Responsiveness of Quality of Life Instruments

A thesis presented in partial fulfilment of the requirements for the degree of

Master of Applied Statistics
in
Statistics

At Massey University, Palmerston North,

New Zealand.

Mark Weatherall

2001

Abstract

Quality of life (QoL) is a phrase that is intuitively meaningful. As a concept it distinguishes between the mere duration of life and a life that is in some sense 'worthwhile'. QoL measurement is thought to be important in the assessment of chronic health conditions and their treatment. It is difficult to create an operational definition of QoL that takes into account different concepts of QoL as well as the heterogeneity of subjects and diseases. Responsiveness is one aspect of instruments which measure QoL. A responsive instrument captures the change in QoL in response to interventions which change underlying health conditions. Internal responsiveness, measured by a variety of standardised mean changes, reflects change in a QoL instrument score measured on subjects who 'should have' changed. External responsiveness relates change in a QoL instrument score to a change in external criteria. Methods of determining external responsiveness include receiver operating characteristic curves, correlation and simple regression. Simple linear regression can be extended using linear mixed models which can estimate parameters either by maximum likelihood or by Markov Chain Monte Carlo methods. This thesis critically examines methods of assessing responsiveness and demonstrates the methodology, including the extension to linear mixed models. The data set used for illustration is based on a study of subjects with rheumatoid arthritis who are assessed before and after a period of inpatient hospital treatment for their condition. Three new QoL instruments, the EuroQol, the Quality of Life Profile and the WHOQoL-Bref were found to be moderately responsive. However the available methodology and the extensions described in this thesis were unable to find any difference in responsiveness. Reasons for this could include that QoL instruments are relatively blunt instruments for the detection of change. The external criteria for change used may not have been ideal. The reasons for a choice of instrument for QoL assessment may be better related to ease of completion, interpretation and analysis, than on sophisticated assessment of responsiveness.

Acknowledgements

Thanks to Dr Will Taylor and Russell Simpson of the Wellington School of Medicine and Health Sciences for access to the data set. Thanks also to Associate Professor Steve Haslett who supervised this thesis. The data set on which the thesis is based was derived from the study 'Measuring change in Quality of Life: Three measures compared', performed at the Rehabilitation Research and Teaching Unit of the Department of Medicine, Wellington School of Medicine and Health Sciences. This study received ethical approval from the Wellington Ethics Committee.

Table of Contents

Figures

Tables

Chapter 1: Health related Quality of Life

Quality of Life and why it may be important to measure

There is no single definition of Quality of Life (QoL). It is a phrase that is often used in the context of health care and health care interventions. However, it is used to define a bewildering array of phenomena. In a study published in 1994 (1) 75 papers were randomly selected from three sources; a bibliographic listing of papers related to quality of life published to 1989 and two searches of the 'Medline' data base through to 1991. To be eligible for selection papers had to include the phrase 'quality of life' in the title and describe or use at least one questionnaire or instrument to measure quality of life. About 390 papers were eligible for random selection from a total of 1131 identified candidate papers. A total of 159 instruments were used to measure QoL in the 75 papers. The mean number of instruments per paper was 3, with a range of 1 to 19. The most frequently used instrument was used in 10 of the papers. An earlier study, published in 1991 (2), identified 67 papers describing randomized trials or observational studies from 6 prominent oncology or cardiology journals between the years 1985 and 1989. In 38 papers which used a 'serious' measure of quality of life 20 different 'validated' instruments were used.

QoL can be measured in relation to a number of contexts. For example social contexts could include the perceived prevalence of crime or the adequacy of housing (3,4). Health related QoL narrows the focus to QoL as it relates to disease states and health. The health contexts relevant for QoL measurement include measuring the health of populations, assessing the benefit of alternative uses of resources, comparing two or more interventions in a clinical trial, making a decision on treatment for an individual patient, communicating with an individual patient, medical audit and cost utility analysis (5,6,7,8). It is this context that allows the abstract concept of QoL to receive an operational definition.

In their discussion of constructing composite measurement scales Coste and colleagues (9) provide a useful summary relevant to QoL measurement. They follow standard definitions to describe *measurement* as the process of assigning a number to an object,

such as a person, in a way as to represent quantities of *attributes*. An *attribute* is a particular characteristic of an object. Attributes may be concrete and easily defined such as height and weight. In this situation measurement may be relatively easily defined. Attributes may also be more abstract, such as QoL or health, and reflect a *construct*, which is a way of conceptualising the relationship between observable or objective phenomena, such as behaviour or responses to questions, and the attribute. Constructs may be more or less hypothetical and may be composed of one or more attributes.

Attributes may be evaluated by a number of elementary criteria or items which, if they are questions, may be answered on a scale. Scales may then be used to rate an attribute, and one or more attributes may be used to determine a construct. Coste and colleagues (9) follow the 'Stevens' typology of measurement levels to state that scales may give rise to numerical information at different levels such as nominal (simple categories), ordinal (categories which are in some sense ordered), interval (the distance between objects is a meaningful concept) or ratio (the distance between objects relates to an absolute level of measurement). Often the numerical outcomes of a number of different scales are combined to give a global measurement for an attribute or construct. Some QoL instruments combine the numerical outcomes for attributes or constructs to give an overall QoL score.

Closely related to this development of the process of QoL measurement is the concept that it is necessary to select areas of health, domains or dimensions, for measurement. These can include physical, psychological and social functioning (10) and symptoms (6,11). In this sense dimensions represent constructs as defined above. Selection of particular dimensions for inclusion in any particular QoL instrument, and one of the reasons why there is no standard definition of QoL, is dependent on the reason for measuring QoL and the group of subjects who are to be measured. The particular concept of health may also differ between different researchers.

Purposes for which QoL instruments are used.

An important framework for producing health status indices distinguishes three main uses relevant to QoL instruments (8,12,13,14).

Discriminative indices aim to distinguish individuals or groups. An important use of this type of index is to quantify the burden of disease across different communities.

Predictive indices aim to determine a likely outcome or performance against a 'gold standard' measurement. For example some indices of health, such as physical function, predict outcomes such as death.

Evaluative indices aim to determine change. For example a QoL instrument might be used to measure change in a clinical trial.

This framework may determine the structure and content of a particular QoL instrument.

How QoL instruments are generated

It is generally accepted that the use of a single carefully framed question to assess quality of life will not meet the measurement needs of users of QoL instruments (8,11). However which items to include in measurement of an attribute, and how to select attributes to reflect underlying constructs is not straight forward (3,4,6,8,15).

Two important health concepts originate from the World Health Organisation.

The first is the World Health Organisation definition of health: Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity (16).

The second is the framework for defining disability, outlined in the International Classification of Impairments, Disability and Handicap, now in its second version, ICIDH-2 (17). This places disease and disease related consequences into a hierarchy. Functioning of a person is classified at the level of the body and body part, at the level of the whole person, and at the level of the whole person in a social context.

The pathological process or disease is the tissue and organ related damage caused by deterioration or an outside agent such as injury or infection.

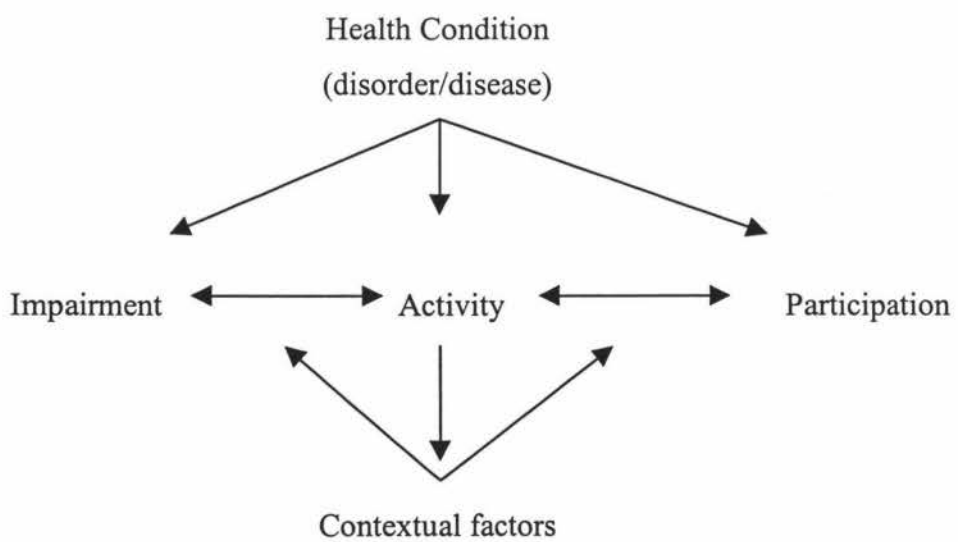
Impairment is a loss or abnormality of body structure or of a physiological function, for example loss of a limb or loss of vision.

An activity is the nature and extent of functioning at the level of the person. Activities may be limited in nature, duration and quality, for example taking care of oneself, maintaining a job. Activity limitation corresponds to the term 'disability' in the earlier version of ICIDH-2.

Participation is the nature and extent of a person's involvement in life situations in relation to Impairment, Activities, Health Conditions and Contextual Factors. Participation may be restricted in nature, duration and quality, for example, participation in community activities, obtaining a driving license. Participation restriction corresponds to the term 'handicap' in the earlier version of ICIDH-2.

Contextual factors include physical conditions such as climate or terrain, as well as aspects of the social and human built environment such as social attitudes, laws, policies, and social and political institutions. The relationships between these concepts of disablement and functioning are shown in Figure 1.

Figure 1: Relationships between health condition and disability



While QoL does not appear in this model of disablement and functioning, this philosophy of disease and disease related consequences, is often at the core of development of QoL instruments. In addition the hierarchy and causal links are sometimes modelled implicitly or explicitly in the development of QoL instruments and in analysis of their psychometric properties.

The process of generating particular items in relation to attributes and constructs is dynamically related to the proposed use of the QoL instrument and the psychometric properties in relation to that use. Psychometric properties include reliability, validity and responsiveness. Without, at present, defining these terms more precisely, an instrument which is used in an evaluative setting must have item scores, which by themselves or when aggregated, give rise to stable intra-subject variation (a part of reliability). The items must also be valid in a longitudinal sense; item scores, or their dimensional aggregate, change if QoL in the particular dimension changes. Responsiveness is also important in this setting; that the change in an item score, or aggregate score, reflects a clinically important difference (12).

