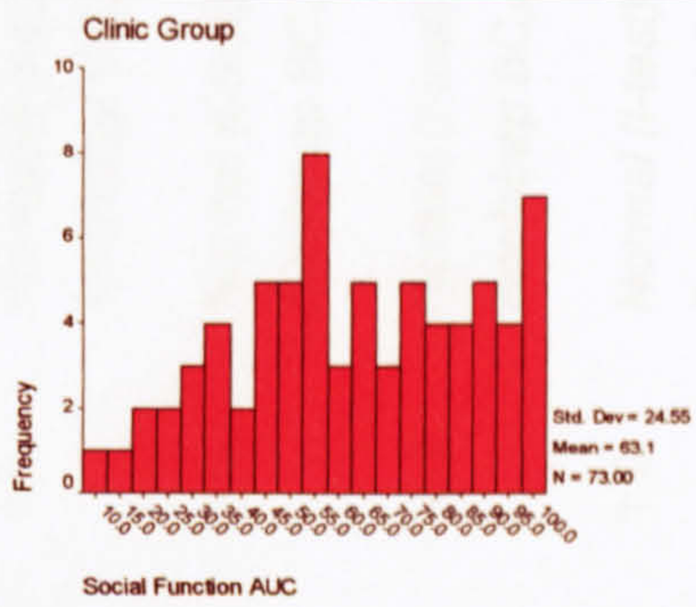
b) Vitality - Clinic



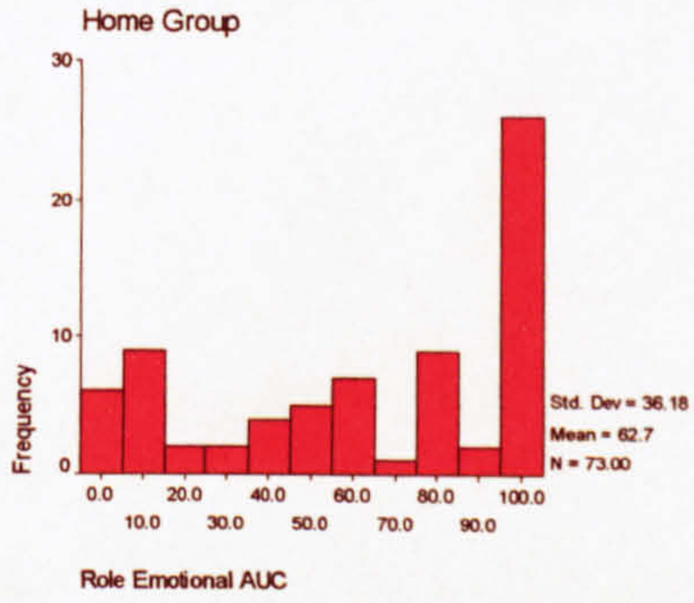
c) Social Function – Home



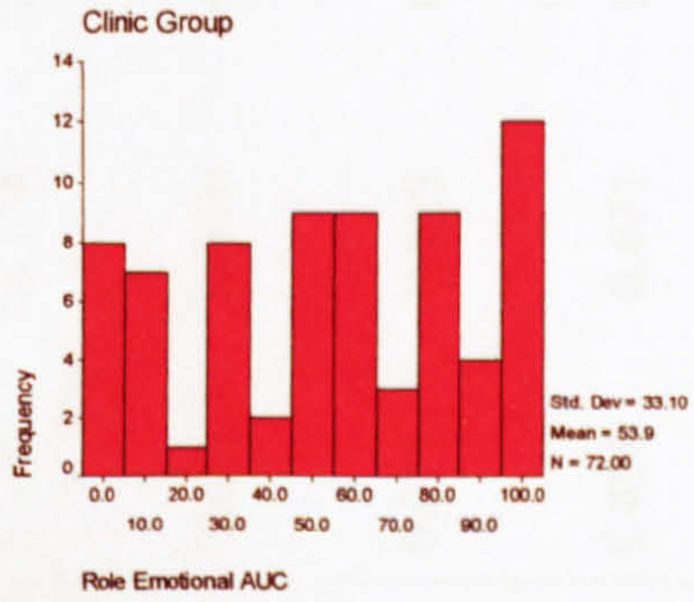
d) Social Function - Clinic



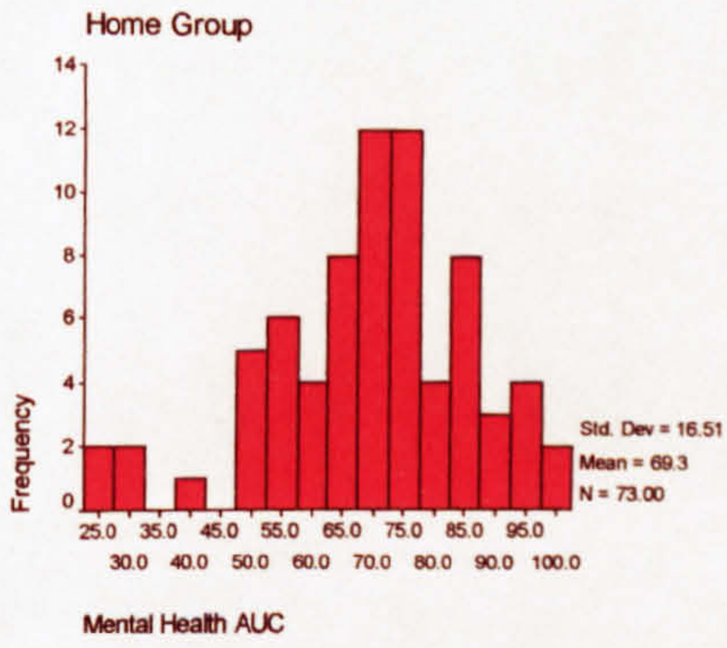
e) Role Emotional – Home



f) Role Emotional - Clinic



g) Mental Health – Home



h) Mental Health – Clinic

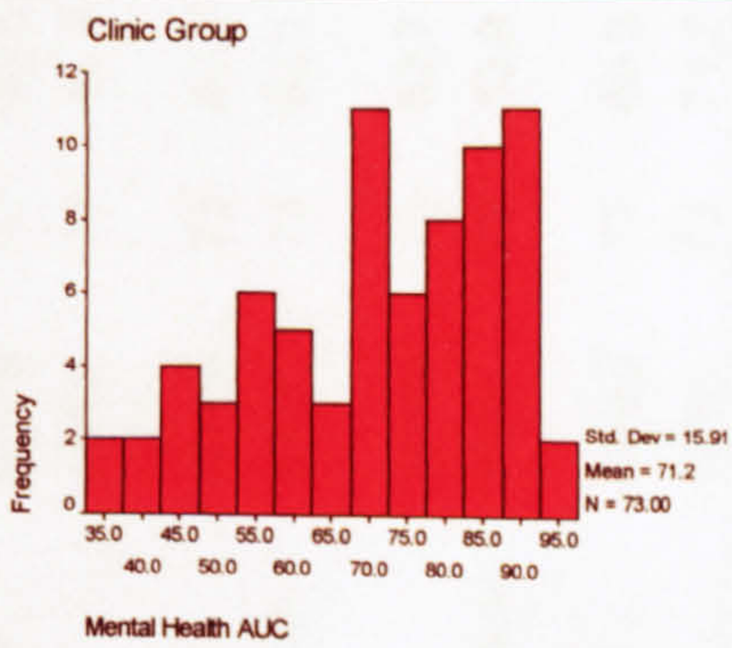


Table 8.4: Leg Ulcer study simple cross-sectional comparison of AUC for Home vs. Clinic Groups

SF-36 Dimension	*Group*	*n*	*mean*	*sd*	*MW-test P-value*	*t-test P-value*	*ASL_{boot}*	*Mean diff*		*95% CI lower*	*upper*	*Interval length*	*shape*
Physical	home	73	45.2	28.3	0.270	0.290	0.291	-5.0	*Normal (t-test)*	-14.4	4.3	-18.7	1.00
Function	clinic	74	40.2	29.1					*Bootstrap BC_A*	-14.5	3.9	-18.4	0.95
Role	home	73	43.8	34.2	0.618	0.598	0.597	-3.0	*Normal (t-test)*	-14.2	8.2	-22.4	1.00
Physical	clinic	73	40.8	34.2					*Bootstrap BC_A*	-13.7	8.3	-21.9	1.06
Bodily	home	72	58.3	25.7	0.677	0.573	0.568	2.3	*Normal (t-test)*	-5.6	10.1	-15.7	1.00
Pain	clinic	74	60.6	22.3					*Bootstrap BC_A*	-5.6	10.3	-16.0	1.02
General	home	73	56.1	22	0.247	0.184	0.188	4.5	*Normal (t-test)*	-2.2	11.2	-13.3	1.00
Health	clinic	73	60.6	18.6					*Bootstrap BC_A*	-2.0	11.0	-13.1	1.00
Vitality	home	72	46.4	18.1	0.150	0.114	0.117	4.6	*Normal (t-test)*	-1.1	10.4	-11.5	1.00
	clinic	73	51.1	17					*Bootstrap BC_A*	-1.2	10.1	-11.3	0.94
Social	home	73	60.2	28.6	0.653	0.516	0.515	2.9	*Normal (t-test)*	-5.8	11.6	-17.4	1.00
Function	clinic	73	63.1	24.6					*Bootstrap BC_A*	-5.6	11.2	-16.8	0.99
Role	home	73	62.7	36.2	0.071	0.129	0.129	-8.8	*Normal (t-test)*	-20.2	2.6	-22.8	1.00
Emotional	clinic	72	53.9	33.1					*Bootstrap BC_A*	-20.3	2.4	-22.7	0.97
Mental	home	73	69.3	16.5	0.454	0.471	0.469	1.9	*Normal (t-test)*	-3.4	7.2	-10.6	1.00
Health	clinic	73	71.2	15.9					*Bootstrap BC_A*	-3.3	7.0	-10.3	0.96

ASL_{boot} based on 5000 bootstrap replications. Mean difference = Clinic mean - Home mean. BC_a confidence intervals based 5000 bootstrap replications.

Leg Ulcer analysis of covariance

The final analysis of the Leg ulcer data involved a multiple regression analysis with the average follow-up HRQoL (the mean of the 3- and 12-month assessments) as the dependent variable, $\bar{Y}_i$, and the baseline HRQoL (x_{Base_i}) and treatment group (x_{Group_i}, coded Home = 0, Clinic = 1) as covariates. The linear regression model for the i^{th} subject was:

$$\bar{Y}_i = \beta_1 + \beta_{Base} x_{Base\_i} + \beta_{Group} x_{Group\_i} + \varepsilon_i, \qquad (8.7)$$

where ε_i is a zero mean error term and β_1 is a constant.

The regression coefficient estimate, $\hat{\beta}_{Group}$, for group represents the difference in mean follow-up HRQoL between the Home and Clinic groups after adjustment for baseline HRQoL. A positive value for the regression coefficient for $\hat{\beta}_{Group}$ indicates the Clinic group has a better mean HRQoL at follow-up than the Home group after adjustment for baseline HRQoL.

We used OLS to estimate the regression coefficients and then applied the bootstrap model and case-based resampling Algorithms (7.3 and 7.4) in S-Plus 2000 (MathSoft, 1999) and STATA (StataCorp, 2003) to estimate bootstrap standard errors and BC_a confidence intervals for the regression coefficients. Again Appendix 4 provides examples of the programs for case and residual resampling and estimation of BC_a confidence intervals.

Table 8.5 compares the OLS and bootstrap standard errors and confidence interval estimates for the group coefficient from the Leg Ulcer data. For the bootstrap methods the standard errors are the standard deviations of the coefficients from the 5000 bootstrap re-samples. Again for ease of interpretation and comparison only the estimates for the group coefficient $\hat{\beta}_{Group}$ are shown.

Table 8.5: Multiple regression, OLS, bootstrap case and model based resampling SE's and Confidence Interval estimates from the Leg Ulcer data with the mean of the two follow-up assessments as the outcome

Dependent Variable	Model	GROUP coefficient N	$\hat{\beta}$	SE	$\hat{\beta}/SE$	p	95% CI Lower	Upper	Interval Length	Shape
Physical Function	OLS	147	**-6.3**	3.31	-1.91	0.058	-12.86	0.21	13.07	1.00
	Case			3.28	-1.93		-13.17	-0.27	12.90	0.89
	Model			3.34	-1.89		-13.09	-0.04	13.04	0.93
Role Physical	OLS	146	**-3.9**	5.43	-0.72	0.474	-14.64	6.84	21.48	1.00
	Case			5.42	-0.72		-14.18	7.15	21.33	1.08
	Model			5.39	-0.72		-14.64	6.52	21.17	0.97
Bodily Pain	OLS	146	**-0.2**	3.69	-0.06	0.954	-7.50	7.07	14.57	1.00
	Case			3.65	-0.06		-7.14	7.34	14.48	1.09
	Model			3.66	-0.06		-7.13	7.06	14.19	1.05
General Health	OLS	146	**0.9**	2.54	0.34	0.733	-4.15	5.89	10.04	1.00
	Case			2.63	0.33		-4.10	6.11	10.21	1.05
	Model			2.60	0.33		-3.97	6.25	10.22	1.11
Vitality	OLS	145	**1.4**	2.52	0.57	0.569	-3.55	6.43	9.98	1.00
	Case			2.51	0.57		-3.56	6.38	9.94	0.99
	Model			2.50	0.58		-3.49	6.41	9.90	1.01
Social Function	OLS	146	**0.7**	4.06	0.18	0.855	-7.27	8.76	16.04	1.00
	Case			4.06	0.18		-6.99	8.75	15.74	1.03
	Model			4.10	0.18		-7.20	8.92	16.11	1.03
Role Emotional	OLS	145	**-5.3**	5.94	-0.90	0.370	-17.10	6.41	23.50	1.00
	Case			5.87	-0.91		-16.54	6.52	23.06	1.06
	Model			5.97	-0.90		-17.05	6.48	23.53	1.01
Mental Health	OLS	146	**-0.4**	2.15	-0.18	0.859	-4.63	3.86	8.49	1.00
	Case			2.16	-0.18		-4.59	3.75	8.34	0.98
	Model			2.14	-0.18		-4.59	3.83	8.42	1.00

The regression analysis shown in Table 8.5 suggests that there is no reliable statistical evidence of a difference in average (3- and 12-month) follow-up HRQoL between Clinic and Home group treated leg ulcer patients on seven out of the eight dimensions of the SF-36 after adjustment for baseline HRQoL dimension score. Only on the PF dimension ($p = 0.058$) is there any suggestion of a difference.

As can be seen from Table 8.5 the standard error estimates are almost identical for the three methods. Similarly the length of the confidence intervals is virtually the same for all three methods, although the bootstrap CIs tend to be asymmetric about the point-estimate of the regression coefficient. Qualitatively seven out eight of the intervals from the three methods either include or exclude zero so the interpretation of the group regression coefficient is the same. The only exception is the PF dimension where both the asymmetric case and model based bootstrap confidence intervals marginally exclude zero. Conversely the upper limit of the OLS estimate is also very close to zero, so the practical interpretation of all three confidence intervals should be the same, i.e. no consistent evidence of a difference in mean follow-up HRQoL between the groups.

Therefore again in this example dataset there appears to be little advantage in using bootstrap case or model based re-sampling to estimate standard errors and confidence intervals compared to conventional methods of confidence interval estimation from the OLS multiple regression model.

Analysis of NAMEIT data

The NAMEIT trial was a 48-week, randomised, double blind RCT to compare Neoral with methotrexate (**Neoral**) versus placebo plus methotrexate (**Placebo**) in patients with early severe rheumatoid arthritis (RA). In order to assess the impact of the treatments on patients' health related quality of life, the SF-36 was completed by subjects at seven time-points, Week 0 (baseline), Weeks 8, 16, 24, 32, 40, and Week 48. Of the 306 subjects

randomised, 223 (72.9%) completed the study, 111 in the Placebo group and 112 in the Neoral group.

We will base our analysis on these 223 patients, but again it should be noted that missing HRQoL assessments may be a serious problem with this dataset and the reasons for the missing data should be thoroughly investigated. We will assume that the data are MCAR, and that this reduced dataset represents a randomly drawn sub-sample of the full dataset and the inferences drawn can be considered reasonable.

We use these data to illustrate various methods for analysing longitudinal data, including ANCOVA, with average follow-up HRQoL as the dependent variable and baseline HRQoL and treatment group as covariates, summary measures (e.g. AUC) and a marginal model analysis, incorporating all seven HRQoL assessments. Again we will compare and contrast conventional methods of standard error and confidence interval estimation with the bootstrap resampling methods.

Graphical presentation of longitudinal data from NAMEIT study

Both Diggle (*et al* 2002) and Fayers and Machin (2000) emphasise the importance of graphical presentation of longitudinal data prior to modelling. Figure 8.3 shows the mean levels of HRQoL in patients with RA, before and during treatment, for the eight dimensions of the SF-36. The curves for some dimensions of the SF-36 overlap (e.g. PF, GH, RE, and MH dimensions) suggesting that it may be unrealistic to assume that the mean difference in HRQoL values on these dimensions remains constant over time. For other dimensions such as BP, V and SF there is some evidence to suggest that for later HRQoL measurements the curves are parallel and that the mean difference between treatments is now fairly constant.

Figure 8.3: Profile of mean SF-36 scores over time by treatment group NAMEIT data

The overlapping lines on some of the graphs in Figure 8.3 imply there may be a 'Treatment x Time' interaction. It is therefore important to test for any such interaction in any regression model. Fortunately, with the marginal model approach this is relatively easy to do and simply involves the addition of an extra regression coefficient to the model. If treatment is coded as a 0/1 variable (i.e. 0 = Placebo and 1 = Neoral) and assessment time as a continuous variable, then the additional interaction term is simply the product of these two variables (which will be 0 for all the Placebo group patients and equal to the HRQoL assessment time in the Neoral Group patients).

Table 8.6 shows the baseline HRQoL and socio-demographic characteristics of the 223 patients in the NAMEIT study. The two groups of RA patients were well matched at baseline with respect to age, gender and HRQoL.

Two simple response feature analyses are calculation of the AUC for the seven repeated HRQoL assessments and the regression of the mean of the six follow-up HRQoL assessments against the baseline HRQoL and treatment group.

NAMEIT AUC analysis

The overall HRQoL of the RA patients over the 48-week study period (and seven HRQoL assessments) can be summarised by the AUC. Again individual AUC were calculated using the trapezium rule as described in section 1 of Appendix 5. If we set the time units for the AUC calculation as a fraction of a year, then an AUC value of 100 implies the RA patient has been in "good health" for the entire 48-week follow-up period. Conversely an AUC value of 0 implies the RA patient has been in "poor health" for the entire 48-week follow-up period. For the NAMEIT study HRQoL was assessed at seven equally spaced time-points eight weeks apart, therefore the AUC calculation is equivalent to the mean of all the seven HRQoL assessments.

Table 8.6: Baseline characteristics of the Placebo and Neoral Group patients from the NAMEIT study

	Placebo			Neoral			Mean	95% CI		
	N	Mean	SD	N	Mean	SD	Diff	Lower	Upper	P-value[a]
Age (years)	112	49.8	11.6	111	48.0	10.9	1.8	-1.2	4.7	0.242
SF-36 Dimensions										
Physical Function	111	40.5	21.5	111	41.6	25.8	-1.1	-7.4	5.2	0.730
Role Physical	110	24.8	35.3	111	26.6	37.7	-1.7	-11.4	8.0	0.725
Bodily Pain	111	39.6	19.3	111	40.3	18.7	-0.7	-5.7	4.3	0.784
General Health	111	39.2	17.0	110	35.4	17.7	3.9	-0.7	8.5	0.098
Vitality	110	39.2	17.7	110	40.5	18.2	-1.3	-6.1	3.5	0.587
Social Function	111	53.4	22.4	111	55.1	22.7	-1.7	-7.7	4.3	0.575
Role Emotional	111	57.4	44.5	110	46.4	44.5	11.0	-0.8	22.8	0.068
Mental Health	111	60.1	19.5	110	56.9	18.3	3.2	-1.8	8.2	0.214
Gender :Female	83		(74.1%)	80		(72.1%)	(2.0%)	(-9.5%)	(13.6%)	0.732[b]
Male	29		(25.9%)	31		(27.9%)				
Total	112		(100.0%)	111		(100.0%)				

a. P-values and CIs for age and HRQoL estimated from unequal variances two independent samples t-test.
b. P-values for gender estimated from Chi-squared test and CIs via Wilson's method.

Table 8.7 gives the results of simple comparisons of differences in mean AUC between the groups using the two independent samples *t*-test, with unequal variances (7.16 and 7.17) and the bootstrap hypothesis test (using Algorithm 7.2). We also show the results of the *MW* test. All analyses were carried out in `STATA v8` (StataCorp, 2003). The ASL from the bootstrap hypothesis test and the BC_a confidence interval estimates are based on 5000 re-samples.

The *p*-values from the *t*-test and the ASL from the bootstrap hypothesis tests are very similar. Only one of the *p*-values using either of these two tests for the eight SF-36 dimensions was less than 0.05. So there is some reliable statistical evidence to suggest a difference in mean RP AUC between the Placebo and Neoral treated RA patients. The results of the *MW* test on the BP dimension of the SF-36 ($p = 0.024$) also provides some evidence of a difference in AUC distributions between the groups, although this *p*-value is not statistically significant using the conventional cut-off of $p < 0.05$ for the *t*-test and bootstrap hypothesis test (with *p*-values of 0.056 and 0.054 respectively).

The Normal theory based CI estimates from the *t*-test and the bootstrap BC_a limits are calculated for a characteristic of the distributions (for example mean difference). The groups may have differences in distributions of HRQoL (e.g. BP) but similar characteristics e.g. means. The lengths of the intervals are very similar, although the bootstrap BC_a intervals tend to have a non-symmetric shape. For seven out of the eight dimensions the estimated CIs include zero, again suggesting no evidence of a difference in mean AUC (HRQoL) between the Neoral and Placebo group patients in the NAMEIT study. The exception is the RP dimension where both the *t*-test and the bootstrap BC_a limits suggest that the Neoral group has a better mean AUC than the Placebo group.

Table 8.7: NAMEIT study simple cross-sectional comparison of AUC for Placebo vs. Neoral Groups

SF-36					*MW test*	*t-test*		*Mean*		*95% CI*		*Interval*	
Dimension	*Group*	*n*	*mean*	*sd*	*P-value*	*P-value*	ASL_{boot}	*diff*		*lower*	*upper*	*length*	*shape*
Physical	Placebo	111	44.8	19.1	0.146	0.186	0.179	3.6	*Normal (t-test)*	-1.8	9.1	-10.8	1.00
Function	Neoral	107	48.5	21.3					*Bootstrap* BC_A	-1.8	8.8	-10.6	0.95
Role	Placebo	107	32.4	27.2	0.020	0.019	0.018	9.1	*Normal (t-test)*	1.5	16.7	-15.1	1.00
Physical	Neoral	105	41.5	28.7					*Bootstrap* BC_A	1.6	16.6	-15.1	1.00
Bodily	Placebo	111	46.7	14.9	0.024	0.056	0.054	3.9	*Normal (t-test)*	-0.1	8.0	-8.1	1.00
Pain	Neoral	108	50.6	15.4					*Bootstrap* BC_A	-0.2	7.8	-8.0	0.92
General	Placebo	107	39.5	16.9	0.495	0.666	0.662	1.0	*Normal (t-test)*	-3.5	5.5	-9.1	1.00
Health	Neoral	102	40.5	16.3					*Bootstrap* BC_A	-3.6	5.5	-9.1	0.97
Vitality	Placebo	109	42.9	14.7	0.186	0.222	0.222	2.4	*Normal (t-test)*	-1.5	6.3	-7.7	1.00
	Neoral	107	45.4	14.1					*Bootstrap* BC_A	-1.4	6.1	-7.5	0.95
Social	Placebo	111	54.0	15.6	0.238	0.255	0.271	2.4	*Normal (t-test)*	-1.8	6.7	-8.4	1.00
Function	Neoral	108	56.4	16.0					*Bootstrap* BC_A	-1.7	6.8	-8.5	1.05
Role	Placebo	105	56.6	29.1	0.989	0.945	0.943	-0.3	*Normal (t-test)*	-8.4	7.8	-16.1	1.00
Emotional	Neoral	101	56.3	29.6					*Bootstrap* BC_A	-8.2	7.8	-16.0	1.01
Mental	Placebo	111	58.6	14.7	0.547	0.612	0.612	-1.0	*Normal (t-test)*	-4.8	2.8	-7.6	1.00
Health	Neoral	107	57.6	13.7					*Bootstrap* BC_A	-4.8	2.8	-7.6	1.01

NAMEIT analysis of covariance

The second analysis of the NAMEIT data involved a multiple regression analysis with the average follow-up HRQoL (the mean of the six follow-up assessments at 8, 16, 24, 32, 40 and 48 weeks) as the dependent variable and the baseline HRQoL and treatment group (Placebo = 0, Neoral = 1) as covariates. The linear regression model was the same as (8.7), but this time the dependent variable ($\bar{Y}_i$) was the average of the six follow-up HRQoL assessments.

The regression coefficient estimate, $\hat{\beta}_{Group}$, for group represents the difference in mean follow-up HRQoL between the Placebo and Neoral groups after adjustment for baseline HRQoL. A positive value for the regression coefficient estimate for $\hat{\beta}_{Group}$ indicates the Neoral group has a better mean HRQoL at follow-up than the Placebo group after adjustment for baseline HRQoL.

We used OLS to estimate the regression coefficients and then applied the bootstrap model and case-based resampling Algorithms (7.3 and 7.4) in S-Plus 2000 (MathSoft, 1999) and STATA (StataCorp, 2003) to estimate bootstrap standard errors and BC_a confidence intervals for the regression coefficients. Again Appendix 4 provides examples of the programs for case and residual resampling and estimation of BC_a confidence intervals. The bootstrap estimates are based on 5000 re-samples.

Table 8.8 compares the OLS and bootstrap standard errors and confidence interval estimates for the group coefficient from the NAMEIT data. For the bootstrap methods the standard errors are the standard deviations of the coefficients from the 5000 bootstrap re-samples. Again for ease of interpretation and comparison only the estimates for the group coefficient are shown.

Table 8.8: Comparison of OLS and bootstrap SE's and CI's - NAMEIT data with the mean of six follow-up assessments as the outcome

Dependent Variable	Model	Group coefficient N	$\hat{\beta}$	SE	$\hat{\beta}/SE$	p	95% CI Lower	Upper	Interval Length	Shape
Physical Function	OLS	222	**2.8**	2.26	1.25	0.213	-1.63	7.28	8.91	1.00
	Case			2.24	1.26		-1.49	7.40	8.89	1.06
	Model			2.24	1.26		-1.63	7.20	8.83	0.98
Role Physical	OLS	221	**9.4**	3.93	2.39	0.018	1.64	17.15	15.51	1.00
	Case			3.87	2.43		1.96	16.96	15.00	1.02
	Model			3.98	2.36		1.51	17.05	15.54	0.97
Bodily Pain	OLS	222	**4.2**	1.97	2.15	0.033	0.36	8.13	7.77	1.00
	Case			1.98	2.14		0.30	8.08	7.78	0.97
	Model			1.97	2.15		0.36	8.07	7.71	0.99
General Health	OLS	221	**4.6**	1.96	2.36	0.019	0.76	8.50	7.73	1.00
	Case			1.99	2.32		0.58	8.35	7.77	0.92
	Model			1.96	2.36		0.82	8.54	7.72	1.03
Vitality	OLS	220	**2.7**	1.80	1.49	0.138	-0.87	6.23	7.10	1.00
	Case			1.81	1.48		-0.69	6.36	7.05	1.09
	Model			1.81	1.48		-0.86	6.29	7.15	1.02
Social Function	OLS	222	**2.4**	2.10	1.15	0.252	-1.73	6.54	8.26	1.00
	Case			2.10	1.14		-1.47	6.87	8.35	1.15
	Model			2.12	1.14		-1.80	6.49	8.29	0.97
Role Emotional	OLS	221	**4.1**	3.91	1.05	0.293	-3.58	11.82	15.40	1.00
	Case			3.99	1.03		-3.86	11.91	15.77	0.98
	Model			3.97	1.04		-3.51	11.99	15.51	1.03
Mental Health	OLS	221	**1.5**	1.64	0.94	0.349	-1.69	4.77	6.47	1.00
	Case			1.69	0.91		-1.88	4.74	6.62	0.94
	Model			1.62	0.95		-1.63	4.66	6.29	0.99

The regression analysis shown in Table 8.8 suggests that there is no reliable statistical evidence of a difference in average follow-up HRQoL between Neoral and Placebo group treated RA patients on five out of the eight dimensions of the SF-36 after adjustment for baseline HRQoL dimension

score. On three other dimensions RP (p = 0.018), BP (p = 0.033) and GH (p = 0.019) there is some suggestion of a difference in average follow-up HRQoL.

As can be seen from Table 8.8 the standard error estimates are almost identical for the three methods. Similarly the length of the confidence intervals is virtually the same for all three methods, although the bootstrap CIs tend to be asymmetric about the point-estimate of the regression coefficient. Qualitatively all the intervals from the three methods either include or exclude zero so the interpretation of the group regression coefficient is the same. For the RP, BP and GH dimensions there is some suggestion that Neoral group has a better average follow-up HRQoL than Placebo group patients.

Therefore again in this example dataset there appears to be little advantage in using bootstrap case or model based re-sampling to estimate standard errors and confidence intervals compared to conventional methods of confidence interval estimation from the OLS multiple regression model. Finally we will now use a marginal model to analyse all seven HRQoL assessments simultaneously.

NAMEIT marginal model analysis

The marginal model we used for the NAMEIT data for analysing the seven HRQoL assessments over time was,

$$Y_{ij} = \beta_1 + \beta_{Base} x_{Base\_i} + \beta_{Age} x_{Age\_i} + \beta_{Sex} x_{Sex\_i} + \beta_{Time} t_{ij} + \beta_{Group} x_{Group\_i} + \varepsilon_{ij}, \qquad (8.8)$$

where Y_{ij} is the HRQoL at time t_{ij} post-baseline; t_{ij} is the time of the QoL assessment, in weeks post baseline, of patient i at visit j; x_{Base_i} is the baseline HRQoL assessment for subject i; x_{Age_i} is the age (in years) of subject i at time 0 (baseline); x_{Sex_i} is the gender of subject i; x_{Group_i} is the treatment group (0 = Placebo, 1 = Neoral) for subject i; β_1 is a constant and ε_{ij} is the residual error. The marginal modelling approach is to assume

(1) $$E(Y_{ij}) = \beta_1 + \beta_{Base} x_{Base\_i} + \ldots + \beta_{Time} t_{ij} + \beta_{Group} x_{Group\_i}, \qquad (8.9)$$

(2) $$Corr(\varepsilon_{ij}, \varepsilon_{ik}) = \rho(t_{ij}, t_{ik}; \alpha). \qquad (8.10)$$

The marginal regression models were fitted in `STATA v8` (StataCorp, 2003) using the `xtgee` command with an identity link function (`link(iden)`) and the `robust` standard errors option. We tried out three autocorrelation structures $Corr(\varepsilon_{ij}, \varepsilon_{ik}) = \rho(t_{ij}, t_{ik}; \alpha)$ for the marginal model:

(1) independent `corr(indep)` $Corr(\varepsilon_{ij}, \varepsilon_{ik}) = \rho(t_{ij}, t_{ik}; \alpha) = 0$;

(2) exchangeable `corr(exc)` $Corr(\varepsilon_{ij}, \varepsilon_{ik}) = \rho(t_{ij}, t_{ik}; \alpha) = \alpha_0$;

(3) autoregressive `corr(ar1)` $Cov(Y_{ij}, Y_{ik}) = \sigma^2 \exp(-\phi|t_j - t_k|)$.

The three autocorrelation structures above assume working correlation matrices $R(\alpha)$ of the identity matrix for the independent structure; a working correlation matrix of $R_{s,t} = \begin{cases} 1 \rightarrow s = t \\ \alpha_0 \rightarrow otherwise \end{cases}$ for the exchangeable structure and finally a working correlation matrix of $R_{s,t} = \begin{cases} 1 \rightarrow s = t \\ \alpha_1^{|s-t|} \rightarrow otherwise \end{cases}$ for the autoregressive model.

The (first) marginal model with the identity link, Normal family and with an independent correlation structure reproduces OLS estimators from the standard multiple linear regression model described in Chapter 7.

The observed correlation matrices in Figure 8.1 clearly show the off-diagonal terms are non-zero and that the assumption of an independent auto-correlation matrix for the marginal model is unrealistic. Therefore we will not consider models with an independent auto-correlation structure any further and will concentrate on reporting the results of models with either an exchangeable (2) or autoregressive correlation (3).

According to the `STATA` reference manual (StataCorp, 2003), the `robust` option specifies that the Huber/White sandwich estimator of variance is to be used in place of the default Iteratively Reweighted Least Squares (IRLS) variance estimator. This produces valid standard errors even if the within subject correlations are not as hypothesised by the specified correlation

structure. It does, however require that the model correctly specifies the mean. As such, the resulting standard errors are labelled "semi-robust" instead of "robust". We will compare the robust SEs and CIs with their bootstrap counterparts.

Treatment x time interactions

In a previous section we mentioned the importance of looking for a 'Treatment x Time' interaction. This can easily be done by creating a new variable x_{Int_ij}, which is the product of the treatment group and time variables and adding this extra term β_{Int} to model (8.8) and fitting this new model

$$Y_{ij} = \beta_0 + \beta_{Base}x_{Base\_i} + \beta_{Age}x_{Age\_i} + \beta_{Sex}x_{Sex\_i} + \beta_{Time}t_{ij} + \beta_{Group}x_{Group\_i} + \beta_{Int}x_{Int\_ij} + \varepsilon_{ij} \quad (8.11)$$

None of the β_{Int} coefficients for the eight SF-36 dimensions were statistically significant (from zero). Thus there was no reliable evidence of a 'Treatment x Time' interaction on any dimension of the SF-36 ($p > 0.05$), irrespective of the autocorrelation structure (independent, exchangeable or autoregressive). Therefore we will only report the results of the simpler model (8.8).

Bootstrap Resampling for the marginal model

We have mentioned that the marginal model approach is robust to miss-specification of the autocorrelation structure particularly as the sample size increases. Furthermore the generation of robust standard errors will produce valid standard errors even if the within subject correlations are not as hypothesised by the specified correlation structure. It does, however require that the model correctly specifies the mean. So will the bootstrap estimation of standard errors and confidence intervals offer any advantage?

The rest of this section will describe a simple bootstrap resampling algorithm for marginal models. The remainder of the chapter will then show how the bootstrap estimates of standard errors and confidence intervals compare with the robust standard errors and confidence intervals from the conventionally fitted marginal model.

Suppose as before we have the linear regression model (8.1). In this simple multiple regression model the ε_{ij} would be mutually independent $N(0, \sigma^2)$ random variables. The longitudinal structure of the data means that we can expect the ε_{ij} to be correlated within subjects.

In Chapter 7 we described two bootstrap resampling algorithms (7.3 and 7.4) for multiple linear regression models – *case* and *model* or residual based resampling. If the repeated HRQoL responses on each subject were assumed to be independent with no autocorrelation (or we choose to ignore the correlation that exists in longitudinal data) then we could use either bootstrap resampling algorithm as the marginal model has been reduced to a simple cross-sectional model.

Unfortunately, it is very unlikely that the successive HRQoL observations (and the corresponding residuals from the regression model) on each subject are independent and uncorrelated (see Table 8.1). Therefore both the case and model based bootstrap resampling Algorithms 7.3 and 7.4 are inappropriate. We must choose a resampling method that takes into account the complex dependences within the observed data.

Suppose we assume a simple uniform or exchangeable correlation between successive HRQoL measurements on the same subject. Then model (8.1) can be simplified to

$$Y_{ij} = \mu_{ij} + U_i + Z_{ij}, \quad i = 1 \text{ to } m; j = 1,\ldots,n \qquad (8.12)$$

where $u_{ij} = E(Y_{ij}) = \beta_1 x_{ij1} + \beta_2 x_{ij2} + ... + \beta_p x_{ijp}$, and the U_i are mutually independent $N(0, \upsilon^2)$ random variables and the Z_{ij} are mutually independent $N(0, \tau^2)$ random variables and the U_i and the Z_{ij} are independent of one another. Then the covariance structure of the data $Cov(Y_{ij}, Y_{ik})$ corresponds to

$$\rho = \frac{\upsilon^2}{(\upsilon^2 + \tau^2)}, \qquad (8.13$$

(which is equivalent to (8.6) with $\rho = \alpha_0$) with variance,

$$\sigma^2 = \upsilon^2 + \tau^2. \qquad (8.14)$$

Equation (8.14) implies the total error or variance can be decomposed into a between subject error (υ^2) and a within subject error term (τ^2).

Suppose we had a fully balanced data set with no missing data and equal numbers $n_i = n$ of HRQoL observations per subject i (with m subjects, i.e. $i = 1$ to m). If we felt an exchangeable autocorrelation structure was appropriate for the model and the residuals, then we could fit the marginal model (with exchangeable autocorrelation) and calculate the n residuals ($\varepsilon_1, \varepsilon_2,\ldots,\varepsilon_n$), for each subject and then use a variation of Algorithm 7.4, this time resampling from the m blocks of residuals (of size n). Unfortunately this approach only has the ability to model the total error or variance σ^2 and assumes the contribution of the between (υ^2) and within subject error (τ^2) to the total error σ^2 is the same across all subjects. Similar concerns would apply if we assumed an autoregressive autocorrelation structure and re-sampled blocks of residuals again.

The beauty of the marginal model and the GEE methodology is that it is very flexible and can in principle deal with all the observed data from a HRQoL study. The subjects are not required to have exactly the same numbers of assessments, and even the assessments can be made at variable times. The latter allows the modelling to proceed even if a subject misses a HRQoL assessment. So it seems unrealistic and unreasonable to use bootstrap resampling methods for marginal models that can only utilise a balanced data set, with equally spaced QoL assessments.

Therefore since we are interested in fitting a marginal model with an autoregressive structure and we are likely to have an unbalanced dataset with unequal observations per subject we will use a simple bootstrap case-resampling Algorithm (8.1) which is a modification of Algorithm 7.3.

Algorithm 8.1

Case (or cluster) -based resampling for a marginal regression model

If Y_{ij} represents the HRQoL response variable and $\boldsymbol{x}_{ij}$ a row vector of length p of explanatory variables observed at time t_{ij}, for observation $j = 1,\ldots,n_i$ on

subject i = 1,...,m. The set of repeated HRQoL outcomes for subject i are collected into a n_i-vector, $\mathbf{Y}_i = \left(Y_{i1}, Y_{i2}, ..., Y_{in_i}\right)$ with a corresponding matrix of explanatory variables $\mathbf{X}_i = \left(X_{i1}, X_{i2}, ..., X_{in_i}\right)$.

We shall fit the marginal linear regression model $Y_{ij} = \beta_1 x_{ij1} + \beta_2 x_{ij2} + ... + \beta_p x_{ijp} + \varepsilon_{ij}$, with one of several autocorrelation structures for the residuals. The data will consist of m blocks, say $\mathbf{W}_i(\mathbf{X}_i, \mathbf{Y}_i)$ = $(\mathbf{X}_1, \mathbf{Y}_1)$, $(\mathbf{X}_2, \mathbf{Y}_2)$,...,$(\mathbf{X}_m, \mathbf{Y}_m)$, corresponding to the data for each subject i.

1. Case-based resampling involves drawing a bootstrap sample of size m, with replacement from these m blocks. The bootstrap data set is of the form:

   $$\mathbf{W}^* = \left\{\left(Y_{i_1}^*, X_{i_1}^*\right), \left(Y_{i_2}^*, X_{i_2}^*\right), ..., \left(Y_{i_m}^*, X_{i_m}^*\right)\right\},$$

   where $i_1, i_2, ..., i_m$ is a random sample of integers 1 through m.
2. A GEE approach (assuming either an independent, exchangeable or autoregressive correlation) using Iteratively Reweighted Least Squares (IRLS) is then used to estimate the regression coefficients $\hat{\beta}_p^*$, for this bootstrap sample of cases.
3. We do this repeatedly, say B times, so we now have B bootstrap samples and B estimates of the regression coefficients, one from each bootstrap sample $\left\{\left(\hat{\beta}_1^*, ..., \hat{\beta}_p^*\right)_1, \left(\hat{\beta}_1^*, ..., \hat{\beta}_p^*\right)_2, ..., \left(\hat{\beta}_1^*, ..., \hat{\beta}_p^*\right)_B\right\}$.
4. The standard error of these estimated coefficients $\text{se}\left(\hat{\beta}_p\right)$ is simply the standard deviation of these B estimates. As before if these estimates are ordered in increasing value, $\left\{\left(\hat{\beta}_p^*\right)_{(1)}, \left(\hat{\beta}_p^*\right)_{(2)}, ..., \left(\hat{\beta}_p^*\right)_{(B)}\right\}$, a simple 95% bootstrap percentile confidence interval for the coefficient would be from the $0.025B^{th}$ to the $0.975B^{th}$ largest values.

Applying the marginal model (8.8) and both the GEE methodology and bootstrap resampling Algorithm 8.1 to the NAMEIT data assuming, exchangeable and autoregressive forms of the autoregressive matrix, gives the results summarised in Tables 8.10 and 8.12 respectively. For ease of

interpretation we only show the estimated coefficients $\hat{\beta}_{Time}$ and $\hat{\beta}_{Group}$ for the Time and Group covariates respectively. All analyses were carried out in `STATA v8` (StataCorp, 2003). The bootstrap standard errors and BC_a confidence interval estimates were based on 1000 bootstrap re-samples using Algorithm 8.1 and the `bootstrap` procedure in `STATA`. (See Appendix 4 for more details).

Results of Marginal Model (2) - exchangeable autocorrelation

Table 8.9 shows the estimated within subject correlation matrices for the eight dimensions of the SF-36 if we fit model (8.8) and assume a compound symmetric structure. The lower diagonal gives the observed matrix before the model fitting. The fitted autocorrelations ranged from 0.43 for the RE dimension to 0.63 for the PF and GH dimensions. On the whole, the model correlation estimates tend to be lower than the actual observed autocorrelations, for HRQoL assessments that are close together. Conversely the model correlation estimates tend to be larger than the observed correlations for HRQoL observations further apart in time. It will usually be the case that after model fitting the autocorrelations will appear to have been reduced (Fayers and Machin, 2000).

Table 8.10 shows the estimated regression coefficients for the group and time variables. There is some evidence that HRQoL increases over time for three dimensions of the SF-36, PF, BP and V. However, we are interested in the effect of treatment and comparing HRQoL over time across the Placebo and Neoral treated groups. Since there is no reliable evidence of a 'Group x Time' interaction the interpretation of the treatment group coefficient is relatively straightforward. The *p*-values for the treatment group regression coefficients in Table 8.10 suggest significant differences in HRQoL between the Neoral and Placebo groups on three dimensions of the SF-36 (RP, GH and BP).

The bootstrap and robust standard errors for the time and group coefficients are different, although the bootstrap SE estimate tends to be the same size or somewhat smaller than its robust counterpart. However both bootstrap and

robust SE estimates are of a similar order of magnitude. More importantly, the ratios of the estimated coefficient to its standard error are of similar size.

Table 8.9 Observed and estimated within-patient auto-correlation matrices (exchangeable model) from RA patients in the NAMEIT study. The lower diagonal gives the observed matrix before model fitting whilst the upper gives the exchangeable form after model-fitting[a]

a) Physical Function

Week	8	16	24	32	40	48
8	1.00	*0.63*	*0.63*	*0.63*	*0.63*	*0.63*
16	0.74	1.00	*0.63*	*0.63*	*0.63*	*0.63*
24	0.69	0.75	1.00	*0.63*	*0.63*	*0.63*
32	0.68	0.80	0.79	1.00	*0.63*	*0.63*
40	0.67	0.77	0.81	0.86	1.00	*0.63*
48	0.64	0.74	0.81	0.81	0.85	1.00

b) Role Physical

Week	8	16	24	32	40	48
8	1.00	*0.48*	*0.48*	*0.48*	*0.48*	*0.48*
16	0.53	1.00	*0.48*	*0.48*	*0.48*	*0.48*
24	0.39	0.57	1.00	*0.48*	*0.48*	*0.48*
32	0.30	0.56	0.67	1.00	*0.48*	*0.48*
40	0.42	0.52	0.60	0.61	1.00	*0.48*
48	0.40	0.59	0.67	0.64	0.71	1.00

c) Bodily Pain

Week	8	16	24	32	40	48
8	1.00	*0.48*	*0.48*	*0.48*	*0.48*	*0.48*
16	0.55	1.00	*0.48*	*0.48*	*0.48*	*0.48*
24	0.47	0.61	1.00	*0.48*	*0.48*	*0.48*
32	0.46	0.51	0.68	1.00	*0.48*	*0.48*
40	0.42	0.57	0.60	0.69	1.00	*0.48*
48	0.46	0.59	0.63	0.68	0.76	1.00

d) General Health

Week	8	16	24	32	40	48
8	1.00	*0.63*	*0.63*	*0.63*	*0.63*	*0.63*
16	0.68	1.00	*0.63*	*0.63*	*0.63*	*0.63*
24	0.67	0.80	1.00	*0.63*	*0.63*	*0.63*
32	0.67	0.77	0.83	1.00	*0.63*	*0.63*
40	0.65	0.72	0.79	0.84	1.00	*0.63*
48	0.65	0.75	0.84	0.82	0.85	1.00

e) Vitality

Week	8	16	24	32	40	48
8	1.00	*0.54*	*0.54*	*0.54*	*0.54*	*0.54*
16	0.58	1.00	*0.54*	*0.54*	*0.54*	*0.54*
24	0.54	0.71	1.00	*0.54*	*0.54*	*0.54*
32	0.59	0.68	0.71	1.00	*0.54*	*0.54*
40	0.49	0.67	0.68	0.77	1.00	*0.54*
48	0.53	0.66	0.72	0.72	0.76	1.00

f) Social Function

Week	8	16	24	32	40	48
8	1.00	*0.53*	*0.53*	*0.53*	*0.53*	*0.53*
16	0.53	1.00	*0.53*	*0.53*	*0.53*	*0.53*
24	0.55	0.63	1.00	*0.53*	*0.53*	*0.53*
32	0.46	0.63	0.70	1.00	*0.53*	*0.53*
40	0.51	0.58	0.64	0.71	1.00	*0.53*
48	0.45	0.58	0.64	0.71	0.71	1.00

g) Role Emotional

Week	8	16	24	32	40	48
8	1.00	*0.43*	*0.43*	*0.43*	*0.43*	*0.43*
16	0.47	1.00	*0.43*	*0.43*	*0.43*	*0.43*
24	0.40	0.59	1.00	*0.43*	*0.43*	*0.43*
32	0.32	0.56	0.62	1.00	*0.43*	*0.43*
40	0.46	0.53	0.56	0.54	1.00	*0.43*
48	0.37	0.49	0.58	0.54	0.69	1.00

h) Mental Health

Week	8	16	24	32	40	48
8	1.00	*0.52*	*0.52*	*0.52*	*0.52*	*0.52*
16	0.62	1.00	*0.52*	*0.52*	*0.52*	*0.52*
24	0.59	0.72	1.00	*0.52*	*0.52*	*0.52*
32	0.55	0.65	0.69	1.00	*0.52*	*0.52*
40	0.54	0.70	0.70	0.74	1.00	*0.52*
48	0.55	0.68	0.72	0.73	0.77	1.00

a) All models contain age, gender, time, baseline HRQoL and group as covariates.