Where particular constructs are complex or abstract, a number of items may be required to capture that concept, and the scores on these items are aggregated in some way to measure the construct. In part this is because different subjects scoring a particular item may have different interpretations of the terms used in the items. For example in measuring the construct of psychological health, as a part of QoL, a rating of anxiety and depression may be important. A number of items might have to be used to rate anxiety and depression to capture what different individuals interpret these words to mean.

An additional issue with item selection for a QoL instrument is that some items might be indicator variables and some causal variables. This distinction is related to the concept of the difference between psychometric scales and clinimetric indices.

Indicator variables are so called because they indicate the level of the underlying construct, such as QoL, or one of the underlying dimensions of QoL, for example psychological well being. Causal variables are so called because they reflect a model of what causes the change in the underlying construct. For example the response to a

question such as 'How are you feeling today' may be an indicator of the psychological dimension of QoL. The response to the question 'Have you vomited today' may be an indicator of nausea, people with nausea don't feel well, and this item reflects the causative sequence of a symptom which leads to distress which leads in turn to a change in QoL. Typically causal items may, by themselves, change QoL. Thus a respondent to a QoL instrument does not have to have positive responses to a complete set of causal items in order to have a reduced QoL.

The importance of this distinction is that correlation techniques are often used in constructing instruments. Correlation between causal variables may not relate to correlation with the underlying constructs, such as overall QoL. The causal variables may change together as a disease changes but this change may not correlate with the change in the underlying, or latent, variable of quality of life.

Fayers (8) relates this distinction between causal and indicator variables to the difference between psychometric scales and clinimetric indices. Psychometric scales are those instruments for which items are selected that reflect the underlying constructs. They are composed of indicator variables. Clinimetric indices are those instruments composed of causal variables, items which are thought to cause changes in QoL and are often used for prognosis or prediction. Often QoL instruments can have a mixture of both types of variables.

A final distinction important in generating QoL instruments is between generic and disease specific QoL (6,8,10,11,14,15). Generic QoL instruments are typically designed to be used in any disease state and also in healthy people. As the phrase indicates disease specific instruments are designed to reflect the particular issues of a particular disease or health state. For example disease specific indices exist for arthritis and urinary incontinence.

Generic instruments typically emphasise subjective and non-physical aspects of QoL such as emotional, social and existential issues. They also usually include one or more items that explicitly enquire about overall QoL. They have the advantage that scores from patients with various diseases may be compared with each other and with the general population.

Disease specific instruments include items that relate to a specific disease or disease related process. These items are likely to be causal items. These instruments are more sensitive to detect differences that arise during clinical trials related to a particular disease.

General issues in analysis of data derived from QoL instruments

1. QoL instruments may accumulate numerical data at different measurement levels. If a single QoL score is produced this can be by a simple or more complex weighting system. Cox and colleagues (5) advocate a simple weighting approach, more or less assigning a simple sequence of integers to the levels of ordinal response variables, and then deriving an arithmetic mean for a set of items which reflect a dimension. This approach has been examined in some detail by Coste and colleagues (18) who found that it can lead to aggregate scores which behave rather like a categorical or ordinal scale rather than a continuous measurement scale data.
2. If multiple dimensions are included in the analysis of QoL data multiple comparisons of possibly correlated data may lead to inflation of the Type I error. An added complication in deriving a global hypothesis test is that some dimensions may be measured on an ordinal scale, such as a Likert scale, and some on a continuous scale, such as a visual analogue scale. Suggestions to deal with these issues include using single global hypothesis tests related to the Hotellings T statistic, using Bonferroni or related Type I error adjustments, concentrating on a few key dimensions, using dimension reduction techniques and using random effects models (5,8,19-23).
3. If global tests of multiple outcome measures are reported it may be difficult to relate the statistical significance to clinical significance or effect size (3,4,8).
4. How to define a meaningful effect size, either in relation to the population, or to similarly diseased individuals (4,14,15).
5. In QoL data collection missing data, both item non-response and complete non-response, may not be missing at random, leading to bias. For example for long instruments items towards the end of the instrument may not be completed by people who are fatigued. If a number of QoL instruments are completed the last

completed instrument may be systematically rated differently than the first completed instrument (2,3,11,15,21).

6. Instruments may be very sensitive to interviewer bias (8).
7. QoL may change in a dynamic sense that attitudes and standards of the respondents may change over the time course that the instrument is administered, particularly in longitudinal studies (24).
8. QoL instruments may not translate very well out of the language and culture for which they were designed (4,7,8).
9. Difficulties in clarity of reporting, particularly for QoL instruments with a variety of dimensions (10).

Some criteria for QoL measurement: Validity and Reliability

Three main criteria have been advocated for development of instruments that measure QoL. These are validity, reliability and responsiveness. Discussion of responsiveness will be deferred until the next chapter.

Validity

Validity refers to whether an instrument measures what it intends to measure, in the context of this discussion, QoL. Validity can be divided into three main aspects; content validity, criterion validity and construct validity. (3,5-8,11,12,14,25)

Content validity concerns the extent to which items are sensible and reflect the intended domain of interest. Evidence for content validity can be obtained through comprehensive sampling of the domain of interest using a number of sources, such as comprehensive review of the relevant literature, consensus expert opinion, qualitative subject interviews, and examination of existing measures of the same or related constructs. A related concept of 'face validity' refers to an instrument appearing to be relevant to the area it purports to measure after the items it contains are put together.

Criterion validity considers whether items and collections of items within the instrument have an empirical association with external criteria. Concurrent validity refers to agreement with a true value, often referred to as a 'gold standard' definition of what the instrument purports to measure. For QoL there usually is no gold standard and in any case the concept of QoL is abstract. A proposed instrument can be measured at the same time as another instrument and correlation or regression used to measure the strength of association. This is perhaps most relevant if a shorter version of a long instrument is being developed. Predictive validity concerns the ability of an instrument to predict future health status, future events or future test results. The implication is that future health status can serve as a criterion against which the instrument can be compared. Predictive validity is probably better considered as a form of construct validity.

Construct validity examines the theoretical relationship of items and collections of items to each other and to the attributes they are intended to measure. It involves first forming a hypothetical model describing the constructs being assessed and postulating their relationships. Then an assessment is made as to the extent to which these relationships are supported by use of the instruments. Two aspects of construct validity are convergent validity and discriminant validity. Some items may be expected to correlate highly with each other, and display convergent validity. Some items may be expected to be measuring different attributes, and display discriminant validity, that is their correlation will be low. Analysis by correlation is, particularly in QoL scale validation, subject to 'third variable' problems. That is, items or collections of items may correlate or not because of their unsuspected relationship to a third variable, which can be an unrecognised underlying construct, or feature of the subjects being measured. Known groups validation is a form of construct validation based on the principle that certain specified groups of subjects will have different scores on the instrument used. The groups chosen for these studies are deliberately chosen to be very different so the 'statistical significance' of scores found between groups is not as relevant as if no difference could be demonstrated. In a similar way predictive validity is a form of construct validity in that people with poor QoL are usually thought of as having a worse prognosis from the underlying health problem because the underlying health problem is worse. Scores on a QoL instrument should predict this poor prognosis if the underlying way the QoL instrument was constructed is valid.

Reliability

This is the property that an instrument yields the same results on repeated administration under the same conditions. There are a number of different uses of the word 'reliability'.

Test-retest reliability refers to measurements which are repeated over time on stable subjects. If an instrument is administered by one interviewer at different times then this is also called intra-rater reliability. If the instrument is administered by different interviewers at the same time this is called inter-rater reliability. If different variants of an instrument are used then the phrases 'equivalent forms', 'parallel forms' or 'alternate

forms' are used to describe reliability. Sometimes internal consistency, which refers to multi-item scales when each item on a particular scale is attempting to measure the same attribute, is referred to as internal reliability. (6,8,12,25)

For binary data that can be put in a two by two table the kappa statistic is usually used to measure reliability (8,26) as illustrated in Table 1. The cells refer to the counts in the particular categories for, say, QoL being positive or negative when measured in the same subjects on two occasions.

Table 1: Construction of the kappa statistic

	First assessment		
Second assessment	Positive	Negative	Total
Positive	x_{11}	x_{12}	R_1
Negative	x_{21}	x_{22}	R_2
Total	C_1	C_2	N

The probability of agreement P(A) is: $(x_{11}+x_{22})/N$

The probability of chance agreements P(C) is: $((C_1\times R_1)+(C_2\times R_2))/(N)^2$

The kappa statistic is: $(P(A)-P(C))/(1-P(C))$

The closer the kappa statistic is to 1, the better the strength of agreement. There is some controversy surrounding the use of the kappa statistic (8,26), as its value can depend on the marginal distributions of the table.

For ordinal data with five or more categories or for continuous data recent reviewers suggest that the intra-class correlation coefficient (ICC) can be used to assess reliability (8,27). There are a number of different ways in which the ICC can be calculated (28-31) although they all attempt to produce a coefficient that represents the proportion of the variance of a measurement that relates to subject variability compared to total measurement variability that can include subject variability, rater variability, and random error. A reliable instrument has most of its variance explained by between subject variability. As an example a set of 'I' subjects, who are in a stable state might each be measured once, using a QoL instrument by the same set of 'K' raters, who are randomly selected from a population of raters. This leads to a total of IK measurements. A random effects model for this data could be (30):

$$Y_{ik} = \mu + s_i + r_k + \varepsilon_{ik}$$

$$s_i \sim N(0, \sigma^2_s)$$
$$r_k \sim N(0, \sigma^2_r)$$
$$\varepsilon_{ik} \sim N(0, \sigma^2)$$
$$Y_{ik} \sim N(\mu, \sigma^2_s + \sigma^2_r + \sigma^2)$$

With s_i the random subject effect and r_k the random rater effect.

The ICC for this model is: $R = \sigma^2_s/(\sigma^2_s + \sigma^2_r + \sigma^2)$

This must be estimated from the data and this is done from an analysis of variance table as illustrated in Table 2.

Table 2: Construction of ICC, first example

Source	**DF**	**MS**	**E(MS)**
Subjects	I-1	MSB	$K\sigma^2_s + \sigma^2$
Raters	K-1	MSR	$I\sigma^2_r + \sigma^2$
Error	(I-1)(K-1)	MSE	σ^2

And the ICC is estimated by:

(MSB-MSE)/[MSB + (K-1)MSE + (K/I)(MSR-MSE)]

If this coefficient is close to one it means that most of the variance of measurement of subjects is due to variation between subjects.