Table 8.10: Comparison of robust and bootstrap SE's and CI's from the NAMEIT data with a Marginal Model and exchangeable autocorrelation

Dependent Variable		Coefficients $\hat{\beta}$	SE	$\hat{\beta}/SE$	p	95% CI Lower	Upper	Interval Length	Shape
Physical Function (n=222)	time	0.11	0.03	3.63	0.001	0.05	0.18	0.12	1.00
			0.03	*3.72*		*0.05*	*0.18*	*0.12*	*0.97*
	group	2.82	2.25	1.25	0.211	-1.60	7.24	8.84	1.00
			1.72	*1.64*		*-0.51*	*6.13*	*6.64*	*0.99*
Role Physical (n=221)	time	-0.06	0.07	-0.90	0.366	-0.19	0.07	0.26	1.00
			0.07	*-0.94*		*-0.19*	*0.06*	*0.25*	*0.91*
	group	9.49	3.93	2.42	0.016	1.79	17.19	15.40	1.00
			3.22	*2.95*		*3.63*	*16.62*	*12.99*	*1.22*
Bodily Pain (n=222)	time	0.16	0.03	4.69	0.001	0.10	0.23	0.14	1.00
			0.03	*4.88*		*0.10*	*0.23*	*0.13*	*1.05*
	group	4.23	1.97	2.14	0.032	0.36	8.10	7.74	1.00
			1.50	*2.82*		*1.44*	*7.25*	*5.81*	*1.08*
General Health (n=221)	time	0.04	0.03	1.67	0.095	-0.01	0.09	0.10	1.00
			0.03	*1.68*		*-0.01*	*0.09*	*0.10*	*0.95*
	group	-4.61	1.96	-2.35	0.019	-8.46	-0.76	7.69	1.00
			1.51	*-3.04*		*-7.36*	*-1.28*	*6.07*	*1.21*
Vitality (n=220)	time	0.09	0.03	3.09	0.002	0.03	0.14	0.11	1.00
			0.03	*3.05*		*0.04*	*0.15*	*0.11*	*1.24*
	group	2.67	1.80	1.48	0.14	-0.87	6.20	7.07	1.00
			1.41	*1.89*		*-0.19*	*5.42*	*5.61*	*0.96*
Social Function (n=222)	time	0.03	0.03	0.77	0.442	-0.04	0.09	0.13	1.00
			0.03	*0.79*		*-0.04*	*0.09*	*0.13*	*0.97*
	group	2.40	2.11	1.14	0.255	-1.73	6.54	8.27	1.00
			1.65	*1.46*		*-0.72*	*5.88*	*6.61*	*1.11*
Role Emotional (n=221)	time	-0.02	0.07	-0.32	0.752	-0.17	0.12	0.29	1.00
			0.08	*-0.31*		*-0.18*	*0.12*	*0.30*	*0.93*
	group	4.14	3.91	1.06	0.29	-3.52	11.81	15.33	1.00
			2.93	*1.41*		*-1.54*	*10.11*	*11.64*	*1.05*
Mental Health (n=221)	time	0.02	0.03	0.69	0.489	-0.03	0.07	0.10	1.00
			0.03	*0.68*		*-0.03*	*0.07*	*0.10*	*1.15*
	group	1.53	1.67	0.92	0.359	-1.74	4.79	6.53	1.00
			1.34	*1.14*		*-0.93*	*4.42*	*5.35*	*1.18*

Note: The bootstrap estimates of SE and BC_a Confidence Intervals are shown in italics below the model based estimates.

A crude test of statistical significance is to examine this ratio, if it is bigger than 2.0 then the estimated regression coefficient is likely to be significantly different from zero. Table 8.10 shows that for all the models where the original

(group or time) regression estimates are significant (i.e. ratios of estimate/SE > 2) then so too is the ratio of the estimate to its bootstrap standard error.

When we compare the bootstrap BC_a confidence intervals with the model-based estimates in Table 8.10 then the length of the bootstrap intervals tend to be the same size or slightly narrower than its robust counterpart. As before the bootstrap estimates are not constrained to be symmetric about the point-estimate of the regression coefficient. Qualitatively both the bootstrap and model based intervals include zero when the estimated regression coefficient is non-significant and exclude zero when the estimated coefficient is significant. Therefore, the actual practical interpretation of the confidence interval estimates is the same. That is for the RP, BP, and GH dimensions there is some evidence that the Neoral group has a better HRQoL than the Placebo group patients over time, after allowing for baseline HRQoL, age and gender.

The observed deviations between the fitted model and observed autocorrelations are not too great, suggesting that the assumption of compound symmetry is not unreasonable (Table 8.9). However, the compound symmetry structure for the correlations implied by exchangeable autocorrelation may not be acceptable on theoretical grounds since it is more likely that correlations between pairs of HRQoL observations widely separated in time will be lower than for observations closer together.

Results of Marginal Model (3) - autoregressive autocorrelation

To allow for a more complex pattern of correlations among the repeated HRQoL observations, we can move to an autoregressive structure. For first order auto-regression specification the correlation between time point's *j* and *k* is assumed to be $\rho^{|j-k|}$ (8.4). This model implies the correlation between a pair of HRQoL assessments on the same patient declines towards zero as the time separation between the assessments increases. The rate of decay is faster for larger values of ϕ.

Table 8.11 Observed and estimated within-patient auto-correlation matrices (exchangeable model) from RA patients in the NAMEIT study. The lower diagonal gives the observed matrix before model fitting whilst the upper gives the autoregressive form after model-fitting[a]

a) Physical Function

Week	8	16	24	32	40	48
8	1.00	*0.69*	*0.48*	*0.33*	*0.23*	*0.16*
16	0.74	1.00	*0.69*	*0.48*	*0.33*	*0.23*
24	0.69	0.75	1.00	*0.69*	*0.48*	*0.33*
32	0.68	0.80	0.79	1.00	*0.69*	*0.48*
40	0.67	0.77	0.81	0.86	1.00	*0.69*
48	0.64	0.74	0.81	0.81	0.85	1.00

b) Role Physical

Week	8	16	24	32	40	48
8	1.00	*0.57*	*0.33*	*0.19*	*0.11*	*0.06*
16	0.53	1.00	*0.57*	*0.33*	*0.19*	*0.11*
24	0.39	0.57	1.00	*0.57*	*0.33*	*0.19*
32	0.30	0.56	0.67	1.00	*0.57*	*0.33*
40	0.42	0.52	0.60	0.61	1.00	*0.57*
48	0.40	0.59	0.67	0.64	0.71	1.00

c) Bodily Pain

Week	8	16	24	32	40	48
8	1.00	*0.58*	*0.34*	*0.19*	*0.11*	*0.07*
16	0.55	1.00	*0.58*	*0.34*	*0.19*	*0.11*
24	0.47	0.61	1.00	*0.58*	*0.34*	*0.19*
32	0.46	0.51	0.68	1.00	*0.58*	*0.34*
40	0.42	0.57	0.60	0.69	1.00	*0.58*
48	0.46	0.59	0.63	0.68	0.76	1.00

d) General Health

Week	8	16	24	32	40	48
8	1.00	*0.70*	*0.49*	*0.35*	*0.24*	*0.17*
16	0.68	1.00	*0.70*	*0.49*	*0.35*	*0.24*
24	0.67	0.80	1.00	*0.70*	*0.49*	*0.35*
32	0.67	0.77	0.83	1.00	*0.70*	*0.49*
40	0.65	0.72	0.79	0.84	1.00	*0.70*
48	0.65	0.75	0.84	0.82	0.85	1.00

e) Vitality

Week	8	16	24	32	40	48
8	1.00	*0.62*	*0.38*	*0.24*	*0.15*	*0.09*
16	0.58	1.00	*0.62*	*0.38*	*0.24*	*0.15*
24	0.54	0.71	1.00	*0.62*	*0.38*	*0.24*
32	0.59	0.68	0.71	1.00	*0.62*	*0.38*
40	0.49	0.67	0.68	0.77	1.00	*0.62*
48	0.53	0.66	0.72	0.72	0.76	1.00

f) Social Function

Week	8	16	24	32	40	48
8	1.00	*0.60*	*0.36*	*0.22*	*0.13*	*0.08*
16	0.53	1.00	*0.60*	*0.36*	*0.22*	*0.13*
24	0.55	0.63	1.00	*0.60*	*0.36*	*0.22*
32	0.46	0.63	0.70	1.00	*0.60*	*0.36*
40	0.51	0.58	0.64	0.71	1.00	*0.60*
48	0.45	0.58	0.64	0.71	0.71	1.00

g) Role Emotional

Week	8	16	24	32	40	48
8	1.00	*0.51*	*0.27*	*0.14*	*0.07*	*0.04*
16	0.47	1.00	*0.51*	*0.27*	*0.14*	*0.07*
24	0.40	0.59	1.00	*0.51*	*0.27*	*0.14*
32	0.32	0.56	0.62	1.00	*0.51*	*0.27*
40	0.46	0.53	0.56	0.54	1.00	*0.51*
48	0.37	0.49	0.58	0.54	0.69	1.00

h) Mental Health

Week	8	16	24	32	40	48
8	1.00	*0.59*	*0.35*	*0.21*	*0.12*	*0.07*
16	0.62	1.00	*0.59*	*0.35*	*0.21*	*0.12*
24	0.59	0.72	1.00	*0.59*	*0.35*	*0.21*
32	0.55	0.65	0.69	1.00	*0.59*	*0.35*
40	0.54	0.70	0.70	0.74	1.00	*0.59*
48	0.55	0.68	0.72	0.73	0.77	1.00

a) All models contain age, gender, time, baseline HRQoL and group as covariates.

Table 8.11 shows the estimated within subject matrix correlation for the eight dimensions of the SF-36 if we fit model (8.8) and assume an autoregressive symmetry structure. The lower diagonal gives the observed matrix before the model fitting. The estimated with-subject correlation matrix (upper diagonal) after model fitting has the expected pattern in which the correlations decrease substantially as the separation between the observations increases.

On the whole, the model correlation estimates tend to be lower than the actual observed autocorrelations, particularly for HRQoL assessments that are further apart in time. For example (Table 8.11a), the observed correlation between the 8 and 16 week PF assessment was 0.74 compared to a model autocorrelation estimate of 0.69. Conversely the model correlation estimates tend to be smaller than the observed correlations for HRQoL observations further apart in time (the observed correlation between the 8 and 48 week PF scores was 0.64 compared to a model fitted estimate of 0.16). Thus the autoregressive model seems to substantially overestimate the decay towards zero in the correlation between pairs of measurements as the time separation between HRQoL assessments increases.

The autoregressive model estimates of the regression coefficients (and their standard errors) shown in Table 8.12 have changed but not substantially from the estimates in Table 8.10. There is some evidence that HRQoL increases over time for four dimensions of the SF-36, PF, BP, V and GH. The GH dimension now has a significant time effect ($p = 0.028$) compared to that obtained with the exchangeable correlation model ($p = 0.095$). However, we are interested in the effect of treatment and comparing HRQoL over time across the Placebo and Neoral treated groups. The *p*-values for the treatment group regression coefficients in Table 8.12 suggest significant differences in HRQoL between the Neoral and Placebo groups on two dimensions of the SF-36 (RP and GH) using a cut-off of $p < 0.05$ for statistical significance. The *p*-value for the group effect for the BP dimension is now marginally non-significant ($p = 0.074$) compared to a *p*-value of 0.032 for the exchangeable model.

Table 8.12: Comparison of robust and bootstrap SE's and CI's from the NAMEIT data with a Marginal Model and autoregressive autocorrelation

Dependent Variable		Coefficients $\hat{\beta}$	SE	$\hat{\beta}/SE$	P	95% CI Lower	Upper	Interval Length	Shape
Physical Function (n=219)	Time	0.13	0.03	3.65	0.001	0.06	0.20	0.14	1.00
			0.04	*2.91*		*0.04*	*0.21*	*0.17*	*0.96*
	group	3.10	2.24	1.39	0.165	-1.28	7.48	8.76	1.00
			3.04	*1.02*		*-2.66*	*9.34*	*12.00*	*1.08*
Role Physical (n=214)	Time	-0.06	0.93	-0.83	0.408	-0.20	0.08	0.28	1.00
			0.10	*-0.61*		*-0.25*	*0.14*	*0.39*	*1.01*
	group	10.11	3.96	2.55	0.011	2.35	17.87	15.52	1.00
			5.52	*1.83*		*-0.57*	*20.86*	*21.42*	*1.01*
Bodily Pain (n=219)	Time	0.18	0.04	5.12	0.001	0.11	0.25	0.14	1.00
			0.05	*3.95*		*0.09*	*0.28*	*0.18*	*0.99*
	group	3.54	1.99	1.78	0.074	-0.35	7.44	7.79	1.00
			2.65	*1.34*		*-1.74*	*8.58*	*10.32*	*0.95*
General Health (n=213)	Time	0.06	0.03	2.19	0.028	0.01	0.12	0.11	1.00
			0.04	*1.72*		*-0.01*	*0.13*	*0.15*	*0.98*
	group	-4.49	1.95	-2.30	0.022	-8.32	-0.66	7.66	1.00
			2.67	*-1.68*		*-10.07*	*0.45*	*10.52*	*0.89*
Vitality (n=217)	Time	0.11	0.03	3.51	0.001	0.05	0.17	0.12	1.00
			0.04	*2.70*		*0.03*	*0.19*	*0.16*	*1.07*
	group	2.59	1.79	1.45	0.148	-0.92	6.11	7.03	1.00
			2.48	*1.05*		*-2.12*	*7.21*	*9.33*	*0.98*
Social Function (n=219)	Time	0.03	0.04	0.80	0.426	-0.04	0.10	0.14	1.00
			0.05	*0.61*		*-0.06*	*0.11*	*0.18*	*0.93*
	group	2.52	2.10	1.20	0.23	-1.59	6.63	8.22	1.00
			2.74	*0.92*		*-2.65*	*7.91*	*10.56*	*1.04*
Role Emotional (n=212)	Time	0.02	0.08	0.21	0.832	-0.14	0.17	0.30	1.00
			0.10	*0.16*		*-0.18*	*0.22*	*0.40*	*1.05*
	group	4.87	3.96	1.23	0.219	-2.89	12.63	15.52	1.00
			5.52	*0.88*		*-6.63*	*15.10*	*21.73*	*0.89*
Mental Health (n=219)	Time	0.02	0.03	0.64	0.522	-0.04	0.07	0.11	1.00
			0.04	*0.48*		*-0.05*	*0.10*	*0.15*	*1.08*
	group	1.24	1.65	0.75	0.454	-2.00	4.48	6.48	1.00
			2.31	*0.54*		*-3.66*	*5.60*	*9.26*	*0.89*

Note: The bootstrap estimates of SE and BC_a Confidence Intervals are shown in italics below the model based estimates.

The bootstrap and robust standard errors for the time and group coefficients are different. This time the bootstrap estimates tend to be the same size or somewhat larger than its robust counterpart. This is turn affects the ratios of

the regression coefficient to its standard error $\hat{\beta}/SE(\hat{\beta})$. Using the bootstrap estimate of SE, none of the ratios are bigger than 2.0 implying that none of estimated regression coefficients are likely to be significantly different from zero.

When we compare the bootstrap BC_a confidence intervals with the model-based estimates in Table 8.12 then the length of the bootstrap intervals tend to be the somewhat longer than its robust counterpart. This in turn effects the interpretation of the treatment group regression coefficients since all the bootstrap based intervals include zero suggesting no evidence of treatment effect. This contrasts with model based robust confidence intervals which suggest that for the RP outcome there is some evidence that the Neoral group has a better HRQoL than the Placebo group patients over time, after allowing for baseline HRQoL, age and gender. Similarly, for the GH outcome, the model based robust confidence intervals suggest that for the GH that the Neoral group has a worse HRQoL than the Placebo group patients over time, after allowing for the covariates. It should be noted that the values for the lower and upper limits for the bootstrap estimates for the RP and GP dimensions respectively are close to zero, implying a treatment group effect is not completely unlikely with the data.

Table 8.11 shows that the observed deviations between the fitted model and observed autocorrelations are quite large, particularly for the HRQoL observations further apart in time. This suggests that the assumption of an autoregressive correlation structure is probably unreasonable for this dataset. Thus, on empirical grounds the compound symmetry structure for the correlations implied by exchangeable autocorrelation may be more suitable for the NAMEIT data. (Although this may not be acceptable on theoretical grounds since it is more likely that correlations between pairs of HRQoL observations widely separated in time will be lower than for observations closer together).

In practice it is often difficult to choose whether an exchangeable or autoregressive autocorrelation structure is appropriate (Fayers and Machin,

2000). We have seen that by examining the initial and subsequent (after model fitting) correlation matrices, (Tables 8.9 and 8.11) that an exchangeable matrix appears to be more likely, although this is not clear cut. The robustness of the inferences about β can be checked by fitting a final model using different covariance assumptions and comparing the two sets of estimates and their robust standard errors. If they differ substantially, a more careful treatment of the covariance model may be necessary (Diggle *et al* 2002).

We have developed models using each of the alternative autocorrelations. The two models are broadly similar with respect to the size of corresponding regression coefficients and their robust standard errors. The empirical evidence seems to suggest that the simpler exchangeable autocorrelation model is a reasonable approximation for the underlying covariance structure. The interpretation of the bootstrap estimates of SE and CI for the regression coefficients from the exchangeable correlation model are in agreement with the interpretation of their robust counterparts. Conversely, the bootstrap SE and CI estimates for the autoregressive model appear to reflect the greater uncertainty in this model (and perhaps the poorer fitting autocorrelation matrix) than the robust estimates.

Summary

In this chapter we have shown how the bootstrap can be used in the analysis of longitudinal HRQoL data collected at three or more time points. The simplest way to analyse repeated HRQoL assessments on individual patients is to reduce the observations to a single summary measure such as the AUC.

In the Leg Ulcer and NAMEIT datasets, hypothesis testing with the bootstrap comparing AUC between the two groups appears to offer no advantage over conventional significance tests such as the *t*-test and *MW* test. The bootstrap also produces CI estimates for the mean difference in AUC between groups that are very similar to their Normal theory (*t*-test) counterparts.

A second response feature analysis involved a multiple regression (or ANCOVA) of average follow-up HRQoL as the dependent variable with

treatment group and baseline HRQoL as covariates. In the two datasets studied, both the case and model based bootstrap re-sampling methods for estimating SEs and CIs for linear regression models gave estimates almost identical to the conventional values estimated by OLS. Therefore in the example datasets there appears to be little advantage in using bootstrap case or model based re-sampling to estimate standard errors and confidence intervals compared to conventional methods of confidence interval estimation from the OLS multiple regression model.

Finally in lieu of reducing the repeated HRQoL responses to a summary statistic we used a marginal model approach to analyse the seven individual HRQoL assessments in the NAMEIT dataset simultaneously. We compared the robust SE and CI estimated from the marginal model with their bootstrap counterparts. Depending on the assumed underlying autocorrelation structure the bootstrap and robust SEs and CIs estimates differed slightly.

If we assume an exchangeable autocorrelation structure for the repeated QoL assessments of the NAMEIT data then bootstrap and robust standard errors for the regression coefficients are different, although the bootstrap estimates tend to be the same size or somewhat smaller than its robust counterpart. Similarly, the length of the bootstrap intervals tended to be the same size or somewhat narrower than its robust counterpart. Despite these subtle differences the practical interpretation of the regression coefficients was the same.

Alternatively, if we assume an autoregressive autocorrelation for the NAMEIT data then again the bootstrap and robust standard errors for the regression coefficients are different, although this time the bootstrap SE estimates tend to be larger than its robust counterpart. Similarly, the length of the bootstrap intervals tended to be the somewhat larger than its robust counterpart. These variations may lead to different practical interpretations of the regression coefficients from the model. The bootstrap SE and CI estimates for the autoregressive model appear to reflect the greater uncertainty in this model

(and perhaps the poorer fitting autocorrelation matrix) than the robust estimates.

Conclusion

The use of the bootstrap to estimate SEs and CIs for marginal longitudinal models appears to offer little advantage (in the NAMEIT data) compared to the conventional robust estimates. The only advantage the bootstrap may have is if there is considerable uncertainty about the autocorrelation structure then the bootstrap may produce larger SE and wider CIs than its robust counterpart and hence lead to a more conservative interpretation of the regression coefficients from the model. As Campbell (2001) notes, "*When the standard and the bootstrap methods agree, we can be more confident about the inference we are making and this is an important use of the bootstrap. When they disagree more caution is needed, but the relatively simple assumptions required by the bootstrap method for validity mean that in general it is to be preferred.*"

Chapter 9: Discussion

Introduction

In the introduction to this thesis we started out with the aim of comparing bootstrap computer simulation methods with standard methods of sample size determination and analysis of HRQoL measures (particularly the SF-36).

We have assumed that there exists an underlying latent continuous HRQoL variable, although most HRQoL measures actually generate data that has bounded, discrete and non-standard distributions. Therefore standard methods of analysis that assume Normality and constant variance may not be appropriate. Computer intensive methods such as the bootstrap that make no distributional assumptions may therefore be more appropriate for estimating sample sizes and analysing HRQoL data.

If we use a hypothesis-testing framework then our null hypothesis (H_0) was that bootstrap methods are not more appropriate for analysing HRQoL outcomes (determining sample size and calculating SEs and CIs) than conventional methods (see Table 9.1).

In the datasets and outcomes studied we have shown that use of the bootstrap does not lead to different sample size estimates or different SE and CI estimates compared to conventional methods. Thus, we cannot effectively reject the null hypothesis and accept the alternative i.e. that in the datasets and outcomes studied the use of the bootstrap for sample size determination and SE and CI estimation is more appropriate than conventional methods.

Table 9.1: potential null and alternative hypotheses

	Hypothesis
H_0	Bootstrap methods **are not more appropriate** for analysing HRQoL outcomes than standard methods.
H_A	Bootstrap methods **are more appropriate** for analysing HRQoL outcomes than standard methods.

There are several limitations or caveats to this rather simple conclusion, which I will discuss further in the rest of this chapter.

Two group comparisons

My analysis has concentrated on using the bootstrap in clinical trials or for two group comparisons. We have focused on a number of measures of statistical accuracy: standard errors, biases, and confidence intervals. All of these are measures of accuracy for parameters (such as the treatment group coefficient, β_{Group}) of a regression model.

Sometimes we are interested in predicting the HRQoL of an individual given various explanatory or prognostic variables. The bootstrap can be used to see how well a model predicts the HRQoL response value of a future observation i.e. to assess prediction error. Efron and Tibshirani (1993, Chapter 17) discuss the use of the bootstrap for cross-validation and prediction error. However, all the predictive models still use the regression coefficients from the original model (estimated by OLS in the linear regression case). So rightly or wrongly we have decided to concentrate on the use of the bootstrap in assessing the statistical accuracy of an estimated parameter, via its SE or CI.

Generalisability

The generalisability of the results of this thesis could be called into question as the results only apply to the limited number of datasets studied (five) and the SF-36 outcome. This means we have only considered selected distributions for the SF-36 outcomes and covariates and assumed an underlying latent continuous HRQoL variable for each of the eight dimensions of the SF-36.

The SF-36 outcome is the most widely used generic HRQoL measure in the world today, so that is one obvious reason to use it (Kaplan, 1998). Secondly, I had easy access to a variety of datasets that had previously used the SF-36 outcome. The five studies (General Population, CPSW, OA Knee, Leg Ulcer and NAMEIT), and datasets were well known to me. They illustrate the use of HRQoL outcomes across a variety of studies including cross-sectional

surveys, RCTs, non-randomised before and after studies and longitudinal designs. So on practical and pragmatic grounds, I felt it was appropriate to use such datasets because of their familiar nature and the analysis was easy to understand and interpret.

The SF-36 is a multi-dimensional outcome with eight dimensions. As described in the Introduction the eight dimensions have a variety of different distributions. I believe these distributions are not atypical of other generic HRQoL measures such as the NHP and EORTC QLQ-C30. The distributions we considered were chosen based on our experiences with HRQoL data in a variety of settings. So I believe that my results about the bootstrap may have generalisability to other HRQoL outcomes (besides the SF-36) used in other studies and populations. Although strictly speaking our results only apply to the SF-36 outcome and the observed datasets, since we have considered only a few distributions for the HRQoL outcomes (and covariates) and therefore we cannot make sweeping generalisations about the impact of the bootstrap on other HRQoL outcomes, used in other studies. Therefore these results need to be replicated with other HRQoL measures on other datasets and populations.

Missing values

I have assumed that any missing HRQoL values in the datasets are Missing Completely at Random (MCAR). This means that the probability of the HRQoL response being missing is independent of the scores on the previous observed questionnaires and independent of the current and future scores had they been observed.

I have assumed that the reduced dataset represents a randomly drawn sub-sample of the full dataset and the inferences drawn can be considered reasonable. This is a strong assumption and needs to be checked. However, there is an extensive literature on the issue of missing values and HRQoL outcomes. The imputation of missing HRQoL scores and the analysis of HRQoL data with missing values is discussed in several papers (Curran *et al* 1998; Fayers *et al* 1998; Troxel *et al* 1998) and book chapters (Fayers and

Machin, 2000 Chapter 11; Fairclough, 2002 Chapters 4 to 8). As the subject of this thesis is the use of the bootstrap I have played down the issue of missing data, although this is an important issue particularly with longitudinal studies.

Sample sizes of datasets used in the thesis

The various datasets used in this study all had a sample size in excess of 100 patients. Some caution should be used in applying my results to smaller sample sizes. However the robustness of the conventional two-sample *t*-test and ANCOVA, for three-, four- and five point ordinal scale data using assigned scores has been demonstrated for sample sizes as small as 20 (Heeren and D'Agostino, 1987; Sullivan and D'Agostino, 2003).

As the various empirical "effect size" estimates calculated in Chapter 5 show, large dramatic differences in HRQoL (using the SF-36 outcome) between groups are unlikely and inconsistent with the observed data. Therefore reasonably large sample sizes will be required to detect significant differences in HRQoL between groups (whatever method of sample-size estimation is used). So it would not seem unreasonable to use studies with sample sizes in excess of 100 patients, as well designed studies (in my opinion) should have at least this number of patients, to detect clinically meaningful and practically important differences in HRQoL between groups.

As an alternative we could try using the bootstrap with smaller sample sizes and then compare its performance with conventional methods. However, this may be slightly unrealistic, as we have already demonstrated that well designed quality of life studies should have reasonably large sample sizes.

Simple bootstrapping may not be very successful in small samples (say < 9 observations), since the observations themselves are less likely to be representative of the study population. As Campbell (2001) states, *"In very small samples even a badly fitting parametric analysis may outperform a non-parametric analysis, by providing less variable results at the expense of a tolerable amount of bias."*

Also we have not compared the bootstrap with alternative models that assume a parametric model distribution for the HRQoL outcome. Again this is a slightly unrealistic scenario as the figures in Chapter 2 show. None of the eight dimensions of the SF-36 appears to follow a Normal distribution and only one, the Vitality dimension, seems to have a symmetrical distribution.

Multi-dimensionality and multiple endpoints

The SF-36 is an example of a profile measure of HRQoL with eight different dimensions. This is a common feature of many other HRQoL outcomes such as the NHP and EORTC QLQ-C30 that can have a number of dimensions. When several HRQoL outcomes are collected on the same people, it is always possible to test each variable separately.

Fairclough (Chapter 11, 2002) extensively discusses the issue of multiple comparisons in clinical trials assessing HRQoL. Multiple comparisons arise from three main sources:

(1) Multiple HRQoL measures (scales or subscales);
(2) Repeated post randomisation assessments;
(3) Multiple (three or more) treatment arms.

Indeed throughout this thesis we have taken a simpler "*univariate*" approach where each individual HRQoL dimension is analysed separately. (Although in some cases the test statistic is derived from a multiple-variable longitudinal marginal analysis). For example, if two groups are compared using the SF-36 then a difference between the means for the two groups can be tested separately for each of the eight dimensions of the SF-36. Unfortunately, there is a drawback to this approach because of the repeated use of significance tests means the probability of falsely finding at least one significant difference accumulates with the number of tests carried out. That is univariate tests of each HRQoL domain and time point can seriously inflate the Type I (false-positive) α error rate for the overall trial such that the analyst is unable to distinguish between true and false positive differences.

Fairclough (2002) mentions three typical strategies to reduce the problem of multiple comparisons and multiple endpoints:

(1) *A priori* specification of a limited number of *confirmatory* tests;

(2) The use of summary measures or statistics (e.g. AUC);

(3) Multiple comparison procedures including α adjustments (e.g. Bonferroni).

In practice, Fairclough believes a combination of all three strategies, that is, focussed hypotheses, summary measures, and multiple comparison procedures are necessary.

Limiting the number of confirmatory tests

One recommended solution (Fayers and Machin, 2000) to the multiple comparison problem are to specify a limited number (≤ 3) of *a priori* endpoints in the design of the trial. While theoretically improving the overall Type I error rate for a study, in practice investigators are reluctant to ignore the remaining data and still present CI and even formal hypothesis test results for the remaining scales/endpoints. A more important critique of this approach is an ethical question about the collection of data that will not be used in the primary analysis (Fairclough, 2002).

Summary measures and statistics

Well chosen summary measures or statistics often have a greater power to detect patterns of consistent HRQoL differences across time or measures (Fairclough, 2002). The use of summary measures such as the AUC (as described in Chapter 8) is a good strategy that both simplifies the presentation of the results and reduces the multiplicity of the repeated assessments over time. For example the use of summary measures such as the AUC in the NAMEIT study can reduce the number of hypothesis tests from six (the number of follow-up HRQoL assessments) to one for each dimension of the SF-36.

Global tests and multiple comparison procedures

Although as we have mentioned above there are ways of adjusting significance levels in order to allow for multiple testing a *single global*

multivariate test that uses the information from all variables together may be preferable. A global test generates a single statistic for testing the overall treatment effect and results in the acceptance or rejection of a set of *K* hypotheses H_0: $H_{0(1)}$, $H_{0(2)}$,...,$H_{0(K)}$. For example, the vector of all eight mean HRQoL dimension scores of the SF-36 in the Intervention group is the same as the vector for all eight mean scores for the Control group. One solution for the global test is to use a multivariate statistic such as Hotelling's T^2 test for the two-group situation or MANOVA (Multivariate Analysis of Variance) for more than two groups (Manly, 1994). The problem with such multivariate methods is that they test very general hypotheses (e.g. does one group differ in some non-specified way in their HRQoL from another) and so consequently have very poor power to detect any real difference. Thus, more often than not they will give a non-significant result (Walters *et al* 2001a). When the overall test of H_0 has been rejected, the question still remains, *"Which of the individual hypotheses can be rejected?"* A global test does not allow inferences to be made about individual endpoints and a series of univariate tests must be performed for these comparisons (Fairclough, 2002).

Tandon (1990) suggested a parametric method that was more specific. For example to compare two groups, calculate *t*-tests for each dimension and then find

$$z = \frac{J'S^{-1}t}{(J'S^{-1}J)^{1/2}}, \qquad (9.1)$$

where J' = (1,...,1, 1), **S** is the estimated correlation matrix and *t* is the vector of *t*-statistics from the separate univariate *t*-tests. The test statistic z has an asymptotic standard Normal distribution. The main drawback of this method is that it does not give an estimate of the treatment effect; it just provides a test statistic (9.1).

Pragmatically, one does not like to use multivariate global hypothesis tests, since the interpretation of the results of such tests is difficult. So I have tended to use univariate methods and analyse each dimension of the SF-36 separately, one at a time. Indeed, Fairclough (2002) advocates the use of

multiple comparison procedures using a *set of univariate test statistics* rather than those using a single multivariate statistic.

Alpha adjustments for K univariate tests

Fairclough (2002) comprehensively describes a number of procedures that can be used to control the Type I error rate for K multiple comparisons using the following notation. If $H_{0(1)}, H_{0(2)},\ldots,H_{0(K)}$ denotes the K null hypotheses that are to be tested and $T_{0(1)}, T_{0(2)},\ldots,T_{0(K)}$ denotes the corresponding K test statistics. The observed p-value, $p_{(k)}$, denotes the *unadjusted* probability of observing the test statistic, $T_{(k)}$, or a more extreme value if the null hypothesis $H_{0(k)}$ is true. The K ordered p-values from smallest to largest can be written as $p_{[1]} \leq p_{[2]} \leq \ldots \leq p_{[K]}$.

Most of the alpha adjustments for K univariate tests described in Fairclough (2002) are a variation of the simple Bonferroni correction described in Chapter 4. The Bonferroni correction adjusts the test statistics on K endpoints. The *global test* is based on the smallest p-value, $p_{[1]}$ for the K HRQoL endpoints. The global null hypothesis $\boldsymbol{H_0}$: $H_{0(1)}, H_{0(2)},\ldots,H_{0(K)}$ is rejected when

$$p_{[1]} \leq \alpha/K\,. \tag{9.2}$$

For individual HRQoL endpoints the Bonferroni procedure is to accept as statistically significant only those tests with p-values that are less than α/K.

The Bonferroni procedure controls the *experiment wise error rate* well but is well known to be quite conservative. (The experiment wise error rate is the probability of incorrectly rejecting at least one true null hypothesis, regardless of which (if any) null hypotheses are true). If the K outcomes are uncorrelated (the tests are independent) and the null hypotheses are all true, then the probability of rejecting at least one of the K hypotheses is approximately[1] αK when α is small. The Bonferroni approach focuses on the detection of large differences in one or more endpoints and is insensitive to a pattern of smaller differences that are all in the same direction.

---

[1] $\text{Prob}[\min(\text{p-value}) \leq \alpha] = 1 - (1 - \alpha)^K \approx \alpha K$

Resampling techniques

The major limitation of all the global tests is that they were developed to control the Type I error rate under the most conservative condition, *K* independent tests. However, in most studies of HRQoL, the *K* endpoints are moderately correlated. As a result, these procedures are very conservative and the power to detect meaningful differences is severely reduced. Fairclough describes a bootstrap algorithm for global hypothesis tests (Fairclough, Chapter 11, p 254; 2002) to address this problem.

The general idea is to obtain an estimate of the distribution of the cut-off test statistic (T_{COT}) for the multiple comparison procedure for endpoints with unknown correlation structure. For example the simple Bonferroni cut-off test statistic and its associated *p*-value is $T_{[1]}$ and $p_{[1]}$ respectively (the largest test statistic or the corresponding smallest *p*-value from the *K* univariate tests). The bootstrap procedure was first proposed by Westfall and Young (1989) and adapted by Reitmeir and Wasser (1999) for multiple comparisons of *K* endpoints between two treatment groups.

The procedure is shown in detail in Algorithm 9.1 taken from Fairclough (2002). The procedure is basically a modification of Algorithm 7.1, although this time the data is a matrix of responses **X** rather than a simple vector **x**. Suppose we have two groups of subjects $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2,\ldots,\mathbf{Z}_n)$ and $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots,\mathbf{Y}_m)$, where the $\mathbf{Z}_i$ (and $\mathbf{Y}_i$) consists of a row vector of *K* HRQoL responses for subject *i* that is $\mathbf{Z}_i = (z_{i1}, z_{i2},\ldots,z_{iK})$. If we let the combined sample of HRQoL responses for **Z** and **Y** be denoted by **X** then:

Algorithm 9.1

Computation of the bootstrap global test statistic for *K* endpoints for a two group study

1. Identify the statistic for the global test T_{COT} or p_{COT} and calculate it from the observed data. For example the simple Bonferroni cut-off test statistic and its associated *p*-value is $T_{[1]}$ and $p_{[1]}$ respectively (the

largest test statistic or the corresponding smallest *p*-value from the *K* univariate tests).

2. Draw a random sample of subjects with replacement (bootstrap sample) from the pooled sample **X** of the same size as the original sample, call the first *n* observations **Z*** and the remaining *m* observations **Y***.
3. Evaluate the global test T^*_{COT} or p^*_{COT} statistic from the data associated with the subjects drawn from the bootstrap sample.
4. Repeat the previous two steps *B* times, *B* = 10,000 is recommended (Reitmeir and Wassmer, 1999). We now have a bootstrap distribution of *B* values of T^*_{COT} and p^*_{COT} respectively i.e.

$$\left(T^{*1}_{COT}, T^{*2}_{COT}, \ldots, T^{*B}_{COT}\right) \text{ and } \left(p^{*1}_{COT}, p^{*2}_{COT}, \ldots, p^{*B}_{COT}\right). \qquad (9.3)$$

5. Since the bootstrap statistics were calculated under the null hypothesis by generating the bootstrap samples from the pooled sample, **X**, the distribution of the global test statistic is approximated by the distribution of the p^{*b}_{COT}.
6. The *p*-value for the global test is the proportion of the *B* bootstrap statistics that are equal to or more extreme than the observed data statistic.

$$\hat{A}SL_{boot} = \frac{\#\left\{p^{*b}_{COT} \le p_{COT}\right\}}{B} \text{ or } \hat{A}SL_{boot} = \frac{\#\left\{T^{*b}_{COT} \ge T_{COT}\right\}}{B} \qquad (9.4)$$

7. If this proportion is less than α, the global test is rejected.

I have not used Algorithm 9.1, since there is no general consensus on what procedure to adopt allow for multiple comparisons (Altman *et al* 2000). I believe in following Altman's recommendation of reporting unadjusted *p*-values (to three decimal places/significant figures) and confidence limits with a suitable note of caution with respect to interpretation. As Perneger (1998) concludes: *"Simply describing what tests of significance have been performed, and why, is generally the best way of dealing with multiple comparisons."*

Pragmatically, one does not like to use multivariate global hypothesis tests, since the interpretation of the results of such tests is difficult. I prefer using a set of univariate test statistics since they are much easier to implement and report. Throughout this thesis, I have favoured a combination of Fayers and Machin's (2001) and Altman *et al* (2000) approaches to multiple comparisons/endpoints, although as I mention above other alternatives are available. My favoured approach is to identify the main study HRQoL endpoints in advance, limit the number of confirmatory hypothesis tests to these outcomes, and report unadjusted *p*-values and confidence limits with a suitable note of caution with respect to interpretation.

Ordinality of HRQoL outcomes

One of the fundamental assumptions we have made throughout this thesis is that there exists an underlying continuous latent variable that measures HRQoL, and that the actual measured outcomes are ordered categories that reflect contiguous intervals along this continuum. If the goal of the analysis is to assess the magnitude of the treatment effect on the ordered outcome, then an appealing approach is to assign numeric scores to the ordered categories and to use a more familiar linear regression method for analysis. Most HRQoL measures such as the SF-36 actually do assign numeric scores to the ordered categories. It is then common practice in medical studies to compare means between groups using conventional linear regression methods. Indeed the comprehensive textbooks on HRQoL analysis by Fairclough (2002) and Fayers and Machin (2000) use this approach. Following on from this we have assumed that our main interest lies in comparing location between treatments i.e. comparing means.

If interest lies elsewhere, for example in comparing the relative frequencies of cumulative probabilities in the ordered categories between treatments, then other techniques such as the proportional odds model would be more appropriate, (described in Chapter 7; Lall *et al* 2001; Walters *et al* 2001a). A limitation of these approaches is in the interpretation of the models.

Heeren and D'Agostino (1987) have demonstrated the robustness of the two independent samples *t*-test when applied to three-, four- and five point ordinal scaled data using assigned scores, in sample sizes as small as 20 subjects per group. Sullivan and D'Agostino (2003) have expanded this work to account for a covariate when the outcome is ordinal in nature. They again assign numeric scores to the distinct response categories and compare means between treatment groups adjusting for a covariate reflecting a baseline assessment measured on the same scale. Their simulation study shows that in the presence of three-, four- and five point ordinal data and small sample sizes (as low as 20 per group) that both ANCOVA and the two independent sample *t*-test on difference scores are robust and produce actual significance levels close to the nominal significance levels.

The bootstrap

Number of bootstrap replications B

How large should we take *B*, the number of bootstrap replications to evaluate the bootstrap estimate of the standard error $\hat{se}_{Boot}$ and estimate BC_a confidence intervals? The ideal bootstrap estimate takes $B = \infty$. This is obviously not possible so the number of bootstrap replications *B* depends mainly how long it takes the computer to evaluate the function $\hat{\theta} = s(x)$.

Time constraints may dictate a smaller value of *B* (e.g. 1000) if $\hat{\theta} = s(x)$ is a very complex function of ***x***, as in the marginal model examples of Chapter 8. We used between 1000 and 5000 bootstrap samples to estimate SE and BC_a CIs. We used more ($B = 10{,}000$) for the simpler sample size algorithm (6.1).

Efron and Tibshirani (1993), describe two rules of thumb based on their experiences for estimating bootstrap SEs.

(1) Even a small number of bootstrap replications, say $B = 25$, is usually informative. $B = 50$ is often enough to give a good estimate of $se_F(\hat{\theta})$.

(2) Very seldom are more than $B = 200$ replications needed for estimating a standard error. (Much bigger values of *B* are required for bootstrap confidence intervals. Efron and Tibshirani (1993) and Davison and

Hinckley (1997) tend to use between 1000 and 2000 replications in the examples throughout their books, for estimating percentile based CIs).

Bootstrap case resampling vs. model based resampling

The results of Chapters 7 and 8 show that there is little to choose from between the case and model based resampling algorithms (7.3 and 7.4) for the multiple linear regression model for estimating SEs and CIs. Table 7.5 provides a summary of the issues involved. Since there was very little difference in the SE and CI estimates from the datasets used, for simplicity one would tend to favour a case based resampling approach. Indeed this was the resampling method for the longitudinal marginal model in Chapter 8.

Bootstrap model based resampling for marginal model

In Chapter 8, for simplicity we described only a simple case based resampling Algorithm (8.1) for the marginal model. In Algorithm 8.1 we effectively carried out a stratified random resampling with replacement. That is we sampled with replacement blocks or clusters of each patients' repeated HRQoL responses. In theory one should be able to adapt the linear regression model based resampling Algorithm 7.4 to the marginal model. The resampling algorithm would be rather complex particularly for autoregressive autocorrelation structures and for unbalanced datasets, with HRQoL assessments at unequally spaced time points. One would have to take into account that the residuals were not independent and uncorrelated, and for the autoregressive correlation structure, that the correlation between residuals within a patient declined over time. This is a very interesting avenue and requires further exploration with other longitudinal datasets.

Bootstrap observed value of the test statistic

The bootstrap is mainly used as a method for assessing statistical accuracy i.e. SE, biases and CIs. Throughout this thesis I have always used the observed value of the test statistic $t(\boldsymbol{x})$ or parameter estimate $\hat{\theta}$ as our best guess at the true value of the unknown parameter θ or statistic. For example if we are interested in estimating the population mean μ (from a random sample) it may seem that the best estimator of the mean of the population is

the mean of the *B* bootstrap estimates. This is turns out not to be the case as the mean of the bootstrap means is biased. The original observed sample mean, $\bar{x}$ from the original data, is always the best estimate of the population mean. The same result applies for other statistics such as the median and regression coefficients.

Use of the Bootstrap for estimating sample size

In Chapter 6 we described and used a bootstrap resampling method (Algorithm 6.1) for estimating power and sample size. A limitation of this chapter was that it only dealt with statistical power (i.e. Type II error). Therefore one could legitimately argue that the issue of Type I error and false positive results has not been adequately addressed.

The bootstrap methodology provides an ideal opportunity to consider Type I error. Resampling Algorithm 6.1 can easily be adapted for this. It simply involves modification of step 1 and not adding δ to the second simulated sample of patients. Under the true null hypothesis of no difference in distributions, the actual Type I error rate can be computed by determining the proportion of simulated cases which had significance levels at or below its nominal value. For a nominal Type I error rate of $\alpha = 0.05$, (i.e. using a cut-off of $p \leq 0.05$ for statistical significance) we would expect 5% of the bootstrap samples to give a (false-positive) significant result under the true null hypothesis of no difference in distributions. The robustness of each test can then be determining by comparing the actual Type I error rates to the nominal Type I error rates.

However, I believe the issue of Type I error is of less importance for sample size estimation, than Type II errors. The Type I error rate can easily be controlled for after the data have been collected by reducing the *p*-value cut-off for statistical significance. Conversely, the Type II error (and power) can only be controlled for at the design stage. Therefore I felt that it was more important to concentrate on the issue of power rather than significance, in Chapter 6, as we were interested in sample size estimation at the study design stage.

For presentational purposes this thesis has separated the methods of data analysis (Chapters 7 and 8) from sample size determination (Chapter 6), although most practitioners would recommend basing sample size directly on the likely method of analysis. For example, if a two independent samples *t*-test is proposed for analysis, then sample size should be based on the *t*-test. Although frequently statisticians will base the sample size calculation on the *t*-test, using Equations (4.3 and 4.4), but actually use multiple regression methods to analyse the data and adjust the outcome variable for other covariates besides the treatment group.

Whitehead's (1993) method for sample determination is derived from the proportional odds model (and the Mann-Whitney test). We did not evaluate the proportional odds model as part of the bootstrap. This was because ordinal regression is equivalent to the *MW* test when there is only a 0/1 variable in the regression (Campbell, 2001). The advantage of the proportional odds model is that it allows the estimation of confidence intervals for the treatment group effect and for the adjustment of the HRQoL outcome for other covariates. Also the odds ratio effect size ($OR_{Ordinal}$) is more readily interpretable than the $p_{Noether}$ effect size, although as we have noted before both are not as easily understandable as a difference in mean HRQoL.

We have demonstrated that the effect size p_{Nother} or Pr(Y > X) is difficult to estimate for non-Normally distributed data. This effect size is also rather difficult to interpret, although we have demonstrated in Chapter 5 that all effect sizes can be reduced to a common metric using the λ and θ statistics. If we use Whitehead's (1993) method to determine sample size (and either a *MW* test or a proportional odds model to analyse the data), then the advantage of Whitehead's method is that we do not need to know the full distribution of the HRQoL outcome variable. Knowledge of the proportions of responses in around five ordered categories for the HRQoL outcome is usually sufficient to estimate the required sample size (given the $OR_{Ordinal}$ effect size).

Specification of the alternative hypothesis.