In a similar way an intra-rater ICC can be calculated. As an example a set of 'I' subjects could be rated by the same rater on 'R' occasions. This leads to a total of IR observations. A random effects model for this data could be:

$Y_{ir} = \mu + s_i + \varepsilon_{ir}$

$s_i \sim N(0, \sigma^2_s)$

$\varepsilon_{ir} \sim N(0, \sigma^2)$

$Y_{ir} \sim N(\mu , \sigma^2_s + \sigma^2)$

With s_i the random subject effect.

The ICC for this model is: $R = \sigma^2_s/(\sigma^2_s + \sigma^2)$

This must be estimated from the data and this is done from an analysis of variance table as illustrated in Table 3.

Table 3: Construction of ICC, second example

Source	**DF**	**MS**	**E(MS)**
Subjects	I-1	MSB	$R_o\sigma^2_s + \sigma^2$
Occasions (Error)	IR_o-I	MSE	σ^2

And the ICC is estimated by:

(MSB-MSE)/(MSB + (R_o-1)MSE)

As taking the expected value of each term yields:

$$(R_o\sigma^2_s + \sigma^2 - \sigma^2)/(R_o\sigma^2_s + \sigma^2 + (R_o - 1)\sigma^2) = \sigma^2_s/(\sigma^2_s + \sigma^2)$$
$$= R$$

In this particular example it is assumed that the covariances between repeated measures on the same subjects are zero. This is unlikely to be true. These covariances could be modelled using a mixed linear model, as will be described later in this thesis.

Which ICC is calculated depends on how the reliability data is collected and which sources of measurement error are to be accounted for. In the review article by Deyo (27) and book by Fayers (8) an ICC calculation is presented that has the form of the ICC from Table 2. This approach seems flawed. Instead of using a random sample of raters to rate the subjects both authors state that QoL is often self assessed and, although not stated by either author, this clearly means that rater effects are completely confounded with subject effects. Both authors then go on to use the occasion on which each subject was measured as a random effect. They both use an 'occasion' effect in the place of the 'rater' effect in Table 2. Neither of the reviews explicitly state which random effects model underlies their calculation for the ICC of I subjects measured on K occasions. Given that they use the form of the ICC from Table 2 it must be of the form:

$$Y_{ik} = \mu + s_i + o_k + \varepsilon_{ik}$$

$$s_i \sim N(0, \sigma^2_s)$$
$$o_k \sim N(0, \sigma^2_o)$$
$$\varepsilon_{ik} \sim N(0, \sigma^2)$$
$$Y_{ik} \sim N(\mu, \sigma^2_s + \sigma^2_o + \sigma^2)$$

Where s_i represents subject effects and o_k represents measurement occasion effect.

The flaws are firstly that the 'occasion' effect is not a random selection of measurement occasions from a larger population of measurement occasions and it therefore should be treated as a fixed effect, thus not contributing to the expected mean square error terms used to generate the ICC. Secondly even if measurement occasions were a random

effect the same argument regarding the likely correlation between successive measurements on the same subjects means that the assumption of zero correlation between error terms is unlikely to be true.

Thus while ICC may have a role to play in establishing reliability close attention must be paid to which sources of variation are important, the underlying statistical model for the measurement variance, and that the mode of data collection reflects both these issues.

Chapter 2: Internal Responsiveness

A third criterion, highlighted by review articles, for the development of QoL instruments is responsiveness. Responsiveness is the ability of an instrument to detect changes (6,8,25,27,32). The phrase 'sensitivity to change' is also used to describe this aspect of QoL instruments (6,8). However the term 'sensitivity' is also used in the analysis of responsiveness, particularly in the context of receiver operating characteristic (ROC) curves. I will use the term 'responsiveness'. Five suggested responsiveness indices are summarised in Table 4. The meanings of the various subscripts are summarised in Table 5. The individual responsiveness indices are then discussed in more detail.

Table 4: Different responsiveness indices

Index	Synonyms	Derivation
RI_1	Guyatt responsiveness index	$CID/(2\times MSE)^{1/2}$
RI_2	Paired t statistic	d_i/s_{di}
RI_3	Relative efficiency index	$(d_i/s_{di})^2/(d_j/s_{dj})^2$
RI_4	Standardised effect size	d_i/s_b
RI_5	Standardised response mean, responsiveness treatment coefficient and the efficiency index	d_i/s_d

Table 5: Abbreviations used in Table 4

Abbreviation	Description
CID	Clinically important difference
MSE	In a random effects model for repeated measures of a QoL instrument the residual mean square error after the random effects of subjects are taken into account (MSE in Table 3)
d_i, d_j	The mean difference for the i^{th} QoL instrument measured on two occasions on the same set of subjects, and the mean difference for the j^{th} QoL instrument measured on the same two occasions on these same set of subjects
s_{di}, s_{dj}	Standard deviation of the differences for the i^{th} QoL instrument measured on two occasions on the same set of subjects, and the standard deviation of the differences for the j^{th} QoL instrument measured on the same two occasions on these same set of subjects, divided by the square root of the sample size
s_b	Standard deviation of the score on a QoL instrument on the 'before' occasion when measured on two occasions, before and after an intervention that improves QoL
s_d	Simple standard deviation of the difference on a QoL instrument measured on two occasions

RI_1: Guyatt responsiveness index

The concept of responsiveness was first raised in the context of QoL instruments in the paper by Kirshner and Guyatt (12) and clarified in subsequent papers (13,33,34). In their initial paper (12) responsiveness is identified as an important aspect of an evaluative instrument, one that is used in a clinical trial as an outcome measure. They use the phrase 'the power of the index to detect a difference when one is present'. It is clear from their subsequent discussion, of whether the index can detect small, medium or large effects in the context of sample size, that responsiveness is related to minimising type II error in a clinical trial. Their suggested strategies for evaluating responsiveness include testing that scores on an instrument improve with application of a treatment of known efficacy and use of the instrument in a clinical trial followed by examination of change scores in those who, by other criteria, improved or deteriorated. This concept was elaborated in a second paper (33). The elaboration in the definition of responsiveness suggested that responsiveness relates to clinically important differences. That is that a more responsive instrument should be capable of detecting clinically important differences, declaring them to be statistically significant, for example between two groups in a clinical trial, with the smallest numbers allocated to the treatment arms. They use a sample size formula to illustrate their discussion. For paired observations, i.e. in the same subjects before and after an intervention where it is the differences in a QoL instrument that are being analysed, the sample size formula used by Guyatt and colleagues (33) was:

$$N = [(Z\alpha + Z\beta)\sigma/\Delta]^2$$

Where N is the total number of paired observations in the intervention group (assumed to be equal), $Z\alpha$ is the normal deviate for the type I error rate, $Z\beta$ is the normal deviate for the type II error rate, σ is the estimate of the standard deviation of the differences, and Δ is the size of the difference to be detected with the nominated type I and type II error rates.

For any nominated type I and type II error rate the sample size depends on the ratio σ/Δ. The smaller the ratio, the smaller the sample size needed to detect a difference. By

inverting the ratio, the larger the value of Δ/σ for any instrument the more responsive that particular instrument.

Guyatt and colleagues (33) suggested a responsiveness index:

$$RI_1 = CID/(2\times MSE)^{1/2}$$

Where CID is the change in an instrument score that is thought to constitute a 'clinically important difference' and represents 'Δ' in their sample size formula. Their use of the term $(2\times MS_E)^{1/2}$ to represent the estimate of σ is ambiguous and may be erroneous. In the original paper (ibid. p174) its derivation is described:

"....-therefore responsiveness is inversely proportional not to within-person standard deviation, but to the between subject variability of the individual changes in score over time."

And from later in the paper (ibid. p175)

"In an analysis of variance model examining repeated test observations in stable subjects, the between subject variability in within-person change in score is represented by the square root of twice the mean square error $[(2\times MSE)^{1/2}]$ (assuming errors are independent of one another)."

No analysis of variance calculations or tables with expected mean square errors are presented to exactly illustrate how these confusing statements relate to an underlying statistical model. In a later paper where the same principal author and other colleagues actually calculate RI_1 (35) they used the standard deviation of the differences of subjects who did not change between consecutive visits at which a QoL instrument was administered, as the denominator in RI_1. This corresponds to the term 's_d' in table 5.

I speculate that the confusion arises when considering an analysis of variance to calculate the standard error of a contrast between the mean effect two treatments in a randomised clinical trial as illustrated in Table 6. This represents analysis of the

differences between I subjects measured on two occasions randomly allocated to one of J treatments with equal numbers in all groups.

Table 6: Speculated derivation of $(2\times MSE)^{1/2}$ for RI_1

Source	**DF**	**MS**	**E(MS)**
Subject differences	I-1	Between subjects	$J\sigma^2_s + \sigma^2$
Treatment	J-1	Between treatments	$[I/(J-1)]\Sigma t^2_j + \sigma^2$
Error	(I-1)(J-1)	MSE	σ^2

The underlying model is:

$Y_{ij} = \mu + s_i + t_j + \varepsilon_{ij}$

$s_i \sim N(0, \sigma^2_s)$

$\varepsilon_{ij} \sim N(0, \sigma^2)$

$Y_{ij} \sim N(\mu + t_j, \sigma^2_s + \sigma^2)$

With the additional constraint that $\Sigma\, t_j = 0$

The difference, or contrast, between the mean values of subject differences for two treatments, say m_1 and m_2, can be tested using a t statistic:

$C = (m_1 - m_2)/((2/n)\times MSE)^{1/2}$ (in this case n = I)

And this can be compared to a t distribution with (I-1)(J-1) degrees of freedom or equivalently the square of the value with an F distribution with 1 and (I-1)(J-1) degrees of freedom. The term $(2\times MSE)^{1/2}$ arises in this context of testing if a contrast is zero, or not, based on the error term in table 6. The confusion in the use of the term 'MSE' in RI_1 appears to be that Guyatt and colleagues use the standard deviation of the differences between subjects as though it is equivalent to the square root of twice the mean square error, in an ANOVA. The two are not equivalent.

If the situation is considered where a QoL instrument score designated by Q_{ij} is measured on I subjects on J occasions, where J equals 1 or 2. The difference between the scores on the two occasions, d_i and the mean difference, d_m , are given by respectively:

$d_i = Q_{i2} - Q_{i1}$ and $d_m = (\Sigma(Q_{i2} - Q_{i1}))/I$

The standard deviation of the difference is given by:

$S = [(\Sigma(d_i - d_m)^2)/(I-1)]^{1/2}$

This can be expressed equivalently as:

$S = [\Sigma d_i^2/(I-1) - (\Sigma d_i)^2/(I(I-1))]^{1/2}$

Now if an ANOVA is carried out as illustrated in Table 7:

Table 7: Derivation of MSE, discussion of RI_1

Source	**DF**	**SS**	**MS**	**E(MS)**
Subjects	I-1	$\Sigma(Q_{i.} - Q_{..})^2$	Between subjects	$I\sigma^2_s + \sigma^2$
Error	I	$\Sigma\Sigma(Q_{ij} - Q_{i.})^2$	MSE	σ^2

Where $Q_{i.}$ is $(Q_{i2} + Q_{i1})/2$ and $Q_{..}$ is $\Sigma\Sigma\ Q_{ij}/2I$

Now $\Sigma\Sigma(Q_{ij} - Q_{i.})^2 = \Sigma[(Q_{i1} - (Q_{i2} + Q_{i1})/2)^2 + (Q_{i2} - (Q_{i2} + Q_{i1})/2)^2]$

$= \frac{1}{4}\ \Sigma[(Q_{i1} - Q_{i2})^2 + (Q_{i2} - Q_{i1})^2]$

$= \frac{1}{2}\ \Sigma d_i^2$

Then with measurement on two occasions MSE is equal to $\Sigma d_i^2/2I$ and hence $(2\times MSE)^{1/2}$ is equal to $[\Sigma d_i^2/I]^{1/2}$. From this it can be seen that the standard deviation of the differences of a set of subjects measured on two occasions will always be less than the term $(2\times MSE)^{1/2}$, based on the analysis of variance because:

$[\Sigma d_i^2/(I-1) - (\Sigma d_i)^2/(I(I-1))]^{1/2} < [\Sigma d_i^2/I]^{1/2}$

Another difficulty in actually applying this index to compare different instruments is to define the CID. Within any single clinical trial which used two or more instruments the ratio of mean treatment difference to $((2/n)\times MS_E)^{1/2}$, when compared to the appropriate t or F distribution, will give information in that trial as to which instrument was more responsive. This is because the putative change in QoL will be the same because subjects were exposed to the same intervention, and the instrument which gave rise to a more statistically significant difference between treatments is the most responsive one. However the difficulty lies in determining between different studies, instruments and interventions what the CID is and furthermore across different studies knowing how comparable the denominator is or indeed how it is calculated if actual details of the statistical models underlying its generation are not given. Overall therefore RI_I does not appear to be a good candidate for a responsiveness index.

An attempt has been made to derive the statistical distribution of RI_I (36) however this was predicated on estimating the difference in a QoL score from a treatment trial as the numerator in the index, which was then nominated as the CID, and using the standard deviation of the differences between the same QoL instrument measured in a different group of stable subjects on two occasions as the denominator. The derivation of this distribution is complex and relies on a number of approximations and assumptions and will not be reproduced here. Tuley and colleagues attempt to illustrate the use of the derived distribution to compare RI_I for two different QoL instruments. The example they use is the difference in score on two QoL instruments, designed to measure QoL in people with hearing impairment, administered 4 months apart in groups randomised to no intervention or provision of a hearing aid. The trial may have been susceptible to bias as it was not placebo controlled and no description is given of blinding of the subjects or researchers to treatment allocation. However given that the same group of subjects were administered the two QoL instruments the straightforward way to assess responsiveness would have been to see, in an ANOVA, which calculation of the statistical significance between active treatment and placebo had the lowest 'P value' for the two measurement instruments.

Calculation of the statistical distribution of RI_1 seems a futile exercise. Firstly the CID is not just the difference between a set of subjects before and after an intervention but a constant nominated difference it is important to detect. Secondly there is confusion even amongst the originators of the index as to what the denominator should be. Finally even if it was calculated in the terms of Tuley and colleagues little reliance can be placed on using the measure of variance for a 'before and after' difference for one set of subjects and applying it to a completely different set of subjects.

RI_2: Paired t test statistic

A second responsiveness index, RI_2, can be derived from a study of a single group undergoing a therapeutic intervention, a one group repeated measure design.

$RI_2 = d_i/s_{di}$,

where d_i is the mean of the differences for each subject measured on the two occasions and s_{di} is the estimate of the standard deviation, i.e. the standard error, of this mean difference adjusted for sample size i.e. divided by the square root of the number of paired subjects, a paired t test statistic. An instrument for which the mean score changed, i.e. the value of RI_2 was greater than the appropriate value of a t distribution, would have some evidence for responsiveness. However in comparing instruments, if for some reason the sample size for those who completed various instruments were different, the 'size' of the paired t statistic would also be different, and this would not be a valid comparison of the responsiveness of different instruments.

RI_3: The 'relative efficiency ' statistic

A third responsiveness index, RI_3 , is a 'relative efficiency' statistic. It is the square of the ratio of paired t statistics for two different instruments, measured on the same group of subjects. The 'relative efficiency' statistic can be used to rank tests. An example of the use of this statistic is in a paper comparing 5 different health status instruments (37). However use of this index is dependent on the same number of subjects completing each of the instruments to be compared. The reason for squaring the t statistic is so that

QoL instruments which 'run' in different directions can be compared. The distribution of RI_3 is also difficult to determine.

RI_4: The 'standardised effect size'

A fourth responsiveness index, RI_4, also known as the 'standardised effect size', can be derived:

$$RI_4 = d_i/s_b,$$

where d_i is the mean of the differences for each subject measured on the two occasions, before and after an intervention, divided by s_b, the simple standard deviation of the measure taken at the 'before' time. Examples of its use are in papers by Beaton (38) and Fitzpatrick (39) . In this second paper the value for d_i was derived only for subjects who improved based on an external criterion, a global question as to whether or not they had improved. The derivation of the value of s_b was not explicitly stated, although implied that it was taken as the standard deviation of the baseline value for all subjects, whether they subsequently improved or not.

RI_5: The 'standardised response mean'

A fifth responsiveness index, RI_5 , also known as the 'standardised response mean', 'responsiveness treatment coefficient' and the 'efficiency index'.

$$RI_5 = d_i/s_d,$$

where d_i is the mean of the differences for each subject measured on the two occasions, before and after an intervention, divided by s_d, the simple standard deviation of the differences. This differs from RI_2 , the paired t statistic, by not adjusting for sample size. Examples of the use of this statistic include the paper by Beaton (38) and Wright (40).

Other issues with internal responsiveness indices

The review paper by Husted and colleagues refers to these measures of responsiveness as internal responsiveness (32). By this they mean the ability of an instrument to detect change over a specified time period when change should have occurred. There are a number of further issues with the use of these responsiveness indices.

1. One group repeated measures designs to evaluate responsiveness

Responsiveness is often measured in a one group repeated measures setting. The mediator of change is usually a therapeutic intervention, demonstrated in a different study to be efficacious. It can be just measuring subjects after sufficient time has passed to justify the assumption that change should have occurred naturally. In this setting a statistically significant change in a QoL instrument score may occur without a corresponding change in clinical or health status. The assumptions that a therapeutic intervention that has been found efficacious in another study in a different group of subjects will also be efficacious in the study of responsiveness, or that the passage of time will cause a change in health status, are difficult to evaluate in the absence of some sort of control group. There are ethical issues regarding subject recruitment to studies of responsiveness. It may be difficult to justify a placebo group where there is already evidence of efficacy, merely to confirm responsiveness of a new QoL instrument.

2. Ceiling and floor effects for QoL instruments

A QoL instrument may have a ceiling or floor effect. By this is meant that there is a limit to how high or low the scores achieved on an instrument can go. This limit may mean that a QoL instrument applied to subjects with a different baseline level of QoL may measure different levels of responsiveness. If an instrument is administered to a group of subjects with poor QoL, and low scores on a particular instrument, the measurement on the instrument can improve. If the same instrument is administered to a group of subjects close to the ceiling score of the instrument it may not be able to improve, there is no where for the score to go. Different studies of the same instrument based on different subjects will then reach different conclusions about its responsiveness.

3. Retrospective evaluation of change to evaluate responsiveness

Some studies evaluate responsiveness by use of a QoL instrument in a single group of subjects before and after a therapeutic intervention (35). The numerator used is the mean difference in the QoL instrument score for those subjects who improved by an external criterion. The denominator is the standard deviation of the difference in the QoL instrument scores for subjects who did not improve. This statistic resembles RI_5, but uses a retrospective evaluation of change to evaluate responsiveness. This methodology leads to quite different judgements about responsiveness compared to a responsiveness assessment that compares the mean change of the whole group to the standard deviation of the mean change of the whole group. Simulation studies by Norman and colleagues (41) used data generated from a normal distribution in two situations. The first where there was no change in the generated scores for an instrument between the two measurements and the second where there was. An external criterion for change was used with a nominated correlation with the QoL instrument. This simulation study found that even when the underlying distribution showed no change between the two measurements because its simulated mean was set at zero, an instrument could be made responsive by using the retrospective criterion. They then went on to show for studies they could locate in which the responsiveness index could be generated in both ways that the retrospective method almost always inflated responsiveness compared to using all the data.

The underlying basis for this seems easy to demonstrate, although this was not done along side the simulation studies described above.

Consider a single group of N subjects who are measured twice, before and after an intervention which improves QoL. Then for the i^{th} subject:

$d_i = \mu + \alpha I_i + \varepsilon_i,$

where,

d_i is the difference in the QoL instrument score measured on two occasions.

μ is the overall mean difference in QoL measured on two occasions.

α is a parameter describing the additional improvement if a subject's QoL has changed based on the external criterion.

I is an indicator variable for whether the external criterion for QoL has changed and is equal to 1 if it has and 0 otherwise.

ε_i is the error term and is distributed as $N(0, \sigma^2)$.

Further let J subjects improve based on the external criterion and K subjects stay the same such that $N = J + K$.

Then similar to the notation of Table 4 and Table 5:

$RI_5 = d_m/s_d$,

where d_m is the mean difference in QoL measurements for all subjects and s_d is the standard deviation of these differences.

Now if RI_5 is derived using as the numerator the mean difference in those that changed and the denominator the standard deviation of the differences for those that didn't change, both based on an external criterion, the index will be inflated if the numerator is larger and the denominator smaller.

If the numerator is considered first. Its expected value for the whole group is $\mu + (J/N)\alpha$, and for those who change by the external criterion $\mu + \alpha$. As the latter will usually be larger the numerator for RI_5 derived in this way will be larger.