Throughout Chapters 5 and 6 we only considered the situation where a single dimension of HRQoL is used at a single endpoint. We have assumed a rather simple form of the alternative hypothesis that the new treatment/intervention would improve HRQoL compared to the control/standard therapy. This form of hypothesis (superiority vs. equivalence) may be more complicated than actually presented. For example, superiority may be due to an improvement in HRQoL in the Intervention group, or due to the one therapy causing a decline due to an adverse experience. Alternatively, the HRQoL superiority for one group may be due to a treatment preventing an adverse clinical event. This may have an important impact for HRQoL outcome distributions that are not symmetrical, especially if they are bounded. All of these considerations are needed for determining study hypotheses and sample sizes. However, the assumption of a simple form of the alternative hypothesis that new treatment/intervention would improve HRQoL compared to the control/standard therapy, is not an unrealistic scenario for most superiority trials and is frequently used for other clinical outcomes.

Clinically meaningful change and the Minimum Important Difference

There is an extensive literature on the important issue of clinically meaningful change and the minimum important difference (MID) for HRQoL outcomes. As the subject of this thesis is the use of computer intensive methods such as the bootstrap we have played down the issue of the MID. Again for brevity and practical purposes of sample size estimation this thesis has assumed the MID for the SF-36 outcome is around five points for each dimension. This is an important issue in sample size estimation. The interested reader is referred to a series of papers from the in the *Mayo Clinic Proceedings* for more detailed discussion (Cella *et al* 2002; Frost *et al* 2002; Guyatt *et al* 2002; Sloan *et al* 2002; Sprangers *et al* 2002 and Symonds *et al* 2002) and Norman *et al* 2003.

SF-36 Version 2.0

We have based our analysis on the original UK version of the SF-36 version 1.0 (a copy of which is Appendix 1). A revised second version of SF-36 has been developed called the SF-36 v2 (Jenkinson, 1999; Ware *et al* 2000). This

has expanded the range of possible responses to the two Role Limitations questions 4 and 5 (see Appendix 1) from two-point (Yes/No) scale to a five-point ordinal scale (ranging from "all of the time" to "none of the time"). This will have the effect of increasing the number of possible discrete scores for the RE and RP dimensions from four (0, 33.3, 66.7, 100) and five (0, 25, 50, 75, 100) values to 13 (0, 8.3, 16.7,...,91.7, 100) and 17 (0, 6.25, 12.5,...,93.75, 100) values respectively.

These changes to the SF-36 will obviously expand the discrete responses for the RP and RE dimensions and perhaps make them have similar distributions to the other SF-36 dimensions such as PF and V. One of the likely consequences of this is that with all eight dimensions of the SF-36 now consisting of seven or more discrete response categories then we can treat these scores as continuous and use statistical methods for comparing means (such as *t*-tests and multiple regression). This does of course rely on the fundamental assumption (as stated in Chapter 1) that there exists an underlying continuous latent variable that measures HRQoL (for each dimension) and that the actual measured outcomes are ordered categories that reflect contiguous intervals along this continuum.

Are the results surprising or unexpected?

Finally are the results of this thesis all that surprising or unexpected? We have shown that the use of bootstrap methods (Algorithm 6.1) for sample size estimation appears to offer little advantage compared to four standard methods in the datasets studied. Pessimistically, it is hard to see how sample size calculations based on bootstrap methods would have general appeal, as they would need to be based on a sample distribution from the population of interest, something we rarely have. However, with a reliable pilot population dataset, bootstrap sample size methods may be used to check the sensitivity of various assumptions and assumed forms of treatment effects.

Firstly, if we are prepared to assume that there exists an underlying continuous latent variable that quantifies the HRQoL response of interest and that the actual measured HRQoL is on an ordered category that reflects

contiguous intervals along this interval. Secondly, if the goal of the analysis is to assess the magnitude of a treatment effect on the (ordered) HRQoL outcome, then an appealing approach is to assign numeric scores to the ordered categories and to compare means between groups. Statistical theory says that if the distribution of *x* is Normal, so will be the distribution of $\bar{x}$. Much more importantly, even if the distribution of *x* is not Normal, that of the sample mean $\bar{x}$ will become closer to the Normal distribution with mean μ and variance σ^2/n as *n* gets larger. This is a consequence of the CLT (Hogg and Tanis, 1988).

The Normal distribution is strictly only the limiting form of the sampling distribution as *n* increases to infinity, but it provides a remarkable good approximation to the sampling distribution even when *n* is small and the distribution of *x* is far from Normal (Armitage *et al* 2002). This implies that the distribution of the sample means for the SF-36 HRQoL data shown in Figure 2.1 is approximately $N(\mu, \sigma^2/n)$ when *n* is sufficiently large and μ and σ^2 are the mean and variance of the underlying HRQoL distribution from which the sample came.

If the sample size is "sufficiently large" the CLT guarantees that the sample means will be approximately Normally distributed (Hogg and Tanis, 1998). Thus, if the investigator is planning a large study and the sample mean is an appropriate summary measure of the HRQoL outcome, then pragmatically there is no need to worry about the distribution of the HRQoL outcome and we can use equation (4.4) and the effect size Δ_{Normal} (4.3) to estimate sample sizes (Walters *et al* 2001a; 2001b). Furthermore, Chapter 5 has shown that the empirical Δ_{Normal} effect sizes were mainly in the "small" to "moderate" (0.30 to 0.50) range for the SF-36 HRQoL outcome using Cohen's (1988) classification. Therefore dramatic effects are unlikely in HRQoL studies using the SF-36 as an outcome and so large samples sizes are likely to be required. So perhaps unsurprisingly, the results of Chapter 6 reflect the robustness of conventional methods with large sample sizes and the application of the CLT

to sample means even for HRQoL data with such bounded, discrete and skewed distributions as shown in Figure 2.1.

Generally, if *n* is greater than 25, these approximations to Normality for sample means will be good. The work of Heeren and D'Agostino (1987) and Sullivan and D'Agostino (2003) described previously certainly supports the robustness of the two independent samples *t*-test and ANCOVA when applied to three-, four- and five point ordinal scaled data using assigned scores, in sample sizes as small as 20 subjects per group. However, if the underlying distribution is symmetric, unimodal and continuous, a value of *n* as small as four can yield a very adequate approximation (Hogg and Tanis, 1998).

The CLT for modelling the sample mean may also apply to the regression techniques in Chapters 7 and 8. This may explain why with "sufficiently large" sample sizes (> 100 in the five datasets studied) the bootstrap estimates of SEs and CIs are very similar to the conventional estimates despite the non-Normal distribution and non-constant variance of the residuals.

So my research using the SF-36 HRQoL outcome and the five datasets has shown that bootstrap methods appear to produce sample size estimates, SE and CIs similar to conventional methods. When the standard and the bootstrap methods agree, we can be more confident about the inference we are making and this is an important use of the bootstrap, (Campbell, 2001). When they disagree more caution is needed, but the relatively simple assumptions required by the bootstrap method for validity mean that in general it is to be preferred. Thus, there appears to be little advantage in using the bootstrap for the analysis of SF-36 data particularly if one is interested in comparing mean HRQoL between treatment groups.

Chapter 10: Summary and Conclusions

In Chapters 1 and 2 we described HRQoL outcomes such as the SF-36, which is one of the most widely used generic multi-dimensional HRQoL outcome measures in the world today. HRQoL outcomes like the SF-36 are usual measured on an ordinal scale. Although most investigators (myself included) assume that there exists an underlying continuous latent variable that measures HRQoL, and that the actual measured outcomes (the ordered categories), reflect contiguous intervals along this continuum.

We demonstrated how this ordinal scaling of HRQoL measures may lead to several problems in estimating sample size and analysing the data. Data from HRQoL outcomes tends to have discrete, bounded and skewed distributions. For this reason non-parametric methods are often used to analyse HRQoL data. The bootstrap (described in Chapter 3) is one such non-parametric method for estimating sample sizes and analysing HRQoL data. The bootstrap is a computer intensive method that involves repeatedly drawing random samples with replacement from the data and repeatedly estimating the statistic of interest.

From an extensive review of the literature (Chapter 4) we found five methods of estimating sample sizes for simple two-group cross-sectional comparisons of HRQoL outcomes. (The fifth method of estimation involved the use of the bootstrap and some of the test statistics involved in the calculation of the sample sizes for the four preceding methods.) All five methods (amongst other factors) require the specification of a *minimum clinically important difference* (MCID) worth detecting and an *effect size*, both of which vary according to the method of sample size estimation.

These effect sizes include: a simple mean difference δ and its standardised counterpart, Δ_{Normal}; a simple absolute difference in proportions δ_{Binary}, for dichotomised outcomes; various odds ratios (OR) for either binary OR_{Binary}, or ordinal ($OR_{Ordinal}$) outcomes; and a probability, $p_{Noether}$ (the probability of a

randomly chosen outcome from one group being larger than a randomly chosen outcome from the second group).

Chapter 5 calculated the six observed effect sizes for the eight dimensions of the SF-36 for various simple two group cross-sectional comparisons across five datasets. In this chapter we also demonstrated how the four seemingly different effect sizes Δ_{Normal}, OR_{Binary}, $OR_{Ordinal}$ and $p_{Noether}$ which are all numerical expressions of treatment efficacy can be combined into a common scale of metric using the A ($A_{XY} = \Pr(X > Y)$ and $A_{YX} = \Pr(Y > X)$), λ and θ statistics.

We showed that for ordinal and continuous outcomes $A_{XY} - A_{YX} = \lambda$ and $A_{XY} / A_{YX} = \theta$ are equivalent to the Absolute Risk Reduction (ARR) and OR for binary outcomes (i.e. δ_{Binary} and OR_{Binary}). Since the Number-Needed-to-Treat (NNT) is the reciprocal of the ARR, the NNT and OR statistics can be generalised to all data types (binary, ordinal and continuous).

The empirical effect sizes calculated in Chapter 5 suggested that large differences in HRQoL (as measured by the SF-36) between groups are unlikely, particularly from the RCT comparisons. Most of the observed effect sizes are mainly in the 'small' to 'moderate' range (0.2 to 0.5) using Cohen's (1988) criteria. Therefore dramatic differences in HRQoL between groups are unlikely using the SF-36 and larger sample sizes may be required to have a reasonable chance of detecting statistically significant differences between groups. We went on to use some of these estimates of effect size as our MCID in the calculation of sample sizes in the next chapter.

Chapter 6 compared the power of various methods of sample size estimation described in Chapter 4 for simple two-group cross-sectional study designs via bootstrap simulation. This chapter showed how under the location shift alternative hypothesis conventional methods (1 to 4) of sample size estimation performed well, particularly Whitehead's (1993) Method (4). Whitehead's method is recommended if the HRQoL outcome has a limited number of

discrete values (or is expected to generate data with a limited number of values) and/or the expected proportion of cases at either of the bounds is high.

If a pilot dataset is readily available (to estimate the shape of the distribution) then bootstrap simulation (Method 5) may provide a more accurate and reliable estimate, than Methods 1 to 4. In the absence of reliable pilot data set, which is frequently the case at the study design stage, bootstrapping is not appropriate and conventional Methods (1 to 4) of sample size estimation or parametric simulation models will need to be used.

The final two results Chapters (7 and 8) describe how the bootstrap can be used for hypothesis testing and the estimation of SEs and CIs for parameters. In both chapters we concentrated on comparing and contrasting the bootstrap with standard methods of analysing HRQoL outcomes as described in Fayers and Machin (2000).

In Chapter 7 we looked at analysing simple cross-sectional HRQoL data or HRQoL data with a baseline and single follow-up assessment. We described two simple resampling algorithms for performing bootstrap hypothesis tests for comparing groups (Algorithms 7.1 and 7.2). In the example dataset studied we show that there appears to be little advantage in using bootstrap hypothesis test compared to conventional non-parametric hypothesis tests such as the *MW* test.

A major limitation of non-parametric hypothesis tests are they do not allow for the estimation of CI for parameters, which is regarded as good statistical practice (Altman *et al* 2000). Nor do they allow for the adjustment of confounding variables such as baseline covariates. Fortunately, the bootstrap is able to estimate SEs and CIs for parameters from a variety of multiple regression models.

Chapter 7 also describes two more bootstrap algorithms for estimating SEs and CIs for regression coefficients from fitting the OLS multiple regression

model, - *case* and *model (residual)* based resampling (Algorithms 7.3 and 7.4 respectively). In the datasets studied, both the case and model based bootstrap resampling methods for estimating SEs and CIs for linear regression models gave estimates almost identical to the conventional values estimated using OLS.

In the final results chapter (Chapter 8) we looked at methods of analysing HRQoL data collected at three or more time points, including the simple analysis of summary measures such as the AUC, and the more complex modelling of longitudinal HRQoL data via marginal models and GEE. In the datasets studied, we used the AUC to summarise the repeated HRQoL assessments into one observation for each subject and then compared mean AUC between treatment groups. The *p*-values from the two independent samples *t*-test and the *ASL* from the bootstrap hypothesis test for comparing mean AUC between the groups were very similar. As were the Normal-theory based CI estimates from the *t*-test compared with their bootstrap BC_a counterparts for comparing the mean difference in AUC between groups.

We also used ANCOVA to regress the mean of the follow-up HRQoL assessments against treatment group and baseline HRQoL. Again, in the datasets studied, both the case and model based bootstrap resampling methods for estimating SEs and CIs for linear regression models gave estimates almost identical to the conventional values estimated using OLS.

Finally, in lieu of reducing the repeated HRQoL responses to a summary statistic we used a marginal model to analyse the individual assessments simultaneously. We used a simple case-based resampling bootstrap algorithm for estimating SEs and CIs for regression coefficients from the marginal model (Algorithm 8.1). We then compared the robust SE and CI estimated from the marginal model (using IRLS) with their bootstrap counterparts. Depending on the assumed underlying autocorrelation the bootstrap SEs and CIs differed slightly. So some caution was needed in interpreting the regression coefficients. When the standard and the bootstrap methods agree, we can be more confident about the inference we are making. However the relatively

simple assumptions required by the bootstrap method for validity mean that in general it is to be preferred (Campbell, 2001)

HRQoL outcome measures frequently generate data with discrete, bounded and skewed distributions. Therefore standard methods of analysis such as the two sample *t*-test and OLS multiple regression which assume Normality and constant variance may not be appropriate. Hence HRQoL outcomes appear to be ideal candidates for the application of non-parametric statistical methods. The bootstrap is one such method and therefore theoretically may be more appropriate for estimating sample sizes and analysing HRQoL outcomes than standard methods.

Overall, in the datasets studied with the SF-36 outcome, the use of the bootstrap for estimating sample sizes and analysing (hypothesis testing, SE and CI estimation) HRQoL data appears to produce results similar to conventional statistical methods. Therefore, the results of this thesis suggest that bootstrap methods are not more appropriate for analysing HRQoL outcome data than standard methods. This result requires replication with other HRQoL outcome measures, interventions and populations.

Appendix 1: The SF-36 health survey questionnaire

HEALTH STATUS QUESTIONNAIRE (SF-36)

The following questions ask you about your health, how you feel and how well you are able to do your usual activities.

If you are unsure how to answer a question, please give the best answer you can.

1. In general, would you say your health is:

Excellent ..... O
Very good ..... O
Good ..... O
Fair ..... O
Poor ..... O

2. Compared to one year ago, how would you rate your health in general now?

Much better than one year ago ..... O
Somewhat better than one year ago..... ..... O
About the same .... O
Somewhat worse now than one year ago..... O
Much worse now than one year ago .... O

HEALTH AND DAILY ACTIVITIES

3. The following questions are about activities that you might do during a typical day. Does your health limit you in these activities? If so, how much?

ACTIVITIES	Yes, limited a lot	Yes, limited a little	No, not limited at all
a. **Vigorous activities**, such as running, lifting heavy objects, participating in strenuous sports	1	2	3
b. **Moderate activities**, such as moving a table, pushing a vacuum cleaner, bowling or playing golf	1	2	3
c. Lifting or carrying groceries	1	2	3
d. Climbing **several** flights of stairs	1	2	3
e. Climbing **one** flight of stairs	1	2	3
f. Bending, kneeling or stooping	1	2	3
g. Walking **more than a mile**	1	2	3
h. Walking **half a mile**	1	2	3
i. Walking **100 yards**	1	2	3
j. Bathing and dressing yourself	1	2	3

4. During the <u>past 4 weeks</u>, have you had any of the following problems with your work or other regular daily activities <u>as a result of your physical health</u>?

	YES	NO
a. Cut down on the **amount of time** you spent on work or other activities	1	2
b. **Accomplished less** than you would like	1	2
c. Were limited in the **kind** of work or other activities	1	2
d. Had **difficulty** in performing the work or other activities (e.g. it took extra effort)	1	2

5. During the <u>past 4 weeks</u>, have you had any of the following problems with your work or other regular daily activities <u>as a result of any emotional problems</u> (such as feeling depressed or anxious)?

	YES	NO
a. Cut down on the **amount of time** you spent on work or other activities	1	2
b. **Accomplished less** than you would like	1	2
c. Didn't do work or other activities as **carefully** as usual	1	2

6. During the <u>past 4 weeks</u>, to what extent have your physical health or emotional problems interfered with your normal social activities with family, friends, neighbours or groups?

Not at all 1
Slightly 2
Moderately 3
Quite a bit 4
Extremely 5

7. How much <u>bodily</u> pain have you had during <u>the past 4 weeks</u>?

None 1
Very mild 2
Mild 3
Moderate 4
Severe 5
Very severe 6

8. During the <u>past 4 weeks</u>, how much did pain interfere with your normal work (including work both outside the home and housework)?

Not at all 1
A little bit 2
Moderately 3
Quite a bit 4
Extremely 5

YOUR FEELINGS

9. These questions are about how you feel and how things have been with you during the past 4 weeks. For each question, please indicate the one answer that comes closest to the way you have been feeling.

How much of the time during the past 4 weeks:	All of the time	Most of the time	A good bit of the time	Some of the time	A little of the time	None of the time
a. Did you feel full of life?	1	2	3	4	5	6
b. Have you been a very nervous person?	1	2	3	4	5	6
c. Have you felt so down in the dumps that nothing could cheer you up?	1	2	3	4	5	6
d. Have you felt calm and peaceful?	1	2	3	4	5	6
e. Did you have a lot of energy?	1	2	3	4	5	6
f. Have you felt down-hearted and low?	1	2	3	4	5	6
g. Did you feel worn-out?	1	2	3	4	5	6
h. Have you been a happy person?	1	2	3	4	5	6
i. Did you feel tired?	1	2	3	4	5	6
j. Has your health limited your social activities (like visiting friends or close relatives)	1	2	3	4	5	6

HEALTH IN GENERAL

10. Please choose the answer that best describes how true or false each of the following statements is for you.

	Definitely true	Mostly true	Not sure	Mostly false	Definitely false
a. I seem to get ill more easily than other people	1	2	3	4	5
b. I am as healthy as anybody I know	1	2	3	4	5
c. I expect my health to get worse	1	2	3	4	5
d. My health is excellent	1	2	3	4	5

Appendix 2: Bootstrap Confidence Intervals

If we let $\hat{\theta}(x)_{obs}$, from now on simplified to $\hat{\theta}_{obs}$ be the estimated observed value of the statistic θ, that is the value of the statistic calculated using the original observed dataset, **x** = (x_1, x_2,…,x_n). Let b = 1, 2,…,B denote the bootstrap samples, and let $\hat{\theta}_b^*$ be the estimated values of the statistic computed using each of these B samples. We now have B bootstrap samples **x**$^*_1$, **x**$^*_2$,…,**x**$^*_B$, and B estimates of the statistic, one from each bootstrap sample $\left(\hat{\theta}_1^*, \hat{\theta}_2^*, ..., \hat{\theta}_B^*\right)$. The bootstrap standard error of $\hat{\theta}$, is estimated as:

$$\hat{se}_{boot}\left(\hat{\theta}\right) = \left\{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \hat{\bar{\theta}}^*\right)^2\right\}^{1/2}, \text{ where } \hat{\bar{\theta}}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*. \quad \text{(A2.1)}$$

Algorithm A2.1 (derived from Efron and Tibshirani, 1993) is a more explicit description of the bootstrap procedure for estimating the standard error of $\hat{\theta} = s(x)$ from the observed data **x**.

Algorithm A2.1

The bootstrap algorithm for estimating standard errors

1. Select B independent bootstrap samples **x**$^*_1$, **x**$^*_2$,…,**x**$^*_B$, each consisting of n data values drawn with replacement from **x**. [For estimating a standard error, the number B will ordinarily be in the range 25-200.]
2. Evaluate $\hat{\theta} = s(.)$ on each bootstrap sample,

$$\hat{\theta}_b^* = s\left(x_b^*\right) \qquad b = 1, 2, ..., B. \quad \text{(A2.2)}$$

3. Estimate the standard error of $\hat{\theta}$, $se_F\left(\hat{\theta}\right)$ by the sample standard deviation of the B bootstrap replications of the statistic $\hat{\theta} = s(x)$.

$$\hat{se}_{Boot} = \left\{\frac{\sum_{b=1}^{B}\left[\hat{\theta}_b^* - \hat{\bar{\theta}}^*\right]}{(B-1)}\right\}, \quad \text{(A2.3)}$$

where $\hat{\bar{\theta}}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$ is the mean of the B bootstrap replications of $\hat{\theta}$.

The *bias* is estimated as

$$\hat{b}ias = \hat{\bar{\theta}}^{*} - \hat{\theta}_{obs}\,. \tag{A2.4}$$

Confidence intervals with nominal coverage rates 1 – α are calculated according to the following formula.

The *Standard Normal-approximation method* yields the confidence intervals:

$$\left[\hat{\theta}_{obs} - \left(z_{1-\alpha/2}\cdot\hat{s}e_{Boot}\left(\hat{\theta}\right)\right), \hat{\theta}_{obs} + \left(z_{1-\alpha/2}\cdot\hat{s}e_{Boot}\left(\hat{\theta}\right)\right)\right], \tag{A2.5}$$

where $z_{1-\alpha/2}$ is the (1-α/2) quantile of the standard Normal distribution ($z_{0.975} = 1.96$).

If *B* estimates of the statistic, one from each bootstrap sample $\left(\hat{\theta}_1^{*},\hat{\theta}_2^{*},...,\hat{\theta}_B^{*}\right)$ are ordered in increasing value, $\left(\hat{\theta}_{(1)}^{*},\hat{\theta}_{(2)}^{*},...,\hat{\theta}_{(B)}^{*}\right)$, a bootstrap 95% confidence interval for the statistic would be from the $0.025B^{th}$ to the $0.975B^{th}$ largest values. This yields the *percentile method* confidence interval. For a 100(1 - α)% percentile interval the limits would the $(\alpha/2)B^{th}$ and $(1 - \alpha/2)B^{th}$ ordered largest values. I.e.

$$\left[\hat{\theta}_{(\alpha/2)(B)}^{*},\hat{\theta}_{(1-\alpha/2)(B)}^{*}\right], \tag{A2.6}$$

where $\hat{\theta}_{(p)}^{*}$ is the p^{th} quantile (the $100p^{th}$ percentile) of the ordered bootstrap distribution $\left(\hat{\theta}_{(1)}^{*},\hat{\theta}_{(2)}^{*},...,\hat{\theta}_{(B)}^{*}\right)$.

Although the percentile method is an obvious choice, it is not the best method for bootstrapping confidence intervals, because it can have a bias, which one can estimate and correct for.

If we let

$$\hat{z}_0 = \Phi^{-1}\left\{\#\left(\hat{\theta}_i^{*} \le \hat{\theta}_{obs}\right)/B\right\}, \tag{A2.7}$$

where # is a count of the number of elements of the bootstrap distribution $\left(\hat{\theta}_1^{*},\hat{\theta}_2^{*},...,\hat{\theta}_B^{*}\right)$ that are less than or equal to the observed statistic $\hat{\theta}_{obs}$, and Φ is the standard cumulative Normal distribution function e.g. $\Phi^{-1}(0.95) = 1.645$. Then $\hat{z}_0$ is known as the *median bias* of $\hat{\theta}_{obs}$. We obtain $\hat{z}_0 = 0$ if exactly half of the $\hat{\theta}^{*}$s are less than or equal to $\hat{\theta}_{obs}$.

Let

$$\hat{a} = \frac{\sum_{i=1}^{n}\left(\hat{\theta}_{(.)} - \hat{\theta}_{(i)}\right)^3}{6\left\{\sum_{i=1}^{n}\left(\hat{\theta}_{(.)} - \hat{\theta}_{(i)}\right)^2\right\}^{3/2}}, \qquad \text{(A2.8)}$$

where $\hat{\theta}_{(i)}$ are the leave-one-out *jackknife* estimates of $\hat{\theta}$, and $\hat{\theta}_{(.)} = \sum_{i=1}^{n}\hat{\theta}_{(i)}/n$, is the mean of the *n* jackknife estimates. Then $\hat{a}$ is known as the jackknife estimate of the *acceleration* for $\hat{\theta}_{obs}$. The quantity $\hat{a}$ is called the acceleration because it refers to the rate of change of the standard error of $\hat{\theta}$ with respect to the true parameter θ. The standard Normal approximation $\hat{\theta} \sim N\left(\theta, se(\theta)^2\right)$ assumes that the standard error of $\hat{\theta}$ is the same for all θ. However, this is often unrealistic and the acceleration constant $\hat{a}$ corrects for this (Efron & Tibshirani, 1993). Let

$$p_1 = \Phi\left\{\hat{z}_0 + \frac{\hat{z}_0 - z_{1-\alpha/2}}{1-\hat{a}\left(\hat{z}_0 - z_{1-\alpha/2}\right)}\right\} \qquad \text{(A2.9)}$$

and

$$p_2 = \Phi\left\{\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1-\hat{a}\left(\hat{z}_0 + z_{1-\alpha/2}\right)}\right\}, \qquad \text{(A2.10)}$$

where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)^{th}$ quantile of the standard Normal distribution (e.g. $z_{0.975} = 1.96$ and $\Phi(1.96) = 0.975$). The *bias-corrected and accelerated* (BC_a) method yields confidence intervals:

$$\left[\hat{\theta}^*_{(p_1)}, \hat{\theta}^*_{(p_2)}\right], \qquad \text{(A2.11)}$$

where $\hat{\theta}^*_{(p)}$ is the p^{th} quantile (the $100p^{th}$ percentile) of the ordered bootstrap distribution $\left(\hat{\theta}^*_{(1)}, \hat{\theta}^*_{(2)}, \ldots, \hat{\theta}^*_{(B)}\right)$ as defined previously. The bias-corrected (but not accelerated) method is a special case of BC_a with $\hat{a} = 0$. If both the bias and the acceleration are zero then $p_1 = \alpha/2$ and $p_2 = 1 - \alpha/2$ and the BC_a method reduces to the percentile method.

Efron and Tibshirani (1993) suggest that very seldom more than $B = 200$ replications are needed for estimating a standard error using Algorithm A2.1. Much bigger values of *B* (of the order 1000 to 5000) are needed for estimating percentile and BC_a bootstrap confidence intervals.

Efron and Tibshirani (1993) show that the BC_a method has two important theoretical advantages. First of all, it is *transformation respecting*. This means that the BC_a

endpoints transform correctly if we change the parameter of interest from 0 to some function of θ. For example, the confidence intervals for $\sqrt{\text{var}(A)} = \sqrt{\theta}$ are obtained by taking the square roots of BC_a endpoints for $\theta = \text{var}(A)$.

The second advantage of the BC_a method concerns its accuracy. A central 1 - $\alpha/2$ confidence interval $\left(\hat{\theta}_{lo}, \hat{\theta}_{up}\right)$ is supposed to have probability α of not covering the true value of θ from above or below,

$$\text{Prob}\left\{\theta < \hat{\theta}_{lo}\right\} = \alpha/2 \text{ and } \text{Prob}\left\{\theta > \hat{\theta}_{up}\right\} = \alpha/2. \qquad \text{(A2.12)}$$

Approximate confidence intervals can be graded on how accurately they match A2.12. The BC_a intervals can be shown to be *second-order accurate*. This means that its errors in matching (A2.12) go to zero at rate $1/n$ in terms of the sample size n,

$$\text{Prob}\left\{\theta < \hat{\theta}_{lo}\right\} \doteq \alpha/2 + \frac{c_{lo}}{n} \text{ and } \text{Prob}\left\{\theta > \hat{\theta}_{up}\right\} \doteq \alpha/2 + \frac{c_{up}}{n}, \qquad \text{(A2.13)}$$

for two constants c_{lo} and c_{up}. (Note we use $\doteq$ to mean "is approximately equal to".) The standard and percentile methods are only *first-order accurate*, meaning that the errors in matching A2.6 are an order of magnitude larger,

$$\text{Prob}\left\{\theta < \hat{\theta}_{lo}\right\} \doteq \alpha/2 + \frac{c_{lo}}{\sqrt{n}} \text{ and } \text{Prob}\left\{\theta > \hat{\theta}_{up}\right\} \doteq \alpha/2 + \frac{c_{up}}{\sqrt{n}}, \text{ (A2.14)}$$

the constants c_{lo} and c_{up} being possibly different from those above. The difference between first and second order accuracy is not just a theoretical nicety. It leads to much better approximations of exact endpoints when exact endpoints exist (Efron & Tibshirani, 1993).

Appendix 3: Published Papers

Copies of:

Walters, S.J., Campbell, M.J., Lall, R. (2001a) Design and Analysis of Trials with Quality of Life as an Outcome: a practical guide. *Journal of Biopharmaceutical Statistics*, 11(3), 155-176.

Walters, S.J., Campbell, M.J., Paisley, S. (2001b) Methods for determining sample sizes for studies involving health-related quality of life measures: a tutorial. *Health Services & Outcomes Research Methodology*, 2, 83-99.

Lall, R., Campbell, M.J., Walters, S.J., Morgan, K., MRC CFAS. (2002) A review of ordinal regression models applied on Health related Quality of Life Assessments. *Statistical Methods in Medical Research*, 11(1), 49-67.

Walters, S.J. and Brazier, J.E. (2003b) Sample Sizes for the SF-6D Preference Based Measure of Health from the SF-36: A Comparison of Two Methods. *Health Services & Outcomes Research Methodology*, 4, 35-47.

JOURNAL OF BIOPHARMACEUTICAL STATISTICS, 11(3), 155–176 (2001)

DESIGN AND ANALYSIS OF TRIALS WITH QUALITY OF LIFE AS AN OUTCOME: A PRACTICAL GUIDE

Stephen J. Walters,[1,*] Michael J. Campbell,[2] and Ranjit Lall[2]

[1]Sheffield Health Economics Group and [2]Institute of General Practice & Primary Care, School of Health and Related Research, University of Sheffield, Sheffield, United Kingdom

ABSTRACT

Health Related Quality of Life (HRQoL) measures are becoming more frequently used in clinical trials, as both primary and secondary endpoints. Investigators are now asking statisticians for advice on how to plan (e.g., sample size) and analyze studies using HRQoL measures.

HRQoL measures such as the SF-36 are usually measured on an ordered categorical (ordinal) scale. In the designing stages and when analyzing, the scales are often scored and the scores treated as if they were continuous and normally distributed. However the ordinal scaling of HRQoL measures leads to problems in determining sample size, and conventional parametric methods of estimation and hypothesis testing may not be appropriate for such outcomes.

We present practical guidelines for the design and analysis of trials with HRQoL measures as outcomes.

We used conventional statistical methods (i.e., *t*-tests and multiple regression), various ordinal regression models (proportional odds, continuation ratio, polytomous and stereotype) and bootstrap methods to analyze an HRQoL dataset. To illustrate the various methods we used HRQoL data on the SF-36 Role Limitations Emotional dimension for two groups of patients with leg ulcers.

The bootstrap, *t*-test, and multiple regression methods gave similar results. The various ordinal regression models also gave similar results.

If the HRQoL measure has a large number of ordered categories, most of which are occupied, and the underlying scale really is continuous but measured imperfectly by an instrument with a limited number of discrete values, then an informal rule of thumb is that this discrete scale should be treated as continuous if it has seven or more categories and as ordinal otherwise.

Key Words: Quality of life; Sample size; Analysis; Ordinal regression

* Corresponding author. School of Health and Related Research, University of Sheffield, Regent Court, 30 Regent St, Sheffield, S1 4DA, UK. Fax: 0114 2225454. E-mail: s.j.walters@sheffield.ac.uk

1. INTRODUCTION

Health-related quality of life (HRQoL) is increasingly used as a primary or secondary outcome measure in clinical trials and Health Services Research (HSR). It is particularly important in HSR as described by the British Medical Research Council in its guidelines to clinical trials (1): "The assessment of likely impact on health services practice should focus on quality of life (including psycho-social impact) morbidity, mortality and economic cost."

Investigators are now asking statisticians for advice on how to plan and analyze studies using HRQoL measures. The analysis of data from quality of life (QoL) measurements requires some basic assumptions. We will assume that (2):

1. QoL is a subjective construct that is not directly observable and measurable.
2. QoL is a multi-dimensional construct consisting of different aspects of physical and psychological well being.
3. QoL is a time-dependent construct reflecting a person's experiences and perceptions over their life history.

The aim of the paper is to provide pragmatic guidance to statisticians on the design and analysis of trials using HRQoL measures as outcomes.

2. HOW IS QUALITY OF LIFE MEASURED?

Most HRQoL instruments are self-completed questionnaires and usually include a number of items to measure the HRQoL concept of interest. Most individual HRQoL items are usually measured or scored on an ordered categorical or ordinal scale. This means that responses to individual questions are usually classified into a small number of response categories, which can be ordered (for example, poor, moderate, and good). In planning and analysis, the question responses are often analyzed by assigning equally spaced numerical scores to the ordinal categories (e.g., 0 = "poor," 1 = "moderate," and 2 = "good") and the scores

240

across similar questions are then summed to generate a HRQoL measurement. These summed scores are treated as if they were from a continuous distribution and normally distributed.

The HADS (Hospital Anxiety and Depression Scale) is an example of a HRQoL measure. The HADS is a brief assessment of anxiety and depression; it consists of 14 items divided into two subscales of seven questions on anxiety and seven questions on depression, and the patient rates each item on a four-point Likert scale (3) An example of one of the items on the scale is: I still enjoy things I used to enjoy: definitely as much {0}; not quite so much {1}; only a little {2}; hardly at all {3}.

The responses to the seven depression-related questions are summed, giving a possible range of scores from 0 to 21. The higher scores indicate the presence of problems. Using psychiatric diagnoses as a gold standard, HADs depression scores of ≤7 were considered normal or noncases, scores of 8 to 10 were considered doubtful or borderline cases, and scores of ≥11 implies definite cases of clinical depression (or anxiety).

3. PROBLEMS WITH HRQoL MEASURES

There are advantages in being able to treat HRQoL scales as continuous (e.g., for sample size calculations and statistical analysis). However the ordinal scaling of HRQoL measures leads to several problems in determining sample size and analysing the data:

1. The apparent continuum hides the fact that only a few discrete values are possible. For example, the Role Limitations Emotional (RLE) dimension of the SF-36 is scored on a 0 to 100 scale (with 100 indicating "good" health), but there are only four possible categories, i.e., 0, 33, 66 and 100 (4).
2. The scale may not be linear. For example, using the SF-36 RLE dimension, is a change of score from 0 to 33 the same as a change from 66 to 100?
3. There is often a floor or ceiling effect: patients cannot be worse than the worst category or better than the best category (e.g., score 0 or 100 on the RLE dimension of the SF-36). For some populations, the level limits are inappropriate, and most people score on either the best category or the worst category. Floor and ceiling effects are more likely to be a problem in longitudinal studies because they limit the ability of the instrument to detect an improvement or deterioration in a patient's HRQoL over time. Table 1 shows that over 56% (130/233) of the combined sample had a score of 100 or were at the ceiling of the distribution.
4. Methods based on the normal distribution (such as linear regression)

Table 1. Descriptive Statistics for the SF-36 Role Emotional Dimension Score at Baseline for Two Groups of Patients with Short- (≤7 yrs) or Long-term (>7 yrs) Duration of Venous Leg Ulcers

	Max Ulcer Duration ≤7 yrs, N = 115		Max Ulcer Duration >7 yrs, N = 118		
	n (%)	Cumulative, %	n (%)	Cumulative, %	Odds Ratio [illegible]
RLE Score					
0.0	23 (20.0)	20.0	29 (24.6)	24.6	0.77
33.3	9 (7.8)	27.8	19 (16.1)	40.7	0.56
66.7	11 (9.6)	37.4	12 (10.2)	50.9	0.58
100.0	72 (62.6)	100.0	58 (49.2)	100.0	
Total	115 (100.0)		118 (100.0)		
Mean	71.6		61.3		
SD	40.8		42.5		
Median	100.0		66.7		
25th percentile	33.3		19.4		
75th percentile	100.0		100.0		
Age (years)					
Mean	74.3		72.7		
SD	12.2		10.2		
Sex (% female)	73.9%		59.3%		

A higher score indicates better health
From Ref. 5

assume that the outcome variable has a constant variance. The variance of ordinal data may not be constant. The variances of changes may depend on initial values. This is a common problem with range-limited values. Patients may enter the study with a wide variety of scores, but tend always to increase their scores. Thus patients who score lower at the start of the study have more range to improve than those who are already close to the maximum

5. Normal approximations may not apply. Because the data are in fact categorical, they may require different techniques of analysis. By definition, no ordinal variable can be normally distributed, although in some cases a normal approximation will suffice
6. Missing values are likely, for example, in response to questionnaires that ask "how far can you walk?" when the patient is in a wheelchair.
7. It is difficult to quantify an effect size (e.g., a difference in mean score between groups) in advance.

[redacted]

4. HOW SHOULD INVESTIGATORS DESIGN STUDIES WITH HRQoL DATA AS OUTCOMES?

When an investigator designs a study to compare the outcomes of an intervention, an essential step is the calculation of sample sizes that will allow a reasonable chance (power) of detecting a predetermined difference (effect size) in the outcome variable at a given level of significance. Sample size is critically dependent on the objectives and outcomes of the study, on the proposed effect size, and on the method of calculating the test statistic. For example, for a given power and significance level, the sample size is inversely proportional to the square of the effect size, so halving the effect size will quadruple the sample size.

In principle, there are no major differences between planning a study using HRQoL assessment and planning one using conventional clinical outcomes. Campbell et al. (6) outlined the ways of calculating sample sizes in two group studies for binary, ordered categorical and continuous outcomes. Further details, examples and tables are given in the book by Machin et al. (7). Sample size is dependent on the outcomes and objectives of the study and the method of analysis. Thus, after deciding on the primary outcome, the investigator must choose an appropriate summary measure of this outcome and then calculate a sample size based on this summary measure.

An appropriate summary measure of the outcome data usually will be the sample mean, median, or a rate or proportion. When comparing two groups or a single group over time, appropriate comparative summary measures may include the difference between sample means, difference in medians, and difference in rates or proportions, the relative risk, or the odds ratio.

The mean is often chosen as a suitable summary measure, although there are several reasons not to use it One reason would be that the HRQoL outcome measure of interest is an ordinal not a continuous variable, and therefore means (and differences in means) are hard to interpret (see points 1 to 7 above).

If the HRQoL outcome is assumed to be continuous and plausibly sampled from a normal distribution then the best summary statistic for a location parameter is the mean, and the usual hypothesis test for a difference or shift in location parameters between two independent samples is the two-sample *t*-test.

For two independent groups, A and B, with continuous and normally distributed data the effect size index is the expected mean value of the intervention outcome minus the expected mean value of the control outcome divided by a standard deviation of the outcomes.

$$\delta = \frac{\mu_A - \mu_B}{\sigma} \tag{1}$$

where δ is the effect size index, μ_A and μ_B are the expected group means of outcome variable under the null and alternative hypotheses, and σ is the standard deviation of outcome variable (assumed the same under the null and alternative hypotheses).

In a two-group study comparing mean HRQoL between the two groups, the required number of subjects per group n for an effect size δ and a two-sided significance level α and power $1 - \beta$ is given by

$$n_{\text{continuous}} = \frac{2[z_{1-\alpha/2} + z_{1-\beta}]^2}{\delta^2} \tag{2}$$

where $z_{1-\alpha/2}$ and $z_{1-\beta}$ are the appropriate values from the standard normal distribution for the $100(1 - \alpha/2)$ and $100(1 - \beta)$ percentiles, respectively.

If the sample size is "sufficiently large," then the Central Limit Theorem guarantees that the sample means will be approximately normally distributed. Thus, if the investigator is planning a large study and the sample mean is an appropriate summary measure of the HRQoL outcome, then pragmatically there is no need to worry about the distribution of the HRQoL outcome and we can use Eq. (2) to calculate sample sizes.

If the HRQoL outcomes are measured on an ordinal scale, then statistical hypothesis test used in this instance (to compare two independent groups) is the Mann-Whitney U test (also known as the Wilcoxon rank sum test) with allowance for ties or chi-squared test for trend.

Whitehead (8) presents the sample size formulae for ordinal data in a key paper. An effect size must be specified to use Whitehead's formulae. For ordinal data Whitehead suggested the odds ratio (OR), which is the odds of a subject being in a given category or lower in one group compared with the odds in the other group. The proportion of subjects in each scale category for one of the groups also must be specified.

Suppose there are two groups A and B and the HRQoL outcome measure of interest Y has c ordered categories y_i denoted by $i = 1, 2, \ldots, c$. Let p_{iA} be the probability of being in category i in Group A and C_{iA} be the expected cumulative probability of being in category i or less in Group A [i.e., $C_{iA} = Pr(Y \leq y_i)$]. For category i, where i takes values from 1 to $c - 1$, the OR is given by

$$\text{OR} = [C_{iA}/(1 - C_{iA})]/[C_{iB}/(1 - C_{iB})] \tag{3}$$

The assumption of proportional odds specifies that the OR will be the same for all categories from $i = 1$ to $c - 1$. Because the derivation of the sample size formulae and analysis of data is based on the Mann-Whitney U test, Whitehead's method can be regarded as a "nonparametric" approach, although it still relies on the assumption of a constant OR for the data. Whitehead's method also assumes a relatively small log odds ratio and a large sample size, which will often be the case in HRQoL studies where dramatic effects are unlikely. Equation (4) gives the number of subjects per group n for a two-sided significance level α and power $1 - \beta$.

$$n_{\text{ordinal}} = \frac{6[(z_{1-\alpha/2} + z_{1-\beta})^2/(\log \text{OR})^2]}{(1 - \sum_i \bar{p}_i^3)} \tag{4}$$

Here π_i is the average proportion of subjects anticipated in category i; that is, $\bar{\pi}_i = (\pi_{A_i} + \pi_{B_i})/2$.

If the number of categories is large it is difficult to postulate the proportion of subjects who would fall in a given category. Both Whitehead (8) and Campbell et al. (6) point out that there is little increase in power (and hence saving in the number of subjects recruited) to be gained by increasing the number of categories beyond five. Categories that are equally likely to occur lead to the greatest efficiency.

But what if the wrong sample size formula is applied? The parametric approach to sample size calculation uses the effect size [Eq. (1)] and is derived from methods assuming an outcome variable with a normal distribution on a continuous scale. The subsequent statistical test, the t-test, is reasonably robust to violations of these assumptions. Larger sample sizes are needed the further the distribution departs from a normal distribution (9).

However, two papers (10,11) have highlighted the discrepancies between sample sizes for intervention studies using HRQoL outcomes (SF-36 and HADS) calculated using parametric and nonparametric approaches.

Using the SF-36, Julious et al. (10) shows that the results given by the parametric and nonparametric methods are similar in some dimensions of the SF-36 but are very different in dimensions where the scores are highly skewed. For such asymmetric distributions the parametric methods give the same sample sizes for effects that are one discrete value above and one discrete value below the population mean. This is because the parametric method assumes a symmetric (normal) distribution. The nonparametric method may give different sample sizes according to the expected direction of the effect.

In both papers (10,11) the authors stress that, "In general, statistics such as means and standard deviations are not suitable summary measures for non-normal distributions, and neither are standardised differences (*effect sizes*) a suitable basis for the calculation of sample sizes."

Julious et al. (11) recommend that the frequency distributions of HRQoL scores should always be given so that one can assess if nonparametric methods should be used for sample size calculations and analysis. Given the skewed/asymmetric distribution of the majority of HRQoL outcomes in general, they recommend nonparametric (ordinal methods) be used for sample size calculations.

5. HOW SHOULD INVESTIGATORS ANALYZE QUALITY OF LIFE DATA?

Suppose the HRQoL measure has a large number of ordered categories, most of which should be occupied if the underlying scale really is continuous, but the scale is measured imperfectly by an instrument with a limited number of discrete values. It is often worth treating this discrete scale as if it were continuous. An

informal rule of thumb is that this discrete scale should be treated as continuous if it has seven or more categories and as ordinal otherwise.

We base this informal rule of thumb on Whitehead's sample size formula for ordinal data. Whitehead (8) illustrates the dependence of sample size n on the number of categories c. It is assumed that all categories are equally probable ($\bar{\pi}_1 = \bar{\pi}_2 = \cdots = \bar{\pi}_i$ for all c). It follows from Eq. (4) that the sample size [denoted by $n(c)$] required when there are c equally probably categories, keeping OR, α, and β constant, is

$$n(c) = \frac{0.75}{1 - 1/c^2} n(2) \tag{5}$$

In the limit as $c \to \infty$, $n(c) \to 0.75n$. The limiting case is approached in large samples in which a full ranking of patient outcomes is achieved. A full ranking is equivalent to a categorization with one patient in each category. So for continuous data the sample size using Eq. (4) tends to be 75% of the binary outcome case. Equation (5) says that little is gained by using more than five categories, as the hypothesis test will be $(0.75/0.78) \times 100 = 96\%$ efficient relative to the use of a full ranking. When the data truly are normally distributed with equal variances, the Mann-Whitney test for untried data is 96% efficient relative to the t-test (12). Thus with five equally probable categories the test is 92% efficient relative to the t-test, when the data are truly normally distributed. These relative efficiencies are all asymptotic and are only valid for moderate to large sample sizes, so it is reasonable to require more than five categories in the assumption. Note that occasionally it will be obvious that the assumption of a continuous scale does not hold, such as when one of the categories is death.

Heeren and D'Agostino (13) demonstrated the robustness of the two-independent samples t-test when applied to ordinal data (with three, four, or five categories) and samples of size 20 or less, although they assumed the scales were equally spaced and cautioned on generalizing the results to scales with more than five values. However, the Central Limit Theorem is likely to ensure the robustness of the t-test if the sample size is large and there are seven or more occupied categories.

Descriptive statistics: The frequency distribution of HRQoL scores should be assessed to see what methods are appropriate for sample size calculations and analysis. It may be preferable to use the median as a summary measure of location (rather than the mean). Similarly, one may want to report a percentile range (25th to 75th) or (10th to 90th) as a measure of spread or variability (rather than the standard deviation). The standard deviation for all data and distributions has the mathematical interpretation as the square root of the average sum of the squared deviations from the mean. Distributions other than 95% of the observations may not be within plus or minus two standard deviations of the mean. However, the mean is sometimes useful if the total sample size is also reported because it enables the total HRQoL to be calculated.