For the denominator the sum of squares for the differences based on the whole group, from which the standard deviation is derived, can be designated SS_T and this can be decomposed as follows:

$SS_T = SS_J + SS_K$

SS_J is the sum of squares for the d_i for the J subjects who improve by the external criterion.

SS_K is the sum of squares for the d_i for the K subjects who do not improve by the external criterion.

Now consider if SS_K is used to derive the denominator. The denominator for the responsiveness index will be smaller, and the value for the responsiveness index larger, under the following condition:

$SS_K /(K-1) < SS_T /(N-1)$

If there are equal numbers in each group then, the responsiveness index will be larger if:

$SS_K < SS_J$

Thus if the subjects who do not improve are less variable than the subjects who do improve this will lead to inflation of the response index.

Returning to the paper by Norman and colleagues (41) where RI_5 was evaluated by the two methods for a selection of studies. Of 29 standard deviations from eight studies only 6 standard deviations based on the subjects in the unchanged group were larger than those for the group as a whole. Empirically subjects who are designated as changed in a therapeutic situation are usually more variable than those who do not, at least for this data.

4. How should the clinically important difference be derived?

For the Guyatt statistic the actual derivation of the clinically important difference or minimal clinically important difference may be quite difficult. The suggestion by Guyatt and colleagues (33) that clinical experience and repeated clinical trials using the same instrument could be used suggest that the clinically important difference may

always be some what subjective. Jaeschke and colleagues (42) presented a study of seven point Likert scales in comparison to global statements by patients with either heart or lung disease, concerning changes in their health in the setting of clinical trials of therapy. Their analysis suggested that a that a 0.5 unit change on their seven point Likert scales may have represented a minimal clinically important difference. However whether their findings could be generalised to other diseases and differently constructed Likert scales is entirely conjectural. Other suggestions in reviews of this area (32,43,44) that changes of 0.2, 0.5 and 0.8 standard deviations of baseline variation represent 'small', 'moderate' and 'large' responsiveness are not capable of proof.

5. Dimensions and summary scores for QoL instruments may disagree

Some comparisons of the responsiveness of different QoL instruments rank instruments by comparing the responsiveness to change of different dimensions of individual instruments (39) and in addition by ranking by global instrument scores as well as dimension scores (37). It is uncertain how to interpret results if these different ways of comparing responsiveness disagree. There are fundamental issues in ranking the responsiveness of dimension scores of different instruments and also in comparing global and dimension scores. These issues largely relate to the construction of instruments. For different instruments the definition of the dimensions may have been arrived at in quite different ways so that although a subsequent researcher may give particular dimensions of different instruments common labels, for example activities of daily living, social function, mobility and so on, it is unlikely that the responsiveness rankings of all the dimensions of different instruments are the same by virtue of their heterogeneous construction. For the global comparison the problem of comparing underlying dimensions and the global score is even more difficult. The global score is a weighted sum of the underlying dimension scores purporting to measure the overall construct of 'quality of life'. The global score depends on which weights are assigned to each dimension and this is often a function of how the individual instrument is constructed, including which sample from which population was used to construct the instrument. Consider the hypothetical situation where each of two instruments, Q_1 and Q_2, have three dimensions, A, B and C. For Q_1 dimensions A and B may be less responsive than the corresponding dimensions for Q_2, and dimension C may be more responsive. If the weights given for the global score for Q_1 for dimensions A and B are

small and that for dimension C is large compared to the same weights for Q_2, then Q_1 will be globally more responsive even though for two of the underlying dimensions, out of three, it is less responsive.

This issue might be less problematic if only the global score is used when QoL instruments are used in clinical trials. The issue of differences in ranking for responsiveness based on dimensions versus the global scores does not then arise.

6. Lack of generalisation of responsiveness assessment

Conclusions about responsiveness, when measured by effect size, may be limited only to the specific groups of subjects and particular interventions carried out on those subjects and may not be able to be generalised to other situations (45).

Chapter 3: External responsiveness: Receiver operating characteristic curves

Husted and colleagues (32) refer to external responsiveness as the ability of an instrument to change over a specified time frame and that this change relates to a change in a reference measure of health status. External responsiveness is related to criterion validity. The distinction is that when assessing validity the comparison with the external criterion occurs in a 'static' situation, that is the QoL instrument and the external criterion are measured once and at the same time in a group of subjects. When external responsiveness is assessed it is in the context of a single group repeated measures design where the QoL instrument and the external criterion are measured before and after intervention and assessed as to whether they change in the same direction. In some studies the external criterion may be a single question such as 'Since last seen has your quality of life or health changed?'. Three methods for assessing external responsiveness are receiver operating characteristic curves, correlation, and regression models.

Receiver operating characteristic curves

Receiver operating characteristic (ROC) curves were first used in the analysis of QoL data by Deyo and colleagues (46). They are well established in medical applications regarding diagnostic testing (47-53).

In the setting of diagnostic testing a test, measured either on a continuous or an ordinal scale, is applied to a set of subjects. By some external criterion the subjects can be in one of two mutually exclusive states, for example diseased and non-diseased or abnormal and normal. The score on the test is used to discriminate between the two states. The higher the score on a test the more likely that a particular subject is abnormal. In an analogy to the terms 'signal' and 'noise' in other measurement applications the distributions of the test scores in the abnormal and normal states are treated as random variables with an overlap in their distributions. Thus for any arbitrary cut off score on the test the subjects can be allocated to a test determined state and a true state, the latter based on the external criterion. The external criterion is often termed the 'gold standard'.

This allocation can be summarised in general terms in a table, illustrated in table 8.

Table 8: Allocation of subjects by diagnostic testing

	True state of subjects by external criterion	
Test determined state of subjects	Abnormal	Normal
Abnormal	A	B
Normal	C	D

The total number of subjects is the sum (A+B+C+D). From this table a number of derived variables are used to determine the performance of the test.

The sensitivity of the test is the proportion A/(A+C), sometimes called the true positive fraction or rate. It is the proportion of subjects who are abnormal by the external

criterion who are labelled by the test as abnormal. It depends only on the truly abnormal subjects.

The specificity of the test is the proportion D/(B+D), sometimes called the true negative fraction or rate. The complement of the specificity, 1-(D/(B+D)), which is equivalent to B/(B+D), is called the false positive fraction or rate. The false positive fraction is the proportion of subjects who are normal by the external criterion who are labelled by the test as abnormal. It depends only on the truly normal subjects.

Sensitivity, specificity (and its complement) range only between 0 and 1. The plot, over all possible cut off values for a test, of sensitivity and the complement of specificity, is called the receiver operating characteristic (ROC) curve. As an illustration the data from a paper by Hanley (54) is used. The scans of 109 subjects were read by a single radiologist. The scans were rated by the radiologist on a five point ordinal scale from definitely normal (1) to definitely abnormal (5) and are illustrated in table 9. By an external criterion, or gold standard, it was known whether each subject was normal or abnormal.

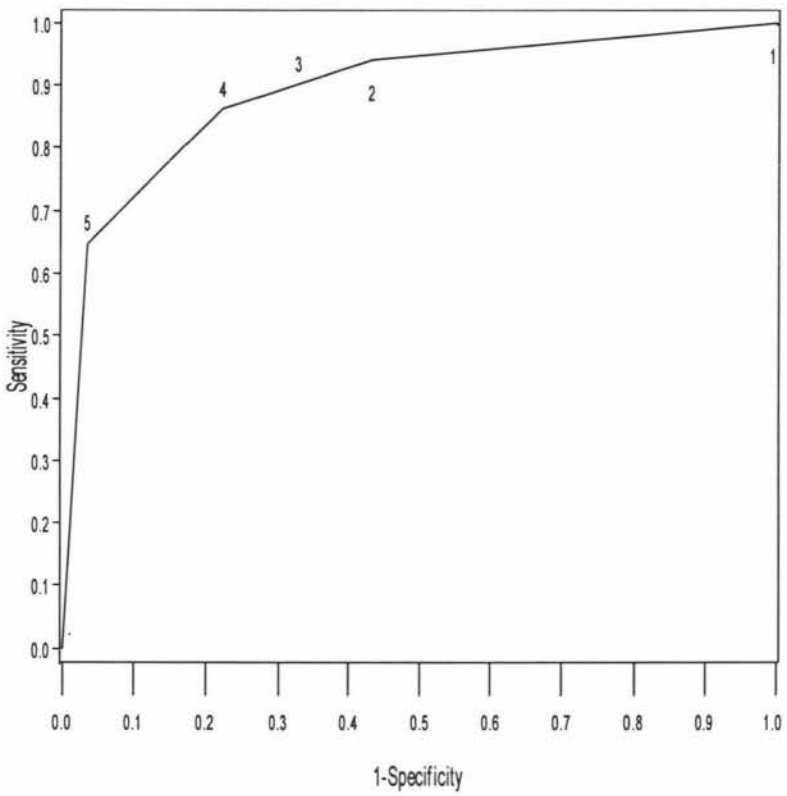
Table 9: Data for illustration of ROC curve

	True state of scans by external criterion	
Radiologists determined state of scans	Abnormal	Normal
Definitely abnormal (5)	33	2
Probably abnormal (4)	11	11
Questionable (3)	2	6
Probably normal (2)	2	6
Definitely normal (1)	3	33

To illustrate sensitivity and specificity at one particular cut off level the rating for definitely abnormal (5) will be used. At this cut off level the sensitivity is 33/(33+11+2+2+3) or 0.65. The specificity at this cut off level is

(33+6+6+11)/(33+6+6+11+2) or 0.97. The ROC curve for all possible cut off values, one through five, is shown in figure 2.

Figure 2: ROC example



This curve is an important summary of the performance of a test. A test which cannot discriminate between normal and abnormal subjects will have a plot close to the line on the unit square joining (0,0) and (1,1). A perfect test is shaped like a rectangle across to the left of the unit square. The most commonly used overall numerical summary of the performance of a test over all possible cut off values is the area under curve (AUC) for the ROC curve. This area is between 0.5 and 1. When the AUC ROC curve is calculated by the trapezoidal rule then there is a close relationship with the Mann Whitney U statistic (55).

A more explicit illustration of this than is provided in the paper by Bamber (55) is provided here. Let X represent a continuous random variable for the value of a diagnostic test for a population of non-diseased subjects, N_X , and let Y represent a continuous random variable for the value of a diagnostic test for a population of diseased subjects, N_Y.

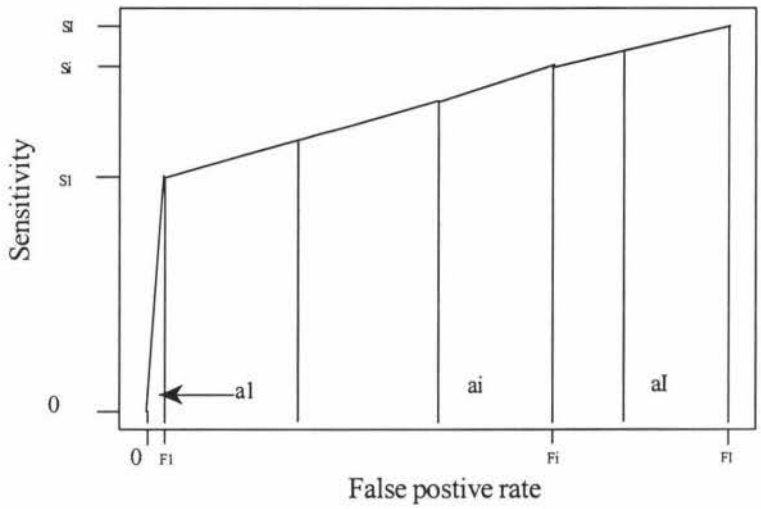
Let C be the set of I discrete values that lie within the range that can be achieved by X and Y with non-zero probability and ordered such that $\{c_1>c_2>,\ldots,> c_{I-1}>c_I\}$ for this diagnostic test.

Now let a random sample from each of the non-diseased and diseased populations is taken of size m , $\{x_1 ,\ldots, x_m\}$, and n, $\{y_1 ,\ldots, y_n\}$, respectively.

The sensitivity of a test at a boundary point, c_i , is the number of times that y , one of the samples from N_Y, is greater than or equal to c_i , divided by n, designated, S_i. The false positive rate for a test at a boundary point, c_i , is the number of times that x , one of the samples from N_X, is greater than or equal to c_i , divided by m, designated, F_i. These sample derived values are unbiased estimators of the corresponding population parameters. The ROC curve constructed from these sample based estimators is called the empirical ROC curve and is in turn an unbiased estimator of the population ROC curve.

The empirical ROC curve has the form illustrated in figure 3.

Figure 3: The form of the empirical ROC Curve



The AUC ROC curve is found by summing the areas of the trapezoids $\{a_1, \ldots, a_I\}$. Consider the i^{th} trapezoid of the AUC ROC curve:

$$a_i = S_{i-1}(F_i - F_{i-1}) + (\tfrac{1}{2})(S_i - S_{i-1})(F_i - F_{i-1})$$

Where:

S_i is the sensitivity at the i^{th} boundary point

$$S_i = \sum_{j=1}^{n} [D_j/n] \qquad D_j = 1 \text{ if } y \geq c_i$$

$$D_j = 0 \text{ if } y < c_i$$

$$S_{i-1} = \sum_{j=1}^{n} [D_j^*/n] \qquad D_j^* = 1 \text{ if } y \geq c_{i-1}$$

$$D_j^* = 0 \text{ if } y < c_{i-1}$$

F_i is the false positive rate at the i^{th} boundary

$$F_i = \sum_{k=1}^{m} [E_k/m] \qquad E_k = 1 \text{ if } x \geq c_i$$

$$E_k = 0 \text{ if } x < c_i$$

$$F_{i-1} = \sum_{k=1}^{m} [E_k^*/m] \qquad E_k^* = 1 \text{ if } x \geq c_{i-1}$$

$$E_k^* = 0 \text{ if } x < c_{i-1}$$

$$(S_i - S_{i-1}) = \sum_{j=1}^{n} [D_j'/n] \qquad D_j' = 1 \text{ if } y = c_i$$

$$Dj' = 0 \text{ otherwise}$$

$$(F_i - F_{i-1}) = \sum_{k=1}^{m} [E_k'/m] \qquad E_k' = 1 \text{ if } x = c_i$$

$$E_k' = 0 \text{ otherwise}$$

Boundary conditions are that:

$F_I = 1$

When $S_i = S_1$ then $S_{i-1} = 0$ and when $F_i = F_1$ then $F_{i-1} = 0$

Now

$$a_i = (1/mn)\left(\sum_{j=1}^{n} [D_j^*] \sum_{k=1}^{m} [E_k'] + (\tfrac{1}{2}) \sum_{j=1}^{n} [D_j'] \sum_{k=1}^{m} [E_k']\right)_i$$

$$= (1/mn)\left(\sum_{j=1}^{n} \sum_{k=1}^{m} [D_j^* E_k' + (\tfrac{1}{2}) D_j' E_k']\right)_i$$

But

$D_j{}^*E_k{}' = 1$ iff $Y_j > X_k$ and 0 otherwise at the i^{th} boundary

And

$D_j{}'E_k{}' = 1$ iff $Y_j = X_k$ and 0 otherwise at the i^{th} boundary

Now let

$Z_{jk} = 1$ if $Y_j > X_k$

$= ½$ if $Y_j = X_k$

$= 0$ if $Y_j < X_k$

Then

$$a_i = (1/mn)(\sum_{j=1}^{n} \sum_{k=1}^{m} [Z_{jk}])_i$$

$$AUC = \sum_{i=1}^{I} a_i$$

$$= (1/mn) \sum_{i=1}^{I} (\sum_{j=1}^{n} \sum_{k=1}^{m} . [Z_{jk}])_i$$

But as the boundaries are discrete and mutually exclusive:

$$AUC = (1/mn) (\sum_{j=1}^{n} \sum_{k=1}^{m} [Z_{jk}])$$

$$mn\ AUC = (\sum_{j=1}^{n} \sum_{k=1}^{m} [Z_{jk}])$$

The term on the right is the Mann Whitney U statistic, using an adjustment for ties to make it suitable for use in the situation of discrete, rather than continuous, underlying variables.

Bamber (55), who is often cited as recognising this relationship, uses a different notation, in part because the initial graph he plots is the so called ordinal dominance graph which has the sensitivity and false positive rate on the X and Y axes respectively, an inverse but equivalent problem.

The correspondence between Bamber's notation and that described above is that:

S_i is referred to as $P(Y \geq c_i)$, the sample probability that $Y \geq c_i$
$(S_i - S_{i-1})$ is referred to as $P(Y=c_i)$, the sample probability that $Y = c_i$

F_i is referred to as $P(X \geq c_i)$, the sample probability that $X \geq ci$
$(F_i - F_{i-1})$ is referred to as $P(X=c_i)$, the sample probability that $X = c_i$

and so he summarises the AUC ROC as:

$P(X<Y) + \frac{1}{2} P(X=Y)$

Which is a calculation for the whole sample of the sample probability that $X_k<Y_j$ plus half the sample probability that $X_k = Y_j$. Taking the term mn out of the denominator of this calculation based on the sample again gives the Mann Whitney U statistic as the number of times X_k is less than Y_j plus a correction factor for ties.

Thus AUC ROC curve = $U/(n \times m)$ and is also an estimator for the probability that for any randomly selected pair of sampled subjects from the non-diseased and diseased population that they will be correctly allocated based on the diagnostic test (54-56). Standard asymptotic theory can then be used to generate confidence intervals for the AUC ROC curve (54,56). The SAS procedure, PROC LOGISTIC, generates the AUC ROC curve by the trapezoidal rule as the 'c' statistic and for the curve in figure 2 this equals 0.893.

The AUC ROC curve can also be estimated by making parametric assumptions, for example that the distributions of test scores in the abnormal and normal subjects are

normally distributed but with different means and variances, and then using maximum likelihood methods (53,57-60).

To demonstrate this I will follow the papers by Metz (59,60). Let X represent a continuous random variable for the value of a diagnostic test for a population of non-diseased subjects, N_X , and let Y represent a continuous random variable for the value of a diagnostic test for a population of diseased subjects, N_Y.

The binormal model assumes that:

$X \sim N(\mu_x, \sigma^2_x)$

$Y \sim N(\mu_y, \sigma^2_y)$

The set C is a set of I discrete values that represent cut off or boundary scores for deciding a subject is non-diseased or diseased, i.e. if the score achieved is below the cut off the test states that the subject is non-diseased, and above then the test states they are diseased, and ordered such that $\{c_1 > c_2 >, \ldots, > c_{I-1} > c_I\}$.

Then the false positive rate (1 – Specificity) at the i^{th} boundary can be defined as:

$$F_i = P(X \geq c_i)$$
$$= \Phi((\mu_x - c_i)/ \sigma_x)$$

And the sensitivity at the i^{th} boundary can be defined as:

$$S_i = P(Y \geq c_i)$$
$$= \Phi((\mu_y - c_i)/ \sigma_y)$$

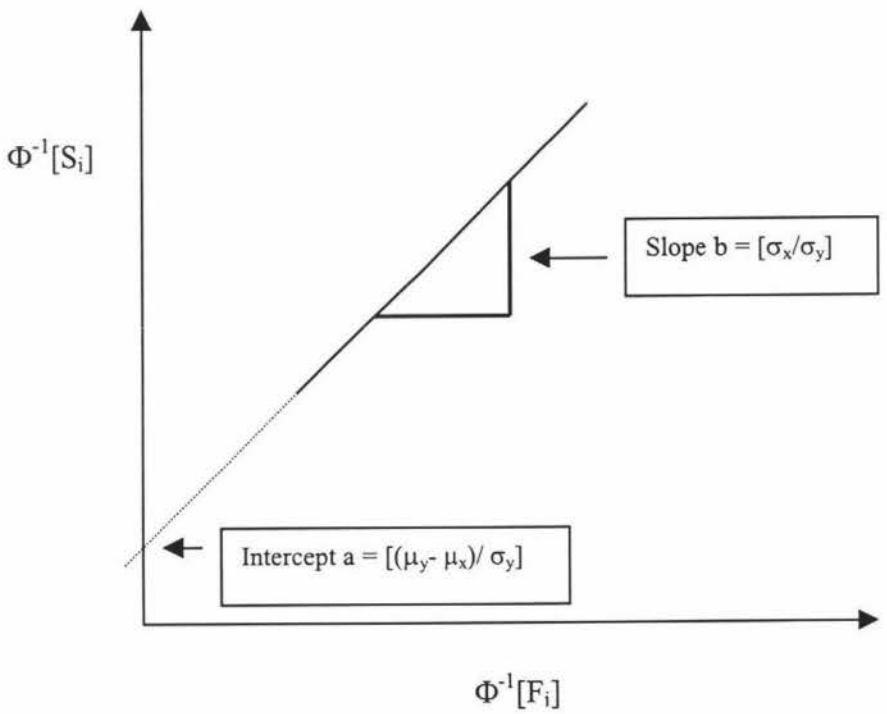
Where 'Φ' is the standard normal deviate.