5.1 Comparing Two Independent Groups—Ordinal HRQoL Measures (with <7 Categories)

If the HRQoL outcomes are measured on an ordinal scale, with less than seven categories, then the statistical hypothesis test used in this instance (to compare two independent groups) is the Mann-Whitney U test with allowance for ties or chi-squared test for trend. In general the Mann-Whitney U test gives very similar *P*-values to the chi-squared test for trend (14). The difficulty with this method is that it does not allow covariates, and does not provide estimates of population parameters.

The simplest approach to analyzing ordinal data is to dichotomize the data and use logistic regression. However this method ignores useful information in the data, may not be very powerful (15), and introduces the problem of where to choose the cut point. If one were to keep the ordinal structure, then there are number of models possible (16,17). These include proportional odds, continuation ratio, polytomous, and stereotype.

Proportional Odds or Cumulative Logit Model

The proportional odds or cumulative logit model is based on the cumulative response probabilities rather than the category probabilities (18).

For example, consider an HRQoL outcome variable Y with c categorical outcomes y_i denoted by $i = 1, \ldots, c$ and let p be a set of covariates $(X_1, X_2, \ldots, X_p)$. The cumulative logit or proportional odds model is

$$C_i = \Pr(Y \le y_i | X_1, X_2, \ldots, X_p) = \frac{\exp(\alpha_i + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}{1 + \exp(\alpha_i + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)} \quad (6)$$
$$i = 1, \ldots, C - 1$$

This can be expressed equivalently in logit form as

$$\text{logit}(C_i) = \log\left[\frac{C_i}{1 - C_i}\right] = \log\left[\frac{\Pr(Y \le y_i | X_1, \ldots, X_p)}{\Pr(Y > y_i | X_1, \ldots, X_p)}\right] \quad (7)$$
$$= \alpha_i + (\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p) \qquad i = 1, \ldots, c - 1$$

where $C_i = \Pr(Y \le y_i | X_1, X_2, \ldots, X_p)$ is the cumulative probability of being in category i or lower given the set of covariates (note that for $i = c$, $\Pr(Y \le y_i | X_1, X_2, \ldots, X_p) = 1$). The $\alpha_i (i = 1, \ldots, c - 1)$ and $\{\beta_1, \beta_2, \ldots, \beta_p\}$ parameters are treated as unknown and the intercept parameters α_i must satisfy the condition $\alpha_1 \le \alpha_2 \le \cdots \le \alpha_{c-1}$ (18). The regression coefficient β_p for a binary explanatory variable X_p (e.g., control or intervention group) is the log-odds ratios for the Y by X_p association controlling for the other covariates in the model. That is the treatment effect on HRQoL after adjusting for prognostic factors such as age, sex, and center.

The $\{\beta_1, \beta_2, \ldots, \beta_p\}$ regression parameters do not depend on the category i, so that the model (7) assumes that the relationship between each of the covariates and Y (HRQoL) is independent of i (the response category). This assumption of identical log-odds ratios across the c categories is the proportional odds assumption.

The proportional odds model is useful when one believes HRQoL is a continuum, which is measured imperfectly by an instrument with a limited number of values. The proportional odds model is invariant when the codes for the response Y are reversed (i.e., y_1 recoded as y_c, y_2 recoded as y_{c-1}, and so on). Also the proportional odds model is invariant under the collapsibility of adjacent categories of the ordinal response [implying that when y_1 and y_2 are combined, the estimate of the odds ratio remains essentially the same as the odds ratios obtained for the individual categories (19)].

Continuation Ratio Model

An alternative method to the proportional odds model is the continuation ratio model. This may be relevant when an ordinal HRQoL scale is thought of as a progression through various stages, so that people start with "excellent" and deteriorate to "poor" and are unlikely to reverse this trend. The cumulative probabilities $C_i = \Pr(Y \le y_i | X_1, X_2, \ldots, X_p)$ of being in category i or lower in the cumulative logit model [Eq. (6)] are replaced by the probability of being in category i [i.e., $p_i = \Pr(Y = y_i)$] divided by the probability of being in a category higher than i [i.e., $\Pr(Y > y_i)$] for the continuation ratio model:

$$\text{logit}\left(\frac{p_i}{1 - C_i}\right) = \log\left[\frac{\left(\frac{p_i}{1 - C_i}\right)}{(1 - p_i/1 - C_i)}\right]$$

$$= \log\left[\frac{\Pr(Y = y_i | X_1, \ldots, X_p)}{\Pr(Y > y_i | X_1, \ldots, X_p)}\right] \quad (8)$$

$$= \alpha_i + (\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p) \qquad i = 1, \ldots, c - 1$$

When the "logit" expansion is replaced by the "complementary log-log" link function in model [Eq. (7)], the resulting model [Eq. (8)] is

$$\log\left[-\log\left(\frac{p_i}{1 - C_i}\right)\right] = \alpha_i + (\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p) \quad (9)$$

which is the Cox proportional-hazards model for survival data in discrete time.

The continuation ratio model is not invariant under the collapsing or reversal of categories. The continuation ratio model is best suited to circumstances in which the individual categories of the HRQoL scale are of interest and a monotonic progression through the individual categories is expected. Armstrong and Sloan (15) have given a useful comparison of the proportional odds and continuation ratio models.

Chi-squared (χ^2) score tests are available for tests of the proportional odds assumption but these lack power (20,21). Also the model is robust to mild, departures from the assumption of proportional odds. A crude test would be to examine the odds ratios and if they are all greater than unity, or all less than unity, then a proportional odds model will suffice. With increasing numbers of categories it is less likely that proportional odds assumption remains true.

The ordinal regression method also allows us to adjust the treatment effect for other prognostic factors and covariates (such as center, sex, and age). The regression coefficients and their standard errors also enable confidence intervals to be calculated. The statistical packages SPSS, SAS, and STATA have procedures for fitting proportional odds or continuation ratio models.

Stereotype Logistic Model

For the cumulative logit model [Eq. (7)], the HRQoL outcome variable Y is assumed to have an unobserved underlying variable (say, Z), which takes on a continuous form. For example, "age" may be represented by ordered categories, which take on the form "young," "middle-aged," "old," and "very old." In this case, there is an underlying variable: calendar age.

HRQoL scales are sometimes constructed in such a way that there is no underlying variable that directly links to the ordered y-response categories. For instance when assessing "pain," one may use a rating scale of the form "none," "mild," "moderate," and "severe." Here pain is rated depending on other factors, such as its severity and type. Although the rating scale is in principle ordered, there is no underlying variable (continuous or otherwise) that directly relates the factors and links these up with the categories on the scale. Anderson (22) recognized these types of ordered categories as being truly discrete and referred to the response as a *judged* or *assessed* variable. As the cumulative logit model would be inappropriate for analyzing such variables, Anderson introduced another model known as the *stereotype* model. One of the main advantages of the stereotype model over other regression models is that it does not assume a priori ordering of the y-response categories.

The stereotype model is based on the polytomous regression model (22), which does not impose any restrictions on the ordering of the categories. The ordinality is in-built into it by imposing a structure on the regression coefficients.

Consider an HRQoL outcome variable Y with c ordered categorical outcomes y_i denoted by $i = 1, 2, \ldots, c$, and let $X_1, X_2, \ldots, X_p$ denote a set of p covariates. The ordinary polytomous regression model can be written as

$$\Pr(Y = y_i | X_1, X_2, \ldots, X_p) = \frac{\exp(\alpha_i + \beta_{i1}X_1 + \beta_{i2}X_2 + \cdots + \beta_{ip}X_p)}{\sum_{i=1}^{c} \exp(\alpha_i + \beta_{i1}X_1 + \beta_{i2}X_2 + \cdots + \beta_{ip}X_p)} \quad (10)$$

where $\alpha_i = 0$ and $\beta_{ik} = 0 (k = 1, \ldots, p)$ to assure identifiability. The log-probability ratios are formed for model Eq. (10) by comparing each response category (y_i) with a reference category (y_1). The choice of the reference category is arbitrary but we shall use the first category. Thus, the log-probability ratio can be represented by a linear model of the form

$$\log\left[\frac{\Pr(Y = y_i | X_1, X_2, \ldots, X_p)}{\Pr(Y = y_1 | X_1, X_2, \ldots, X_p)}\right] = \alpha_i + \beta_{i1}X_1 + \beta_{i2}X_2 + \cdots + \beta_{ip}X_p \quad (11)$$
$$i = 2 \text{ to } c$$

The regression coefficient β_{ip} for the pth covariate X_p corresponds to the log-probability ratio comparing ($Y = y_i$) versus ($Y = y_1$) for a unit increase in X_p.

From model in Eq. (11), it is clear that the ordinal nature is not accounted for in any way. The ordinality can be built into this model by imposing a structure on the regression coefficients $\beta_{ik} (k = 1, \ldots, p)$. Anderson proposed modeling the regression coefficients, β_{ik}, by imposing the relationship

$$\beta_{ik} = \phi_i \beta_k \qquad i = 2, \ldots, c; k = 1, \ldots, p \quad (12)$$

where β_k is a list of new parameters and the ϕ_i values can be thought of as the scores attached to the response y_i. Note that since $\beta_{1k} = 0$, we have $\phi_1 = 0$, and a further constraint, $\phi_c = 1$ (in order to uniquely identify the parameters when using estimated scores). Substituting Eq. (12) into Eq (11) yields the stereotype model

$$\log\left[\frac{\Pr(Y = y_i | X_1, X_2, \ldots, X_p)}{\Pr(Y = y_1 | X_1, X_2, \ldots, X_p)}\right] = \alpha_i + \phi_i(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p) \quad (13)$$
$$i = 2 \text{ to } c$$

Thus, it can be seen that the stereotype model determines a set of parameters $\{\phi_i\}$ for the dependent variable and a single parameter β_k for each covariate. The ϕ_i values are decided on for the response variable and are directly tied up with the effect of the explanatory variables. Thus, with a positive β_k, when the log probability ratios $\{\phi_i \beta_k\}$ form a decreasing trend, the ϕ_i parameters also become ordered such that:

$$\phi_c \beta_k \geq \cdots \geq \phi_2 \beta_k \geq \phi_1 \beta_k = 0 \text{ then } 1 = \phi_c \geq \cdots \geq \phi_2 \geq \phi_1 = 0 \quad (14)$$

Here we can say that the effect of the covariates on the first log probability ratio is greater than their effect upon the second and so on (or the effect is the same on the consecutive log probability ratios), and that provided Eq. (14) holds, then the model in Eq. (13) is an ordered regression model.

The model fitted does not necessarily require the ϕ_i values to be ordered; whether there is ordering or not is purely determined by the empirical evidence provided by the data. Two categories denoted by c_1 and c_2 are indistinguishable with respect to the covariates if $\phi_{c1} = \phi_{c2}$, that is, the effect of the covariates is the same in the two categories. The product $\phi_i\beta_p$ for the pth covariate X_p corresponds to the log-probability ratio comparing $(Y = y_i)$ versus $(Y = y_1)$ for a unit increase in X_p.

Greenland (19) argues strongly in favour of the stereotype model when there is no underlying continuum that is directly related to the response categories, but where each state is assessed. The statistical package STATA has a procedure for performing stereotype regression via constrained polytomous logistic regression. The stereotype model can also be fitted in SAS using PROC CATMOD.

Table 2 shows the results of fitting a binary logistic model, cumulative logit model, continuation ratio model, polytomous model, and stereotype model to the RLE data with group and sex as factors and age as a covariate.

For the logistic model the outcome variable is dichotomised into a score of 100 "good health" and less than 100 ("less than good health"). The OR of 0.54 (95% CI: 0.32 to 0.93) for the binary logistic model implies that the odds of patients with long-standing ulcer problems (>7 years) having "good" health is 0.54 times that of patients with short-term ulcer problems (≤7 years) after allowing for age and sex.

Similarly the proportional odds model implies that having a long-standing ulcer problem (>7 years) carried with it an odds ratio of 0.58 compared to that of patients with short-term ulcer problems (≤7 years) for being in a given category or below (i.e., better HRQoL) after allowing for age and sex. As the proportional odds model assumes a constant OR for all categories, Table 1 shows how the three $(k - 1)$ observed ORs compare with the estimated common OR of 0.58 from the model. All observed ORs are less than 1 and seem similar to the model estimate. A chi-squared score test of proportional odds was $\chi^2 = 4.1$ on 6 df, $p = 0.67$. Thus, there is no statistical evidence to reject the assumption of proportional odds.

Similarly, the continuation ratio model OR estimate implies that having a long-standing ulcer problem (>7 years) carried with it an odds ratio of 0.60 (95% CI: 0.38 to 0.93) compared to that of patients with short-term ulcer problems (≤7 years) for better HRQoL after allowing for age and sex. As the model assumes a constant hazard for all categories, a formal statistical test of a constant hazard was $\chi^2 = 2.88$ on 6 df, $p = 0.82$. Again, there is no statistical evidence to reject the assumption of constant hazard ratios.

The polytomous logistic model implies that the probability of having a RLE score of 33 compared to 0 is 1.66 times the probability for subjects with a long-

Table 2. Results of Fitting Various Binary and Ordinal Models to the SF-36 Role Emotional Dimension Score at Baseline for Two Groups of Patients with Short (≤7 yrs) or Long-term (>7 yrs) Duration of Venous Leg Ulcers

Model[a]	B	SE(b)	P-value	OR[b]	95% CI for OR	χ^2 Goodness of Fit[c]
Logistic regression (binary)[d]	−0.608	0.274	0.027	0.54	0.32	0.93	9.53 on 3 df, $p = 0.02$
Proportional odds	−0.546	0.26	0.036	0.58	0.35	0.96	7.95 on 3 df, $p = 0.05$
Continuation ratio	−0.512	0.226	0.024	0.60	0.38	0.93	9.45 on 3 df, $p = 0.02$
Polytomous Model[e]							
33 vs 0	0.504	0.498	0.311	1.66	0.62	4.39	11.46 on 9 df, $p = 0.25$
66 vs 0	−0.128	0.549	0.831	0.85	0.32	2.39	
100 vs 0	−0.507	0.338	0.133	0.60	0.31	1.17	
Stereotypical model[f]			0.015	0.58	0.38	0.90	
b	−0.543	0.22					
33 vs 0 Φ_{c1}	0.345			1.41			
66 vs 0 Φ_{c2}	0.039			1.04			
100 vs 0 Φ_{c3}	−0.543			0.58			

[a] All models include age and sex as covariates.
[b] The response variable was the estimated ORs (0 or 33 or 66 vs. 100.
[c] Response variable is RLE score, i.e., 0, 33, 66, 100.
[d] The response variable is RLE score with 0 as the reference category.
[e] With $\phi_1 = 0$, $\phi_2 = -0.64$, $\phi_3 = 0.07$, and $\phi_4 = 1$.
[f] Odds ratios except for the polytomous and stereotype models where it is a ratio of probability ratios.
[g] Likelihood-ratio statistics for testing null model (no covariates) against the extended model (with covariates).
From Ref. 5

term (as opposed to short-term) ulcer problem; the probability of having a RLE score of 66 as opposed to 0 is 0.88 times the probability for subjects with long-term ulcers; and the probability of having a RLE score of 100 (as opposed to 0) is 0.60 times the probability for subjects with long-term ulcers. It is clear from this that the trend in the probability ratios is monotonic as the health status goes from the "poor" stage (i.e., score of 0) through to the better stages (i.e , RLE scores of 66 or 100).

Attaching a set of scores to the beta parameters in the polytomous model, leads to the formation of the stereotype model (Table 2). In this model the probability of having an RLE score of 33 compared to a score of 0 is 1.41 times the probability for subjects with a long-term (as opposed to short-term) ulcer problem; the probability of having an RLE score of 66 as opposed to 0 is 1.04 times the probability for subjects with long-term ulcers; and the probability of having an RLE score of 100 (as opposed to 0) is 0.58 times the probability for subjects with long-term ulcers.

5.2 Comparing Two Independent Groups—HRQoL Scales with More Than Seven Categories

When the HRQoL scale has more than seven categories, it is important to check in any particular situation that most of the categories are occupied, to rule out having only a few of a potentially large number of categories occupied. Where the distribution is spread over a number of categories, then it is useful to assume that the data were generated from a continuous distribution, especially if there is reason to believe the underlying scale is linear In this case the usual parametric tests such as the *t*-test or a nonparametric test such as the Mann-Whitney can be used. A further advantage to this assumption is that multiple regression can be used to adjust for confounding variables such as baseline covariates.

Estimation and Confidence Intervals (CI)

Journals such as the *British Medical Journal* and *Lancet* now expect scientific papers to contain confidence intervals when appropriate Indeed several statisticians have argued strongly for a change in emphasis in statistical analysis from hypothesis testing to estimation (14,23).

One way to estimate CIs is via the bootstrap method. The bootstrap is a computer intensive method for statistical analysis (24). There are several methods of estimating bootstrap confidence intervals. These include the percentile method, bias corrected, bias corrected and accelerated (BC_a), and approximate bootstrap confidence interval (ABC) method A full discussion of these methods is given in Efron and Tibshirani (24) or Davison and Hinkley (25). Efron and Tibshirani imply that the "best" bootstrap CIs are the BC_a ones.

Table 3. Comparison of Parametric, Bootstrap, and Multiple Regression Estimates of Confidence Intervals for the SF-36 Role Emotional Dimension Score at Baseline for Two Groups of Patients with Short- (<7 yrs) or Long-term (>7 yrs) Duration of Venous Leg Ulcers

Method	Mean Difference[a,b]	95% Confidence Interval	
		Lower Limit	Upper Limit
Normal (t-test)	10 3	-0 5	21.0
Bootstrap[c] percentile	10 3	0 3	20.8
CI bias corrected	10.3	-0 4	20.7
Bca	10 3	-0 6	20.5
Regression[d]	11 1	0 2	22.0

[a] *t*-test, $t = 1.8$ on 231 df, $p = 0.060$.
[b] Mann-Whitney *U*, $Z = -1.93$, $p = 0.053$
[c] $\chi^2_{\text{mod}} = 3.52$ on 1 df, $p = 0.061$
Bootstrap estimates are based on 10 000 random samples with replacement.
[d] Ordinary least squares multiple regression estimate ($p = 0.045$) adjusted for age and sex
From Ref. 5

Table 3 shows the observed difference in mean RLE scores between the short- and long-term ulcer duration groups and various estimates of 95% CIs from parametric, bootstrap, and regression methods.

The normal and bootstrap CIs give similar results, although the bootstrap limits are not constrained to be symmetrical about the mean difference estimate. Only the multiple regression estimates of the confidence limits (after allowing for age and sex) exclude zero.

6. OVERALL ANALYSIS OF MULTIPLE OUTCOME MEASURES: UNIVARIATE VS. MULTIVARIATE TESTS

HRQoL outcomes may have a number of dimensions. When several HRQoL outcomes are collected on the same people then it is always possible to test each variable separately. For example, if two groups are compared using the SF-36 then a difference between the means for the two groups can be tested separately for each of the eight dimensions of the SF-36. Unfortunately, there is a drawback to this approach because of the repeated use of significance tests, which means that the probability of falsely finding at least one significant difference accumulates with the number of tests carried out.

There are ways of adjusting significance levels in order to allow for multiple testing [e.g., Bonferroni correction (26)] but a single test that uses the information from all variables together may be preferable. One solution is to use multivariate methods such as Hotelling's T^2 test The problem with such multivariate methods is that they test very general hypotheses (e.g., does one group differ in some

nonspecified way in their HRQoL from another) and so consequently have very poor power to detect any real difference Thus, more often than not they will give a nonsignificant result. Tandon (27) suggested a parametric method that was more specific.

For example, to compare two groups, calculate t-tests for each dimension and then find

$$z = \frac{J'S^{-1}t}{(J'S^{-1}J)^{1/2}} \tag{15}$$

where $J' = (1, \ldots, 1, 1)$, S is the estimated correlation matrix, and t is the vector of t-statistics from the separate univariate t-tests. The test statistic z has an asymptotic standard normal distribution. The main drawback of these methods is that they do not give an estimate of the treatment effect, they just provide a test statistic.

7. CHANGES FROM BASELINE

Although it is common to take the patients' status at the end of the study period as the outcome of interest, sometimes it is more appropriate to take the *change* (or difference) from the pretreatment, or baseline, measurement as the prime outcome measure.

When changes from baseline are analyzed, it is misleading to perform separate analyses (either hypothesis tests or CIs) within each treatment group. A better approach is to calculate each patient's change from baseline, and then directly compare the changes in the different groups.

Analysis of Changes from Baseline: Given the distribution of the changes or differences in outcome measures are more likely to be symmetric and normally distributed, parametric tests can be used to compare differences in changes between groups, especially if we assume that for a seven-point HRQoL scale going from to 2 to 1 is the same as a change from, say, 7 to 6. Parametric CIs for means and their differences can then be calculated Multiple regression can be used (with change from baseline as the dependent Y variable) to compare groups and adjust for other covariates and factors (such as baseline HRQoL, age, sex, and treatment center).

Bajorski and Petkau (28) describe a nonparametric method of comparing changes from baseline on ordinal responses for two independent groups. (This method is based on performing several Wilcoxon rank-sum tests on the follow-up of HRQoL scores stratified by baseline HRQoL score These separate Wilcoxon test statistics for each strata are then weighted and summed together to produce an overall test statistic). However, again this method is purely a hypothesis test

and does not allow estimation of CIs and so does not appear particularly useful in the analysis of HRQoL.

8. REPEATED MEASURES

A common study design entails recording serial measurements of the same variable(s) on the same individual at several points in time. Such data are often analyzed by calculating means and SDs at each time and comparing groups. Repeated measurements of the same variable on one individual should not be treated as independent observations when comparing groups of individuals. The correct statistical method, which takes account of the relationships between the observations at different times is to use some form of repeated measures or multivariate analysis of variance.

However, it may be more valuable to analyze some characteristic of the individual response profiles, such as the time taken to reach a peak or the length of time above a certain level or the area under the response profile. The area under the curve (AUC) is a useful way of summarizing the information from a series of measurements on one individual (29). Parametric CIs for the mean difference in AUC between groups can also be calculated as again the AUCs are more likely to be a fairly good fit to the normal.

9. OTHER ISSUES

Missing values are a common problem, particularly when the patients are suffering from diseases with a high mortality. Several papers discuss the problem of missing data in quality of life studies and provide overviews of methods of analyzing incomplete longitudinal HRQoL data (30–32).

Billingham et al (33) critically reviewed the methods for analyzing survival and quality of life data in health technology assessment and found these methods fell into three broad categories: (i) QoL analysis in the presence of informative drop-out, (ii) analysis of survival data adjusting for QoL, and (iii) simultaneous analysis of QoL and survival data.

Ribaudo et al. (34) discusses two methods of analysing longitudinal ordinal data Both methods involve the use of the proportional odds model described previously [Eqs (5) and (6)] One analysis uses summary statistics and the second uses a multilevel model for tracking patients over time (35).

10. FUTURE DEVELOPMENTS

There is an increasing use of HRQoL outcomes in clinical trials, and it is therefore necessary to standardize methods for reporting HRQoL, from the point of view of both pharmaceutical companies and regulatory authorities. Staquet et

al. (36,37) have produced guidelines for reporting HRQoL studies although these do not appear to be widely used.

With health care resources becoming scarce and with many competing health technologies, choices will need to be made on what is the most effective use of those health care resources. Therefore HRQoL measures that have utilities attached will become more widespread and will be used in cost-utility analysis to compare different health technologies.

Further research is required to establish what is a realistic and clinically meaningful effect size for a number of HRQoL measures. If a nonparametric approach is used then the appropriate "effect size" is the odds ratio. Therefore further research is required on what are plausible ORs and what is the meaning of such effects.

The nonparametric method relies on the assumption of a common OR. This assumption of proportional odds should be checked. Only one paper (38) appears to have checked the assumption of proportional odds. Furthermore it is not clear how robust the nonparametric method is to departures from the proportional odds assumption.

Alternatively, given the robustness of the parametric (t-test) approach to violations of the assumptions of normality and equality of variances, when can an ordinal HRQoL measure be treated as approximately continuous (e.g., how many categories)? Similarly, how skewed/asymmetric does the distribution of a HRQoL measure have to be to markedly affect the parametric approach to statistical analysis and sample size calculation? As we mentioned previously the central limit theorem guarantees the distribution of sample means will be nearly normally distributed whatever the distribution of the measurement amongst individuals, provided the sample size is large enough.

With the rapid development in personal computers and software it is surprising that bootstrap methods and computer simulation are not used more widely in analyzing HRQoL outcomes. This is an interesting avenue and requires further exploration.

11. RECOMMENDATIONS FOR THE DESIGN AND ANALYSIS OF TRIALS WITH HRQoL OUTCOME MEASURES

It is important to make maximum use of the information available from other related studies or extrapolation from other unrelated studies. The more precise the information, the better we can design the trial. This information or clinical experience may suggest a reasonable effect size or odds ratio. A relevant sample size is essential to the design of any study.

If an investigator is uncomfortable with the assumptions required for a particular sample size calculation, then it is good practice to calculate sample sizes under a variety of scenarios so that the sensitivity to assumptions can be assessed.

If there is little prior knowledge of the full distribution of scores for the

HRQoL outcome this may not be too problematical. Using the nonparametric (ordinal) approach to sample size calculation, knowledge of the anticipated distribution within four or five broad categories is usually sufficient to determine the required number of subjects.

On the basis of work presented here, we would recommend that researchers planning a study with HRQoL as the primary outcome pay careful attention to any evidence of the validity and frequency distribution of the proposed HRQoL instrument.

The frequency distribution of HRQoL scores should be plotted. If there is extreme skewness, or it is bimodal, or modes at the top or bottom of the ranges then nonparametric methods should be used for sample size calculations and analysis.

As a rule of thumb, we suggest that if HRQoL outcome has seven or more occupied categories, and it is plausible that there is an underlying continuum, then the scale can be treated as if it were continuous; if less than seven then treat it as if it were ordinal. For HRQoL measures with less than seven categories, ordinal regression models may be appropriate. These regression models enable treatment estimates and CIs to be calculated and adjust outcomes for other covariates.

When comparing the change from baseline these differences are more likely to be normally distributed, therefore one may be able to use parametric tests to compare changes and estimate CIs. For repeated measures (taken at, say, three or more time points) calculate a summary measure such as the AUC and use parametric tests (e.g., t-tests) to compare summary measures and estimate CIs.

The guidance presented here is not meant to imply that other more fundamental design factors, such as whether a randomized controlled design can be used, are not important or should not be considered. However, to date, the points made about calculating sample sizes for HRQoL measures and analyzing HRQoL outcomes have not been well recognized. Perhaps the adoption of some of the above recommendations by the developers of HRQoL instruments and in guidelines used by medical journals for refereeing HRQoL studies would help facilitate change.

REFERENCES

1 Medical Research Council. *Clinical Trials Guidelines*. MRC. London, 1993.
2 Olschewski, M.; Schumacher, M. Statistical Analysis of Quality of Life Data in Cancer Clinical Trials. Statistics in Medicine 1990, 9, 749–763.
3 Zigmond, A.S., Snaith, R.P. The Hospital Anxiety and Depression Scale. Acta Psychiatrica Scandinavica 1983, 67, 361–370.
4 Ware, J.E., Snow, K.K., Kosinski, M., Gandek, B. *SF-36 Health Survey Manual & Interpretation Guide*. New England Medical Centre. Boston, MA, 1993.
5. Walters, S.J., Morrell, C.J., Dixon, S. Measuring Health-Related Quality of Life in Patients with Venous Leg Ulcers. Quality of Life Research 1999, 8, 327–336.
6 Campbell, M.J., Julious, S.A., Altman, D.G. Estimating Sample Sizes for Binary,

249

Ordered Categorical, and Continuous Outcomes in 2 Group Comparisons. British Medical Journal **1995**, *311*, 1145–1148.
7 Machin, D , Campbell, M.J., Fayers, P.M ; Pinol, A.P.Y. *Sample Size Tables for Clinical Studies*, 2nd Edition. Blackwell: Oxford, 1997.
8. Whitehead, J Sample Size Calculations for Ordered Categorical Data. Statistics in Medicine **1993**, *12*, 2257–2271; erratum appears in Stat Med **1994** *13*, 871.
9 Hogg, R.V ; Tanis, E A. *Probability and Statistical Inference*, McMillan. New York, 1987.
10. Julious, S.A.; George, S.; Campbell, M J Sample Sizes for Studies Using the Short Form 36 (SF-36) Journal of Epidemiology & Community Health **1995**, *49*, 642–644.
11. Julious, S.A., George, S.; Machin, D , Stephens, R.J Sample Sizes for Randomized Trials Measuring Quality of Life in Cancer Patients Quality of Life Research **1997**, *6*, 109–117.
12. Armitage, P., Berry, P. *Statistical Methods in Medical Research*, 3rd Edition; Blackwell Oxford, 1984.
13. Heeren, T , D'Agostino, R Robustness of the Two Independent Samples t-Test When Applied to Ordinal Scaled Data Statistics in Medicine **1987**, *6*, 79–90.
14. Altman, D.G. *Practical Statistics for Medical Research*; Chapman & Hall: London, 1991.
15. Armstrong, B G ; Sloan, M Ordinal Regression Models for Epidemiologic Data. American Journal of Epidemiology **1989**, *129*, 191–204.
16. Ananth, C.V ; Kleinbaum, D.G. Regression Models for Ordinal Responses: A Review of Methods and Applications. International Journal of Epidemiology **1997**, *26* (6), 1323–1333.
17. Manor, O ; Matthews, S., Power, C. Dichotomous or Categorical Response? Analysing Self-Rated Health and Lifetime Social Class. International Journal of Epidemiology **2000**, *29*, 149–157.
18 McCullagh, P ; Nelder, J.A *Generalised Linear Models*, 2nd Edition; Chapman & Hall London, 1989
19 Greenland, S Alternative Models for Ordinal Logistic Regression. Statistics in Medicine **1994**, *13*, 1665–1677.
20. Brant, R Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression Biometrics **1990**, *46*, 1171–1178.
21. Peterson, B , Harrell, F.E Partial Proportional Odds Models for Ordinal Response Variables Applied Statistics **1990**, *39*, 205–217.
22. Anderson, J A Regression and Ordered Categorical Variables (with Discussion). J. Roy. Stats. Soc Series B **1984**, *46*, 1–30
23. Altman, D.G , Machin, D , Bryant, T N , Gardner, M J *Statistics with Confidence*, 2nd Edition; BMJ. London, 2000.
24 Efron, B., Tibshirani, R J *An Introduction to the Bootstrap*, Chapman & Hall: New York, 1993.
25. Davison, A C , Hinkley, D V *Bootstrap Methods and Their Applications*; Cambridge University Press Cambridge, 1997.
26. Bland, J M , Altman, D G Statistics Notes. Multiple Significance tests The Bonferroni Method British Medical Journal **1995**, *310*, 170.
27. Tandon, P K. Applications of Global Statistics in Analysing Quality of Life Data. Statistics in Medicine **1990**, *9*, 749–763

28. Bajorski, P., Petkau, J Non-parametric Two-sample Comparisons of Changes on Ordinal Responses J American Statist. Assoc. **1999**, *94* (447), 970–978.
29 Matthews, J N S , Altman, D G , Campell M.J , Royston P Analysis of Serial Measurements in Medical Research British Medical Journal **1990**, *300*, 230–235.
30. Curran, D ; Molenberghs, G.; Fayers, P M , Machin, D Incomplete Quality of Life Data in Randomised Trials Missing Forms Statistics in Medicine **1998**, *17*, 697–709.
31 Fayers, P.M., Curran, D ; Machin, D Incomplete Quality of Life Data in Randomised Trials: Missing Items Statistics in Medicine **1998**, *17*, 679–696
32 Troxel, A.B ; Fairclough, D.L , Curran, D , Hahn, E A Statistical Analysis of Quality of Life with Missing Data in Cancer Clinical Trials Statistics in Medicine **1998**, *17*, 653–666.
33 Billingham, L.J.; Abrams, K R., Jones, D.R Methods for the Analysis of Quality-of-Life and Survival Data in Health Technology Assessment Health Technol. Assess **1999**, *3* (10), 1–151.
34. Ribaudo, H J , Thompson, S.G A Multilevel Analysis of Longitudinal Ordinal Data: Evaluation of the Level of Physical Performance of Women Receiving Adjuvant Therapy for Breast Cancer J Roy Statist Soc Series A **1999**, *163* (3), 349–360.
35 Beacon, H.J., Thompson, S Multi-Level Models for Repeated Measurement Data: Application to Quality of Life Data in Clinical Trials Statistics in Medicine **1996**, *15*, 2717–2732.
36. Staquet, M , Berzon, R , Osoba, D ; Machin, D. Guidelines for reporting results of quality of life assessments in clinical trials Quality of Life Research **1996**, *5*, 496–502.
37 Staquet, M.J ; Hays R D , Fayers, P M *Quality of Life Assessment in Clinical Trials: Methods and Practice*, Oxford University Press Oxford, 1998.
38 Bolland, K , Soonyarachchi, M R , Whitehead, J Sample Size Review in a Head Injury Trial with Ordered Categorical Responses Statistics in Medicine **1998**, *17*, 2835–2847.

Received August 2000
Revised February 2001, July 2001
Accepted July 2001

Health Services & Outcomes Research Methodology 2: 83–99, 2001

Methods for Determining Sample Sizes for Studies Involving Health-Related Quality of Life Measures: A Tutorial

STEPHEN J. WALTERS, M.Sc.*
Lecturer in Medical Statistics, Sheffield Health Economics Group, School of Health and Related Research, University of Sheffield

MICHAEL J. CAMPBELL, Ph.D.
Professor of Medical Statistics, Institute of General Practice & Primary Care, School of Health and Related Research, University of Sheffield

SUZY PAISLEY, M.A.
Information Officer, Information Resources, School of Health and Related Research, University of Sheffield

Received August 15, 2000, revised October 3, 2001; accepted January 2, 2002

Abstract. Health Related Quality of Life (HRQoL) measures are becoming more frequently used in clinical trials and health services research, both as primary and secondary endpoints. Investigators are now asking statisticians for advice on how to plan and analyse studies using HRQoL measures, which includes questions on sample size. Sample size requirements are critically dependent on the aims and objectives of the study, the proposed summary measure and effect size and the method of calculating the test statistic.

We present a tutorial on methods of sample size calculation for HRQoL outcomes. We also briefly review the HRQoL literature to see what has been done in practice. The aim of this tutorial is provide pragmatic guidance to researchers on the methods of calculating sample sizes when using HRQoL measures as outcomes.

HRQoL measures such as the SF-36, NHP and QLQ-C30 are usually measured on an ordered categorical (ordinal) scale. We argue that it is often incorrect to treat the scales as if they were continuous and normally distributed and that the mean score may not be a good summary measure of HRQoL data. However the ordinal scaling of HRQoL measures leads to problems in determining sample size, and we suggest that the odds ratio (OR) may be a more suitable summary measure for comparing groups (rather than the mean difference) and therefore methods suitable for ordinal data be used for analysis.

The frequency distribution of HRQoL scores should be assessed to see if parametric assumptions are satisfied and whether or not the sample mean is a good summary measure of the data. Given the non-normal distribution of the majority HRQoL outcome measures, summary measures such as means and standard deviations are difficult to interpret. Thus standardised differences (effect sizes) and parametric methods may not be a suitable basis for calculation of sample size. Finally we argue, that any sample size calculation (with all its attendant assumptions) leads to better research than no sample size calculation at all.

*To whom correspondence should be addressed at: School of Health and Related Research, University of Sheffield, Regent Court, 30 Regent St, Sheffield, S1 4DA.
Tel.: 0114 2220730, Fax: 0114 2225454, E-mail: s.j walters@sheffield.ac.uk

1. Introduction and Background

Sample size calculations are now mandatory for many research protocols and are required to justify the size of clinical trials in papers before they will be accepted by journals (Altman et al., 2000). Thus, when an investigator is designing a study to compare the outcomes of an intervention, an essential step is the calculation of sample sizes that will allow a reasonable chance (power) of detecting a predetermined difference (effect size) in the outcome variable, at a given level of significance. Sample size is critically dependent on the summary measure, the proposed effect size and the method of calculating the test statistic. For example, for a given power and significance level, the sample size is inversely proportional to the square of the effect size, so, halving the effect size will quadruple the sample size.

Health Related Quality of Life (HRQoL) measures are being used more frequently in clinical trials and health services research, both as primary and secondary endpoints. Investigators are now asking statisticians for advice on how to plan and analyse studies using HRQoL measures, and this includes questions on the sample size.

HRQoL measures such as the SF-36, Nottingham Health Profile (NHP) and QLQ-C30 are described in Fayers and Machin (2000) and are usually measured on an ordered categorical (ordinal) scale. This means that responses to individual questions or items are usually classified into a small number of response categories, which can be ordered, for example, poor, moderate and good. In planning and analysis, the question responses are often analysed by assigning equally spaced numerical scores to the ordinal categories (e.g. 0 = 'poor', 1 = '*moderate*' and 2 = 'good') and the scores across similar questions are then summed to generate a HRQoL measurement. These 'summated scores' are treated as if they were from a continuous distribution and normally distributed.

However the ordinal scaling of HRQoL measures may lead to several problems in determining sample size and analysing the data. To illustrate this we use some HRQoL data from a randomised controlled trial that aimed to compare the difference in health status in a group of women who were offered postnatal support (intervention) from a community midwifery support worker compared with a control group of women who were not offered support (Morrell et al., 2000).

1. The apparent continuum hides the fact that only a few discrete values are possible. For example the Role Limitations Physical (RLP) dimension of the SF-36 (Brazier et al., 1992) is scored on a 0 to 100 scale but there are only five possible categories/scores e.g. 0, 25, 50, 75 and 100 (see Figure 1).
2. It is unlikely that the scale is linear. For example using the SF-36 RLP dimension is a change of score from 0 to 25 the same as a change from 75 to 100?
3. There is often a floor or ceiling effect (patients cannot be worse than the worst category or better than the best category). For some populations the level is wrong and most people either score on the best category or the worst category. Floor and ceiling effects are more likely to be a problem in longitudinal studies because this limits the ability of the instrument to detect an improvement or deterioration in a patient's HRQoL over time. Figure 1 shows that over 55% (291/525) of the combined sample had scored 100 and were at the ceiling of the distribution.

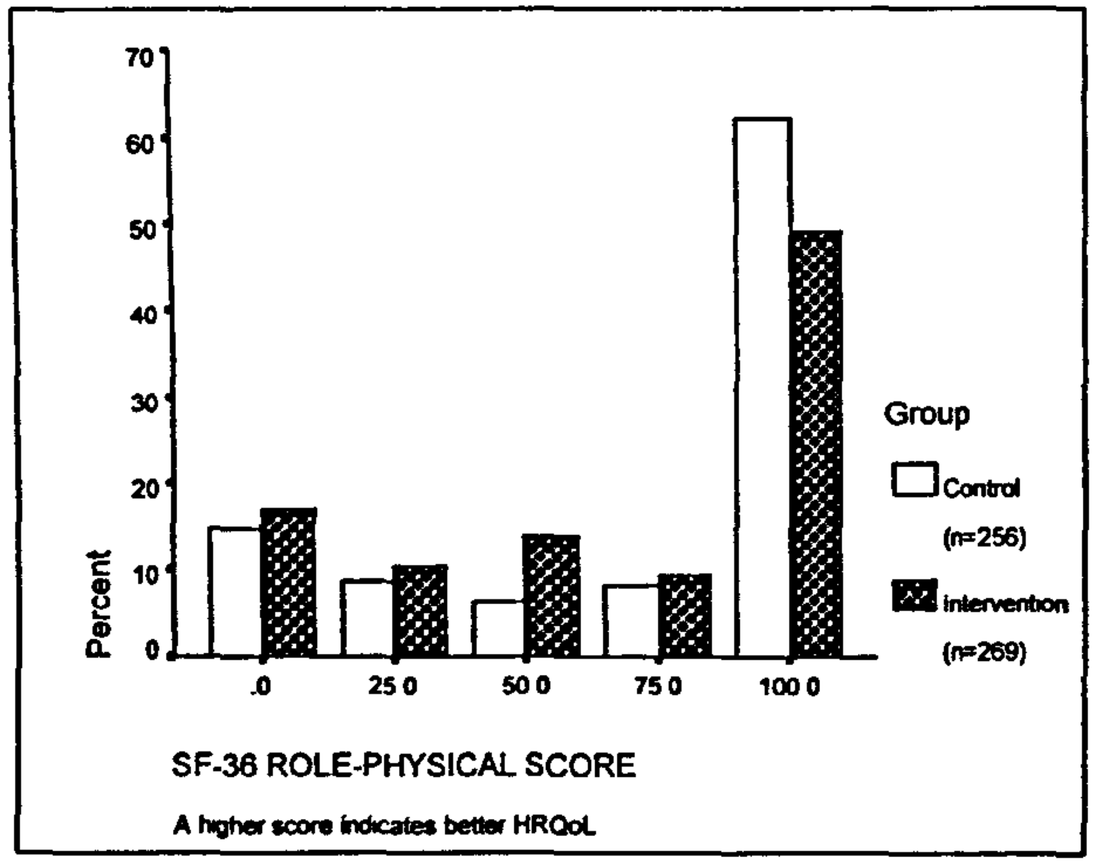


Figure 1. Distribution of SF-36 Role Limitations Physical outcome by group (Morrell et al., 2000).

4. Methods based on the normal distribution (such as linear regression) assume that the outcome variable has a constant variance. The variance of ordinal data is not constant. Variances of changes may depend on initial values. This is a common problem with range-limited values. Patients may enter the study with a wide variety of scores, but tend to always increase their scores. Thus patients who score lower at the start of the study have more range to improve than those who are close to the maximum already.
5. Normal approximations may not apply. Since the data are in fact categorical, they may require different techniques of analysis. By definition, no ordinal variable can be *normally* distributed, although in some cases a normal approximation will suffice.
6. Missing values are likely, for example questionnaires that ask 'how far can you walk?' when patient is in a wheel chair.
7. It is difficult to quantify an effect size (e.g. a desirable difference in mean score between groups), in advance.

There are advantages in being able to treat HRQoL scales as continuous (e.g. for statistical analysis and economic evaluations) and, therefore, it is important to examine such discrepancies for different instruments and their scales.

The aim of this article is to provide a tutorial on methods of calculating sample sizes when using HRQoL measures as outcomes and to provide pragmatic guidance to researchers on what method to use.

The remainder of this paper is structured into the following sections. Section 2 summarises the methods and the sample size formulae. What researchers actually do in practice is discussed in Section 3. The consequences if different sample size formulas are applied are explored in Section 4. Section 5 discusses multiple end-points. The final sections (6 and 7) talk about the choice of sample size method with HRQoL outcomes and conclusions.

2. Which Sample Size Formulae?

In principle, there are no major differences in planning a study using HRQoL assessment to those using conventional clinical outcomes. Sample size is dependent on the outcomes and objectives of the study and the method of analysis. Thus, after deciding on the primary outcome, the investigator must choose an appropriate summary measure of this outcome and then calculate a sample size based on this summary measure.

An appropriate summary measure of the outcome data will usually be the sample mean, median, or a rate or proportion. When comparing two groups or a single group over time appropriate comparative summary measures may include the difference between sample means, difference in medians, difference in rates or proportions, the relative risk (RR) or the odds ratio (OR).

The mean is often chosen as a suitable summary measure, although there are several reasons against using it. One reason would be that the HRQoL outcome measure of interest is an ordinal not a continuous variable, and therefore means are hard to interpret (see points 1 to 7 above) Also the HRQoL outcome may have a skewed distribution and the median may be a more useful summary of HRQoL outcome.

For individual patients the outcome of treatment is usually dichotomous (the treatment either works or the treatment does not work) or ordinal (the effect of treatment worsens the patients' HRQoL, has no effect, or improves HRQoL). In this case given the probability of a successful outcome (improved HRQoL) on the control treatment (p_C), and the OR that the new treatment is beneficial (compared to the control treatment) then the probability that the new treatment will work for an individual patient (p_T) is:

$$p_T = \frac{\mathrm{OR}p_C}{\mathrm{OR}p_C + 1 - p_C} \tag{1}$$

Therefore the OR may be a suitable comparative summary measure for the effect of treatment at an individual level. We can calculate the risk difference ($p_T - p_C$) and hence the reciprocal of the risk difference $1/(p_T - p_C)$ which is the Number Needed to Treat (NNT).

$$\mathrm{NNT} = \frac{1}{p_T - p_C} \tag{2}$$

The NNT is the number of patients who need to be treated with the new treatment rather than the standard control treatment in order for one additional patient to benefit. Thus, the NNT is a useful summary measure for clinicians to compare two treatments, although it

does require the HRQoL outcome to be dichotomised (Laupacis, Sackett and Roberts, 1988).

The mean (and mean difference) is a more suitable summary measure for the effect of treatment on average (i.e. at a hospital level). The mean indicates the effect of treatment on average in this group of patients. This summary measure is useful for health care providers (or hospitals) in deciding whether to offer a new treatment or not to its population.

Campbell et al. (1995) outline the ways of calculating sample sizes in two group studies for binary, ordered categorical and continuous outcomes. Further details, examples and tables are given in the book by Machin et al. (1997).

2.1. Continuous Data – Comparing Two Means

If the HRQoL outcome is assumed to be continuous and plausibly sampled from a normal distribution then the best summary statistic for a location parameter is the mean, and the usual hypothesis test for a difference or shift in location parameters between two independent samples is the two-sample t test.

For two independent groups with continuous and normally distributed data the standardised effect size is the expected mean value of the intervention outcome minus the expected mean value of the control outcome divided by a standard deviation of the outcomes. I.e.

$$\delta_{\text{Continuous}} = \frac{\mu_T - \mu_C}{\sigma} \tag{3}$$

where $\delta_{\text{Continuous}}$ is the standardised effect size index, μ_T and μ_C are the expected group means of outcome variable under the null and alternative hypotheses and σ is the standard deviation of outcome variable (assumed the same under the null and alternative hypotheses).

In a two-group study comparing mean HRQoL between the two groups the number of subjects per group n for a two-sided significance level α and power $1-\beta$ is given by equation (4)

$$n_{\text{Continuous}} = \frac{2[z_{1-\alpha/2} + z_{1-\beta}]^2}{\delta^2_{\text{Continuous}}} \tag{4}$$

where, $z_{1-\alpha/2}$ and $z_{1-\beta}$ are the appropriate values from the standard normal distribution for the $100(1-\alpha/2)$ and $100(1-\beta)$ percentiles respectively.

If the sample size is "sufficiently large" then the Central Limit Theorem (CLT) guarantees that the sample means will be approximately normally distributed (Hogg and Tanis, 1988). Thus, if the investigator is planning a large study and the sample mean is an appropriate summary measure of the HRQoL outcome, then pragmatically there is no need to worry about the distribution of the HRQoL outcome and we can use equation (2) to calculate sample sizes. Although the normal distribution is strictly only the limiting form of the sampling distribution of the sample mean as the sample size n increases to infinity, but it provides a remarkably good approximation to the sampling distribution even when n is small and the distribution of the data is far from normal (Armitage and Berry, 1994).

Generally, if n is greater than 25 or 30, these approximations will be good. However, if the underlying distribution is symmetric, unimodal, and of the continuous type, a value of n as small as 4 or 5 can yield a very adequate approximation (Hogg and Tanis, 1988). Figure 1 clearly illustrates the discrete and skewed nature of HRQoL data and that a large sample size may be required for the assumption of normality to be valid.

The skewed distribution of the HRQoL data in Figure 1 also implies that the sample mean and mean difference may not be a suitable summary measures to compare the two groups. The mean score of the control group was 73.5 (SD 38.4) compared with a mean score of 65.7 (SD 39.2) in the intervention group at six weeks postnatally. A mean difference of 7.8 (95% CI: 1.2 to 14.2; $t = 2.31$ on 523 df, $p = 0.02$). The median scores were 100 and 75 respectively, with over 62% of the control group and 49% of the intervention group scoring 100.

Suppose we are planning a two-group study comparing HRQoL (using the RLP as the primary outcome) between the groups. We believe that the mean difference in HRQoL scores between the two groups is an appropriate comparative summary measure. Assuming a standard deviation σ of 38 and that a mean difference $(\mu_A - \mu_B)$ of 8 or more points is clinically and practically relevant gives a standardised effect size (from equation (3)) of 0.21. Using this effect size in equation (4) with a two-sided 5% significance level and 80% power gives the estimated number of subjects per group as 363.

2.2. Transformations

If the HRQoL outcome data is continuous but has a skewed distribution it may be transformed using a logarithmic transformation. The transformed variable may have more symmetric distribution that approximates better to the normal form. The problem is that certain HRQoL measures are scored on 0 to 100 scales and the natural logarithm of zero does not exist.

If we recode the RLP dimension so that a score of $0 = 1$, $25 = 2$, $50 = 3$, $75 = 4$ and $100 = 5$. Then the mean RLP score in the control group (on a 1 to 5 scale) is now 3.94 (SD 1.54). If we take natural logarithms ($\log_e$) of the recoded RLP score then the mean log-transformed score is 1.24 (SD 0.59). Equation (4) can now be applied to the log-transformed scale once the standardised effect size $\delta_{\text{Continuous}}$ is specified. Unfortunately, there is no simple interpretation for the log-transformed RLP scale, and so the inverse transformation is used to obtain scores corresponding to the recoded (1 to 5) RLP scale. The mean RLP score (on the 1 to 5 scale) using the inverse transformation is now $\exp(1.24) = 3.46$ compared to the original value of 3.94.

If one third of a category or unit change on the recoded RLP scale is considered the minimum clinically important difference to detect. This is approximately equivalent to an eight-point difference on the original 0 to 100 scale of the SF-36 RLP dimension as a one category change corresponds to 25 points. Then the untransformed effect size from equation (3) is: $(4.27 - 3.94)/1.55 = 0.21$. Using equation (4) this leads to a sample size of 363 patients per group.

Using the log-transformed scale of the RLP a third of a unit increase is approximately from 3.46 to 3.79. This is then expressed as an anticipated effect on the log transformed

scale as $\delta_{Continuous} = (\mu_T - \mu_C)/\sigma = [\log_e(3.46 + 0.33) - \log_e(3.46)]/0.59 = 0.15$. Using equation (4) with $\delta_{Continuous} = 0.15$ gives $n_{Continuous} = 711$ patients per group.

We have used a logarithmic transformation for non-normal data and made the sample size calculations accordingly. Other possible transformations for this purpose are the reciprocal or square root. A difficulty with the use of transformations is that they distort HRQoL scales and make interpretation of treatment effects difficult (Fayers and Machin, 2000). In fact, only the logarithmic transformation gives results interpretable on the original scale (Bland and Altman, 1996). The logarithmic transformation expresses the effect as a ratio of the geometric mean for patients in the treatment group to the geometric mean for patients in the control group. This is because the difference between two logarithms is the logarithm of the ratio: log (T) − log (C) = log (T/C).

However, this ratio will vary in a way that depends on the geometric mean value of the control treatment C (Fayers and Machin, 2000). For example, if the geometric mean for the control treatment C is 2 and treatment T induces a change in RLP of 1 unit compared to this level, then this implies an effect size of $\log_e(3/2) = 0.41$. On the other hand, for geometric mean of 4 for the treatment C but the same numerical change of one unit implies an effect size of $\log_e(5/4) = 0.22$. Thus, although in this example the effect size is a one unit difference in HRQoL in both cases when expressed on the untransformed RLP scale, the logarithmic transformation results in a second effect size which is almost half (0.22/0.41 = 0.54) the first. This makes interpretation difficult.

2.3. Dichotomous Categorical Data – Comparing Two Proportions

If the HRQoL outcomes are measured on a binary or dichotomous categorical scale, then an appropriate summary measure of the outcome data will usually be the sample rate or proportion. When comparing two groups or a single group over time appropriate comparative summary measures may include the difference in rates or proportions, the relative risk or the odds ratio.

The statistical hypothesis test used to compare two independent groups when the outcome is binary is the Pearson chi-squared test for a 2 × 2 contingency table. In this situation the anticipated effect size is $\delta_{Binary} = (\pi_T - \pi_C)$, where π_T and π_C are the proportions in the two treatment groups. In a two-group study comparing differences in rates or proportions between the groups the number of subjects per group n_{Binary} for a two-sided significance level α and power $1-\beta$ is given by equation (5).

$$n_{Binary} = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2[\pi_T(1-\pi_T) + \pi_C(1-\pi_C)]}{(\pi_T - \pi_C)^2} \tag{5}$$

Alternatively, the same difference between treatments may be expressed through the odds ratio (OR), which is defined as:

$$OR_{Binary} = \frac{\dfrac{\pi_T}{1-\pi_T}}{\dfrac{\pi_C}{1-\pi_C}} = \frac{\pi_T(1-\pi_C)}{\pi_C(1-\pi_T)} \tag{6}$$

This formulation leads to an alternative for equation (5) for the sample size. Thus,

$$n_{OR} = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2/(\log OR_{Binary})^2}{\bar{\pi}(1-\bar{\pi})} \tag{7}$$

where $\bar{\pi} = (\pi_T + \pi_C)/2$.

Equations (5) and (7) are quite dissimilar, but Julious and Campbell (1996) show that they give for all practical purposes very similar sample sizes, with divergent results only occurring for relative large (or small) OR_{Binary}.

Figure 1 indicates that approximately 60% of patients in the control group scored 100, i.e. "good health". Suppose it is anticipated that this may improve to 70% having good health with treatment T. The anticipated treatment effect is thus, $\delta_{Binary} = (\pi_T - \pi_C) = 0.70 - 0.6 = 0.10$. This equates to a sample size of $n_{Binary} = 353$ per group from equation (5).

Alternatively, this anticipated treatment effect can be expressed (using equation (6)) as $OR_{Binary} = (0.70/0.30)/(0.6/0.4) = 1.556$. Using this in equation (7) with $\bar{\pi} = (0.70 + 0.60)/2 = 0.65$ gives a sample size per group of $n_{OR} = 354$ patients. As we indicated previously, there is usually only a small and inconsequential difference between the calculations from the alternative formulae.

2.4. Ordered Categorical (Ordinal) Data

If the HRQoL outcomes are measured on an ordinal scale, then statistical hypothesis test used in this instance (to compare two independent groups) is the Mann-Whitney U test with allowance for ties or a chi-squared test for trend (Altman, 1991).

Whitehead presents the sample size formulae for ordinal data in a paper (Whitehead, 1993). To use Whitehead's formulae we need to specify an effect size. For ordinal data Whitehead suggested the odds ratio ($OR_{Ordinal}$), which is the odds of a subject being in a given category or lower in one group compared with the odds in the other group.

Suppose we have two groups treatment (T) and control (C) and the HRQoL outcome measure of interest Y has k ordered categories y_i denoted by $i = 1,2,\ldots,k$. Let p_{iT} be the probability of being in category i in group T and C_{iT} be the expected cumulative probability of being in category i or less in group T (i.e. $C_{iT} = Pr(Y \leq y_i)$). For category i, where i takes values from 1 to $k-1$, the $OR_{Ordinal}$ is given by

$$OR_{Ordinal} = \frac{\dfrac{C_{iT}}{(1-C_{iT})}}{\dfrac{C_{iC}}{(1-C_{iC})}} \tag{8}$$

The assumption of proportional odds specifies that the $OR_{Ordinal}$ will be the same for all categories from $i = 1$ to $k-1$, and is equal to $OR_{Ordinal}$. This is the proportional odds assumption which underlies the proportional odds model and hence the derivation of the formulae.

Figure 1 illustrates that the HRQoL outcome has five categories, which implies four cut-offs (RLP scores = 0, ≤ 25, ≤ 50 and ≤ 75) and therefore four separate ORs. As the proportional odds model assumes a constant OR for all categories, Figure 2 shows the four observed ORs compared to estimated common OR of 1.56 (95% CI: 1.12 to 2.17) from the proportional odds model (Whitehead, 1993). All observed ORs are greater than 1 and seem similar to the model estimate. A chi-squared score test of proportional odds was $\chi^2 = 6.27$ on 3 df, $p = 0.10$. This suggests that the proportional odds assumption is plausible. Although in other cases the test may lack sufficient power to detect meaningful departures from proportional odds (Peterson and Harrell, 1990; Brant, 1990). The model is robust to mild departures from the assumption of proportional odds. A crude test would be to examine the odds ratios and if they are all greater than unity, or all less than unity, then assume a proportional odds model will suffice. With increasing numbers of categories it is less likely that proportional odds assumption remains true.

The proportional odds model OR estimate implies that patients in the intervention group have 1.56 times the odds of being in a given category or below (i.e. have worse HRQoL) than patients in the control group.

Whitehead's method can be regarded as a 'non-parametric' approach as the derivation of the sample size formulae and analysis of data is based on the Mann-Whitney U test, although it still relies on the assumption of a constant odds ratio for the data. Whitehead's method also assumes a relatively small log odds ratio and a large sample size, which will often be the case in HRQoL studies where dramatic effects are unlikely. Equation (9) gives the number of subjects per group n for a two-sided significance level α and power $1-\beta$.

$$n_{\text{Ordinal}} = \frac{6[(z_{1-\alpha/2} + z_{1-\beta})^2/(\log \text{OR}_{\text{Ordinal}})^2]}{\left[1 - \sum_{i=1}^{k} \bar{\pi}_i^3\right]} \tag{9}$$

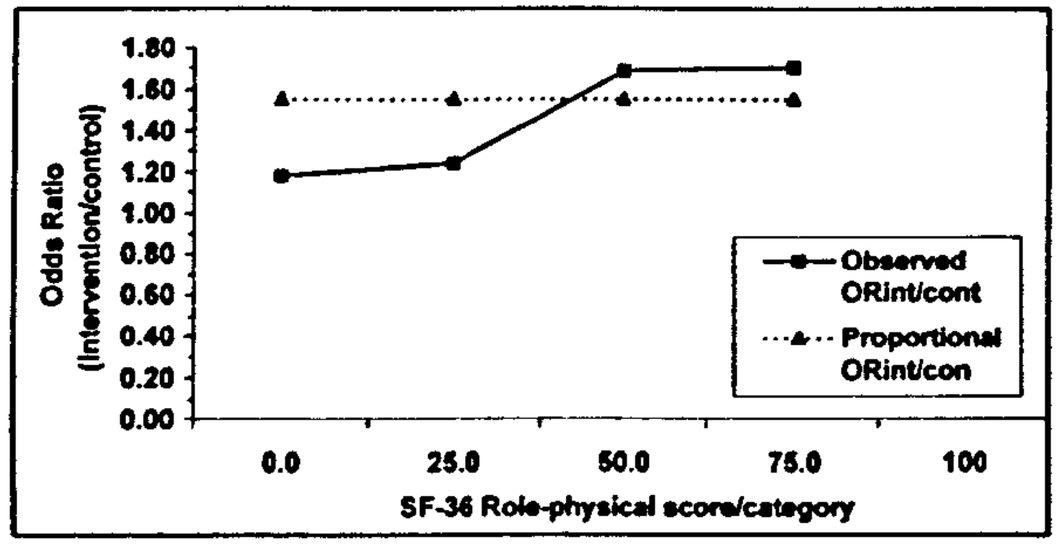


Figure 2. Odds Ratios for SF-36 Role Limitations Physical categories based on observed data and proportional odds model (Morrell et al., 2000).

Here $\bar{\pi}_i$ is the average proportion of subjects anticipated in category i, that is, $\bar{\pi}_i = (\pi_{Ti} + \pi_{Ci})/2$.

Suppose (as before) we are planning a two-group study to compare HRQoL (using the RLP as the primary outcome) between the groups. We believe that the mean difference in HRQoL scores between the two groups is *not* an appropriate comparative summary measure. However the odds of patient in the invention group having an HRQoL score in a given category or below compared to the odds for a patient in the control group is felt to be an appropriate comparative summary measure.

Approximately 60% of patients in the control group scored 100 i.e. "good health", with 40% scoring less than good health. As before suppose it is anticipated that this may improve to 70% having good health with treatment T, implying that 30% have less than good health. Using equation (8)

$$\text{OR}_{\text{Ordinal}\,i} = \frac{\dfrac{0.3}{(1-0.3)}}{\dfrac{0.4}{(1-0.4)}} = \frac{0.43}{0.67} = 0.64$$

leads to an $\text{OR}_{\text{Ordinal}} = 0.64$ which is the reciprocal of $\text{OR}_{\text{Binary}}$.

If we assume proportions of patients of 0.15, 0.09, 0.06, 0.08 and 0.62 respectively in the five RLP categories, 0, 25, 50, 75 and 100 in the control group. The cumulative proportions C_{iC} in category i for the control treatment C ($i = 1$ to 5) are 0.15, 0.24, 0.30, 0.38 and 1.0. Then, for a given constant $\text{OR}_{\text{Ordinal}} = 0.64$ the anticipated cumulative proportions (C_{iT}) for each category of treatment T are given by:

$$C_{iT} = \frac{\text{OR}_{\text{Ordinal}} C_{iC}}{\text{OR}_{\text{Ordinal}} C_{iC} + (1 - C_{iC})} \quad i = 1 \text{ to } k-1 \tag{10}$$

After calculating the cumulative proportions (C_{iT}), the anticipated proportions falling in each treatment category, π_{iT} can be determined from the difference in successive C_{iT}s. Finally, the combined mean ($\bar{\pi}_i$) of the proportions of treatments C and T for each category is calculated.

Using equation (10) with this OR and ($\bar{\pi}_i$) with a two-sided 5% significance level and 80% power gives the estimated number of subjects per group as 340. With a sample size of 340 and proportions of 0.10, 0.06, 0.05, 0.07 and 0.72 scoring 0, 25, 50, 75 and 100 respectively in the treatment group leads to a mean RLP score of 81.1 in the treatment group compared to a mean RLP score of 73.5 in the control group. This is a mean difference of 7.6 points, which is slightly smaller than the eight point mean difference used in equation (4) to calculate $n_{\text{Continuous}}$.

If the number of categories is large it is difficult to postulate the proportion of subjects who would fall in a given category. Both Whitehead (1993) and Julious et al. (1995) point out that there is little increase in power (and hence saving in the number of subjects recruited) to be gained by increasing the number of categories beyond five. Categories that are equally likely to occur lead to the greatest efficiency.

Julious and Campbell (1996) show that with two categories only, the method given by Whitehead is approximately equivalent to one described by Machin et al. (1997) for the binary case, even though at first sight the equations are very dissimilar. They state that the practical importance of this is to give the choice of two alternative measures of differences between groups: differences in proportions or odds ratios.

2.5. Alternative Approaches: Exact Methods and Simulation

Hilton and Mehta (1993) describe methods for sample size determinations based on either exact power or a very precise Monte Carlo estimate of it. This involves the use of an algorithm for computing the exact probabilities of each marginal table for a given fixed sample size. The conditional probability of a particular permutation of the data can be obtained by using the generalised hypergeometric distribution. Even with an efficient algorithm, processing the number of permutations is a considerable computational task.

Lesaffre et al. (1993) method involves the use of pilot data to estimate power and sample size. They used a variety of computer simulations including Monte Carlo and Bootstrap methods to estimate power for a fixed sample size.

Limitations of the formulae of Whitehead for small samples and when model assumptions are violated have been pointed out by Kolassa (1995) and Hilton (1996) respectively.

Julious and Campbell (1998) discuss the problem of calculating the number of subjects required in a matched or paired study in which the outcome variable is ordinal. In the two category (binary) case the sample size is dependent on the expected number of discordant pairs. They suggest that as a rule of thumb the required discordant sample size for the binary/two category case can be used as an approximation to the total sample size when the number of categories is greater than two.

2.6. Clustered Data

In some circumstances treatment is allocated to groups of patients rather than to individuals. Subjects within the same cluster (i.e. hospital or primary care facility) cannot be regarded as independent of each other and so the sample-size calculation must be modified and inflated by the design effect (DE). Let $\bar{c}$ be the mean number of patients per cluster, n the sample size for individual randomisation (obtained using equations (3), (5), (7), and (9)) and ρ_{Intra} the intra-class correlation i.e. the correlation within clusters with the outcome. Then the sample size per group (n_{Cluster}) for a two-armed cluster trial is given by:

$$n_{\text{Cluster}} = \text{DE}n \tag{11}$$

where

$$\text{DE} = 1 + (\bar{c} - 1)\rho_{\text{Intra}} \tag{12}$$

3. What do Authors/Researchers Actually Do in Real Life?

King (1996) mentions the importance of effect sizes in calculating sample sizes for clinical trials and also discusses the alternative parametric and non-parametric approaches, although does not give a recommendation for either one. King notes that there can be quite marked differences between sample sizes calculated from parametric and non-parametric methods particularly for HRQoL outcome measures that have a highly skewed distribution.

A few papers (Fayers and Machin, 2000; Julious and Campbell, 1996; Julious et al., 1997) appear to have used Whitehead's (1993) non-parametric method for ordinal outcomes. Bolland et al. (1998) applied Whitehead's method to a three category ordinal outcome (good recovery (GR), moderate disability (MD), severe disability/vegetative state/dead (SD/V/D)) in a randomised trial of patients suffering from severe head injury. They assumed a common odds ratio (proportional odds) of 1.84; proportions of patients of 0.17, 0.30 and 0.53 respectively in the three categories GR, MD and SD/V/D in the control group; and no effect of prognostic factors on outcome. This lead to an initial sample size of 400 patients.

Due to the uncertainty about these assumptions the authors planned a blinded sample size after approximately 100 patients were recruited. The review was performed on the first 93 patients to respond and led to an increase in sample size from 400 to 450. On completion of the study the authors note that the proportional odds assumption was, "whilst not fully valid, was not misleading".

Roset et al. (1999) apply parametric and non-parametric sample size methods to two datasets that have used the EQ-5D. They recommend parametric methods when the outcome variables are thought to be reasonably symmetrical and non-parametric methods when the data are skewed.

4. What Happens When Different Sample Size Formulas are Applied?

The sample sizes per group with similar anticipated treatment effects calculated for our example using equations (4), (5) and (9) respectively were $n_{\text{Cont}} = 363$, $n_{\text{Binary}} = 354$ and $n_{\text{Ordinal}} = 340$. The binary and ordinal calculations gave lower estimated sample sizes than for the continuous case, which may reflect the skewed nature of the RLP outcome data. Although for practical purposes the sample size estimates are broadly similar.

Three papers (Julious and Campbell, 1996; Julious et al., 1997; Machin and Fayers, 1998) have highlighted the discrepancies between sample sizes for intervention studies using HRQoL outcomes (HADS and SF-36) calculated using conventional parametric techniques and non-parametric approaches. In order to make the alternative hypotheses comparable, the authors used the distribution of the outcome for the control, and shifted it by one category for the intervention. The odds ratios between categories and groups formed by such a shift were calculated, so that the parametric and non-parametric methods are calculated using the same alternative hypothesis.

Using the SF-36 Julious et al. (1996) show that the results given by the parametric and non-parametric methods are similar in some dimensions of the SF-36 but are very different

in dimensions where the scores are highly skewed. For such asymmetric distributions the parametric methods give the same sample sizes for effects that are one unit above and one unit below the population mean. This is because the parametric method assumes a symmetric (normal) distribution, whereas the non-parametric method may give different sample sizes according to the expected direction of the effect. For example they show that the non-parametric estimate of the sample size to detect a change of one category for the General Health Perception dimension (which is quite symmetric) is similar for one category up or down, but for the Mental Health dimension (which is asymmetric) the sample size required to detect a change of one category down is three times that to detect one category up.

In all three articles the authors stress that "In general, statistics such as means and standard deviations are not suitable summary measures for non-normal distributions, and neither are standardised differences (*effect sizes*) a suitable basis for the calculation of sample sizes."

Julious et al. (1997) recommend that the frequency distributions of HRQoL scores should always be given so that one can assess if non-parametric methods should be used for sample size calculations and analysis. Given the skewed/asymmetric distribution of the majority of HRQoL outcomes in general they recommend ordered categorical methods be used for sample size calculations.

Prieto et al. (1996) in a letter to the editor about Julious et al. (1996) paper strongly disagreed with this recommendation. Firstly, Prieto and colleagues argue that between an ordinal and continuous scales there are a number of instruments (such as the SF-36) that can be labelled as 'summated scales' and in which the total score is the sum of a set of ordinal rankings and so these scales are 'between' ordinal and continuous. They do not claim that equal increments in the observed score along the summated scale represent equal increments in the underlying latent variable being measured, but the mode of construction of the instrument suggests that deviations from interval properties may not be extreme. Secondly, they argue that failure to meet the assumptions required for the use of parametric methods does not appear to have serious consequences in most instances. Therefore they suggest parametric techniques should be used for SF-36 sample size calculations. They note, however, that the minimum clinically important difference (MCID) for the SF-36 scales are still unknown and further research is needed to clarify the clinical significance of score changes on the SF-36 scales.

In reply Campbell et al. (1996) stated that the parametric method requires one to specify an effect size based on the standard deviation of the outcome. It is the distribution of the population not the estimate, that is important, and that the standard deviation for data that are not normally distributed is uninterpretable in terms of the distribution of the data. Thus one cannot expect that 95% of the observations to be within plus or minus two standard deviations of the mean. The problems are exacerbated when there are a limited number of categories, for example one dimension of the SF-36 (role limitations emotional (RLE)) can take only four values (0, 33, 67, 100), and in one study most of the population scored 100 (Brazier et al., 1992). In practice an apparent continuous scale is composed of several correlated binary responses and effectively the final response scale is binary (<100 or =100). In this case, Julious' methods demonstrate that the required sample size approaches the size required for a binary variable.

There is general agreement that further research is required to establish what are realistic and clinically meaningful effect sizes for the SF-36 and other HRQoL measures.

5. Multiple End Points

We have based the above calculations on the assumption that there is a single identifiable endpoint or HRQoL outcome, upon which treatment comparisons are based. Sometimes there is more than one endpoint of interest; HRQoL outcomes are multi-dimensional (e.g. the SF-36 has eight dimensions including RLP). If one of these dimensions is regarded as more important than the others it can be named as the primary endpoint and the sample size estimates calculated accordingly. The remainder should be consigned to exploratory analyses or descriptions only.

A problem arises when there are several outcome measures that are all regarded as equally important. One approach is to repeat the sample-size estimates for each outcome measure in turn and then select the largest number as the sample size required to answer all the questions of interest. Here, it is essential to note the relationship between significance tests and power: it is well recognised that P-values become distorted if many endpoints are each tested for significance and that adjustments should be made.

To guard against false statistical significance as a consequence of multiple hypothesis testing it is a sensible precaution to examine the consequences of replacing the significance level α in the various equations by a significance level adjusted using the Bonferroni correction. The Bonferroni correction is:

$$\alpha_{\text{Bonferroni}} = \alpha/t \tag{13}$$

where t is the number of endpoints or hypothesis tests to be performed. Such a correction will clearly lead to larger sample sizes. The Bonferroni approach to adjusting for multiple comparisons tends to be conservative as it assumes all the different endpoints are uncorrelated (Altman et al., 2000). In the case of HRQoL outcomes there is likely to be a strong correlation between the different dimensions. This conservativeness implies utilising criterion (13) will lead to failure to reject the null hypothesis on too many occasions.

6. Choice of Sample Size Method with HRQoL Outcomes

It is important to make maximum use of the information available from other related studies or extrapolation from other unrelated studies. The more precise the information the better we can design the trial. We would recommend that researchers planning a study with HRQoL as the primary outcome pay careful attention to any evidence on the validity and frequency distribution of the proposed HRQoL instrument.

The frequency distribution of HRQoL scores from previous studies should be assessed to see if means, rates or proportions are appropriate summary measures for the data, and hence whether parametric or non-parametric methods should be used for sample size

calculations and analysis. Given the skewed distribution of the majority HRQoL outcome measures, summary measures of central location such as means and summary measures of variability such as standard deviations may not be appropriate; and so standardised differences (effect sizes) and parametric methods may not be a suitable basis for calculation of sample size. It is difficult to interpret an effect defined by equation (3) when the data are skewed. We would suggest that investigators consider clinically meaningful effect sizes, and not rely on generic 'small', 'medium' or 'large' ones as suggested by Cohen (1988).

There maybe considerable uncertainties in estimates of such quantities as the standard deviation and the treatment effect. Sample size calculations are sometimes based on estimates "pulled out of thin air". If an investigator is uncomfortable with the assumptions then it is good practice to calculate sample sizes under a variety of scenarios so that the sensitivity to assumptions can be assessed (Julious and Campbell, 1996). We would recommend that various anticipated benefits be considered, ranging from the optimistic to the more realistic, with sample sizes being calculated for several scenarios within that range. It is a matter of judgement, rather than an exact science, as to which of the options is chosen for the final study size (Fayers and Machin, 2000).

If there is little prior knowledge of the full distribution of scores for the HRQoL outcome sample size calculation may not be too problematical. Using the ordinal approach to sample size calculation, knowledge of the anticipated distribution within four or five broad categories is usually sufficient determine the required number of subjects (Julious and Campbell, 1996; Whitehead, 1993).

The guidance presented here is not meant to imply that other more fundamental design factors such as whether a randomised controlled design can be used are not important or should not be considered. However, to date, the points made about calculating sample sizes for HRQoL measures have not been well recognised. Perhaps the adoption of some of the above recommendations by the developers of HRQoL instruments and in guidelines used by medical journals for refereeing HRQoL studies would help facilitate change.

7. Conclusions

Given that the end goal of using HRQoL outcomes in research studies is to assess a patient's health and well being, using the right type of HRQoL outcome in the right setting with an appropriate sample size calculation is crucial. Much time and energy is devoted to developing and validating HRQoL measures. Developers and researchers need to complement this effort with clearer descriptions of the distribution of such outcomes and what is an appropriate summary measure, the mean or the proportion with a certain score.

Finally we would stress the importance of a sample size calculation (with all its attendant assumptions), and that any such estimate is better than no sample size calculation at all, particularly in a trial protocol (Williamson et al., 2000). The mere fact of calculation of a sample size means that a number of fundamental issues have been thought about: what is the main outcome variable, what is a clinically important effect, and how is it measured? The investigator is also likely to have specified the method and frequency of data analysis.

Thus protocols that are explicit about sample size are easier to evaluate in terms of scientific quality and the likelihood of achieving objectives.

References

D. G. Altman, *Practical Statistics for Medical Research*, London: Chapman & Hall, 1991.

D. G Altman, D Machin, T. N. Bryant and M J. Gardner, *Statistics with Confidence. Confidence intervals and statistical guidelines, 2nd Ed.*, London: British Medical Journal, 2000.

P. Armitage and G. Berry, *Statistical Methods in Medical Research, 3rd Ed*, Oxford. Blackwell Science, 1994.

J. M. Bland and D. G Altman, "The use of transformation when comparing two means," *British Medical Journal*, 312, p. 1153, 1996.

K. Bolland, M. R. Sooriyarachchi and J. Whitehead, "Sample size review in a head injury trial with ordered categorical responses," *Statistics in Medicine*, 17, pp. 2835–2847, 1998.

R. Brant, "Assessing proportionality in the proportional odds model for ordinal logistic regression," *Biometrics*, 46, pp. 1171–1178, 1990.

J. E. Brazier, R. Harper, N M. B. Jones, A. O'Cathain, K. J. Thomas, T. Usherwood and L. Westlake, "Validating the SF-36 health survey questionnaire: new outcome measure for primary care," *British Medical Journal*, 305, pp. 160–164, 1992.

M. J. Campbell, S. A Julious and D G Altman, "Estimating sample sizes for binary, ordered categorical, and continuous outcomes in 2 group comparisons," *British Medical Journal*, 311, pp. 1145–1148, 1995.

M. J. Campbell, S. A. Julious and S. L. George, "Estimating sample sizes for studies using the SF-36 health survey – Reply," *Journal of Epidemiology & Community Health*, 50, pp. 473–474, 1996.

J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, New Jersey: Lawrence Earlbaum, 1988.

P. M. Fayers and D. Machin, *Quality of Life Assessment, Analysis and Interpretation*, Chichester: Wiley, 2000.

J. F. Hilton, "The appropriateness of the Wilcoxon test in ordinal data," *Statistics in Medicine*, 15, pp. 631–645, 1996.

J. F. Hilton and C. R. Mehta, "Power and sample size calculations for exact conditional tests with ordered categorical data," *Biometrics*, 49, pp. 609–616, 1993.

R. V Hogg and E. A. Tanis, *Probability and Statistical Inference*, 3rd Ed. New York: McMillan, 1988.

S. A. Julious and M. J Campbell, "Sample sizes calculations for ordered categorical data," *Statistics in Medicine*, 15, pp. 1065–1066, 1996.

S. A. Julious and M. J. Campbell, "Sample size calculations for paired or matched ordinal data," *Statistics in Medicine*, 17, pp. 1635–1642, 1998.

S. A. Julious, S. George, D. Machin and R. J. Stephens, "Sample sizes for randomized trials measuring quality of life in cancer patients," *Quality of Life Research*, 6, pp. 109–117, 1997.

S. A. Julious, S. George and M. J. Campbell, "Sample sizes for studies using the short form 36 (SF-36)," *Journal of Epidemiology & Community Health*, 49, pp. 642–644, 1995.

M. T. King, "The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30," *Quality of Life Research*, 5, pp. 555–567, 1996.

J. E. Kolassa, "A comparison of size and power calculations for the Wilcoxon statistic for ordered categorical data," *Statistics in Medicine*, 14, pp. 1577–1581, 1995.

A. Laupacis, D. L. Sackett and R. S. Roberts, "An assessment of clinically useful measures of the consequences of treatment," *N. Engl. J. Med.*, 317, pp. 1728–1733, 1988.

E. Lesaffre, I. Scheys, J. Frohlich and E. Bluhmki, "Calculation of power and sample size with bounded outcome scores," *Statistics in Medicine*, 12, pp. 1063–1078, 1993.

D. Machin, M. J. Campbell, P M. Fayers and A. J. Y Pinol, *Sample Sizes Tables for Clinical Studies*, Oxford: Blackwell Science, 1997.

D. Machin and P M. Fayers, "Sample sizes for randomised trials measuring quality of life," In: M. J. Staquet, R. D. Hays and P M. Fayers, Eds., *Quality of Life Assessment in Clinical Trials: Methods and Practice*, Oxford: Oxford University Press, pp. 37–50, 1998.

C. J. Morrell, H. Spiby, P. Stewart, S. Walters and A. Morgan, "Costs and effectiveness of community postnatal support workers: randomised controlled trial," *British Medical Journal*, 321, pp. 593–598, 2000.

B. Peterson and F. E. Harrell, "Partial proportional odds models for ordinal response variables," *Applied Statistics*, 39, pp. 205–217, 1990.

L. Prieto, J. Alonso and J. M. Anto, "Estimating sample sizes for studies using the SF-36 health survey," *Journal of Epidemiology & Community Health*, 50, p. 473, 1996.

M. Roset, X. Badia and N. E. Mayo, "Sample size calculations in studies using the EuroQol 5D," *Quality of Life Research*, 8, pp. 539–549, 1999.

J Whitehead, "Sample size calculations for ordered categorical data [published erratum appears in Stat Med 1994 Apr 30; 13(8): 871]," *Statistics in Medicine*, 12, pp. 2257–2271, 1993.

P. Williamson, J. L. Hutton, J. Bliss, J. Blunt, M. J. Campbell and R. Nicholson, "Statistical review by research ethics committees," *J. Roy. Statist. Soc. A*, 163, pp. 5–13, 2000.

Statistical Methods in Medical Research 2002; **11**: 49–67

A review of ordinal regression models applied on health-related quality of life assessments

R Lall, MJ Campbell Institute of General Practice and Primary Care, University of Sheffield, Community Sciences Centre, Northern General Hospital, Sheffield, UK, **SJ Walters** University of Sheffield, Sheffield Health Economics Group, Sheffield, UK, **K Morgan** Department of Human Sciences, Loughborough University, Loughborough, Leicestershire, UK and **MRC CFAS Co-operative** Institute of Public Health, Cambridge, UK

There has been increasing emphasis in medical research on the design and analysis of quality of life scales. Many quality of life scales are ordinal and statistical methods such as ordinal regression models have been reviewed on a number of occasions. However, when such models are applied, the way the data have been generated is often overlooked. In this paper we illustrate the use of ordinal regression models, in particular the proportional odds model, the partial proportional odds model and the stereotype model in the MRC Cognitive Function and Ageing Study (MRC CFAS). The partial proportional odds and the stereotype models are often under-utilized, perhaps due to the lack of available software. However, in this paper, analysis based on these models has been carried out using the popular statistical software package SAS and macros devised in SAS. Furthermore, bootstrapping techniques have been applied to obtain valid estimates of the standard errors of the parameters in the stereotype model. Strikingly different results were obtained using the different ordinal regression models. We conclude that the way the data have been generated is particularly important for the analysis of quality of life assessments. Different methods of generating scores yield data with different properties. It is now possible to fit a variety of ordinal regression models and so select the appropriate one that correctly models the data.

1 Introduction

There has been an increasing recognition that medical outcomes are not necessarily the most important results in studies that examine the effect of health interventions. This is particularly true for diseases that are presently incurable, such as advanced cancer and chronic diseases of the elderly. It is often the case that two interventions will have very similar medical outcomes, but have different effects on other aspects of people's lives. For this reason there has been increasing emphasis on the use of scales that measure quality of life. It is important that investment in healthcare delivers not only a longer life, but also an improved and maintained quality of life. In conjunction with economic and clinical measures, quality of life outcomes have provided a broader and more accurate assessment of the health status and well-being of patients. In addition, quality of life assessments have provided a means of examining the quality of care given and also have provided utilities such as quality-adjusted life years (QALYs) that aid in policy-making decisions. Quality of life assessments are often measured using questionnaires, and the choice is often between a standard (*generic*) one, which asks about

Address for correspondence: R Lall, Institute of General Practice and Primary Care, University of Sheffield, Community Sciences Centre, Northern General Hospital, Herries Road, Sheffield S5 7AU, UK.

10.1191/0962280202sm271ra

general health and which normally has a history of successful use, and a disease-specific one that has been specifically developed within the therapeutic area in question. The quality of life data can be summarized into a categorical scale that is often either nominal or ordinal in nature. For the analysis of nominal data, standard methods such as the Pearson's chi-squared test, logistic regression models etc, exist which quite adequately provide results and summarize the data. For ordinal scales, more complex statistical methods such as *ordinal regression models*[1] can be employed. When such models are applied to the analysis of quality of life data, the way this data have been generated is often overlooked. Other authors have also highlighted this point. For example, Greenland[6] emphasized that the type of ordinal regression model used for analysis should depend on the way the data have been processed and generated. This is particularly important in the case of quality of life and health status assessments, as different types of data are obtained depending on the biological and sampling processes that generated the data. For instance, in the case of the HADS (Hospital Anxiety and Depression Scale), a response to a question 'I still enjoy things I used to enjoy' can be recorded as: '(0) Definitely as much', '(1) Not as much', '(2) Only a little' or '(3) Hardly at all'. There are in total seven questions and the scores are summed and the final score ranges from 0 to 21. This score is divided into a three-category ordinal scale: 'Normal (<7)'; 'Borderline (8–10)' and 'Clinical depression (or anxiety) (11+)'. The categories on this scale are related to an underlying continuum, which is the score ranging from 0 to 21, and the ordinal variable is termed a *'grouped continuous'* variable.[7,8] The quality of life data obtained in this way are different to that, for instance, in some of the dimensions of the SF-36 quality of life questionnaire.[9] This questionnaire assesses the general health of individuals, and there is a question on health status that asks 'In general would you say your health is 'Excellent', 'Very good', 'Good', 'Fair', 'Poor'?'. Here the rank of the categories is known to exist on a single dimension. Although we can assume the categories are ordered, we do not know the structure of this ordering with respect to a given explanatory variable. Also, although an underlying variable exists when the categories are ordered, it is not directly related to the ordinal categories on the quality of life scale. For this reason Anderson[10] recognized these types of ordered categories as being discrete and referred to the outcome response as a *judged* or an *assessed* variable. Another example of an assessed variable is social class of different occupations, in which the ordering may depend on covariates such as income or level of education.

In general, assessed variables are likely to have greater observer error, and in most cases there is more subjectivity associated with them compared to the grouped continuous variables. It is therefore important to distinguish between the different ordinal quality of life variables, as this has consequences on the choice of ordinal regression models used to analyse the data.

The purpose of this paper is to illustrate the use of ordinal regression models, in particular the *proportional odds model*, the *partial proportional odds models* and the *stereotype model*, in the MRC Cognitive Function and Ageing Study (MRC CFAS).[2] The models are described in Section 2, the data in Section 3 and the fit of the models to the data in Section 4. We conclude with a discussion. Analysis has been carried out using the statistical software package SAS[4] (SAS Institute, Cary, NC; version 6.10 for Windows 95) and macros devised in SAS.[5]

2 Ordinal regression models

Prior to fitting any ordinal regression models, we assessed the general association of the response variable with respect to the covariates using the Cochran–Mantel–Haenszel (row mean scores) statistic as presented by Mantel.[11] This statistic examines the association between the ordinal response variable and one given covariate, while adjusting for the effect of the other covariate by treating it as a stratification variable. The ordinality of the response variable is taken into account by assigning scores to the response categories, forming means, and then examining location shifts of the means across the levels of the rows or *sub-populations* (which result when the levels of the covariates are cross-classified). Furthermore, as it is not clear whether the y-response categories are equally spaced, modified ridit scores were assigned to account for the ordinality. The formulation of the statistic is complex, and the computational details have been omitted as the statistic can be obtained in standard software; in this case PROC FREQ in SAS[12] was used.

2.1 Proportional odds model

The proportional odds model[7] is also known as the *cumulative logit* model. It is the most appropriate method of analysis when one is presented with a grouped continuous response variable. Consider the HADS scale mentioned in the introduction. Let Y denote the response and y_1, y_2 and y_3 indicate the categories of the HADS score: 'Normal (<7)'; 'Borderline (7–10)' and 'Clinical depression (or anxiety) (11+)'. Thus $\Pr(Y = y_j)$ is the probability that a randomly selected individual is in category j. The points '7' and '10' are the *cut-off points*. In generalizing this to a c-point scale, let the response categories be denoted by $y_1, y_2, \ldots, y_c$ and $X_1, X_2, \ldots, X_p$ be a set of explanatory variables or covariates. Taking the proportions $\Pr(Y = y_j) = \pi_j (j = 1, \ldots, c)$ which are based on the marginal totals of a sub-population, one can form cumulative probabilities. Thus for a given sub-population, let $\Pr(Y \leq y_j)$ denote the probability of a response in category y_j or below, ie $\Pr(Y \leq y_j) = \pi_1 + \pi_2 + \cdots + \pi_j$, then $\Pr(Y \leq y_1) \leq \Pr(Y \leq y_2) \cdots \Pr(Y \leq y_c) = 1$ exists, and these cumulative probabilities reflect the ordering in the response categories. The proportional odds model is based on such cumulative probabilities, and this model can be written as:

$$\log\left[\frac{\Pr(Y \leq y_j / X_1 \cdots X_p)}{\Pr(Y > y_j / X_1 \cdots X_p)}\right] = \alpha_j - \left(\beta_1 X_1 + 1\beta_2 X_2 + \cdots + \beta_p X_p\right) \qquad j = 1, 2, \ldots, c-1 \tag{1}$$

Note that the negative sign in the systematic component of (1) makes the sign of each component of β have the usual interpretation in terms of whether the effect is positive or negative. In (1) the regression parameters $\beta_k (k = 1, \ldots, p)$ and α_j are unknown and therefore estimated. The ordinal response categories are monotonically related to an underlying continuous latent variable Z. The relationship between Y and Z is such that the parameters α_j are the division points on the continuum Z and a response in category

y_2, for example, is observed if Z lies between α_1 and α_2. If two adjacent y-response categories are pooled together, or Y is changed by moving a cut-off point, the regression parameters β_k in the model remain unchanged. This property, known as *invariance*, is an attractive feature of the proportional odds model. In practice, however, one uses observed Y values and pooling them will lead to different estimates and inferences of β.

Although the proportional odds model is primarily used in the case where one is presented with a y-response which has an underlying grouped continuous variable, it can also be used in circumstances, for example, when the categories y_j are not directly related to an underlying continuum. In this case, however, the interpretation of the parameters, particularly the $\{\alpha_j s\}$ becomes difficult[7].

The proportional odds model is the most commonly used regression model in the context of analysing ordinal scales,[3] mainly because it provides a single estimate of the log odds ratio over the cut-off points. This estimate is not a weighted average of the cut-off point-specific log odds ratios, but is the optimum estimate obtained using the maximum likelihood or weighted least squares methods. It is ideal in terms of the ease of interpretation of the data and in terms of model parsimony. The $\beta_k (k = 1, \ldots, p)$ parameters in model (1) represent the constant cumulative log odds across all the cut-off points for the covariate X_k, having accounted for all the other covariates in the model. The cumulative log odds ratio, λ_k, is obtained by subtracting the log odds (also known as the *logit*) of one row from the log odds of another row.

In SAS, PROC LOGISTIC and PROC CATMOD can be used to fit the proportional odds model. For the analysis presented here, PROC CATMOD was used and the proportional odds model was fitted using the *clogit* link function and an appropriate design matrix.[13] The cumulative logits formed in PROC CATMOD are the reciprocal of those in (1). Note that this procedure only provides the estimates of the $\beta_k (k = 1, \ldots, p)$, and the log odds ratios have to be obtained by subtracting the logits of the appropriate rows of interest. The assumption of a constant odds ratio across the cut-off points is assumed for each covariate in model (1), and the $\{\beta_k s\}$ are calculated with this in mind. Prior to fitting a proportional odds model, it is important to carry out a preliminary check of whether the assumption of proportionality is satisfied by each covariate. One way of doing this is to fit a different β_k for each level of the outcome. A different slope model is a starting point for the analysis, and for each covariate the cut-off point-specific β_k parameters together with their standard errors are observed. The homogeneity of the proportional odds ratios over all the cut-off points can be tested using the χ^2-score test statistic.[14] This test is anti-conservative, as it lacks power for moderate departures from the proportional odds assumption, but does highlight major departures. It is also a global test of non-proportionality and does not distinguish heterogeneity associated with individual covariates. In cases where the proportional odds model does not hold for some of the covariates, then alternative models are considered (see below).

2.2 Partial proportional odds models

There is general consensus that the assumptions underlying the proportional odds assumption is quite stringent.[1] This is exacerbated when one considers more than one covariate, and in practice, the chance of all the covariates in the model having proportional odds is likely to be quite rare. The partial proportional odds model

permits some covariates to be modelled with the assumption of proportional odds, whilst allowing others to have odds ratios which vary by cut-off point. In general there are two types of partial proportional odds model: the Unconstrained Partial Proportional Odds model and the Constrained Partial Proportional Odds model.[14,15] For the former model the cut-off point-specific odds ratios are obtained for the variables where the odds are thought not to be proportional and a constant odds ratio is obtained for variables where the odds are believed to be proportional. For the latter model, for covariates where the proportional odds assumption does not hold, one may expect a certain 'pattern' in the cut-off point-specific odds ratios, eg a linear trend may be expected in the log odds ratios over the cut-off points. In such a case a set of linear constraints may be placed on the parameter in the model, such that an adequate fit be obtained.

These models are an extension of the proportional odds model. Invariance exists in these models for variables where there are proportional odds and the quality of life scale has an underlying continuum, which is directly related to the y-response categories. Again, due to the way the cumulative probabilities are formed, ordering is inherent in the model irrespective of the relationship of the y-response and the covariate.

2.2.1 *Unconstrained partial proportional odds model*

Let Y be the response variable that has a similar form to that presented in Section 2.1. Then a partial proportional odds model where there are p predictor variables, some of which have proportional odds and some of which have non-proportional odds (say q of them), takes the form:

$$\log\left[\frac{\Pr(Y \leq y_j/X_1 \cdots X_p)}{\Pr(Y > y_j/X_1 \cdots X_p)}\right] = \alpha_j - \Big([\beta_1 X_1 + \gamma_{j1} T_1] + [\beta_2 X_2 + \gamma_{j2} T_2] + [\beta_q X_q + \gamma_{jq} T_q] + \cdots [\beta_p X_p]\Big) \qquad j = 1, 2, \ldots, c-1 \tag{2}$$

Here $X_1, X_2, \ldots X_p$ are the complete set of covariates, and q of these are known to have non-proportional odds, with the remaining having proportional odds. The $\beta_1 \cdots \beta_p$ parameters are the components of each of the covariate-specific log odds, for which proportionality over the cut-off points can be assumed. The $T_1 \cdots T_q$ $(= Y_1 \cdots X_Q)$ exist only for the q variables that have non-proportional odds. Thus $\gamma_{j1} \cdots \gamma_{jq}$ are non-zero for the q-covariates and zero otherwise and are the components of the log odds that vary over the cut-off points.

For model (2), we estimate the $c-1$ intercept parameters, p beta regression parameters that are independent of the cut-off points, and a further $(c-1) \times q$ γ-parameters which are associated with each covariate and cut-off point. For a variable X_m where non-proportional odds exist in relation to the response, $\alpha_j - \beta_m X_m$ is incremented by a regression coefficient $\gamma_{jm} T_m$, which is the effect associated with each jth cumulative logit, having accounted for all the covariates. Note that $\gamma_{1m} = 0$, such that the logit associated with the first cut-off point is based on $\alpha_1 - \beta_m X_m$.

2.2.2 *Constrained partial proportional odds model*

Given the relationship of a covariate and the response is represented with non-proportional odds, then for the individual cut-off point-specific odds ratios, often a certain type of trend may be anticipated, e.g., a linear trend may be expected. In such a case, a set of constraints can be placed on the parameters in the model, so that the trend is taken into account. When the constraints are incorporated into the unconstrained partial proportional odds model, this model becomes:

$$\log\left[\frac{\Pr(Y \leq y_j/X_1 \cdots X_p)}{\Pr(Y > y_j/X_1 \cdots X_p)}\right] = \alpha_j - \Big([\beta_1 X_1 + \beta_2 X_2 + \cdots \beta_p X_p] + \tau_j[\gamma_1 T_1 + \gamma_2 T_2 + \cdots \gamma_q T_q]\Big) \qquad j = 1, 2, \ldots, c-1 \tag{3}$$

The τ_j are fixed scalars that take the form of constraints placed on the parameters. Note that for a given covariate X_m, γ_m does not depend on the cut-off points, but is multiplied by the τ_j for the jth logit. As $\tau_1 = 0$ the first logit is always equal to $\alpha_1 - \beta_m X_m$.