Thus at the i^{th} boundary point:

$$\Phi^{-1}[S_i]\, \sigma_y - \mu_y = \Phi^{-1}[F_i]\, \sigma_x - \mu_x$$
$$\Phi^{-1}[S_i] = [(\mu_y - \mu_x)/ \sigma_y] + [\sigma_x/\sigma_y]\, \Phi^{-1}[F_i]$$

When this is plotted on binormal graph paper this is a straight line with an intercept: $a = [(\mu_y - \mu_x)/ \sigma_y]$ and slope $b = [\sigma_x/\sigma_y]$ as illustrated in figure 4.

Figure 4: Sensitivity and False positive rate plotted on binormal scaled axes



Now this can also be interpreted that the false positive rate and sensitivity arise from underlying latent distributions:

$X \sim N(0, 1)$

$Y \sim N(a, b)$

Now define:

p_i as the probability $P(c_i \leq X \leq c_{i-1})$ and

q_i as the probability $P(c_i \leq Y \leq c_{i-1})$ and

then

$p_i = \Phi(c_{i-1}) - \Phi(c_i)$ and

$q_i = \Phi(a-(b \times c_{i-1})) - \Phi(a-(b \times c_i))$

Now the likelihood function for a given set of data based on a random sample of size m and n drawn from these populations of non-diseased and diseased subjects, where at the i^{th} cut off point there are k_i non-diseased cases and l_i diseased cases, is based on the multinomial distribution and has the form:

$$L = m!n!\Pi^{I}_{i=1} ((p_i^{ki} \times q_i^{li})/(k_i! \times l_i!)),$$

where p_i and q_i are estimated by the false positive rate and true positive rate at each cut off value.

Now if the log likelihood value is chosen this has the form:

$$LL \propto \sum_{i=1}^{I} k_i \ln p_i + \sum_{i=1}^{I} l_i \ln q_i$$ and then substituting the two equations:

$p_i = \Phi(c_{i-1}) - \Phi(c_i)$ and

$q_i = \Phi(a-(b \times c_{i-1})) - \Phi(a-(b \times c_i))$

and differentiating the LL at each of the I+2 parameters, gives a set of non-linear equations of the form:

$\partial LL/\partial a = 0$

$\partial LL/\partial b = 0$

$\partial LL/\partial c_i = 0$.

These equations do not have a closed form solution and are solved using an iterative process, using the simple least squares fit on normal-deviate axes as an initial estimate. A variety of software is available that includes algorithms to solve for a and b. The AUC ROC curve estimated by this method is given by:

AUC ROC curve = $\Phi(a/\sqrt{(1+b^2)})$

Demonstration of this relationship is outlined here following the description by Thompson (61).

Consider the relationship between the sensitivity and the false positive rate as described by the binormal model:

$\Phi^{-1}[S_i] = [(\mu_y - \mu_x)/ \sigma_y] + [\sigma_x/\sigma_y] \Phi^{-1}[F_i]$

where

$a = [(\mu_y - \mu_x)/ \sigma_y]$ and slope $b = [\sigma_x/\sigma_y]$,

from the underlying latent distributions:

$X \sim N(0, 1)$
$Y \sim N(a, b)$.

Then this relationship can be expressed:

$$[S_i] = \Phi\{a + b\Phi^{-1}[F_i]\},$$

where for notational convenience F_i will be denoted by x and the total area under the ROC curve, A, can be expressed:

$$A = \int_{x=0}^{1} \Phi\{a + b\Phi^{-1}[x]\}dx.$$

Now making the substitution $x_1 = \phi(x)$:

$$A = \int_{x1=-\infty}^{\infty} \Phi\{a + b[x_1]\}\phi(x_1)dx_1,$$

$$A = \int_{x1=-\infty}^{\infty} \int_{x2=-\infty}^{a+bx1} \phi(x_2)\phi(x_1)dx_2dx_1.$$

Now making the substitution:
$z_1 = (x_2\text{-}bx_1)/\sqrt{(1+b^2)}$ and $z_2 = x_2$ yields the bi-variate normal integrand:

$$A = \int_{-\infty}^{\infty} \int_{x2=-\infty}^{a/\sqrt{1+b**2}} \phi(z_1, z_2; \rho)dz_1dz_2,$$

where ρ corresponds to $-b/\sqrt{(1+b^2)}$.
This is equivalent to the marginal density of z_1, that is the standard uni-variate Normal probability density function:

$$A = \int_{-\infty}^{a/\sqrt{1+b**2}} \phi(z_1)dz_1,$$

Thus $A = \Phi(a/\sqrt{(1+b^2)})$.

ROC curves and QoL instrument evaluation

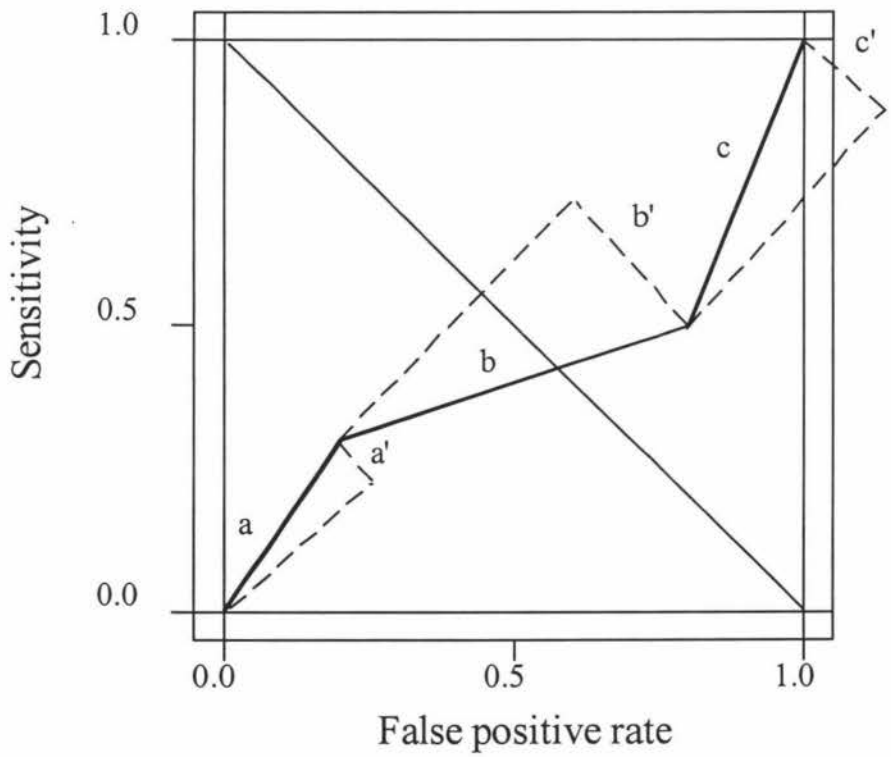
In the context of evaluating QoL instruments the 'diagnostic test' is the change in QoL score of a particular instrument, or dimension of an instrument, and this is compared to an external criterion. The external criteria used can be another 'established' QoL instrument or, more commonly, the response to a simple question such as 'Have things improved for you since we last tested you?'. Thus subjects must be administered the instrument on two occasions, typically done before and after an intervention which on other grounds is thought to improve QoL. In the paper by Deyo (46) two different external criteria were used, one was the response by the patient 'Have you resumed full activities?' and the second was where the doctor caring for the patient (all the patients had back pain) and the patient agreed that the patient had improved. Other examples of studies of responsiveness that have used a global index, as an external criterion, to rate the subjects as improved or not improved, include two studies of condition specific and generic QoL measures in back pain using a single item ordinal scale to rate satisfaction (62,63), and a study of QoL measures after stroke using the criterion 'living at home 26 weeks after the stroke' (64). Examples of use of an existing QoL score to rate subjects as improved or not include a study of low back pain (65) and rheumatoid arthritis (66).

Problems with the use of ROC curves

- There is often no gold standard external criterion for improvement in QoL.
- Subjects are sometimes excluded from the ROC analysis on the basis that they have deteriorated. This may lead to an over estimate of responsiveness based on similar arguments used for measures of internal responsiveness (41).
- Measurement error in QoL is not taken into account and bias may be present. The bias is typically downwards, that is that the AUC may be underestimated. (67-69).
- Whether to use parametric or non-parametric methods. Under some circumstances the normality assumption of distribution of the diagnostic markers in abnormal and normal subjects leads to very different, and incorrect, calculations of AUC (70).
- Complexity of current methods, is a simpler method needed? (71).
- Other methods may be better, although none have been used in the context of QoL studies of external responsiveness. These methods include:

1. The projected length of the ROC curve (72) which is the projection of the empirical ROC curve onto a line on the unit square joining the points (0,1) and (1,0). It is illustrated in figure 5. In this figure the hypothetical ROC curve consists of the three lines labelled 'a', 'b', and 'c'. The projected length of the ROC curve is found by summing the lengths of the lines parallel to the diagonal connecting the points (0,1) and (1,0) 'a'', 'b'' and 'c''.

Figure 5: The projected length of the ROC curve



2. The area swept out by the ROC curve (72). This is the sum of the areas swept by a ray emanating from (0,0) to each point of the empirical ROC curve. This is illustrated in Figures 6 and 7. The 4 points of the hypothetical ROC curve are 'a' at co-ordinates (0,0), point 'b', 'c' and 'd', the latter at co-ordinates (1,1). The first ray, Ray 1, radiating from (0,0) to point 'b' sweeps no area under the ROC curve. The second ray, Ray 2, sweeping to the next point, 'c' sweeps the area designated 'Area 1' in Figure 6. The third ray, Ray 3, sweeping back to point 'd' sweeps the area designated 'Area 2' in Figure 7. The area swept out by the ROC curve is the sum of the areas 'Area 1' + 'Area 2'.