The choice of constraints can be decided upon in several ways; ideally they should be determined using pilot data or one can choose some values which are based on some *a priori* knowledge of the way the odds ratios are likely to behave. However, this is not always possible and some authors[1] have examined the odds ratios obtained from the unconstrained model to determine a set of constraints for the constrained model. This is problematic as one is using the data to estimate the constraints, but may be the only way to obtain the required constants. Regardless of the way the $\{\tau_j s\}$ are obtained, the crucial point is that the same set of constraints has to be assigned to each covariate.

In either model (2) or (3), if we assume that the relationship of the response categories and X_m is represented by non-proportional odds ratios, then the constant component β_m of the log odds ratios and the γ_{jm} (or γ_m) are obtained by fixing and conditioning on all the remaining covariates in the model. $\beta_{jm} = \beta_m + \gamma_{jm}$ refers to the log odds for the jth cut-off point based on the unconstrained model. Similarly $\beta^*_{jm} = \beta^*_m + \tau_j \gamma^*_m$ refers to the log odds based on the constrained partial proportional odds model. The log odds ratios are obtained by subtracting the logits for the appropriate levels (or rows) of the covariate. In the case where X_m is a continuous or ordinal covariate, and there is unit spacing, then for a fixed cut-off point there is a constant log odds ratio when comparing all pairs of adjacent rows, and there are in total $c-1$ log odds ratios. In the case where X_l is a nominal variable, then, given all the other covariates, the log odds ratios do not only vary by cut-off point, but also vary for the two rows compared, resulting in $(c-1)(r-1)$ log odds ratios. Where a covariate of interest has proportional odds, the $\{\beta_k s\}(k = 1, \ldots, p)$ are interpreted in exactly the same way as those in model (1).

The partial proportional odds models can be fitted in SAS using the PROC CATMOD procedure and the *clogit* link function. The program code differs for each model in the way the design matrix is specified.[13] One can assess whether the proportional odds model is as good a fit as the unconstrained partial proportional odds model by comparing the -2log-likelihoods for the two models. Unfortunately SAS uses weighted least squares estimation for the analysis, and therefore no values for the log-likelihoods are readily available. Thus comparison of the models is made using

contrast statements. In the unconstrained model, for a given parameter where different log odds are fitted over the cut-off points the null hypothesis $H_0 : \gamma_{1k} = \gamma_{2k} = \cdots \gamma_{(c-1)k} = 0$ is incorporated into the contrast statement, and this assesses whether the proportional odds model is as good a fit as the unconstrained partial proportional odds model. Likewise, for a covariate where a trend is apparent in the beta parameters and a set of constraints are considered, the test of whether a model using the constraints is as good a fit as a model using the individually estimated log odds can also be obtained using the contrast statements. This test is set up in the unconstrained partial proportional odds model and for a given covariate one assesses the null hypothesis $H_0 : \gamma_{1k} = \tau_1\gamma_k; \gamma_{2k} = \tau_2\gamma_k; \cdots \gamma_{(c-1)k} = \tau_{(c-1)}\gamma_k$.

2.3 Stereotype model

Consider a quality of life scale that assesses 'pain' with respect to some treatment, and assume that the response is recorded on an ordinal scale as 'none', 'mild', 'moderate' or 'severe'. Although the categories are scaled on a single dimension, they are not a discrete version of some continuous variable. A model which assesses the ordinality of the response by looking at the ordering of log odds ratios of the categories is the *stereotype model*.[10] One of the features of this model is that ordering of the response categories with respect to a set of covariates is no longer an assumption but becomes part of a more general model.

The stereotype model is based on the *polytomous regression model*.[1,16] This model is an extension of the logistic regression model and is designed to analyse nominal scales where there are several categories. The logits are formed for this model by comparing each response category with a baseline one, the choice of which is arbitrary and for the analysis presented here, is the first category. Thus the log odds ratio can be represented by a linear model of the form:

$$\log\left[\frac{\Pr(Y = y_j/X_1 \cdots X_p)}{\Pr(Y = y_1/X_1 \cdots X_p)}\right] = \alpha_j + \beta_{j1}X_1 + \beta_{j2}X_2 + \cdots + \beta_{jp}X_p \qquad j = 2, \ldots, c \quad (4)$$

From this model it is clear that the ordinal nature is not accounted for. The ordinality is built into this model by imposing a structure on the log odds $\beta_{jk}(j = 2, \ldots, c; k = 1, \ldots, p)$ such that

$$\beta_{jk} = \phi_j\beta_k \qquad \begin{matrix} j = 2, \ldots, c \\ k = 1, \ldots, p \end{matrix} \quad (5)$$

(note: $\phi_1 \equiv 0$ since $\beta_{1k} = 0$).

This results in the stereotype model that takes the form:

$$\log\left[\frac{\Pr(Y = y_j/X_1 \cdots X_p)}{\Pr(Y = y_1/X_1 \cdots X_p)}\right] = \alpha_j + \phi_j\left(\beta_1X_1 + \beta_2X_2 + \cdots + \beta_pX_p\right) \qquad j = 2, \ldots, c \quad (6)$$

Thus, it can be seen that the stereotype model determines a set of weights, the $\{\phi_j s\}$ for the dependent variable and a single parameter β_k for each independent variable. The weights are decided upon for the response variable and are directly related to the effect of the covariates. Thus, when the odds ratios form an increasing trend, the weights can be constrained to be ordered such that

$$0 = \phi_1 \subseteq \phi_2 \subseteq \cdots \subseteq \phi_c \quad (7)$$

Here we can say that the effect of the covariates upon the first odds ratio is less than their effect upon the second and so on, and that, provided constraint (7) holds, model (6) is an ordered regression model. The model fitted does not necessarily require the $\{\phi_j s\}$ to be ordered; whether there is ordering or not is purely determined by the empirical evidence provided by the data.

The weights can be determined in a number of ways. Some authors[6] suggest that they be decided upon *a priori*, either estimated from some pilot data or a suitable set of values be chosen (fixed scores). With such predetermined set of weights, the stereotype model can be fitted in SAS using PROC CATMOD, and the stereotype model remains of a generalised linear form. Hendrickx,[5] however, has designed macros (suitable for use in SAS and in Stata), that fit the stereotype model and estimate the $\{\phi_j s\}$ as a set of parameters in the model (estimated scores). In this case the stereotype model is intrinsically non-linear, but is easily fitted by performing a series of generalized-linear model fits in which the β_k and ϕ_j parameters are alternatively held fixed while the other is estimated. The model produces $(c - 1)$ standard multinomial intercept parameters for the y-response, $(c - 2)$ independent $\{\phi_j s\}$ and a single beta parameter for each independent variable. Although the stereotype model is more flexible than the proportional odds model, it is less parsimonious with the extra weighting parameters. In the case where the weights are estimated, the β_k and ϕ_j parameters are conditional on the estimates, and thus the estimates of the standards errors of these β_k parameters, which assume the $\{\phi_j s\}$ are known, are invalid. Likewise any inference based on the standard errors or the likelihood-based tests is also not correct, and instead we used bootstrap techniques (random sampling with replacement from the original data) to obtain the correct standard errors and tests. Thus using 100 bootstrap samples we refitted the stereotype model to each bootstrap sample and estimated the $\{\beta_k s\}$ and $\{\phi_j s\}$. Using these estimates, six joint distributions for the $\{\beta_k\phi_j\}(j = 1, 2, 3; k = 1, 2)$ were obtained. From each distribution, the mean and standard error were calculated to give the estimate of the log odds ratio and its standard error based on each cut-off point and covariate.

The change in the −2log-likelihood for the polytomous model and the stereotype model provides a way of establishing whether a model with weights is as good a fit as a model where no weights are imposed on a set of covariates. The −2log-likelihoods for these models were also obtained using bootstrap techniques. The null hypothesis was based on the fact that the polytomous model was a good fit to the data. The observed change in the polytomous model and the stereotype model was compared by examining its position in the distribution of the 100 changes in the −2log-likelihoods obtained using the bootstrap samples.

For nominal covariate X_l, β_{2l} is the log odds ratio which is based on comparing the second category with the referent (first) one. To obtain the subsequent cut-off point-specific log odds ratios, $\beta_{jl}(j = 3, \ldots, c)$ the parameter β_l is multiplied by the weights $\phi_j(j = 3, \ldots, c)$ In the case where X_l is a continuous or ordinal and there are r rows with unit spacing, then for a fixed cut-off point j, the log odds ratio is constant when comparing each consecutive pair of rows. Also for a given pair of rows there are a total of $(c - 1)$ log odds ratios. In the case where X_l is nominal, the log odds ratio will vary depending on the two rows compared, as well as over the cut-off points.

3 The data

3.1 The survey

Data (version 4.1) were provided by the Medical Research Council Cognitive Function and Ageing study (MRC CFAS).[2] MRC CFAS commenced as a longitudinal, two-wave (prevalence/incidence), two-stage (screening/assessment) epidemiological study of dementia conducted in six centres throughout England and Wales (urban Liverpool, Newcastle, Nottingham and Oxford, and rural Gwynedd and Cambridgeshire). As the study design was slightly different for the Liverpool centre,[2,17] this centre was omitted in the analysis. At the first visit all respondents were screened with a basic interview covering socio-demographic details, activities of daily living, physical health measures cognitive function and medication. Subsequently further interviews were carried out (annual follow-up visits and the incidence screen and assessment visits), but these are not detailed here as the analysis only used the first visit data.

At each centre random samples were selected of sufficient size to yield 2500 interviews from individuals aged 65 years and over, with equal numbers in the age groups 65–74 years old and 75 years old and above. The total sample available at baseline was 20 234 for the five centres and there were 17 608 respondents identified as being eligible. Of these 13 006 were interviewed at the initial visit and were regarded as the achieved sample. An outcome that measured the health status of an individual using an ordinal scale was selected for the purpose of the analysis. This outcome was in a form of a question and was asked by an interviewer: 'Would you say that for someone of your age, your own health in general is: 'Excellent', 'Good', 'Fair', 'Poor', 'No answer', 'Not asked' and 'missing'?'. Two categorical covariates 'Have you ever suffered from a heart attack? 'Yes', 'No', 'No answer', 'Not asked' and 'missing'' and 'Do you smoke? 'Yes', 'No', 'No answer', 'Not asked' and 'missing'' were chosen as the independent variables to be used in the models, as these provided a good example of discrimination between the different ordinal regression models outlined.

3.2 Response rates

The number of missing observations for the outcome response was 309 (2.4%), the number of observations where no answer was provided was 61 (0.5%) and the number of respondents who were not asked the health status question was 14 (0.1%). These response categories were ignored as they could not be incorporated into the ordinal scale and could not be analysed using the ordinal regression models. For the 'heart attack' question, the number of subjects who had missing observations was 340 (2.6%),

Table 1 Frequency table for the data from MRC Cognitive and Function Ageing Study (MRC CFAS)

Do you smoke?	Have you had a heart attack?	Rating of health status				
		Excellent	Good	Fair	Poor	Total
Yes	Yes	27 (0.11)	76 (0.31)	101 (0.42)	39 (0.16)	243
Yes	No	402 (0.19)	1050 (0.50)	522 (0.25)	145 (0.07)	2119
No	Yes	83 (0.08)	406 (0.39)	442 (0.42)	114 (0.11)	1045
No	No	1959 (0.21)	4521 (0.50)	2243 (0.25)	405 (0.04)	9128

Parentheses reference the proportions based on the marginal totals.

six (0.05%) respondents did not provide an answer and a further two (0.02%) subjects did not answer the question. For the 'smoke' covariate, there were 392 (3.0%) respondents who had missing data, and four (0.03%) respondents were not asked the question. Missing observations were eliminated from the analysis. Similarly the number of respondents who had 'no answer' or were 'not asked' the question were few and these too were eliminated from the analysis. Thus, although approximately 13 000 elderly people were presented in the 'achieved' sample, complete observations on the response and the covariates were available on 12 535 subjects. The data for the y-response (health status) were cross-tabulated with the two covariates of interest to form four sub-populations and are shown in Table 1.

From Table 1, the majority of patients rate their health as 'good' or 'fair' irrespective of whether or not they smoke and whether or not they have had a heart attack. Regardless of whether a respondent smokes or not, there is a tendency for those who have had a heart attack to rate their health lower than those who have not had a heart attack. Regardless of whether a respondent has had a heart attack or not, provided he/she is a smoker there is a greater chance of rating his/her health as 'poor'.

Using the Cochran–Mantel–Haenzel (row mean score) statistic a significant association was found between the general health of the respondent and whether he/she has had a heart attack (after controlling for whether he/she smokes or not—$Q_{SMH} = 190.767$ on 1 d.f.; $P = 0.001$). Likewise there was evidence of a notable association between the health status and whether or not a respondent smokes (after having accounted for the fact that a respondent may or may not have had a heart attack—$Q_{SMH} = 4.212$ on 1 d.f.; $P = 0.04$). Furthermore, there was no evidence that the two covariates of interest were associated ($\chi_1^2 = 0.001$; $P = 0.982$). The main drawback of this method, however, is that it is only capable of displaying the general association. No estimates for the general or cut-off point-specific odds ratios are readily available. Ordinal regression models are a superior way of assessing the relationship between the ordered response and a set of covariates.

4 Fitting the models

4.1 Preliminary analysis and the proportional odds model

Before fitting the proportional odds model, one should check the assumption of proportionality, and so the individual cut-off point-specific cumulative odds ratios were

Table 2 Different slopes models (single cumulative model and three separate logistic regression models)

Variable	Cut-off points					
	(Good, fair, poor) vs excellent		(Fair, poor) vs (excellent, good)		Poor vs (excellent, good, fair)	
	ln(O.R.)	s.e. ln(O.R.)	ln(O.R.)	s.e. ln(O.R.)	ln(O.R.)	s.e. ln(O.R.)
Suffered from a heart attack (yes/no)?	1.0208	0.1024	1.0374	0.0596	0.9656	0.0966
	(1.0459[a]	0.1024)	(1.0348[a]	0.0596)	(0.9664[a]	0.0965)
Do you smoke (yes/no)?	0.1222	0.0590	0.1304	0.0488	0.4592	0.0894
	(0.1246[a]	0.0591)	(0.1242[a]	0.049)	(0.4561[a]	0.0895)

[a]Refers to the analysis carried out by the logistic regression model.

$$\log\left[\frac{\Pr(Y \le y_j/X_1.X_2)}{\Pr(Y > y_j/X_1.X_2)}\right] = \alpha_j - \beta_{j1}(\text{heart attack}) - \beta_{j2}(\text{smoke}); \qquad j = 1, 2, 3.$$

computed. Table 2 illustrates the results where different slopes (based on each cut-off point) were fitted using cumulative probabilities in a single model, and the results of three separate logistic regression models (based on the three cut-off points). It is interesting to note that the results are very similar for the single cumulative model and the three logistic regression models. When comparing the results from these models it can be seen that the standard errors of the log odds estimates are almost identical and there is only a slight variation in the estimates.

By observing the adjusted log odds ratios from the single cumulative model in Table 2, we conclude that there is little difference in the probability of lower ratings of health as opposed to higher in those who may or may not have had a heart attack. For the 'smoke' covariate there appears to be much more variation in the odds ratios. This would suggest that the constant odds assumption is unlikely to be satisfied for the proportional odds model and indeed this was the case as suggested by the χ^2-score test statistics in Table 3 (χ^2-score test statistics: overall, ${\chi_4}^2 = 16.2192$, $P = 0.0027$; 'Heart attack', ${\chi_2}^2 = 0.5804$, $P = 0.7481$; 'smoke', ${\chi_2}^2 = 16.0482$, $P = 0.0003$). Furthermore the proportional odds model was found to be a poor fitting model (chi-squared residual test, ${\chi_7}^2 = 22.76$; P-value = 0.0019).

4.2 Unconstrained partial proportional odds model

As the proportional odds assumption does not hold for one of the two covariates in the model, a partial proportional odds model, initially using no constraints, was fitted and the results of this are illustrated in Tables 4 and 5. For the subjects who may have had a heart attack (as opposed to not having had one), a constant adjusted log odds ratio could be assumed across the health status categories. However, the estimated adjusted log odds for those who were/were not smokers varied by cut-off point, and in relation to model (2) are denoted by β_2 (corresponding to the first cut-off point), $\beta_2 + \gamma_{22}$ (corresponding to the second cut-off point) and $\beta_2 + \gamma_{32}$(corresponding to the third cut-off point). Table 4 displays the weighted least-squares estimates and these have been used to obtain the estimates of the log odds ratios together with their standard errors (Table 5). This model accommodates the proportional odds present in

Table 3 Model fitting the proportional odds model

Variable	χ^2-score test statistic	d.f.	P-value	ln(O.R.)	s.e. ln(O.R.)	ln(O.R.)	s.e. ln(O.R.)
Heart attack (yes/no)	0.5804	2	0.7481	1.0241	(0.0553)		
Smoke (yes/no)	16.0482	2	0.0003			0.1460	(0.0426)
Heart attack x Smoke	16.2192	4	0.0027	1.0222	(0.0554)	0.1542	(0.0428)

$$\log\left[\frac{\Pr(Y \le y_j/X_1.X_2)}{\Pr(Y > y_j/X_1.X_2)}\right] = \alpha_j - \beta_1(\text{heart attack}) - \beta_2(\text{smoke}); \qquad j = 1, 2, 3.$$

Table 4 Unconstrained partial proportional odds model: weighted least squares parameter estimates

Parameters	Estimate ± s.e.	Adjusted heart attack covariate Estimate (s.e.)		Adjusted smoke covariate Estimate (s.e.)		Wald's test statistic	
						$\gamma_1 = \gamma_2 = \gamma_3 = 0$	$\gamma_3 = 40\gamma_2$
α_1	1.8752 ± 0.0379						
α_2	−0.3199 ± 0.0323						
α_3	−2.3404 ± 0.0481						
β		0.5115	(0.0277)	0.0609	(0.0295)		
γ_1				0		${\chi_1}^2 = 5.69$	
γ_2				0.00411	(0.0314)	$P = 0.02$	${\chi_1}^2 = 0.00$
γ_3				0.1691	(0.0503)		$P = 0.9970$

$$\log\left[\frac{\Pr(Y \le y_j/X_1.X_2)}{\Pr(Y > y_j/X_1.X_2)}\right] = \alpha_j - \beta_1(\text{heart attack}) - [\beta_2(\text{smoke}) + \gamma_{12}(\text{smoke: health status} > \text{excellent}) + \gamma_{22}(\text{smoke: health status} > \text{good}) + \gamma_{32}(\text{smoke: health status} > \text{fair})]; \qquad j = 1, 2, 3.$$

Table 5 Log odds ratios for unconstrained partial proportional odds model

Variable			(Good, fair, poor) vs excellent		(Fair, poor) vs (excellent, good)		Poor vs (excellent, good, fair)	
	ln(O.R.)	s.e. ln(O.R.)	ln(O.R.)	s.e. ln(O.R.)	ln(O.R.)	s.e. ln(O.R.)	ln(O.R.)	s.e. ln(O.R.)
	Constant component of log odds ratio across cut-off points		*Increment at cut-off points*					
Suffered from a heart attack (yes/no)?	1.023	0.0554	—	—	—	—	—	—
Do you smoke (yes/no)?	0.1218	0.059	0		0.00822	(0.0628)	0.3382	(0.1006)
			Log odds ratios at cut-off points					
Do you smoke (yes/no)?	—	—	0.1218	(0.059)	0.1300	(0.0991)	0.4600	(0.1281)

the 'heart attack/no heart attack' covariate, and for the non-proportional odds present in the 'smoke/no smoke' covariate. The log odds ratios obtained for this model are very similar to those presented in Table 2. In terms of interpretation, the respondents who have had a heart attack (as opposed to not having had one) are three times as likely to have lower ratings of health as opposed to higher ratings. However, as an underlying continuum directly related to the response categories does not exist for this quality of life scale, the invariance property does not apply. Given a respondent is a smoker (as opposed to not being a smoker), then after adjusting for the fact that he/she may or may not have had a heart attack, the odds of having ('good', 'fair', 'poor') health is 1.1 times that of having ('excellent') health, and this is similar to the odds of having ('fair', 'poor') health versus ('excellent', 'good') health. The adjusted odds of having ('poor') health is 1.6 times that of having ('excellent', 'good', 'fair') health for a smoker (compared to a non-smoker). It is evident that the non-proportionality in the 'smoke' covariate is as a result of the odds ratio obtained at cut-off point 3. The unconstrained partial proportional odds model was found to be a good fitting model to the data (chi-squared test of residuals, $\chi_5^2 = 7.14$, P-value = 0.2102) and was a better fit than the proportional odds model ($H_0, \gamma_{1k} = \gamma_{2k} = \cdots \gamma_{3k} = 0$; $\chi_2^2 = 15.62$; P-value = 0.01). Furthermore, it is a more parsimonious model than the model that allowed for separate slopes for the cut-off points for each covariate (since seven parameters are estimated in the unconstrained partial proportional odds model and nine parameters are estimated in the different slopes model).

4.3 Constrained partial proportional odds model

In fitting the constrained partial proportional odds model, the cut-off point-specific log odds ratios are observed from the unconstrained partial proportional odds model. A monotonic trend is apparent in the log odds ratios across the health status categories in relation to the covariate that assesses whether respondents smoke or not smoke. In order to simplify the interpretation, a constraint can be placed on the parameters (leading to the formation of the constrained partial proportional odds model). Thus whilst a proportional odds is apparent in the variable assessing 'heart attack', the 'smoke' variable has odds ratios which follow an increasing trend over the cut-off points. Based on these, the following constraints were chosen: $\tau_{12} = 0$; $\tau_{22} = 1$; $\tau_{32} = 40$. These formed the following log odds: $\beta_2 + 0 \cdot \gamma_2 T_2$, $\beta_2 + 1 \cdot \gamma_2 T_2$ and $\beta_2 + 40 \cdot \gamma_2 T_2$. The parameter estimates and the log odds ratios are presented in Tables 6 and 7. The interpretation of the parameter estimates for this model is very similar to that for the unconstrained partial proportional odds model. The constrained partial proportional odds model was found to be a good fit model (test of residuals: $\chi_6^2 = 7.16$; P-value = 0.3036). This model was found not to be significantly different from the unconstrained model ($H_0, \gamma_3 = \tau_3\gamma_2$; $\chi_1^2 = 0.00$; P-value = 0.970) and therefore in terms of model parsimony was the preferred model (as only six parameters were estimated, as the constraints are not considered model parameters in this case).

4.4 Polytomous model

Before fitting the stereotype model, we examined the log odds ratios (and their standard errors) from the polytomous model. In both models the referent category was chosen to be 'excellent', and therefore the cut-off point-specific odds ratios are based on

Table 6 Constrained partial proportional odds model: weighted least squares parameter estimates

Parameters	Estimates (s.e.)		Adjusted heart attack covariate		Adjusted Smoke covariate	
			Estimates (s.e.)		Estimates (s.e.)	
α_1	1.8751	(0.0357)				
α_2	−0.3199	(0.0315)				
α_3	−2.3404	(0.0479)				
β			0.5115	(0.0277)	0.0608	(0.0217)
γ (constraint parameter)					0.00423	(0.00107)

$$\log\left[\frac{\Pr(Y \le y_j/X_1.X_2)}{\Pr(Y > y_j/X_1.X_2)}\right] = \alpha_j - \beta_1(\text{heart attack}) - [\beta_2(\text{smoke}) + \tau_1\gamma_2(\text{smoke: health status>excellent}) + \tau_2\gamma_2(\text{smoke: health status>good}) + \tau_3\gamma_2(\text{smoke: health status>fair})]; \quad j = 1, 2, 3.$$

Constraints: $\tau_1 = 0$; $\tau_2 = 1$; $\tau_3 = 40$.

Table 7 Log odds ratios for constrained partial proportional odds model

Variable			(Good, fair, poor) vs excellent		(Fair, poor) vs (excellent, good)		Poor vs (excellent, good, fair)	
	ln(O.R.)	s.e. ln(O.R.)	ln(O.R.)	s.e. ln(O.R.)	ln(O.R.)	s.e. ln(O.R.)	ln(O.R.)	s.e. ln(O.R.)
	Constant component of log odds ratio across cut-off points		*Increment at cut-off points*					
Suffered from a heart attack (yes/no)?	1.023	(0.0554)	—		—	—	—	—
Do you smoke (yes/no)?	0.1216	(0.0434)	0		0.00846	(0.00214)	0.3384	(0.0856)
			Log odds ratios at cut-off points					
Do you smoke (yes/no)?	—	—	0.1216	(0.0434)	0.1300	(0.0435)	0.4600	(0.0455)

'good' versus 'excellent' (cut-off point 1), 'fair' versus 'excellent' (cut-off point 2) and 'poor' versus 'excellent' (cut-off point 3). The weighted least squares estimates and the adjusted odds ratios for both covariates using the polytomous model are displayed in Tables 8 and 9, respectively. Using the test statistic values, we found that the cut-off point-specific adjusted log odds ratios for the 'heart attack' variable are significantly different from one another (Wald's test statistics, $\beta_1 = \beta_2$, $\chi_1^2 = 150.78$, $P < 0.0001$; $\beta_2 = \beta_3$, $\chi_1^2 = 11.68$, $P = 0.0006$). A constant adjusted log odds ratio can be assumed for the 'good' versus 'excellent' and 'fair' versus 'excellent' cut-off points for the smokers versus non-smokers (Wald's test statistics, $\beta_1 = \beta_2$, $\chi_1^2 = 0.12$, $P = 0.73$). However, a different odds ratio has to be assumed for the 'poor' versus 'excellent' categories (Wald's test statistics, $\beta_2 = \beta_3, \chi_1^2 = 19.43$, $P < 0.0001$). Note that this is in contradiction to the conclusions drawn from the proportional odds model and the partial proportional odds models. Thus using the polytomous model, for the respondent who has suffered from a heart attack, the adjusted odds of having 'good' health is

Table 8 Polytomous logistic regression model: weighted least squares parameter estimates

Parameters	Estimate	(s.e.)	Adjusted heart attack covariate		Adjusted smoke covariate	
			Estimate	(s.e.)	Estimates	(s.e.)
α_1	1.1957	(0.0580)				
α_2	0.9110	(0.0587)				
α_3	−0.4108	(0.0724)				
β_1			0.3098	(0.0542)	0.0429	(0.0313)
β_2			0.7197	(0.0541)	0.0524	(0.0329)
β_3			0.8955	(0.0669)	0.2656	(0.0509)

$$\log\left[\frac{\Pr(Y = y_j/X_1.X_2)}{\Pr(Y = y_1/X_1.X_2)}\right] = \alpha_j + \beta_{j1}(\text{heart attack}) + \beta_{j2}(\text{smoke}); \quad j = 2, 3, 4.$$

Table 9 Polytomous model

Variable	Cut-off points						Wald's test statistic	
	Good vs excellent		Fair vs excellent		Poor vs excellent			
	ln(O.R.)	s.e. ln(O.R.)	ln(O.R.)	s.e. ln(O.R.)	ln(O.R.)	s.e. ln(O.R.)	$\beta_1 = \beta_2$	$\beta_2 = \beta_3$
Suffered from a heart attack (yes/no)?	0.6196	(0.1084)	1.4394	(0.1082)	1.791	(0.1338)	$\chi_1^2 = 32.61$ $P = 0.0000$	$\chi_1^2 = 150.78$ $P = 0.0000$
Do you smoke (yes/no)?	0.0858	(0.0626)	0.1048	(0.0698)	0.5312	(0.1018)	$\chi_1^2 = 1.88$ $P = 0.1706$	$\chi_1^2 = 0.12$ $P = 0.7323$

approximately twice that of having 'excellent' health. The adjusted odds of having 'fair' health is approximately four times that of having 'excellent' health and the adjusted odds of having 'poor' health is approximately six times of having 'excellent' health. For the smokers the adjusted odds of having 'good' or 'fair' health is approximately 1.1 times that of having 'excellent' health, and the adjusted odds of having 'poor' health is approximately 1.7 times of having 'excellent' health. The polytomous model was found to be a good fit model (test of residuals, $\chi_3^2 = 6.37$, P-value = 0.0949), although it lacks parsimony.

The difference in the results produced by the proportional odds models and the polytomous model can be explained by assessing the proportions in Table 1. It is evident that the smoker *vs* non-smoker contrast within each of the 'heart attack' sub-populations is very similar across the health status categories. Thus for the 'smoke' covariate, the difference in the two types of models is due to the way the logits are formed. In fact, the log odds ratios are quite similar for this covariate given the two cumulative models and polytomous model. When assessing the proportion of those who have had a heart attack *vs* not having had one, within each of the levels of the smoke covariate, there is a notable difference. The ratio of those who have had a heart attack compared to those who have not had a heart attack increases over the health status categories for both levels of the 'smoke' covariate and this is manifested in the

polytomous model. The effect of accumulating the probabilities over the cut-off points removes the increase in the odds ratios to reveal proportional odds.

4.5 Stereotype model

As stated earlier, the polytomous model can be simplified to become the stereotype model. The weights were estimated as parameters using maximum likelihood estimation in model (6), together with the α_j and β_k parameters are presented in Table 10. Differences between the ϕ_j scale values indicates how the log odds of one health status category versus another is affected by having/not having a heart attack and whether respondents smoke or not. The impact of the independent variables on a log odds between adjacent categories is largest for 'good' versus 'poor' health status where the difference between the scale values is 0.439. The smallest impact is on the 'fair' versus 'poor' health status categories, with a difference of only 0.223. The impact of having a heart attack on the logit of having 'excellent' versus 'good' health status is 0.6147 ($\phi_1\beta_1$), resulting in the log odds ratio. Using the remaining weights and the β_k values, the cut-off point-specific log odds ratios can be obtained in a similar way for both covariates. The odds ratios were found not to be much different to those obtained using the polytomous model (see Table 11). The interpretation of these log odds is also similar to that of the polytomous model. As the weights are ordered in a monotone fashion we can assume that there is an ordering in the y-response with respect to the covariates. The observed −2log-likelihood for the nine-parameter polytomous model was 29329.16 compared to a −2log-likelihood of 29343.73 for the seven-parameter stereotype model. The observed change in the −2log-likelihood values of these two models was 14.57. Comparing this with its position in the bootstrap distribution of 100 changes in −2log-likelihoods of the two models, it was evident that the constrained model was not as good a fit as the polytomous model. Assessing the distribution of the change in the −2log-likelihood values, it was found that 45.5% of the bootstrap sample values lay below the observed value and 54.6% lay above. This implied that the P-value was approximately 0.5, indicating that the null hypothesis (which was based on the fact that the polytomous model was a good fit to the data) could not be rejected. Thus the stereotype model was found to be a poor fit compared to the polytomous model.

Table 10 Parameter estimates using the bootstrap techniques: stereotype model

Parameters	Adjusted heart attack covariate		Adjusted smoke covariate	
	Estimate	(s.e.)	Estimate	(s.e.)
β	1.8468	(0.1346)	0.2703	(0.0805)
ϕ_1	0		0	
ϕ_2	0.3407	(0.0426)	0.3407	(0.0426)
ϕ_3	0.7770	(0.0534)	0.7770	(0.534)
ϕ_4	1		1	

$$\log\left[\frac{\Pr(Y = y_j/X_1.X_2)}{\Pr(Y = y_1/X_1.X_2)}\right] = \alpha_j + \phi_j(\beta_1(\text{heart attack}) + \beta_2(\text{smoke})); \quad j = 2, 3, 4.$$

Table 11 Stereotype model

Variable	Cut-off points					
	Good vs excellent		Fair vs excellent		Poor vs excellent	
	ln(O.R.)	s.e. ln(O.R.)	ln(O.R.)	s.e. ln(O.R.)	ln(O.R.)	s.e. ln(O.R.)
Suffered from a heart attack (yes/no)?	0.6305	(0.0980)	1.4328	(0.1189)	1.8468	(0.1346)
Do you smoke (yes/no)?	0.0920	(0.0295)	0.2082	(0.0575)	0.2703	(0.0805)

5 Discussion

An alternative ordinal regression model, which was not considered in the analysis is the *Fienberg's continuation ratio model.*[18] This model is usually relevant when an ordinal quality of life scale may be thought of as a progression through various stages, so that people start with 'excellent' and deteriorate to 'poor' and are unlikely to reverse this trend. Such data usually resemble failure-time data or outcomes measuring threshold points. The data in this paper were not of this type and therefore the continuation ratio model was considered as being irrelevant.

Irrespective of the modelling techniques, the response variable in a quality of life scale can essentially arise in one of two ways: (a) where there are clearly ordered categories for which there is a single underlying latent variable; or (b) where the categories are discrete and for which ordering may depend on covariate information.

Having established how the data are generated, then one is in a position to decide which model will be most appropriate in terms of analysing the data. Given the *y*-response is a grouped continuous response variable, the proportional odds and partial proportional odds models are often the most applicable due to the assumptions these models make. In the case of the data presented, the proportional odds model was found to be a poor fitting model and one could have used other forms of strategies. These would include using a different link function in the model, eg the log–log function would produce a response function that was non-symmetric, or alternatively one could include additional terms in the model, such as the interaction term (although in this case the interaction was found to be non-significant). Generally, however, as the covariates increase in a proportional odds model, the lack of fit increases but is compensated by the parsimony of the model. The unconstrained partial proportional odds model is a better fitting model than the proportional odds model, although the parameters increase at a drastic rate as the number of covariates and number of *y*-response categories increase. The constrained partial proportional odds model would probably be the most ideal model, given a set of covariates and a *k*-ordered group continuous response variable. However, obtaining the constraints is somewhat problematic, especially if there are a large number of covariates with non-proportional odds. Presently there is no method available for estimating the constraints, and one can only use fixed constraints that have been determined prior to fitting the model.

When the ordered categories in the response variable are of a discrete nature, and there is no directly related underlying continuum, the interpretation of the parameters in the proportional odds and partial proportional odds models becomes difficult. Ideally one would fit the polytomous model when presented with such response data. However, although this produces a good fitting model, it is at the cost of estimating a large number of parameters. As the number of covariates increases in the model, the stereotype model becomes more parsimonious and the estimation of the weights is not problematic, regardless of the number of covariates presented.

The stereotype model would ideally be the most favourable of all these models given that one is presented with an assessed response. However, results from our data demonstrate that the stereotype model was not as good a fit as the polytomous model and the fit of the stereotype model was further illustrated using the Akaike Information Criterion[4,19] (AIC = $-2 \times$ log-likelihood $+ 2 \times p$, where p is the number of parameters in the model). This criterion adjusts the -2log-likelihood statistic for the number of terms in the model and the number of observations used. It is clear that there is not much difference in fit for the proportional odds or stereotype model (AIC for proportional odds model = 29353.837; AIC for the stereotype model = 29353.73 and the AIC for the polytomous model = 29339.163).

Given the results above, despite the number of parameters estimated, the polytomous model is taken to be the most appropriate model that summarized the data. It is found to fit the data well and, more importantly, it allows for the processes that generate the data. Thus using this model we can conclude that smoking has a greater impact on health status than does having had a heart attack. The odds of having a lower rating of health increase dramatically if one smokes (adjusting for whether they have had a heart attack or not) compared to a non-smoker.

It is now computationally possible to fit most, if not all, the different ordinal regression models using routine statistical packages. There is therefore no reason why one should not account for as much information as possible regarding the data. In this paper, we have attempted to illustrate that the way the data have been generated can be accommodated in a given ordinal regression model and this provides more accurate and refined results.

Acknowledgement

MRC Cognitive Function and Aging study is supported by major awards from the Medical Research Council and the Department of Health. The MRC CFAS website can be obtained on: www.mrc-bsu.cam.ac.uk/cfas/

References

1 Ananth C, Kleinbaum D. Regression models for ordinal responses: a review of methods and applications. *International Journal of Epidemiology* 1997; 26: 1323–33.

2 MRC CFAS Study. The MRC Multicentre Study of Cognitive Function and Ageing: a EURODEM incidence study in progress. *Neuroepidemiology* 1992; 11: 37–43.

3 Scott S, Goldberg M, Mayo N. Statistical assessment of ordinal outcomes in comparative studies. *Journal of Clinical Epidemiology* 1997; 50: 45–55.

4 *SAS/STAT User's Guide, Version 6*, 4th edn, Vol 2. Cary, NC: SAS Institute.
5 Hendrickx J. Special restrictions in multinomial logistic regression. *Stata Technical Bulletin* 2000; STB-56: 18–26.
6 Greenland S. Alternative models for ordinal logistic regression. *Statistics in Medicine* 1994; 13: 1665–77.
7 McCullagh P. Regression models for ordinal data (with discussion). *Journal of Royal Statistical Society Series B* 1980; 42: 109–42.
8 Engel J. Polytomous logistic regression. *Stat Neerlandica* 1988; 42: 233–52.
9 Ware J, Snow K, Kosinski M, Gandek B. *SF-36 health survey manual and interpretation guide*. Boston, MA: New England Medical Centre, the Health Institute, 1993.
10 Anderson JA. Regression and ordered categorical variables (with discussion). *Journal of Royal Statistical Society Series B* 1984; 46: 1–30.
11 Mantel N. Chi-squared tests with one degree of freedom: extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association* 1963; 58: 690–700.
12 *SAS Procedures Guide, Version 6*, 4th edn, Vol 2. Cary, NC: SAS Institute.
13 Stokes M, Davis C, Koch G. *Categorical data analysis using the SAS system*. Cary, NC: SAS Institute, 1995.
14 Peterson B, Harrell F. Partial proportional odds model for ordinal response variables. *Applied Statistics* 1990; 39: 205–17.
15 Peterson B, Harrell F. Partial proportional odds models and the LOGIST procedure. In *SAS Users Group International, Thirteenth Annual Conference*, Orlando FL. Cary, NC: *SAS* Institute, 1988, pp 952–56.
16 Agresti A. Tutorial on modelling ordered categorical response data. *Psychological Bulletin*, 1989; 105: 290–301.
17 MRC CFAS. Cognitive function and dementia in six areas of England and Wales: the distribution of MMSE and prevalence of GMS organicity level in the MRC CFA Study. *Psychological Medicine* 1998; 28: 319–35.
18 Everitt BS, Dunn G. *Advanced Methods of Data Exploration and Modelling*. London: Heinemann Educational, 1983: 164–75.
19 *STATA manual, Version 7*. College Station, TX: Statacorp, 2001.

Health Services & Outcomes Research Methodology 4: 35–47, 2003

Sample Sizes for the SF-6D Preference Based Measure of Health from the SF-36: A Comparison of Two Methods

STEPHEN J. WALTERS* s.j.walters@sheffield.ac.uk
JOHN E. BRAZIER
Sheffield Health Economics Group, School of Health and Related Research, University of Sheffield, Regent Court, 30 Regent St., Sheffield, S1 4DA

Received September 10, 2002; Revised May 5, 2003; Accepted June 3, 2003

Abstract. *Background*: Health Related Quality of Life (HRQoL) measures are becoming more frequently used in clinical trials. Investigators are now asking statisticians for advice on how to plan and analyse studies using HRQoL measures, which includes questions on sample size. Sample size requirements are critically dependent on the aims of the study, the outcome measure and its summary measure, the effect size and the method of calculating the test statistic. The SF-6D is a new single summary preference-based measure of health derived from the SF-36 suitable for use in clinical trials.

Objectives: To describe and compare two methods of calculating sample sizes when using the SF-6D in comparative clinical trials and to give pragmatic guidance to researchers on what method to use.

Methods: We describe two main methods of sample size estimation. The parametric (t-test) method assumes that the SF-6D data is continuous and Normally distributed and that the effect size is the difference between two means. The non-parametric (Mann-Whitney or MW) method makes no distributional assumptions about the data and the effect size is defined in terms of the probability that an observation drawn at random from population Y would exceed an observation drawn at random from population X. We used bootstrap computer simulation to compare the power of the two methods for detecting a shift in location.

Results: Computer simulation suggested that if the distribution of the SF-6D is reasonably symmetric then the t-test appears to be more powerful than the MW test at detecting differences in means. If the distribution of the SF-6D is skewed then the MW test appears to be more powerful at detecting a location shift (difference in means) than the t-test. However the differences in power (between the t and MW tests) are small and decrease as the sample size increases.

Conclusions: Computer simulation has suggested that parametric methods work reasonably well. Therefore pragmatically we would recommend that parametric methods be used for sample size calculation and analysis when using the SF-6D.

Keywords: sample size, health-related quality of life, SF-36, preference-based measures of health, bootstrap simulation

File Mismatch
Disk Followed

1. Introduction

Health Related Quality of Life (HRQoL) measures are becoming more frequently used in clinical trials and health services research, both as primary and secondary endpoints. Investigators are now asking statisticians for advice on how to plan and analyse studies using

*Author to whom correspondence should be addressed.

P1: GIM
Health Services & Outcomes Research Methodology KL2076-04 July 23, 2003 20:42

HRQoL measures, which includes questions on sample size. Sample size calculations are now mandatory for many research protocols and are required to justify the size of clinical trials in papers before they will be accepted by journals [1].

Thus, when an investigator is designing a study to compare the outcomes of an intervention, an essential step is the calculation of sample sizes that will allow a reasonable chance (power) of detecting a predetermined difference (effect size) in the outcome variable, at a given level of statistical significance. Sample size is critically dependent on the purpose of the study, the outcome measure and how it is summarised, the proposed effect size and the method of calculating the test statistic.

Whatever type of study design is used the problem of sample size must be faced. Sometimes we may wish to show that a new treatment is clinically equivalent in efficacy to the standard treatment. Machin et al. [15] describe statistical methods for calculating the appropriate sample sizes for demonstrating equivalence between two treatments. For simplicity in this paper we will assume that we are interested in comparing the effectiveness (or superiority) of a new treatment compared to a standard treatment.

HRQoL outcome data may not meet the distributional requirements (usually that the data have a Normal distribution) of parametric methods of analysis. Therefore non-parametric methods are most often used to analyse HRQoL data. The main aim of this paper is to describe and compare two methods of sample size estimation (parametric and non-parametric) when using the SF-6D as an outcome in comparative clinical trials and to provide pragmatic guidance to researchers on what method to use.

The remainder of this paper is structured into the following sections. Section 2 briefly describes the SF-36 measure and the single preference weighted SF-6D index. Section 3 summarises the methods and the sample size formulae. The next Section 4 compares the different methods of sample size calculation using computer simulation. The final Sections 5 and 6 talk about the choice of sample size method with the SF-6D and conclusions.

2. SF-36 health survey and the SF-6D health state classification

The SF-36 originated in the USA [27], but it has been anglicised for use in the UK [2]. It contains 36 questions measuring health across eight dimensions—physical functioning, role limitation because of physical health, social functioning, vitality, bodily pain, mental health, role limitation because of emotional problems and general health. Responses to each question within a dimension are combined to generate a score from 0 to 100, where 100 indicates "good health." Two further summary components, the Mental Component Summary (MCS) and Physical Component Summary (PCS) have also been derived from the eight dimensions using factor analysis [26]. The PCS and MCS scales of the SF-36 are standardised such that a mean score of 50 (standard deviation 10) reflects the mean score of a standard population. Thus, the SF-36 generates a profile of HRQoL outcomes (on up to 10 dimensions), which makes statistical analysis and interpretation difficult [11].

The simple scoring algorithm for the eight dimensions assumes equal intervals between the response choices, and that all items are of equal importance, which may not be appropriate. Brazier et al [3] have derived a preference-based or utility measure of health from the SF-36, called the SF-6D, which reduces all the outcomes to a single summary

measure for use in clinical trials and economic evaluations. All responders to the original SF-36 questionnaire can be assigned SF-6D score provided the 11 items used in the six dimensions of the SF-6D have been completed. The SF-6D preference-based measure can be regarded as a continuous outcome scored on a 0.29 to 1.00 scale, with 1.00 indicating "full health."

In this paper we will assume the SF-6D is being used as the primary HRQoL endpoint in a two group comparative clinical study, at a single time point, to assess the superiority (not equivalence) of a new treatment over a control treatment.

3. Which sample size formulae?

In principle, there are no major differences in planning a study using the SF-6D as an outcome to those using conventional clinical outcomes. Pocock outlines five key questions regarding sample size [18]:

1. What is the main purpose of the trial?
2. What is the principal measure of patient outcome?
3. How will the data be analysed to detect a treatment difference?
4. What type of results does one anticipate with standard treatment?
5. How small a treatment difference is it important to detect and with what degree of certainty?

Thus, after deciding on the purpose of the study and the principle outcome measure, the investigator must decide how the data are to be analysed to detect a treatment difference. We must also identify the smallest treatment difference that is of such clinical value that it would be very undesirable to fail to detect it. Given answers to all of the five questions above, we can then calculate a sample size.

Machin et al. [15] outline the ways of calculating sample sizes in two group studies for binary, ordered categorical and continuous outcomes. We describe two methods of sample-size estimation when using the SF-6D in the comparative clinical trials of two health technologies (Table 1). The first method (Method 1) assumes the SF-6D is continuous and Normally distributed and the second method (Method 2) makes no distributional assumptions about the SF-6D.

Figure 1 shows the overall distribution of the SF-6D in a general population sample aged 16 to 74 years [2]. The SF-6D does not appear to be Normally distributed and appears to be negatively skewed, with more people reporting better health in this general population sample. Conversely, figure 2 shows the distribution of the SF-6D in a group of patients with venous leg-ulcers [25]. The distribution of the SF-6D in this group is more symmetric with patients reporting poorer health than the general population sample.

Method 1: Normally distributed continuous data—comparing two means

Suppose we are planning a two-group study comparing HRQoL (using the SF-6D as the primary outcome) between the groups. We believe that the mean difference in SF-6D scores between the two groups is an appropriate comparative summary measure. Therefore using

Table 1 Effect size and sample size formulae.

	Method 1	Method 2
Assumptions	Normally distributed continuous data	No distributional assumptions
Summary measure	Mean and mean difference	Median
Hypothesis test	Two-independent samples t-test	Mann-Whitney U (also known as the Wilcoxon rank sum test)
Effect size	$\Delta_{Normal} = \frac{\mu_T - \mu_C}{\sigma}$ (Eq (1))	$p_{Noether} = \Pr(Y > X)$ (Eq. (3))
Sample size formulae	$n_{Normal} = \frac{2[z_{1-\alpha/2} + z_{1-\beta}]^2}{\Delta^2_{Normal}}$ (Eq. (2))	$n_{Non\text{-}normal} = \frac{[z_{1-\alpha/2} + z_{1-\beta}]^2}{6(p_{Noether} - 0.5)^2}$ (Eq. (4))

Δ_{Normal} is the standardised effect size index, μ_T and μ_C are the expected group means of outcome variable under the null and alternative hypotheses and σ is the standard deviation of outcome variable (assumed the same under the null and alternative hypotheses).
$p_{Noether}$ is an estimate of the probability that an observation drawn at random from population Y would exceed an observation drawn at random from population X.
$z_{1-\alpha/2}$ and $z_{1-\beta}$ are the appropriate values from the standard Normal distribution for the $100(1-\alpha/2)$ and $100(1-\beta)$ percentiles respectively.
Number of subjects per group n for a two-sided significance level α and power $1-\beta$.

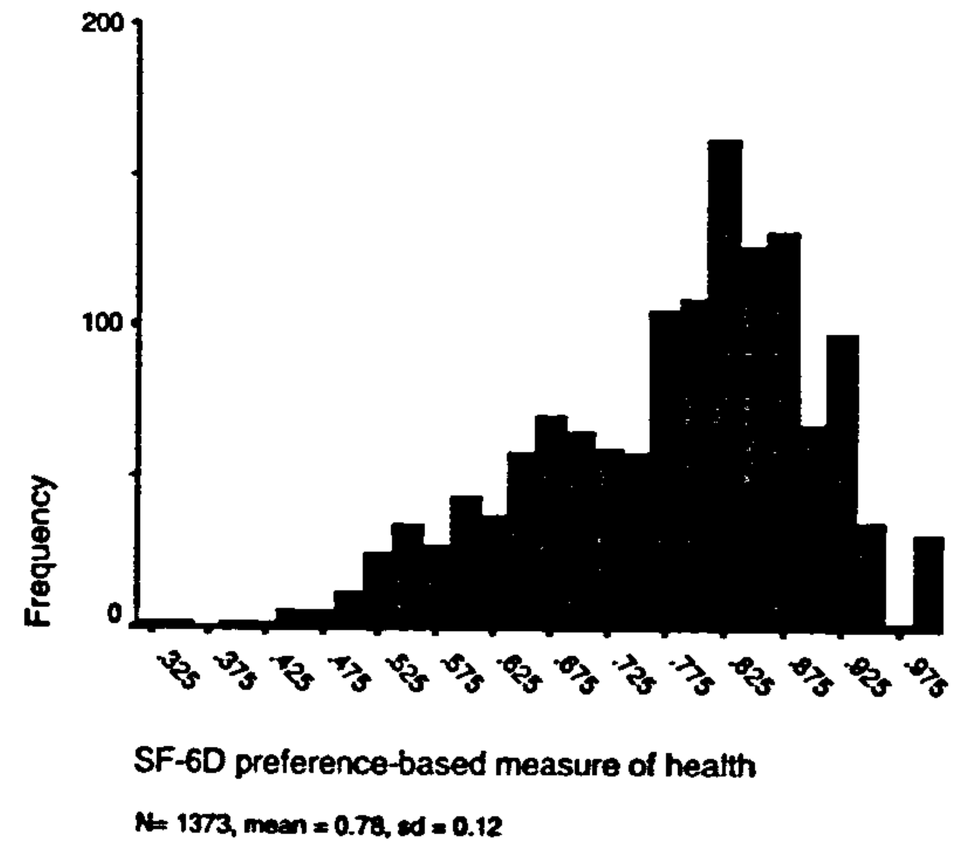


Figure 1 Histogram of the SF-6D in the Sheffield population aged 16–74.

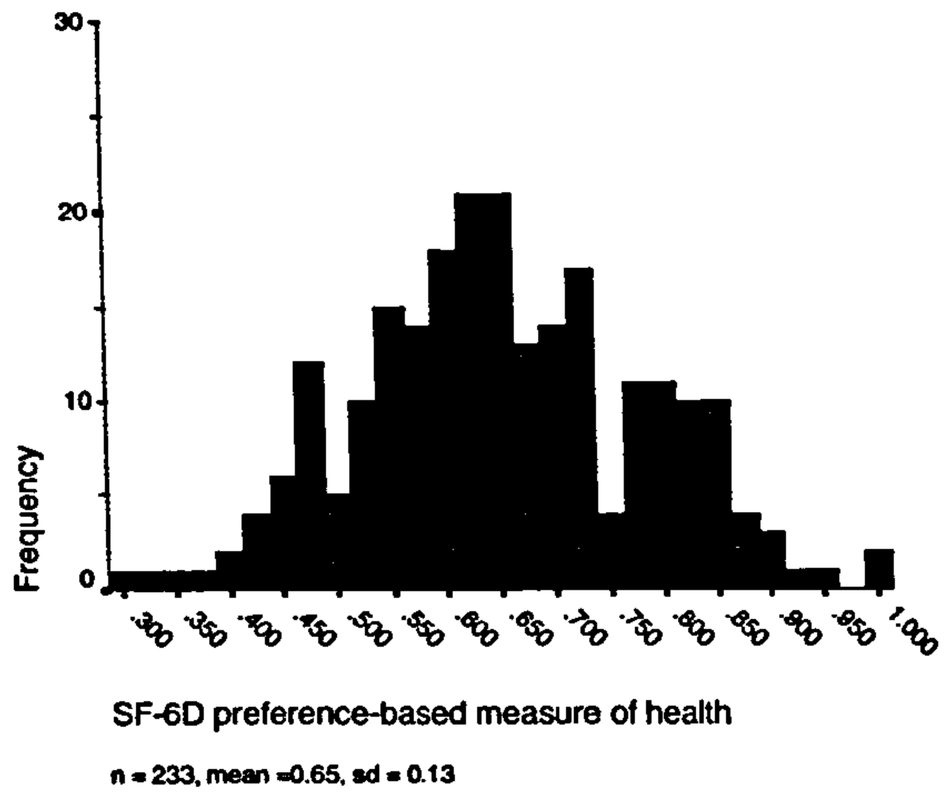


Figure 2. Histogram of the SF-6D in patients with leg-ulcers.

Method 1 and assuming a standard deviation σ of 0.12 and that a mean difference ($\mu_{ET} - \mu_{EC}$) of 0.05 or more points between the two groups is clinically and practically relevant gives a standardised effect size Δ_{Normal} (from Eq. (1)) of 0.417. Using this standardised effect size in Eq. (2) with a two-sided 5% significance level and 80% power gives the estimated number of subjects per group as 93.

Transformations

If the SF-6D outcome data were continuous but had a skewed distribution they may be transformed using a logarithmic transformation. The transformed variable may have a more symmetric distribution that is better approximated by the Normal form. One problem with transforming data is that some preference-based utility measures are scored on 0.0 to 1.0 scales and the natural logarithm of zero does not exist. Unfortunately log-transforming the general population data in figure 2 did not make the distribution of the data more symmetric.

Method 2: No distributional assumptions

If the SF-6D outcome is assumed to be continuous and plausibly not sampled from a Normal distribution then the most popular (not necessarily the most efficient) non-parametric test

for comparing two independent samples is the two-sample *Mann-Whitney U* (also known as the Wilcoxon rank sum test) [13].

Suppose we have two independent random samples $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ and we want to test the hypothesis that the two samples have come from the same population against the alternative that the Y observations tend to be larger than the X observations. As a test statistic we can use the Mann-Whitney (MW) statistic U, i.e.,

$$U = \#(Y_j > X_i), \quad i = 1, \ldots, m; \quad j = 1, \ldots, n,$$

which is a count of the number of times the Y_js are greater than the X_is. The magnitude of U has a meaning, because U/nm is an estimate of the probability that an observation drawn at random from population Y would exceed an observation drawn at random from population X.

Noether [16] derived a sample size formula for the Mann-Whitney test (see Eq. (3) in Table 1), using an effect size $p_{Noether}$, that makes no assumptions about the distribution of the data (except that it is continuous), and can be used whenever the sampling distribution of the test statistic U can be closely approximated by the Normal distribution, an approximation that is usually quite good except for very small n [5].

Thus to determine the sample size, we have to find the 'effect size' $p_{Noether}$. There are several ways of estimating $p_{Noether}$ [20], under various assumptions, one possibility is $p_{Noether} = U/nm$ [14]. Let μ_X, σ_X^2, μ_Y, and σ_Y^2 be the mean and variance of the X and Y variables respectively. Then if $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ then Simonoff et al. [20] show that the maximum likelihood estimator of Prob $(Y > X)$ using the sample estimates of the mean and variance $(\hat{\mu}_X, \hat{\sigma}_X^2, \hat{\mu}_Y, \hat{\sigma}_Y^2)$ is:

$$p = \text{Prob}(Y > X) = \Phi\left(\frac{\hat{\mu}_Y - \hat{\mu}_X}{\left(\hat{\sigma}_X^2 + \hat{\sigma}_Y^2\right)^{1/2}}\right), \tag{5}$$

where Φ is the Normal cumulative distribution function.

If we assume the SF-6D is Normally distributed and $\sigma_X = \sigma_Y = \sigma$ then Eq. (5) allows the calculation of two comparable 'effect sizes' $p_{Noether}$ and Δ_{Normal} thus enabling the two methods of sample size estimation (Eqs. (2) and (4)) to be directly contrasted. If this SF-6D is not Normally distributed then we cannot use Eq. (5) to calculate comparable effect sizes and must rely on the empirical estimates calculated post hoc from the data.

Suppose we are planning a two-group study comparing HRQoL (using the SF-6D as the primary outcome) between the groups. We believe the SF-6D to be continuous, but not Normally distributed and are intending to compare SF-6D scores in the two groups with a Mann-Whitney U test. Therefore Noether's method will be appropriate. As before if we assume a mean difference of 0.05 and a standard deviation of 0.12 for the SF-6D, then using Eq. (5) this leads to an effect size $p_{Noether} = \text{Prob}(Y > X)$ of 0.616. Substituting $p_{Noether} = 0.616$ in Eq. (4) with a two-sided 5% significance level and 80% power gives the estimated number of subjects per group as 98.

The two methods have given similar sample size estimates. The two methods can be regarded as equivalent when the two distributions have the same shape and equal variances.

When the two distributions are Normally distributed with equal variances, the MW test will require about 5% more observations than the two-sample t-test to provide the same power against the same alternative. For non-Normal populations, especially those with long tails, the MW test may not require as many observations as the two-sample t-test [10].

Effect sizes

There is general agreement that further research is required to establish what a realistic and clinically meaningful effect size is for the SF-6D. To illustrate the various methods of sample size calculation we assumed a mean difference of 0.05 in SF-6D scores was the minimum clinically important difference (MCID) worth detecting [22]. Research on another preference-based measure the Health Utilities Index (HUI) has suggested that a difference of 0.03 is considered important [7].

4. Comparison of the two methods of sample size estimation

We used bootstrap methods to compare the power of the t-test and Mann-Whitney for detecting a shift in location using the SF-6D as an outcome [5, 6]. The bootstrap is a computer intensive method for statistical analysis [9]. It involves repeatedly drawing random samples from the original data, with replacement. It seeks to mimic in an appropriate manner the way the sample is collected from the population in the bootstrap samples from the observed data. The 'with replacement' means that any observation can be sampled more than once.

Suppose (as before) we have two independent random samples $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$. The Xs are Ys are random samples from continuous distributions having cumulative distribution functions (cdf), F_X and F_Y respectively. We will consider situations where the distributions have the same shape, but the locations may differ. Thus if δ denotes the location difference (i.e. mean (Y) − mean $(X) = \delta$), then $F_Y(y) = F_X(y - \delta)$, for every y. We shall focus on the null hypothesis H_0: $\delta = 0$ against the alternative H_A: $\delta > 0$. We can test these hypotheses using an appropriate significance test (e g. Mann-Whitney or t-test), and will let $\pi(F, \delta, \alpha, n)$ denote the power function of the test.

The bootstrap strategy is to use pilot data to a provide a non-parametric estimate, $\hat{F}$ of F and to use a simulation method for finding the power of the test associated with any specified sample size n if the data follow the estimated distribution function. If we denote the distribution function estimate by $\hat{G}$, under the alternative hypothesis δ, we can estimate the approximate power, $\hat{\pi}(G, \delta, \alpha, n)$ by the following computer simulation procedure [5, 6].

1. Draw a random sample with replacement of size $2n$ from $\hat{F}$. The first n observations in the sample form a simulated sample of X's, denoted by $X_1^*, \ldots, X_n^*$, with estimated cdf $\hat{F}^*$. Then δ is added to each of the other n observations in the sample to form the simulated sample of Y's, denoted by $Y_1^*, \ldots, Y_n^*$, with estimated cdf $\hat{G}^*$. (The Y^*'s and X^*'s have been generated from the same distribution except that the distribution of the Y^*'s is shifted d units to the right.)

2. The test statistic (Mann-Whitney or t-test) is calculated for the X^*'s and Y^*'s, yielding T^*. If $T^* \geq T_{1-\alpha/2}$, (where $T_{1-\alpha/2}$ is the critical value of the test statistic) a success is recorded; otherwise a failure is recorded.
3. Steps 1 and 2 are repeated J times. The estimated power of the test, $\hat{\pi}(G, \delta, \alpha, n)$, is approximated by the proportion of successes among the J repetitions. (In all cases discussed in this paper, $J = 10{,}000$).

The software Resampling Stats was used for the bootstrapping [19]. The bootstrap computer simulation procedure involved separately using two datasets. The first used SF-6D data from a general population survey based on 1373 people aged 16–74 years as the pilot dataset [2]. Figure 1 shows the non-symmetric distribution of the SF-6D. The second pilot data used SF-6D data from a sample of 232 patients with venous leg-ulcers [25]. Figure 2 shows the more symmetric distribution of the SF-6D in the leg ulcer sample.

Figure 3 shows the estimated power curves for the t and Mann-Whitney tests at the 5% two-sided significance level for detecting a location shift (mean difference) $\delta = 0.05$ in the SF-6D general population data for sample sizes per group varying from 20 to 240. For these general population data a location shift of $\delta = 0.05$ is equivalent to a standardised effect size $\Delta_{\text{Normal}} = 0.42$ and $p_{\text{Noether}} = \text{Prob}(Y > X) = 0.63$. For a sample size per group of 100 the Mann-Whitney test has an estimated power of 0.89 compared to an estimated power of 0.83 for the t-test. The Mann-Whitney test appears to be more powerful at detecting a location shift of $\delta = 0.05$ than the t-test. So for the general population data (with its skewed SF-6D distribution) the MW test is preferable to the t-test. Therefore for a fixed power, significance level and effect size using Noether's method would produce the smaller sample size estimates. However the differences in power are small and decrease as the sample size increases.

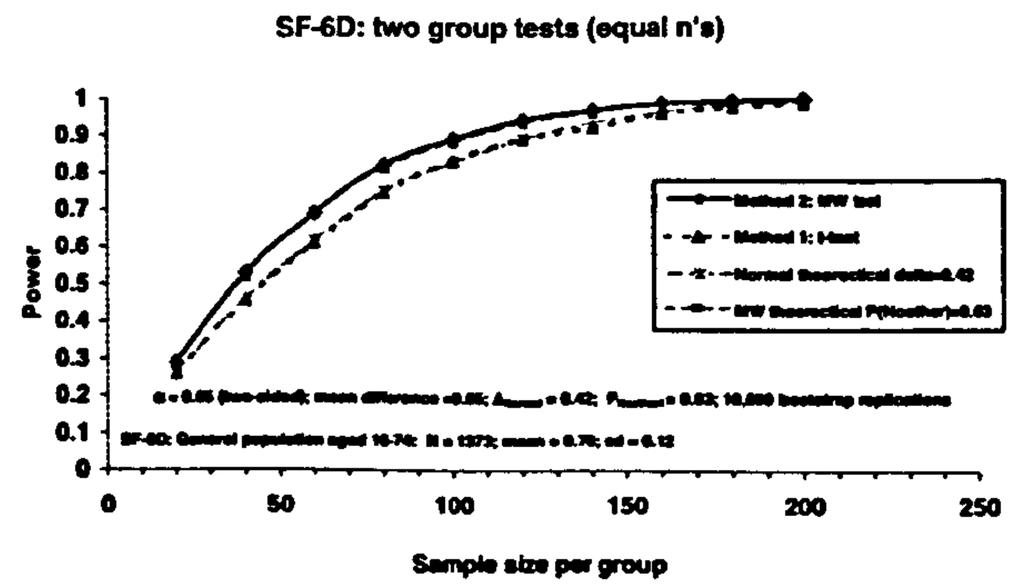


Figure 3. Estimated power curve for the SF-6D using general population data.

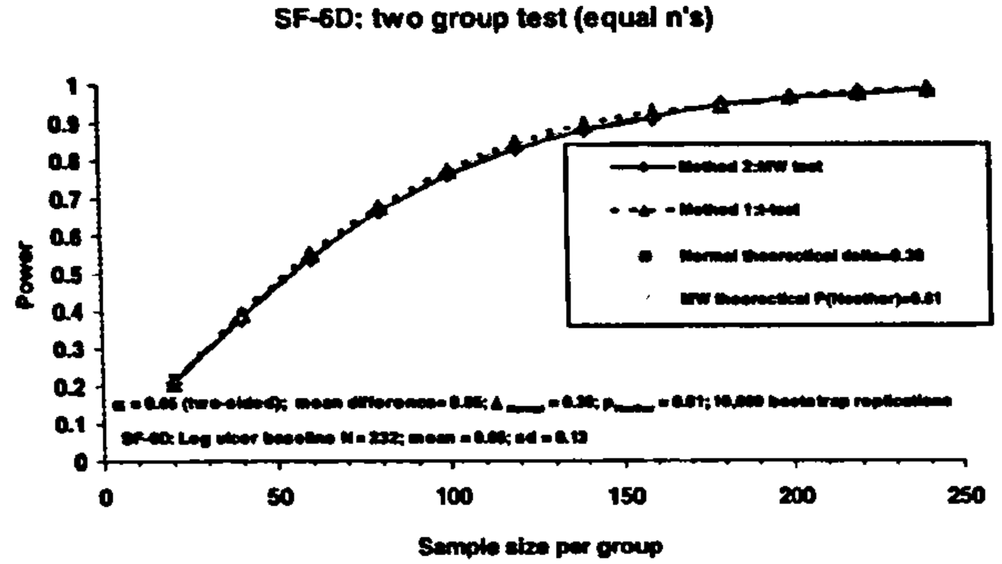


Figure 4. Estimated power curve for the SF-6D using leg ulcer data.

Figure 3 also shows the theoretical power curves calculated directly from the formulae (Eqs. (2) and (4)). Almost identical results were obtained compared to the bootstrap simulated power estimates. This illustrates that the standard methods work well.

Figure 4 shows that for the leg-ulcer data (with their more symmetric SF-6D distribution), the t-test appears to be slightly more powerful at detecting a location shift of $\delta = 0.05$ than the MW test. For these data a location shift of $\delta = 0.05$ is equivalent to a standardised effect size $\Delta_{\text{Normal}} = 0.38$ and $p_{\text{Noether}} = \text{Prob}(Y > X) = 0.61$. However again the differences in power between the t-test and MW tests are small and decrease as the sample size increases.

5. Choice of sample size method with the SF-6D outcomes

It is important to make maximum use of the information available from other related studies or extrapolation from other unrelated studies. The more precise the information the better we can design the trial. We would recommend that researchers planning a study with SF-6D as the primary outcome pay careful attention to any evidence on the validity and frequency distribution of the SF-6D.

The frequency distribution of SF-6D scores from previous studies should be assessed to see whether parametric or non-parametric methods should be used for sample size calculations and analysis. Computer simulation has suggested that if the distribution of the SF-6D is reasonably symmetric then the parametric method performs reasonably well at detecting differences in means. (Although the increase in the power of the t-test is very minor compared to the MW test). Therefore if the distribution of the SF-6D is symmetric or expected to be reasonably symmetric then parametric methods should be used for sample size calculations and analysis. The use of parametric methods for analysis (i.e. t-test) also

enables the relatively easy estimation of confidence intervals, which is regarded as good statistical practice [1].

If the distribution of the SF-6D is skewed then the Mann-Whitney test appears to be more powerful at detecting a location shift (difference in means) than the t-test. So in these circumstances the MW test is preferable to the t-test and non-parametric methods could be used for sample size calculations and analysis. The use of non-parametric methods for sample size estimation requires the effect size to be defined in terms of $P(Y > X)$, which is difficult to quantify and interpret. The arithmetic mean and mean difference is a better summary measure for heath care providers in deciding whether to offer a new treatment or not to its population. The mean provides information about the total benefit or utility from treating all patients, which is needed as the basis for health care policy decisions [21]. Therefore, since parametric methods have performed reasonably well, pragmatically we would recommend that parametric methods be used for sample size calculation and analysis when using the SF-6D in clinical trials and economic evaluations.

If the sample size is "sufficiently large" then the Central Limit Theorem (CLT) guarantees that the sample means will be approximately Normally distributed [12]. Thus, if the investigator is planning a large study and the sample mean is an appropriate summary measure of the SF-6D outcome, then pragmatically there is no need to worry about the distribution of the SF-6D outcome and we can use Eq. (3) to calculate sample sizes. Although the Normal distribution is strictly only the limiting form of the sampling distribution of the sample mean as the sample size n increases to infinity, but it provides a remarkably good approximation to the sampling distribution even when n is small and the distribution of the data is far from Normal. Generally, if n is greater than 25, these approximations will be good. However, if the underlying distribution is symmetric, unimodal, and of the continuous type, a value of n as small as 4 can yield a very adequate approximation [12] If after data collection one is still concerned about the validity of the CLT for small samples then one may use non-parametric bootstrap methods to estimate confidence intervals [9].

More work is required on what is a clinically meaningful effect sizes for the SF-6D. To illustrate the various methods of sample size calculation we assumed a mean difference of 0.05 in SF-6D scores was the MCID worth detecting. Retrospectively calculating the SF-6D for a variety of studies that had previously used the SF-36 has shown mean differences between groups varying between 0.025 and 0.12 [22]. So large differences between groups in SF-6D scores are unlikely. Therefore larger sample sizes may be required to detect statistically significant differences between groups in mean SF-6D scores.

There maybe considerable uncertainties in estimates of such quantities as the standard deviation and the treatment effect. Sample size calculations are sometimes based on estimates "pulled out of thin air." If an investigator is uncomfortable with the assumptions then it is good practice to calculate sample sizes under a variety of scenarios so that the sensitivity to assumptions can be assessed. We would recommend that various anticipated benefits be considered, ranging from the optimistic to the more realistic, with sample sizes being calculated for several scenarios within that range. It is a matter of judgement, rather than an exact science, as to which of the options is chosen for the final study size [11].

In this paper we have concentrated on the issue that HRQoL outcome data (such as the SF-6D) may not meet the distributional requirements of parametric methods of sample size

estimation and statistical analysis. There are other equally important problems with HRQoL measures such as ordinal scaling, linearity of the scale, floor/ceiling effects, non-constant variance and missing data which are discussed more fully in Walters et al. [23, 24].

6. Conclusions

Given that the end goal of using HRQoL outcomes in research studies is to assess a patient's health and well being, using the right type of HRQoL outcome in the right setting with an appropriate sample size calculation is crucial. Much time and energy is devoted to developing and validating HRQoL measures. We have provided a clear description of the distribution of the SF-6D and believe that the mean is an appropriate summary measure for the SF-6D when it is to be used in comparative clinical trials and the economic evaluation of new health technologies. Therefore pragmatically we would recommend that parametric methods be used for sample size calculation and analysis when using the SF-6D.

Quality adjusted life years and cost-utility analysis

Preference-based health state scores or utilities do not have natural units. Since health is a function of both length of life and quality of life, the Quality-adjusted life year (QALY) has been developed in an attempt to combine the value of these attributes into a single index number. If utilities are multiplied by the amount of time spent in that particular health state then they become QALYs (and are measured in units of time). QALYs allow for varying times spent in different states by calculating an overall score for each patient.

If information on the resources consumed and the cost of the resources is collected then an economic evaluation may be performed alongside the clinical trial. Cost-effectiveness analysis (CUA) is one form of full economic evaluation, where both the costs and consequences of heath programmes or treatments are examined [8]. In cost-utility analysis (CUA) the incremental cost of a programme, from a particular viewpoint, is compared to the incremental health improvement attributable to the programme, where the health improvement is measured in QALYs gained The results are usually expressed as a cost per QALY gained.

In the case of preference-based measures, such as the SF-6D, one might argue that the ultimate objective is to influence resource allocation decisions [7]. Therefore, it is the difference in cost-effectiveness (e.g., incremental cost per QALY) that is important not the change in HRQoL. Hence changes in the HRQoL measure alone may not be of interest without also considering the cost of bringing about those changes Thus, the sample size calculation if one was performed would be designed such that it would be possible to assess whether the incremental cost per QALY for the new treatment, compared with the existing one, is within an acceptable interval (e g., less than £30,000 per QALY). There are several statistical methods for constructing confidence intervals for incremental cost-effectiveness ratios (e.g. Taylor series approximation, Fieller's Method and the bootstrap) [4].

If decision makers at the design stage of a study, can specify their maximum threshold willingness to pay for an additional unit of effectiveness R_{Max}, then using formulae developed by Willan and O'Brien [29] we can determine the required sample size n, such that the upper limit of the 95% CI for the cost-effectiveness ratio, R is less than R_{Max}.

A likely consequence of designing studies to test hypotheses jointly about costs and effects is that the sample required may be larger than that to show differences in effects only. O'Brien et al. [17] raise an important ethical question: would it be ethical to continue a clinical trial to reach sufficient power to test a cost-effectiveness question when the number to show efficacy has been reached? They suggest a pragmatic way forward in that both the clinical and economic questions can be assessed by the same sample size (n for efficacy), but the investigator must simply accept greater uncertainty and wider 95% confidence intervals for the economic outcomes.

Finally we would stress the importance of a sample size calculation (with all its attendant assumptions), and that any such estimate is better than no sample size calculation at all, particularly in a trial protocol [28] The mere fact of calculation of a sample size means that a number of fundamental issues have been thought about: what is the main outcome variable, what is a clinically important effect, and how is it measured? The investigator is also likely to have specified the method and frequency of data analysis. Thus protocols that are explicit about sample size are easier to evaluate in terms of scientific quality and the likelihood of achieving objectives.

References

1 Altman, D G Machin, D . Bryant, T N . and Gardner M J . *Statistics with confidence Confidence intervals and statistical guidelines*, 2nd edn . Britsh Medical Journal. London. 2000.
2 Brazier, J E., Harper, R Jones, N M B . O'Cathain, A . Thomas, K J . Usherwood, T., and Westlake, L., "Validating the SF-36 health survey questionnaire New outcome measure for primary care," *British Medical Journal* 305, 160–164, 1992.
3 Brazier, J E., Roberts J F . and Deverill, M D . "The estimation of a preference based measure of health from the SF-36," *Health Economics* 21, 271–292, 2002.
4 Briggs, A H., Mooney, C Z . and Wonderling, D E., "Constructing confidence intervals for cost-effectiveness ratios An evaluation of parametric and non-parametric techniques using monte carlo simulation," *Statistics in Medicine* 18, 3245–3262, 1999.
5 Collings, B J and Hamilton, M A . "Determining the appropriate sample size for nonparametric tests for location shift," *Technometrics* 3(33), 327–337, 1991.
6 Collings, B J and Hamilton, M A., "Estimating the power of the two-sample wilcoxon test for location shift," *Biometrics* 44, 847–860, 1998.
7 Drummond, M F., "Introducing economic and quality of life measures into clinical studies," *Ann Med.* 33, 344–349, 2001.
8 Drummond, M F., Stoddard G L . and Torrance, G W., *Methods for the economic evaluation of health care programmes*, 2nd edn . Oxford University Press, Oxford, 1997.
9 Efron, B and Tibshirani R J . *An Introduction to the bootstrap* Chapman & Hall, New York, 1993.
10 Elashoff J D *nQuery advisor version 3 0 user's guide* Statistical Solutions, Los Angeles, 1999.
11 Fayers, P M and Machin, D . *Quality of life Assessment analysis & interpretation*. Wiley, Chichester, 2000.
12 Hogg R V and Tanis, E A *Probability and statistical inference*, 3rd edn. Macmillan, New York, 1988.
13 Lehman, E L., *Nonparametric statistical methods based on ranks* Holden-Day, San Francisco, 1975.
14 Lesaffre E., Scheys, I., Frohlich, J and Bluhmki E., "Calculation of power and sample size with bounded outcome scores," *Statistics in Medicine* 12, 1063–1078, 1993.
15 Machin, D . Campbell, M J Fayers, P M., and Pinol, A J Y . *Sample sizes tables for clinical studies*. 2nd edn., Blackwell Science. Oxford, 1997.
16 Noether, G E., "Sample size determination for some common nonparametric tests," *J American Statistical Association* 82(398), 645–647, 1987.

17 O'Brien, B.J , Drummond, M F., Labelle, R.J , and Willan, A , "In search of power and significance: Issues in the design and analysis of stochastic cost-effectiveness studies in health care *Medical Care* 32(2), 150–163, 1994.
18. Pocock, S.J , *Clinical trials A practical approach*, Wiley, Chichester, 1983.
19. Simon, J L , *Resampling stats Users guide v5 02* Resampling Stats Inc., Arlington, 2000.
20 Simonoff, J S., Hochberg, Y , and Reiser, B , "Alternative estimation procedures for Pr($X < Y$) in categorised data," *Biometrics* 42, 895–907, 1986.
21. Thompson, S G. and Barber, J A , "How should cost data in pragmatic randomised trials be analysed?" *British Medical Journal* 320, 1197–1200, 2000.
22. Walters, S.J. and Brazier, J.E., "What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D," *Health & Quality of Life Outcomes* 1(4), 1–8, 2003.
23 Walters, S.J., Campbell, M J , and Lall, R , "Design and analysis of trials with quality of life as an outcome: A practical guide." *Journal of Biopharmaceutical Statistics* 11(3), 155–176, 2001.
24. Walters, S.J . Campbell, M.J , and Paisley, S . "Methods for determining sample sizes for studies involving quality of life measures· A tutorial," *Health Services & Outcomes Research Methodology* 2, 83–99, 2001.
25. Walters, S.J., Morrell, C.J., and Dixon, S , "Measuring health-related quality of life in patients with venous leg ulcers," *Quality of Life Research* 8(4), 327–336, 1999.
26 Ware, J E Jr., Kosinski, M , and Keller, S D , *SF-36 Physical and mental health summary scales A user's manual* Health Institute, Boston, 1994.
27 Ware, J E , Jr and Sherbourne, C D , "The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection," *Medical Care* 30, 473–483, 1992.
28 Williamson, P., Hutton, J L., Bliss, J , Blunt, J , Campbell, M J , and Nicholson, R , "Statistical review by research ethics committees." *J Roy Statist. Soc A* 163, 5–13, 2000.
29. Willan, A R and O'Brien, B.J , "Sample size and power issues in estimating incremental cost-effectiveness ratios from clinical trials data," *Health Economics* 8(3), 203–211, 1999.

Appendix 4: Bootstrap Programs

Examples of Resampling Stats code for implementing Algorithm 6.1

t-test

```
'-------------------------------------------------
'Bootstrap simulation for estimating power for comparing two means
via a 2-sample t-test
' Taking into account the bounded nature of the outcome
' i.e. recodes values above 100 to 100
' Under the location shift alternative hypothesis
' Author: Stephen Walters
' Date:   17 April 2001
'-------------------------------------------------

MAXSIZE default 10000

READ FILE "C:      \Documents and Settings      \stephen walter   \My
Documents\PhD\Bootstrap data\other datafiles\females.dat" id age sex
phys social rlp rlm mental energy pain ghp pcs mcs sf6d

'Remember to enter the following parameters
LET samsize = 0          'Sets initial sample size to samsize
LET step    = 50         'Defines the sample size increment
LET t= 1.98              'Defines critical value for Test Statistic
LET delta = 5
LET testdata = ghp       'Defines testdata as the data vector to be
bootstrap
LET boot =10000          'Defines the number of BOOTSTRAP repetitions
LET loop = 12            'Defines the number of LOOPs starting at
                         'SAMPSIZE and increasing by STEP

' t alpha 0.05    8  df = 2.306     38 df = 2.024      80 df = 1.99
'              18 df = 2.101     48 df = 2.011     100 df = 1.984
'              28 df = 2.048     58 df = 2.002     150 df = 1.976

' Calculates means and standard deviations for observed control data
MEAN testdata mn_cont
STDEV testdata sd_cont
PRINT mn_cont sd_cont

LET n1 = samsize
LET n2 = samsize
LET n = n1 + n2

REPEAT loop
'adds STEP to starting sample size and repeats LOOP times
ADD step n1 n1
ADD step n2 n2

      REPEAT boot
      'Draws a random sample of size n with replacement
            SAMPLE n2 testdata treat$
            SAMPLE n1 testdata cont$

            'Adds the treatment effect delta to one sample
            ADD treat$ delta treat_d1
```

```
' Recodes values above 100 to 100
RECODE treat_d1 between 101 200 100 treat_d2
RECODE treat_d2 between -100 0 0 treat_d

MEAN cont$ mean_c
MEAN treat_d mean_t
VARIANCE cont$ var_c
VARIANCE treat_d var_t
STDEV treat_d sd_t
STDEV cont$   sd_c
LET s2=((n1-1)*var_c) + ((n2-1)*var_t)
LET denom = n1 + n2 - 2
              DIVIDE s2 denom s3
              LET sp=s3^0.5
              LET se=sp*((1/n1) + (1/n2))^0.5              'Calculates SE
              SUBTRACT mean_t mean_c diff
              DIVIDE diff se z                'Calculates Test Statistic z
              LET z_abs = abs(z)

              'Stores Test Statistic, means & mean difference
              SCORE z_abs scrboard
              SCORE mean_c cmean_sc
              SCORE mean_t tmean_sc
              SCORE diff diff_sc
              SCORE sd_t sdt_sc
              SCORE sd_c sdc_sc
       END

'Counts the number of significant tests
COUNT scrboard >= t  k
MEAN cmean_sc bt_cmean
MEAN tmean_sc bt_tmean
MEAN diff_sc  bt_diff
MEAN sdt_sc   bt_sdtrt
MEAN sdc_sc   bt_sdcon
DIVIDE k boot power              'Calculates the power
PRINT n1 n2 power bt_cmean  bt_sdcon bt_tmean bt_sdtrt bt_diff
CLEAR scrboard
CLEAR cmean_sc
CLEAR tmean_sc
CLEAR diff_sc
CLEAR sdt_sc
CLEAR sdc_sc
END
```

MW test

```
'-------------------------------------------------------------------
' Bootstrap simulation for estimating power for comparing two means
via a Mann-Whitney U test
' Adjusted for ties
' And corrected for the bounded outcome
' I.e. scores of 100 or more are set to 100
' Assumming a location shift alternative hypothesis
' Author: Stephen Walters
' Date:   17 April 2001
'-------------------------------------------------------------------

MAXSIZE default 10000
```

```
READ FILE "C:      \Documents and Settings      \stephen walter   \My
Documents\PhD\Bootstrap data\other datafiles\females.dat" id age sex
phys social rlp rlm mental energy pain ghp pcs mcs sf6d

'Remember to enter the following parameters.............
LET samsize  = 0         'Sets initial sample size to samsize
LET t        = 1.96      'Defines critical value for Test Statistic
LET delta    = 5         'Defines treatment effect
LET testdata = mental    'Defines testdata as the data vector   to be
bootstrap
LET boot     = 10000     'Defines the number of bootstrap resamples
LET loop     = 12        'Defines the number of LOOPS
LET step     = 50        'Defines the increments

' Calculates means and standard deviations for observed control data
MEAN testdata mn_cont
STDEV testdata sd_cont
PRINT mn_cont sd_cont

LET n1= samsize
LET n2 =samsize
LET n = n1 + n2
COPY samsize n3

REPEAT loop
'adds STEP to starting sample size and repeats LOOP times
ADD step n1 n1
ADD step n2 n2

      REPEAT boot
      'Draws a random sample of size n1 with replacement
      SAMPLE n1 testdata treat$

      'Draws a random sample of size n2 with replacement
      SAMPLE n2 testdata cont$

      'Adds the treatment effect delta to one sample
      ADD treat$ delta treat_d1

      ' Recodes values above 100 to 100
      RECODE treat_d1 between 101 200 100 treat_d
      MEAN cont$ mean_c
      MEAN treat_d mean_t
      STDEV cont$    sd_cont
      STDEV treat_d  sd_trt
      SUBTRACT mean_t mean_c diff

      'Combines the two datasets into one vector
      CONCAT treat_d cont$ c

      'Calculates ranks for the combined dataset
      RANKS c rnk

'Calculates an indicator variable with n1 zeros and n2 ones
      URN n1#1 n2#0 ind_01

      MULTIPLY rnk ind_01 rk_prod
      SUM rk_prod rk1_sum
      LET rnk_2 = rnk^2
      SUM rnk_2 rk_2_sum
      LET u = rk1_sum - (n1*(n1+1))/2
```

```
LET e_u = (n1* n2)/2
LET var1 = (n1*n2)/((n1+n2)*(n1+n2-1))
LET var2 = ((n1*n2)*(n1+n2+1)^2)/(4*(n1+n2-1))
LET var_u = (var1* rk_2_sum) - var2
LET se_u = var_u^0.5
'calculates Mann-Whitney/Wilcoxon test statistic
LET w= (u - e_u)/se_u
LET w_abs=abs(w)

'Calculates the Probability (X_cont < Y_trt)
LET prob_xy = u/(n1*n2)

                'Stores Test Statistic, means & mean difference
                SCORE w_abs scrboard
                SCORE mean_c cmean_sc
                SCORE mean_t tmean_sc
                SCORE sd_cont sdc_sc
                SCORE sd_trt  sdt_sc
                SCORE diff diff_sc
                SCORE prob_xy pr_xy_sc
        END

'Counts the number of significant tests
COUNT scrboard >= t  k
MEAN cmean_sc bt_cmean
MEAN tmean_sc bt_tmean
MEAN diff_sc  bt_diff
MEAN pr_xy_sc bt_prbxy
MEAN sdc_sc bt_sdcon
MEAN sdt_sc bt_sdtrt
DIVIDE k boot power                             'Calculates the power
PRINT n1 n2 power bt_prbxy bt_cmean bt_sdcon bt_tmean bt_sdtrt
bt_diff
CLEAR scrboard
CLEAR cmean_sc
CLEAR tmean_sc
CLEAR diff_sc
CLEAR pr_xy_sc
CLEAR sdc_sc
CLEAR sdt_sc
END
```

Example of SPSS syntax code for implementing Algorithm 7.1

SPSS F = G

```
*=============================================================.
* TITLE:    BOOTSTRAP TEST STATISTIC FOR TESTING F=G.
* Bootstrap testing F = G.sps
* AUTHOR:   S.J.Walters.
* DATE:     30/7/2003.
* COMMENTS:.

*-------------------------------------------------------------------------------------------.
* SPSS Syntax for generating bootstrap random samples AND .
* COMPUTATION OF BOOTSTRAP TEST STATISTIC FOR TESTING F=G.

* I.e. equality of distributions (means and variances).
* nsamples = no of random samples (1000-2000) to estimate CIs.
* ncases = no of cases.

GET
  FILE='C:\Documents and Settings\stephen walter\My Documents\PhD\Bootstrap'+
 ' data\SPSS datafiles\cmsw summary data n=495.sav'.

* calculates the plug-in estimates from the empirical distribution.
T-TEST
  GROUPS=group(0 1)
  /MISSING=ANALYSIS
  /VARIABLES=tmental
  /CRITERIA=CIN(.95) .

SORT CASES BY group .
SPLIT FILE
  BY group .

SPLIT FILE
  OFF.

COMPUTE ID=$CASENUM.

SAVE OUTFILE 'C:\Documents and Settings\stephen walter\My Documents\PhD\Bootstrap
data\SPSS datafiles\ORIGDATA.SAV'.

* REMEMBER to alter ncases EVERY TIME you change variables.
INPUT PROGRAM.
*ncases=495 number of valid cases.
LOOP SAMP=1 to 5000.
LOOP V= 1 to 495.
COMPUTE ID=TRUNC(UNIFORM(495))+ 1.
END CASE.
LEAVE SAMP.
END LOOP.
END LOOP.
END FILE.
END INPUT PROGRAM.

SORT CASES BY ID.

* Calls the first n observations Group 1 and the remaining m observations 2.
* V > 241 ie number in first group.
```

```
COMPUTE GROUP=1.
DO IF (V > 241).
COMPUTE GROUP=2.
END IF.

SAVE OUTFILE 'C:\Documents and Settings\stephen walter\My Documents\PhD\Bootstrap'+
 ' data\SPSS datafiles\BOOTSAM.SAV'.

MATCH FILES /FILE * /TABLE 'C:\Documents and Settings\stephen walter\My
Documents\PhD\Bootstrap'+
 ' data\SPSS datafiles\ORIGDATA.SAV'/BY ID.
SORT CASES BY SAMP.

SAVE OUTFILE 'C:\Documents and Settings\stephen walter\My Documents\PhD\Bootstrap'+
 ' data\SPSS datafiles\BOOTDATA.SAV'.


* Creates bootstrap mean and bootstrap standard deviation.
* Remember to alter tmental to name of outcome variable.

* selects the FIRST group.
SELECT IF (GROUP = 1).
AGGREGATE
 /OUTFILE='C:\Documents and Settings\stephen walter\My Documents\PhD\Bootstrap'+
 ' data\SPSS datafiles\BOOTAGG1.SAV'
  /BREAK=samp
  /btmean1 'Mean of Bootstrap Replications' = MEAN(tmental)
  /btsd1 'SD of bootstrap replications' = SD(tmental).

* selects the SECOND group.
GET FILE 'C:\Documents and Settings\stephen walter\My Documents\PhD\Bootstrap'+
 ' data\SPSS datafiles\BOOTDATA.SAV'.

SELECT IF (GROUP = 2).
AGGREGATE
 /OUTFILE='C:\Documents and Settings\stephen walter\My Documents\PhD\Bootstrap'+
 ' data\SPSS datafiles\BOOTAGG2.SAV'
  /BREAK=samp
  /btmean2 'Mean of Bootstrap Replications' = MEAN(tmental)
  /btsd2 'SD of bootstrap replications' = SD(tmental).

GET FILE='C:\Documents and Settings\stephen walter\My Documents\PhD\Bootstrap'+
 ' data\SPSS datafiles\BOOTAGG1.SAV'.

MATCH FILES /FILE * /TABLE 'C:\Documents and Settings\stephen walter\My
Documents\PhD\Bootstrap'+
 ' data\SPSS datafiles\BOOTAGG2.SAV'
/ BY SAMP.
SORT CASES BY SAMP.

SAVE OUTFILE 'C:\Documents and Settings\stephen walter\My Documents\PhD\Bootstrap'+
 ' data\SPSS datafiles\GROUP.SAV'.

* number in each group.
* No of cases in GROUP 1.
COMPUTE n = 241.
 * No of cases in GROUP 2.
COMPUTE m = 254.

* Computes variances.
```

```
COMPUTE btvar1= btsd1**2.
COMPUTE btvar2= btsd2**2.

* Computes test statistics.
COMPUTE txboot=btmean1-btmean2.
COMPUTE sigma = sqrt(((n-1)*btvar1 + (m-1)*btvar2)/(n + m -2)).
COMPUTE txstud =(btmean1-btmean2)/(sigma*sqrt((1/n) + (1/m))).

VARIABLE LABELS txboot 'Difference between Means'/
        txstud 'Studentised Difference between Means'/
        n 'No of Cases in Group 1'/
        m 'No of Cases in Group 2'/
        sigma 'Pooled SD'.

SAVE OUTFILE 'C:\Documents and Settings\stephen walter\My Documents\PhD\Bootstrap'+
 ' data\SPSS datafiles\GROUP.SAV'.

* Note STANDARD DEVIATION of bootmean is an estimate of the STANDARD ERROR of
the statistic.

DESCRIPTIVES
  VARIABLES txboot txstud
  /FORMAT=LABELS NOINDEX
  /STATISTICS=MEAN SUM STDDEV .
FREQUENCIES
  VARIABLES=txboot txstud.

FREQUENCIES
  VARIABLES=txboot txstud
  /FORMAT=NOTABLE
  /PERCENTILES= 0.01 0.1 0.5 1 2.5 5 10 25 50 75 90 95 97.5 99 99.5 99.9 99.99
  /STATISTICS=STDDEV SEMEAN MEAN
  /HISTOGRAM .
```

Example of STATA commands for implementing Algorithm 7.2

STATA ASL_{boot} and generating BC_a CIs

```
* STATA program for testing equality of means
* and calculating ASL-boot
* based on unequal variances t-test

use "C:\Documents and Settings\stephen walter\My
Documents\PhD\Bootstrap data\STATA datafiles\cmsw summary data
n=495.dta", clear
log using "C:\Documents and Settings\stephen walter\My
Documents\PhD\Bootstrap output\Stata output\CMSW pf analysis
log.log",replace

ttest tphys, by(group) unequal

scalar tobs=r(t)
summarize tphys, mean
scalar omean =r(mean)

summarize tphys if group==0, mean
replace tphys = tphys - r(mean) + scalar(omean) if group ==0

summarize tphys if group==1, mean
replace tphys = tphys - r(mean) + scalar(omean) if group ==1
sort group
by group: summarize tphys
log close

keep tphys group

bootstrap "ttest tphys, by(group) unequal" t=r(t), rep(5000)
strata(group) notable replace saving(bsauto2) nowarn

use "C:\Documents and Settings\stephen walter\My
Documents\bsauto2.dta", clear

generate indicator =abs(t) >=abs(scalar(tobs))
summarize indicator, mean
display "ASLboot = "r(mean)
```

Example of S-PLUS commands for Implementing Algorithm 7.3 and 7.4

S-PLUS -Regression Case and Model Resampling

```
#####################################################################
# OA Knee study surgery vs rheumatology patients
# PHYSICAL FUNCTION DIMENSION


# GROUP coded as 0 = rheumatology clinic 1 = surgery
# SEX coded as   0 = female   1 = male

# Bootstrap regression stratified by GROUP
# strata option
#####################################################################

library(boot)
attach(OA.Knee.pf.data)


# case resampling (Algorithm 7.3)

tphys.lm <- glm(FTPHYS~TPHYS+AGE+SEX+GROUP, data=OA.Knee.pf.data)
summary(tphys.lm)
tphys.diag <-glm.diag.plots(tphys.lm,ret=T)
tphys.fit <-function(data) coef(glm(data$FTPHYS~data$TPHYS + data$AGE +
   data$SEX + data$GROUP))
tphys.case <- function(data, i) tphys.fit(data[i,])
tphys.boot1 <-boot(OA.Knee.pf.data, tphys.case,
   strata=OA.Knee.pf.data$GROUP,R=4999)

tphys.boot1

# regression coefficients from original data

tphys.boot1$t0

# calculating bootstrap 95 CIs for regression coefficient estimates

tphys.boot1ci1  <- boot.ci(tphys.boot1,conf=0.95,
   type=c("norm","basic","perc","bca"), index=1)
tphys.boot1ci1
tphys.boot1ci2  <- boot.ci(tphys.boot1,conf=0.95,
   type=c("norm","basic","perc","bca"), index=2)
tphys.boot1ci2
tphys.boot1ci3  <- boot.ci(tphys.boot1,conf=0.95,
   type=c("norm","basic","perc","bca"), index=3)
tphys.boot1ci3
tphys.boot1ci4  <- boot.ci(tphys.boot1,conf=0.95,
   type=c("norm","basic","perc","bca"), index=4)
tphys.boot1ci4
tphys.boot1ci5  <- boot.ci(tphys.boot1,conf=0.95,
   type=c("norm","basic","perc","bca"), index=5)
tphys.boot1ci5

#####################################################################

# Model (residual) based resampling stratified by GROUP (Algorithm 7.4)
# Note: you need to perform the case-based resampling first to get values
# for the tphys.lm, tphys.diag & tphys.fit functions
```

```
# calculating modified residuals

tphys.res <- tphys.diag$res*tphys.diag$sd
tphys.res <- tphys.res - mean(tphys.res)

tphys.df <- data.frame(OA.Knee.pf.data, res=tphys.res, fit=fitted(tphys.lm))
tphys.model <- function(data, i)
{ d <- data
d$FTPHYS <- d$fit + d$res[i]
tphys.fit(d)}
tphys.boot2 <- boot(tphys.df, tphys.model,strata=OA.Knee.pf.data$GROUP,
   R=4999)

tphys.boot2

# regression coefficients from original data

tphys.boot2$t0

# calculating bootstrap 95 CIs for regression coefficient estimates
tphys.boot2ci1  <- boot.ci(tphys.boot2,conf=0.95,
   type=c("norm","basic","perc","bca"), index=1)
tphys.boot2ci1
tphys.boot2ci2  <- boot.ci(tphys.boot2,conf=0.95,
   type=c("norm","basic","perc","bca"), index=2)
tphys.boot2ci2
tphys.boot2ci3  <- boot.ci(tphys.boot2,conf=0.95,
   type=c("norm","basic","perc","bca"), index=3)
tphys.boot2ci3
tphys.boot2ci4  <- boot.ci(tphys.boot2,conf=0.95,
   type=c("norm","basic","perc","bca"), index=4)
tphys.boot2ci4
tphys.boot2ci5  <- boot.ci(tphys.boot2,conf=0.95,
   type=c("norm","basic","perc","bca"), index=5)
tphys.boot2ci5

##########################################################################
```

Example of STATA commands for GEE case resampling Algorithm 8.1

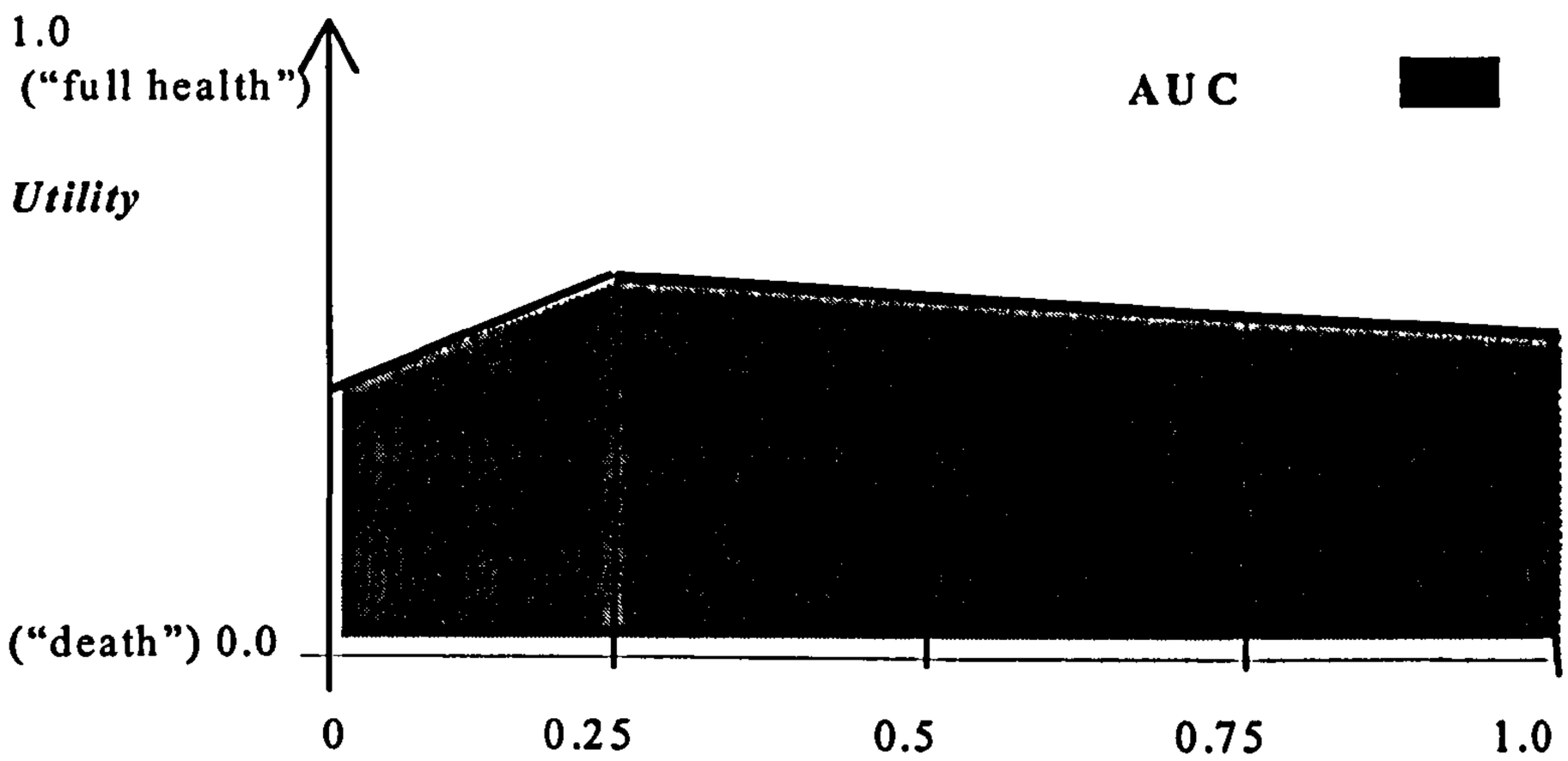
```
. xtgee  pf   base age sex time group, i(id) t(time) corr(exc)
link(iden) fam(gauss) robust


. bootstrap "xtgee  pf  base age sex time group, i(id) t(time)
corr(exc) link(iden) fam(gauss)  " _b,  reps(1000) bca  nobc nonormal
nopercentile  cluster(id)  strata(group)

. xtgee  pf  base age sex time group, i(id) t(time) corr(ar1)
link(iden) fam(gauss)robust


. bootstrap "xtgee  pf  base age sex time group, i(id) t(time)
corr(ar1) link(iden) fam(gauss)  " _b,  reps(1000) bca  nobc nonormal
nopercentile  cluster(id)  strata(group)
```

Appendix 5: Statistical Background

1. The area under the curve (AUC)

The AUC is a useful way of summarising the information from a series of measurements on one individual (Matthews *et al* 1990). The AUC can also be used to summarize repeated HRQoL scores over time into a single measure of health for each patient.

Figure A5.1: Summary measure of HRQoL: the AUC



Calculating the AUC

The area (see Figure A5.1 above) can be split into a series of shapes called trapeziums. The areas of the separate individual trapeziums are calculated and then summed for each patient. The mean AUC in each group can then be calculated.

If Y_{ij} represents the HRQoL response variable observed at time t_{ij}, for observation j = 1, ...n_i on subject i = 1, ...,m. The set of repeated HRQoL outcomes for subject i are collected into a n_i-vector, $\mathbf{Y}_i = \left(Y_{i1}, Y_{i2}, ..., Y_{in_i}\right)$,. The AUC is for the i^{th} subject is calculated by

$$AUC_i = \frac{1}{2}\sum_{j=1}^{n_i}\left(t_{j+1} - t_j\right)\left(Y_j + Y_{j+1}\right). \qquad \text{(A5.1)}$$

The units of AUC are the product of the units used for Y_{ij} and t_{ij}, and may not be easy to understand, since HRQoL outcomes have no natural units. So it may be useful to divide the AUC by the total time to get a weighted average level over the time period.

We can calculate AUC even when there are missing data, except when the first and final observations are missing.

The AUCs calculated from the Leg Ulcer study and NAMEIT studies were based on 12 months and 48 week follow-up respectively (i.e. approximately one year follow-ups). If the time t_{ij} for each HRQoL assessment is represented as a fraction of a year then the AUCs represent the weighted average level of HRQoL over the year. An AUC of 100, corresponds to "good health" over the year, conversely an AUC of 0, corresponds to "poor health" over the period.

2. Likelihood, Generalised Linear models, and robust standard errors

Likelihood inference

Likelihood inference is based on a specification of the probability or probability density of the observed data, *y*. This expression, $f(y;\theta)$, is indexed by a vector of unknown parameter(s) θ. Once the data are observed, the only quantities in $f(\cdot)$ that are unknown to the investigators are θ. Then, the likelihood function for θ is the function

$$L(\theta \mid y) = f(y;\theta). \qquad \text{(A5.2)}$$

The likelihood is interpreted as a function of θ, with *y* held fixed at its observed value. The *maximum likelihood estimate* of θ is the value, $\hat{\theta}$, which maximises the likelihood function or equivalently, its logarithm. That is, for any value of θ,

$$L(\theta \mid y) \leq L(\hat{\theta} \mid y) \qquad \text{(A5.3)}$$

According to the likelihood principle, $\hat{\theta}$ is then regarded as the value of θ which is most strongly supported by the observed data. In practice, $\hat{\theta}$ is obtained by the direct maximisation of log *L*, or by solving the set of equations

$$S(\theta) = \frac{\partial \log L}{\partial \theta} = 0. \qquad \text{(A5.4)}$$

The function $S(\theta)$ is known as the *score function* for θ. Very often, numerical methods are required to evaluate the maximum likelihood estimate (A5.4).

The asymptotic variance matrix of $\hat{\theta}$ is given by the expression

$$V = \left\{ -E\left(\frac{\partial^2 \log L}{\partial \theta^2} \right) \right\}^{-1} \qquad \text{(A5.5)}$$

The matrix V^{-1} is also known as the *Fisher information matrix* for θ.

Generalized linear models

Regression models for independent (binary, discrete and continuous) responses have been unified under the class of *generalized linear models*, or GLMs (McCullagh and Nelder, 1989), thus providing a common body of statistical methodology for different types of response.

Linear, logistic and Poisson regression models are all special cases of GLMs, which share the following features. First, the mean response, $\mu_i = E(Y_i)$, is assumed to be related to a vector of covariates, ***x***, through

$$h(\mu_i) = x_i^T \beta \,. \tag{A5.6}$$

For linear regression, $h(\mu_i) = \mu_i$; for logistic regression $h(\mu_i) = \log\left\{\frac{\mu_i}{1-\mu_i}\right\}$; and for Poisson regression $h(\mu_i) = \log(\mu_i)$. The function $h(\cdot)$ is called the *link function.*

Second, the variance of Y_i is a special function of its mean, μ_i namely,

$$Var(Y_i) = v_i = \phi v(\mu_i). \tag{A5.7}$$

In this expression, the known function $v(\cdot)$ is referred to as the *variance function*; the scaling factor ϕ is a known constant for some members of the GLM, whereas in others it is an additional parameter to be estimated. For the linear model $\phi = \sigma^2$, whereas for the Poisson model $\phi = 1$.

Third, each class of GLMs corresponds to member of the exponential family of distributions, with a likelihood function of the form

$$f(y_i, \theta_i, \phi) = \exp\left[\frac{\{y_i\theta_i - \psi(\theta_i)\}}{\phi} + c(y_i, \phi)\right]. \tag{A5.8}$$

The parameter θ_i is known as the *natural parameter*, and is related to μ_i through $\mu_i = \frac{\partial \psi(\theta_i)}{\partial \theta_i}$. For example, the Normal distribution is a special case of the exponential family, with

$$f(y_i,\theta_i,\phi)=\frac{1}{\sqrt{(2\pi\sigma^2)}}\exp\left\{\frac{-(y_i-\mu_i)^2}{2\sigma^2}\right\}$$

$$=\exp\left[\frac{\{y_i\mu_i-(\mu_i^2/2)\}}{\sigma^2}-\frac{1}{2}\left\{\frac{y_i^2}{\sigma^2}+\log(2\pi\sigma^2)\right\}\right]$$

$$\theta_i=\mu_i,\ \psi(\theta_i)=\frac{\theta_i}{2},\ c(y_i,\phi)=\frac{1}{2}\left\{\frac{y_i^2}{\phi}+\log(2\pi\phi)\right\}\ \text{and}\ \phi=\sigma^2. \quad \text{(A5.9)}$$

Similarly, the Poisson distribution is another example from the exponential family, with

$$\theta_i=\log\mu_i,\ \psi(\theta_i)=\exp(\theta_i),\ c(y_i,\phi)=-\log(y_i!)\ \text{and}\ \phi=1. \quad \text{(A5.10)}$$

In any GLM the regression coefficients, β, can be estimated by solving the same estimating equation,

$$S(\beta)=\sum_{i=1}^{m}\left(\frac{\partial\mu_i}{\partial\beta}\right)^T Var(Y_i)^{-1}\{Y_i-\mu_i(\beta)\}=0. \quad \text{(A5.11)}$$

$S(\beta)$ is the derivative of the logarithm of the likelihood function. The solution $\hat{\beta}$, which is the maximum likelihood estimate, can be obtained by iteratively reweighted least squares (IRLS) as described in McCullagh and Nelder (1989).

Finally, in large samples $\hat{\beta}$ follows a Normal distribution with mean β and variance

$$V=\left(\sum_{i=1}^{m}\left(\frac{\partial\mu_i}{\partial\beta}\right)^T Var(Y_i)^{-1}\frac{\partial\mu_i}{\partial\beta}\right)^{-1}. \quad \text{(A5.12)}$$

This variance can be estimated by $\hat{V}$ which is obtained by β replacing with $\hat{\beta}$ in the expression (A5.12).

Quasi-likelihood

One important property of the GLM family is that the score function, $S(\beta)$ depends only on the mean and variance of the Y_i. Wedderburn (1974) was the first to point out that the estimating equation (A5.11) can be used to estimate the regression coefficients for any choices of link and variance functions, whether or not they correspond to a particular form of the exponential family. The name *quasi-score function* was coined for $S(\beta)$ in (A5.11) since it's integral with respect to β can be thought of as a 'quasi-likelihood' even if it does not constitute a proper likelihood

function. This suggests an approach to statistical modelling in which we make assumptions about the link and variance functions without attempting to specify the entire distribution of Y_i or its likelihood. This is desirable, since we often do not understand the precise details of the probabilistic mechanism by which the data were generated (Diggle *et al* 2002). McCullagh (1983) showed that the solution, $\hat{\beta}$, of the quasi-score function has a sampling distribution which, in large samples, is approximately Normal with mean β and variance given by equation (A5.12).

Generalised estimating equations (GEEs) for longitudinal data (adapted from Zeger and Liang, 1986)

Consider the observations (Y_{ij}, $\boldsymbol{x}_{ij}$) for times t_{ij}, $j = 1,\ldots,n_i$ and subjects $i = 1,\ldots,m$. Here Y_{ij} is the outcome variable and $\boldsymbol{x}_{ij}$ is a p x 1 vector of covariates. Let $\boldsymbol{Y}_i$ be the n_i x 1 vector $\left(Y_{i1}, Y_{i2},\ldots Y_{in_i}\right)^T$ and $\boldsymbol{x}_i$ be the n_i x p matrix $\left(x_{i1}, x_{i2},\ldots, x_{in_i}\right)^T$ for the i^{th} subject.

Quasi-likelihood has previously been applied in the regression context (A5.11) where $n_i = 1$ for all i. Hence in discussing the results in this section, we drop the subscript j and treat each subject's data as a scalar.

Define μ_i to be the expectation of Y_i and suppose that

$$\mu_i = h(x_i\beta) \qquad \text{(A5.13)}$$

where β is a p x 1 vector of parameters. The inverse of h is referred to as the *link* function (McCullagh and Nelder, 1989). In quasi-likelihood, the variance, v_i of Y_i is expressed as a known function, g, of the expectation, μ_i, i.e.,

$$v_i = g(\mu_i)/\phi \qquad \text{(A5.14)}$$

where ϕ is a scale parameter. The focus of quasi-likelihood is on methods for inference about β. Hence, ϕ is treated as a nuisance parameter.

The quasi-likelihood estimator is the solution of the score-like equation system

$$S_k(\beta) = \sum_{i=1}^{m}\left(\frac{\partial\mu_i}{\partial\beta_k}\right)^T v^{-1}\{Y_i - \mu_i(\beta)\} = 0\,,\ k = 1,\ldots p. \qquad \text{(A5.15)}$$

Equation (A5.15) is in fact the score equation (A5.11) for β when Y_i has a distribution from the exponential family.

To apply the quasi-likelihood approach to the analysis of longitudinal data we must consider the mean and covariance of the vector of responses, $\boldsymbol{Y}_i$, for the i^{th} subject. In addition let $\mathbf{R}_i(\alpha)$ be the n_i x n_i "working" correlation matrix for each $\boldsymbol{Y}_i$. Note that

the observation times and the correlation matrix can differ from subject to subject. $\mathbf{R}_i(\alpha)$, however, is assumed to be fully specified by the s x 1 vector of unknown parameters, α, which is the same for all subjects. Then following the quasi-likelihood approach, the working covariance matrix for $\mathbf{Y}_i$ is given by

$$\mathrm{V}_i = \mathrm{A}_i^{1/2}\mathrm{R}_i(\alpha)\mathrm{A}_i^{1/2}/\phi\,, \qquad \text{(A5.16)}$$

where A_i is an n_i x n_i diagonal matrix with $g(\mu_{ij})$ as the j^{th} diagonal element. We refer to $\mathbf{R}_i(\alpha)$ as a "working" correlation matrix because we do not expect it to be correctly specified. We would like estimators that are consistent and have consistent variance estimates even when $\mathbf{R}_i(\alpha)$ is incorrect. (A5.16 will be equal to cov(Y_i) if is indeed $\mathbf{R}_i(\alpha)$ the true correlation matrix for the Y_i's.) Our extension of equation (A5.15) to the longitudinal case is given by

$$S_\beta(\beta,\alpha)=\sum_{i=1}^{m}\left(\frac{\partial\mu_i}{\partial\beta}\right)^T Var(Y_i)^{-1}\{Y_i-\mu_i\}=0\,. \qquad \text{(A5.17)}$$

Or simplified to

$$\sum_{i=1}^{m}\mathrm{D}_i^T\mathrm{V}_i^{-1}\mathrm{S}_i=0\,, \qquad \text{(A5.18)}$$

where $\mathrm{S}_i = \mathbf{Y}_i - \mu_i$, $\mu_i = (\mu_{i1},\ldots,\mu_{in})^T$ and $\mathrm{D}_i = \partial\mu_i/\partial\beta$.

Equations (A5.18) reduce to the quasi-likelihood estimating equations (A5.15) when n_i = 1 for all i. More generally $\mathrm{U}_i(\beta,\alpha)=\mathrm{D}_i^T\mathrm{V}_i^{-1}\mathrm{S}_i$ is equivalent to the estimating function suggested by Wedderburn (1974) except that the V_i's here are functions of α as well as β.

While the estimating equations (A5.18) now depend on α as well as β, they can be re-expressed as a function of β alone by first replacing α in equations (A5.16) and (A5.18) by a $m^{1/2}$-consistent estimator, $\hat{\alpha}(Y,\beta,\phi)$, then replacing ϕ in $\hat{\alpha}$ by a $m^{1/2}$-consistent estimator, $\phi(Y,\beta)$. Consequently, for any given $\mathbf{R}_i(\alpha)$, the estimate $\hat{\beta}_R$, of β is defined as the solution of the "generalized estimating equation" (GEE)

$$U_i(\beta,\alpha)=\sum_{i=1}^{m}\left(\frac{\partial\mu}{\partial\beta}\right)^T V_i^{-1}(\alpha)(Y_i-\mu_i)=0\,,\ \text{or simplified to}$$

$$\sum_{i=1}^{m}\mathrm{U}_i\left\{\beta,\hat{\alpha}\left[\beta,\hat{\phi}(\beta)\right]\right\}=0. \qquad \text{(A5.19)}$$

Under mild regularity conditions, Liang and Zeger (1986) show that as $m \to \infty$, $\hat{\beta}_R$ is a consistent estimator of β and that $m^{1/2}\left(\hat{\beta}_R - \beta\right)$ is asymptotically multivariate Normal with covariance matrix $\mathbf{V_R}$ given by

$$\mathrm{V_R} = \lim_{m\to\infty} m\left(\sum_{i=1}^{m} D_i^T V_i^{-1} D_i\right)^{-1}\left[\sum_{i=1}^{m} D_i^T V_i^{-1}\,\mathrm{cov}(Y_i) V_i^{-1} D_i\right]\left(\sum_{i=1}^{m} D_i^T V_i^{-1} D_i\right)^{-1}$$

$$\mathrm{V_R} = \lim_{m\to\infty} m\left(\mathrm{V_1^{-1}V_0V_1^{-1}}\right), \qquad \text{(A5.20)}$$

where the covariance of Y_i is the actual rather than the assumed covariance. $\mathbf{V_R}$ can be estimated consistently without evaluating cov(Y_i) directly. This is achieved by simply replacing cov(Y_i) by $S_i S_i^T$ and α, β and ϕ by their estimates in (A5.20).

To solve the GEE for $\hat{\beta}_R$, we iteratively solve for the regression coefficients and the correlation and scale parameters, α and ϕ. Given an estimate of $\mathbf{R}_i(\alpha)$ and of ϕ, we can calculate an updated estimate of β by IRLS. Given an estimate of β, we calculate standardised residuals, $r_{ij} = \left(Y_{ij} - \hat{\mu}_{ij}\right)/\sqrt{\left[\hat{\mathrm{V}}_1^{-1}\right]_{jj}}$, which are used to consistently estimate α and ϕ. These two steps are iterated until convergence. Details on computing $\hat{\beta}_R$ and $\hat{\mathrm{V}}_\mathrm{R}$ are provided by Liang and Zeger (1986). As in many quasi-likelihood problems, it is often possible to estimate β without estimating ϕ directly. We require only that the elements of **R** be multiples of the parameters, α.

Robust standard errors

We estimate β by using IRLS to solve the GEE (A5.19). A robust variance estimate is given by (A5.20) which can be simplified to (Zeger *et al* 1988):

$$V_{\hat{\beta}_R} = M_0^{-1} M_1 M_0^{-1}, \qquad \text{(A5.21)}$$

where

$$M_0 = \sum_{i=1}^{m}\left(\frac{\partial\hat{\mu}_i}{\partial\hat{\beta}_R}\right)^T \hat{V}_i^{-1}\left(\frac{\partial\hat{\mu}_i}{\partial\hat{\beta}_R}\right)$$

and

$$M_1 = \sum_{i=1}^{m}\left(\frac{\partial\hat{\mu}_i}{\partial\hat{\beta}_R}\right)^T \hat{V}_i^{-1}(Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)^T \hat{V}_i^{-1}\left(\frac{\partial\hat{\mu}_i}{\partial\hat{\beta}_R}\right)$$

(A5.21) is also consistent even when cov(Y_i) ≠ V_i. The marginal regression models were fitted in `STATA v8` (StataCorp, 2003) using the `xtgee` command with an identity link function (`link(iden)`) and the `robust` standard errors option.

References

Allard, S., and the NAME IT Study Group. (2000) *Phase IIIB.IV Clinical Study Report of a Double-Blind, Randomised, Controlled Study to Compare Methotrexate plus Neoral® versus Methotrexate plus Placebo in Subjects with Early Severe Rheumatoid Arthritis*. Basel, Switzerland, Novartis Pharma AG.

Altman, D.G. (1991) *Practical Statistics for Medical Research*. London, Chapman & Hall.

Altman, D.G., Machin, D., Bryant, T.N., and Gardner, M.J. (2000) *Statistics with Confidence: Confidence intervals and statistical guidelines.* 2nd edition. London, British Medical Journal.

Ananth, C., and Kleinbaum, D. (1997) Regression Models for Ordinal Responses: A Review of Methods and Applications. *International Journal of Epidemiology*, 26(6), 1323-1333.

Anderson, J.A. (1984) Regression and ordered categorical variables (with discussion). *Journal of Royal Statistical Society Series B*, 46, 1-30.

Angst, F., Aeschlimann, A., and Stucki, G. (2001) Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the lower extremities. *Arthritis & Rheumatism*, 4, 384-391.

Armitage, P., Berry, G., and Matthews, J.N.S. (2002) *Statistical Methods in Medical Research.* 4th edition. Blackwell Science, Oxford.

Armstrong, B.G. and Sloan, M. (1989) Ordinal regression models for epidemiologic data. *American Journal of Epidemiology*, 129, 191-204.

Bajorski, P., and Petkau, J. (1999) Non-parametric Two-sample Comparisons of Changes on Ordinal Responses. *Journal of the American Statistical Association,* 94(447), 970-978.

Bland, J.M., and Altman, D.G. (1995) Statistics Notes: Multiple significance tests: the Bonferroni method. *British Medical Journal, 310*, 170.

Bland, J.M., and Altman, D.G. (1996) The use of transformation when comparing two means. *British Medical Journal*, 312, 1153.

Bolland, K., Sooriyarachchi. M.R., and Whitehead, J. (1998) Sample Size Review in a Head Injury Trial with Ordered Categorical Responses. *Statistics in Medicine*, 17, 2835-2847.

Bowling, A. (1995) *Measuring Disease: A review of Disease-Specific quality of life measurement scales.* Buckingham, Open University Press.

Bowling, A. (1997) *Measuring Health: A review of quality of life measurement scales.* 2nd edition. Buckingham, Open University Press.

Brant, R. (1990) Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, 46, 1171-1178.

Brazier, J.E, Harper, R., Jones, N.M.B., O'Cathain, A., Thomas, K.J., Usherwood, T., and Westlake, L. (1992) Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *British Medical Journal*, 305, 160-164.

Brazier, J.E., Harper, R., Munro, J.F., Walters, S.J., and Snaith, M.L. (1999) Generic and condition-specific outcome measures for people with osteoarthritis of the knee. *Rhemautology,* 38, 870-877.

Campbell, M.J., Julious, S.A., and Altman, D.G. (1995) Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. *British Medical Journal*, 311, 1145-1148.

Campbell, M.J., Julious, S.A., and George, S.L. (1996) Estimating sample sizes for studies using the SF-36 health survey - Reply. *Journal of Epidemiology & Community Health*, 50, 473-474.

Campbell, M.J., and Machin, D. (1999) *Medical Statistics: A Commonsense Approach.* 3rd Edition. Wiley, Chichester.

Campbell, M.J. (2001) *Statistics at Square Two: Understanding Modern Statistical Applications in Medicine.* London, British Medical Journal.

Carpenter, J., and Bithell, J. (2000) Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, 19, 1141-1164.

Cella, D., Bullinger, M., Scott, C., Barofsky, I., and the Clinical Significance Consensus Meeting Group. (2002) Group vs individual approaches to understanding the clinical significance of differences or changes in quality of life. *Mayo Clinic Proceedings,* 77(4), 384-392.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioural Sciences.* 2nd edition. New Jersey, Lawrence Earlbaum.

Collings B.J., and Hamilton, M.A. (1988) Estimating the Power of the Two-Sample Wilcoxon Test for Location Shift. *Biometrics* , 44, 847-860.

Curran, D., Molenberghs, G., Fayers, P.M., and Machin, D. (1998) Incomplete quality of life data in randomised trials: missing forms. *Statistics in Medicine*, 17, 697-709.

Davison, A.C., and Hinkley, D.V. (1997) *Bootstrap Methods and their Applications.* Cambridge, Cambridge University Press.

Diggle, P.J., Heagerty, P., Liang, K-Y., and Zeger, S.L. (2002) *Analysis of Longitudinal Data*. 2nd edition. Oxford, Oxford University Press.

Dobson, A.J. (1990) *An Introduction to Generalized Linear Models.* London, Chapman & Hall.

Efron, B., and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap.* New York, Chapman & Hall.

Elashoff, J.D. (1999) *nQuery Advisor Version 3.0 User's Guide.* Los Angeles, Statistical Solutions.

Everitt, B.S. (1995) *The Cambridge Dictionary of Statistics in the Medical Sciences.* Cambridge, Cambridge University Press.

Everitt, B.S. (2001) *Statistics for Psychologists.* Mahwah, New Jersey, Lawrence Erlbaum Associates.

Everitt, B.S. (2002) *A Handbook of Statistical Analyses using S-Plus.* 2nd edition. Boca Raton, Florida, Chapman & Hall/CRC.

Fairclough, D.L. (2002) *Design and Analysis of Quality of Life Studies in Clinical Trials.* New York, Chapman & Hall.

Fayers, P.M., and Machin, D. (2000) *Quality of Life Assessment, Analysis and Interpretation.* Chichester, John Wiley.

Fayers, P.M., Curran, D., and Machin, D. (1998) Incomplete quality of life data in randomised trials: missing items. *Statistics in Medicine*, 17, 679-696.

Frison, L., and Pocock, S.J. (1992) Repeated Measures in Clinical Trials: Analysis Using Mean Summary Statistics and Its Implications for Design. *Statistics in Medicine*, 11, 1685-1704.

Frost, M.H., Bonomi, A.E., Ferrans, C.E., Wong, G.Y., Hays, R.D., and the Clinical Significance Consensus Meeting Group. (2002) Patient, clinician, and population perspectives on determining the clinical significance of quality-of-life scores. *Mayo Clinic Proceedings*, 77(5), 488-494.

Goldstein, H., Rasbash, L., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse , G., and Healy, M. (1998) *A user's Guide to MLwiN.* London, Institute of Education.

Greenland, S. (1994) Alternative models for ordinal logistic regression. *Statistics in Medicine*, 13, 1665-77.

Guyatt, G.H., Walter, S., and Norman, G. (1987) Measuring change over time: assessing the usefulness of evaluative instruments. *Journal of Chronic Disease*, 40, 171-178.

Guyatt, G.H., Osoba, D., Wu, A.W., Wyrwich, K.W., Norman, G.R. and the Clinical Significance Consensus Meeting Group. (2002) Methods to explain the clinical significance of health status measures. *Mayo Clinic Proceedings,* 77(4), 371-383.

Hamilton, M.A., and Collings, B.J. (1991) Determining the Appropriate Sample Size for Nonparametric Tests for Location Shift. *Technometrics,* 3(33), 327-337.

Hand, D., Crowder, M. (1996) *Practical Longitudinal Data Analysis.* Chapman & Hall, London.

Hays, R.D., Anderson, R.T., and Revicki, D. (1998) Assessing reliability and validity of measurement in clinical trials. *In:* Staquet, M.J., Hays, R.D., Fayers, P.M. (eds) *Quality of Life Assessment in Clinical Trials: Methods and Practice.* Oxford, Oxford University Press. p. 167-182.

Hays, R.D., and Morales, L.S. (2001) The RAND-36 measure of health-related quality of life. *Annals of Medicine,* 33(5), 350-357.

Heeren, T. and D'Agostino, R. (1987) Robustness of the two independent samples *t*-test when applied to ordinal scaled data. *Statistics in Medicine*, 6, 79-90.

Hendrickx, J. (2000) Special restrictions in multinomial logistic regression. *Stata Technical Bulletin*, STB-56: 18-26.

Hilton, J.F., and Mehta, C.R. (1993) Power and sample size calculations for exact conditional tests with ordered categorical data. *Biometrics* , 49, 609-616.

Hilton, J.F. (1996) The appropriateness of the Wilcoxon test in ordinal data. *Statistics in Medicine,* 15, 631-645.

Hogg, R.V., and Tanis, E.A. (1988) *Probability and Statistical Inference.* 3rd edition. New York, McMillan.

Jaeschke, R., Singer, J., Guyatt, G.H. (1989) Measurement of Health Status. Ascertaining the Minimal Clinically Important Difference. *Controlled Clinical Trials,* 10, 407-415.

Jenkinson, C., Stewart-Brown, S., Petersen, S., and Paice, C. (1999) Assessment of the SF-36 version 2 in the United Kingdom. *Journal of Epidemiology & Community Health*, **53**, 46-50.

Julious, S.A., George, S., and Campbell, M.J. (1995) Sample sizes for studies using the short form 36 (SF-36). *Journal of Epidemiology & Community Health*, 49, 642-644.

Julious, S.A., and Campbell, M.J. (1996) Sample sizes calculations for ordered categorical data. *Statistics in Medicine*, 15, 1065-1066.

Julious, S.A, George, S., Machin, D., and Stephens, R.J. (1997) Sample sizes for randomized trials measuring quality of life in cancer patients. *Quality of Life Research*, 6, 109-117.

Julious, S.A., and Campbell, M.J. (1998) Sample Size Calculations for Paired or Matched Ordinal Data. *Statistics in Medicine*, 17, 1635-1642.

Juniper, E.F., Guyatt, G.H., Willan, A., and Griffith. L.E. (1994) Determining a minimal important change in a disease-specific Quality of Life Questionnaire. *Journal of Clinical Epidemiology,* 47(1), 81-87.

Kaplan, R.M. (1998) Profile versus utility based measures of outcome for clinical trials. *In:* Staquet, M.J., Hays, R.D., Fayers, P.M. (eds) *Quality of Life Assessment in Clinical Trials: Methods and Practice.* Oxford, Oxford University Press. p. 69-90.

Kazis, L.E., Anderson, J.J. and Meenan, R.F. (1989) Effect Sizes for Interpreting Changes in Health Status. *Medical Care*, 27, (3), S178-S189.

King, M.T. (1996) The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. *Quality of Life Research,* 5, 555-567.

Kolassa, J.E. (1995) A comparison of size and power calculations for the Wilcoxon statistic for ordered categorical data. *Statistics in Medicine*, 14, 1577-1581.

Lall, R., Campbell, M.J., Walters, S.J., Morgan, K., and MRC CFAS. (2002) A review of ordinal regression models applied on Health related Quality of Life Assessments. *Statistical Methods in Medical Research*, 11(1), 49-67.

Laupacis, A., Sackett, D.L., and Roberts, R.S. (1988) An Assessment of clinically useful measures of the consequences of treatment. *New England Journal of Medicine*, 317, 1728-33.

Lehman, E.L. (1975) *Nonparametric Statistical Methods Based on Ranks.* San Francisco, Holden-Day.

Lesaffre, E., Scheys, I., Frohlich, J., and Bluhmki, E. (1993) Calculation of power and sample size with bounded outcome scores. *Statistics in Medicine*, 12, 1063-1078.

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

Liang, M.H., Larson, M.G., Gullen, K.E., and Schwartz, J.A. (1985) Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. *Arthritis & Rheumatology*, 28, 545-547.

Liang, M.H., Fossel, A.H., and Larson, M.G. (1990) Comparisons of Five Health Status Instruments for Orthopedic Evaluation. *Medical Care*, 28(7), 632-642.

Machin, D., Campbell, M.J., Fayers, P.M., and Pinol, A.P.Y. (1997) *Sample size tables for clinical studies.* 2nd Edition. Oxford, Blackwell.

Manly, B.F.J. (1994) *Multivariate Statistical Methods: A primer.* 2nd edition. London, Chapman & Hall.

Machin, D., and Fayers, P.M. (1998) Sample sizes for randomised trials measuring quality of life. *In*: M.J. Staquet, R.D. Hays and P.M. Fayers (eds) *Quality of Life Assessment in Clinical Trials: Methods and Practice*. Oxford, Oxford University Press. p. 37-50.

Manor, O., Matthews, S., and Power, C. (2000) Dichotomous or categorical response? Analysing self-rated health and lifetime social class. *International Journal of Epidemiology*, *29*, 149-157.

MathSoft. (1999) *S-PLUS 2000 Guide to Statistics, Volume 2.* Seattle, Washington, Data Analysis Products Division, MathSoft.

Matthews, J.N.S., Altman, D.G., Campbell, M.J., and Royston, P. (1990) Analysis of serial measurements in medical research. *British Medical Journal*, 300, 230-235.

McCullagh, P. (1983) Quasi-likelihood functions. *Annals of Statistics* 11, 59-67.

McCullagh, P., and Nelder, J.A. (1989) *Generalized Linear Models.* 2nd edition. London, Chapman & Hall/CRC.

Morrell, C.J., Walters, S.J., Dixon, S., Collins, K.A., Brereton, L.M.L., Peters, J., and Brooker, C.G.D. (1998) Cost-effectiveness of community leg ulcer clinics: randomised controlled trial. *British Medical Journal,* 316, 1487-1491.

Morrell, C.J., Spiby, H., Stewart, P., Walters, S., and Morgan, A. (2000) Costs and effectiveness of community postnatal support workers: randomised controlled trial. *British Medical Journal,* 321, 593-598.

Noether, G.E. (1987) Sample Size Determination for Some Common Nonparametric Tests. *Journal of the American Statistical Association*, 82(398), 645-647.

Norman, G.R., Sridhar, F.G., Guyatt, G.H., and Walter, S.D. (2001) The Relation of Distribution- and Anchor-Based Approaches in Interpretation of Changes in Health Related Quality of Life. *Medical Care*, 39(10), 1039-1047.

Norman, G.R., Sloan, J.A., and Wyrwich, K.W. (2003) Interpretation of Changes in Health Related Quality of Life: The Remarkable Universality of Half a Standard Deviation. *Medical Care*, 41(5), 582-592.

O'Brien, P.C. (1988) Comparing two samples, extensions of the *t*, rank-sum and log-rank tests. *Journal of the American Statistical Association*, 83, 52-61.

Olschewski M., and Schumacher, M. (1990) Statistical analysis of quality of life data in cancer clinical trials. *Statistics in Medicine*, 9(7), 749-763.

Perneger, T.V. (1998) What's wrong with Bonferroni Adjustments. *British Medical Journal,* 316, 1236–1238.

Peterson, B., and Harrell, F. (1990) Partial Proportional Odds Model for Ordinal Response Variables. *Applied Statistics*, 39(2), 205-217.

Phillips, A., Campbell, M. (1997) Using aspects of study design in sample size estimation. *Journal of Biopharmaceutical Statistics*, 7(2), 215-226.

Pocock, S.J. (1983) *Clinical Trials*: *A Practical Approach.* Chichester, Wiley.

Prieto, L., Alonso, J., and Anto, J.M. (1996) Estimating sample sizes for studies using the SF-36 health survey. *Journal of Epidemiology & Community Health*, 50, 473.

Quenouille M. (1949) Approximate tests of correlation in time series. *Journal of the Royal Statistical Society Series B,* 11, 18-44.

Rabbee, N., Coull, B.A., Mehta, C., Patel N., and Senchaudhuri P. (2003) Power and sample size for ordered categorical data. *Statistical Methods in Medical Research*, 12, 73-84.

Rabe-Hesketh, S. and Everitt, B.S. (2000) *A Handbook of Statistical Analyses using Stata.* 2nd edition. Boca Raton, Florida, Chapman & Hall/CRC.

Roset, M., Badia, X., and Mayo, N.E. (1999) Sample size calculations in studies using the EuroQol 5D. *Quality of Life Research*, 8, 539-549.

Satterthwaite, F.E. (1946) An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2, 110-114.

Shepstone, L. (2001) Re-conceptualising and Generalising the Absolute Risk Difference: A unification of Effect Sizes, Odds Ratios and Number-Needed-to-Treat. *Journal of Epidemiology & Community Health,* 55(Suppl 1) 1a: A7.

Simon, J.L. (2000) *Resampling Stats: Users Guide. v5.02.* Arlington, Resampling Stats Inc.

Simonoff, J.S., Hochberg, Y., and Reiser, B. (1986) Alternative Estimation Procedures for Pr(X < Y) in Categorised Data. *Biometrics*, 42, 895-907.

Sloan, J.A., Aaronson, N., Cappelleri, J.C., Fairclough, D.L., Varricchio, C. and Clinical Significance Consensus Meeting Group. (2002) Assessing the clinical significance of single items relative to summated scores. *Mayo Clinic Proceedings*, 77(5), 479-487.

Sloan, J.A., Cella, D., Frost, M., Guyatt, G.H., Sprangers, M., Symonds, T., and the Clinical Significance Consensus Meeting Group. (2002) Assessing clinical significance in measuring oncology patient quality of life: introduction to the symposium, content overview, and definition of terms. *Mayo Clinic Proceedings*, 77(4), 367-370.

Smirnov, N.V. (1939) On the estimation of the discrepancy between empirical distribution curves for two independent samples. *Bulletin de l'Université de Moscou, Série Internationale (Mathématiques)*, 2, 3-4.

Sprangers, M.A., Moinpour, C.M., Moynihan, T.J., Patrick, D.L., Revicki, D.A. and the Clinical Significance Consensus Meeting Group. (2002) Assessing meaningful change in quality of life over time: a users' guide for clinicians. *Mayo Clinic Proceedings*, 77(6), 561-571.

SPSS. (2001) *SPSS 11.0 Syntax Reference Guide.* Chicago, IL, SPSS Inc.

Staquet, M.J., Hays, R.D., and Fayers, P.M. (1998) *Quality of Life Assessment in Clinical Trials: Methods and Practice.* Oxford, Oxford University Press.

StataCorp. (2003) *Stata Statistical Software: Release 8.0.* College Station, Texas, Stata Corporation.

Strickland, P.A.O., and Lu, S-E. (2003) Estimates, power and sample size calculations for two-sample ordinal outcomes under before-after study designs. *Statistics in Medicine*, 22: 1807-1818.

Sullivan, I.M., and D'Agostino, R.B. (2003) Robustness and power of analysis of covariance applied to ordinal scaled data as arising in randomized controlled trials. *Statistics in Medicine*, 22, 1317-1334.

Symonds, T., Berzon, R., Marquis, P., Rummans, T.A., and the Clinical Significance Consensus Meeting Group. (2002) The clinical significance of quality-of-life results: practical considerations for specific audiences. *Mayo Clinic Proceedings*, 77(6): 572-583.

Tandon, P.K. (1990) Applications of global statistics in analysing quality of life data. *Statistics in Medicine*, 9, 749-763.

Thompson, S.G., and Barber, J.A. (2000) How should cost data in pragmatic randomised trials be analysed? *British Medical Journal* , 320, 1197-1200.

Troendle, J.F. (1999) Approximating the Power of Wilcoxon's Rank-Sum Test Against Shift Alternatives. *Statistics in Medicine*, 18, 2763-2773.

Troxel, A.B., Fairclough, D.L., Curran, D., and Hahn, E.A. (1998) Statistical analysis of quality of life with missing data in cancer clinical trials. *Statistics in Medicine*, 17, 653-666.

Tsodikov, A., Hasenclever, D., and Loeffler, M. (1998) Regression with Bounded Outcome Score: Evaluation of Power by Bootstrap and Simulation in a Chronic Myelogenous Leukaemia Clinical Trial. *Statistics in Medicine,* 17, 1909-1922.

Tukey, J.W. (1958) Bias and confidence in not quite large samples. *Annals of Mathematical Statistics* 29, 614.

Walters, S.J., Campbell, M.J., and Paisley, S. (2000) Systematic review of literature on methods for determining sample sizes for studies involving health-related quality of life measures. *Sheffield Health Economics Group Discussion Paper Series 00/3.* Sheffield, ScHARR, University of Sheffield. Available from: http://www.shef.ac.uk/~sheg/discussion/00_3FT.pdf).

Walters, S.J., Campbell, M.J., and Lall, R. (2001a) Design and Analysis of Trials with Quality of Life as an Outcome: a practical guide. *Journal of Biopharmaceutical Statistics*, 11(3), 155-176.

Walters, S.J., Campbell, M.J., and Paisley, S. (2001b) Methods for determining sample sizes for studies involving health-related quality of life measures: a tutorial. *Health Services & Outcomes Research Methodology*, 2, 83-99.

Walters, S.J., Munro, J.F., and Brazier, J.E. (2001c) Using the SF-36 with older adults: cross-sectional community based survey. *Age & Ageing*, 30, 337-343.

Walters, S.J, and Brazier, J.E. (2002) Sample sizes for the SF-6D preference based measure of health from the SF-36: a practical guide. *Sheffield Health Economics Group Discussion Paper Series 02/3.* Sheffield, ScHARR, University of Sheffield. Available from: http://www.shef.ac.uk/~sheg/discussion/02_3FT.pdf).

Walters, S.J., and Brazier, J.E. (2003a) What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. *Health & Quality of Life Outcomes*, 1(4), 1-8.

Walters, S.J., and Brazier, J.E. (2003b) Sample Sizes for the SF-6D Preference Based Measure of Health from the SF-36: A Comparison of Two Methods. *Health Services & Outcomes Research Methodology*, 4, 35-47.

Ware, J.E. Jr., and Sherbourne, C.D. (1992) The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care*, 30, 473-483.

Ware, J.E. Jr, Snow, K.K., Kosinski, M., and Gandek, B. (1993) *SF-36 Health Survey Manual and Interpretation Guide*. Boston MA: The Health Institute, New England Medical Centre.

Ware, J.E, Kosinski, M., and Keller, S.D. (1994) *SF-36 Physical and Mental Health Summary Scales: A User's Manual.* Boston, Health Institute.

Ware, J.E. Jr., Kosinski, M., and Dewey, J.E. (2000) *How to Score Version Two of the SF-36 Health Survey.* Lincoln, RI, QualityMetric Incorporated.

Wedderburn, R.W.M. (1974) Quasi-likelihood functions, generalised linear models and the Gaussian method. *Biometrika*, 61, 439-447.

Welch, B.L. (1947) The generalization of Student's problem when several different population variances are involved. *Biometrika*, 34, 28-35.

Westfall, P.H., and Young, S.S. (1989) P-value adjustment for multiple testing in multivariate binomial model. *Journal of the American Statistical Association*, 84, 780-786.

White, I.R., and Thomson, S.G. (2003) Choice of test for comparing two groups, with particular application to skewed outcomes. *Statistics in Medicine*, 22, 1205-1215.

Whitehead, J. (1993) Sample size calculations for ordered categorical data. [published erratum appears in *Statistics in Medicine* 1994 13(8): 871]. *Statistics in Medicine*, 12, 2257-2271.

Williamson, P., Hutton, J.L., Bliss, J., Blunt, J., Campbell, M.J., and Nicholson, R. (2000) Statistical review by research ethics committees. *Journal of the Royal Statistical Society Series A*, 163, 5-13.

Zeger, S.L., and Liang, K.-Y. (1986) Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics*, 42, 121-130.

Zeger, S.L., Liang, K.-Y., and Albert, P.S. (1988) Models for Longitudinal Data: A Generalized Estimating Equation Approach. Analysis for Discrete and Continuous Outcomes. *Biometrics*, 44, 1049-1060.