Figure 6: Example of the area swept out by the ROC curve 1

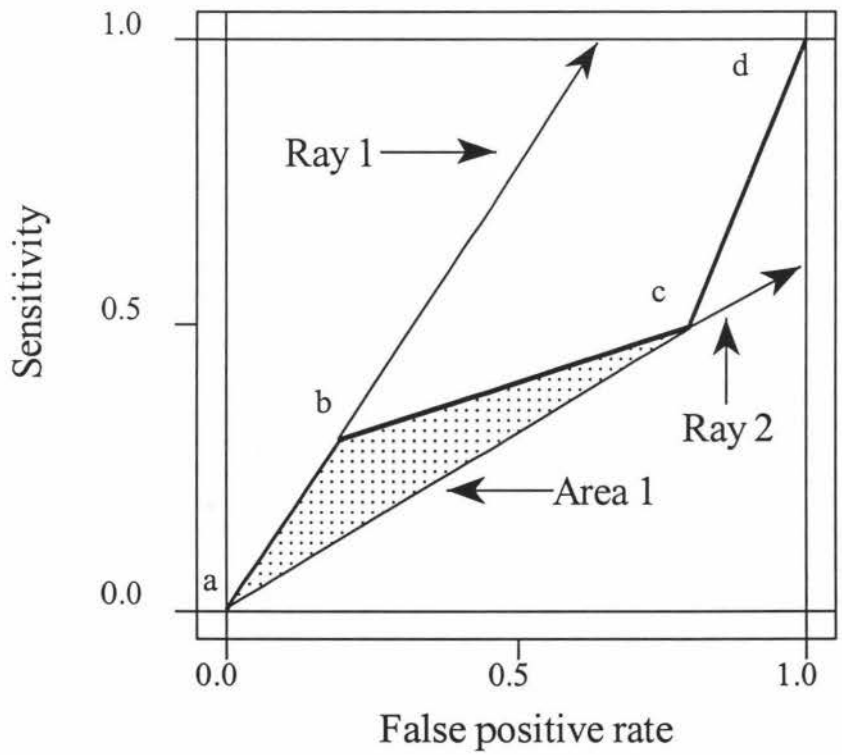
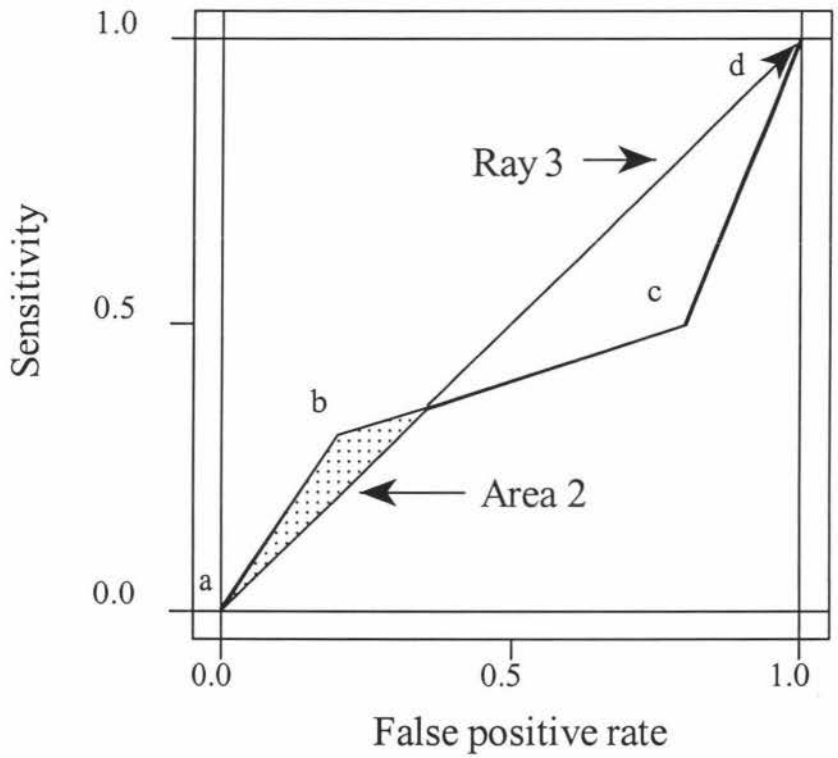


Figure 7: Example of the area swept out by the ROC curve 2



Both the projected length of the curve and the area swept out by the curve are related, in probabilistic terms, to the likelihood ratio for each cut off point for a diagnostic test. The likelihood ratio, for a test which labels a subject as diseased, is the ratio of the sensitivity and the false positive rate, at any particular cut off point. The advantages of these two approaches are in settings where a high or low value for the diagnostic test predicts that a disease is present, or where the distribution of the diseased subjects is skewed, for example to the right, compared to the distribution of non-diseased subjects. In the setting of QoL assessment the first scenario is unlikely to arise, that is the QoL instrument is very unlikely to be constructed such that a small change in a QoL instrument score, as well as a large change are both associated with an external criterion that predicts change. The second might arise where the distributions of changes in the QoL instrument scores are not the same in the groups labelled improved and not improved by an external criterion.

3. Indices derived from a plot of the cumulative proportions of subjects with a particular score for a diagnostic marker in diseased and non-diseased groups, the Lorenz curve (73). These derived indices are the Gini index, illustrated in Figure 8, which is twice the area between the Lorenz curve and the diagonal on the unit square, and the Pietra index, illustrated in Figure 9, which is the area of the largest triangle that can be drawn between the diagonal on the unit square with its apex on the Lorenz curve corresponding to a particular cut off score. The author (ibid.) demonstrates that the Gini index is the 'average absolute difference in post-test probabilities of two randomly selected subjects' and the Pietra index is an estimate of the 'average absolute change in disease probability provided by testing'. In the setting of testing a QoL instrument for responsiveness these sorts of interpretations may not be particularly useful, for determining if a QoL instrument score change is related to an external criterion change.

Figure 8: The Lorenz curve and the area used for the Gini index

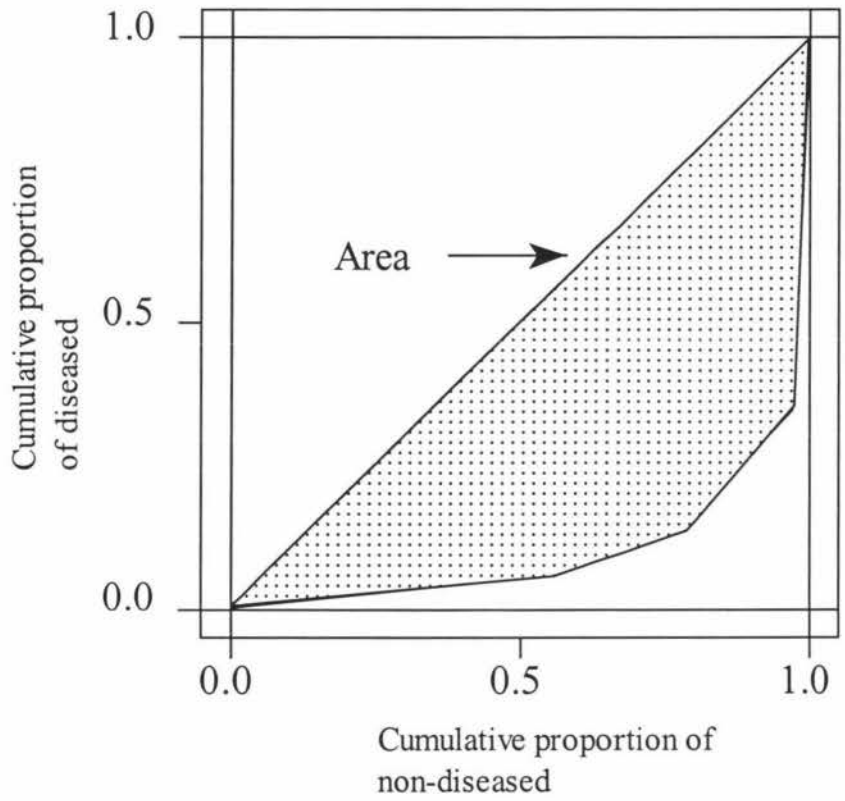
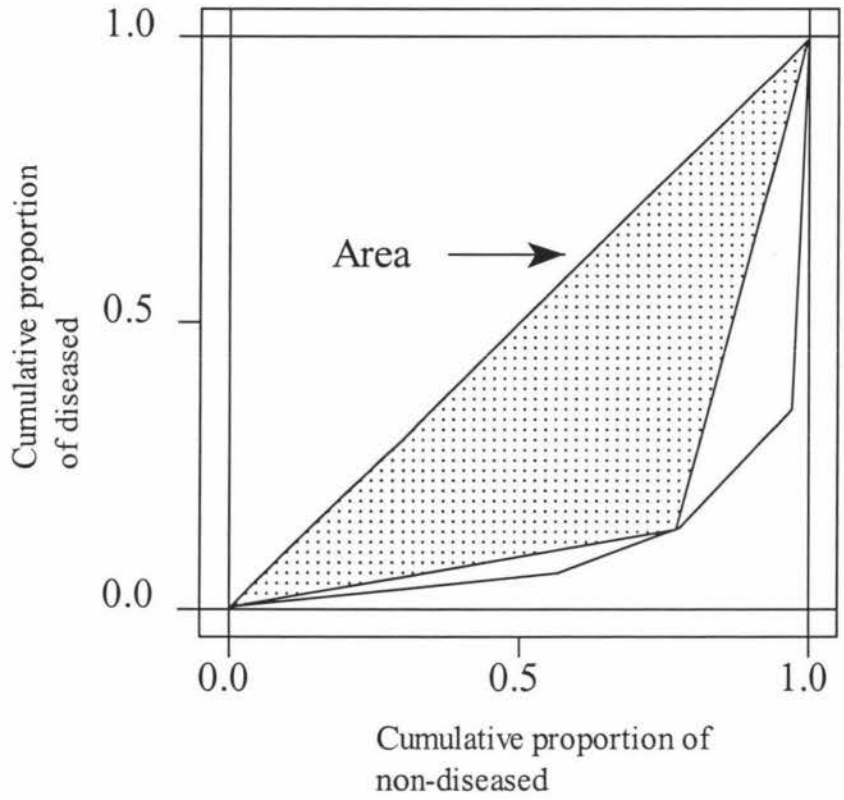


Figure 9: The Lorenz curve and the area used for the Pietra index



4. Concentrating on portions of the ROC curve rather than the whole curve (61). In this setting the AUC ROC curve used in the comparison of more than one diagnostic tests is an area of interest, for example only those parts of the curve where the false positive rate is less than 20%. This corresponds to a more realistic use of diagnostic tests because users will typically only want to use a test which has a low false positive rate. Conclusions based on the whole ROC curve may differ from those based on a part of the curve over a limited area. The authors suggest that under the binormal model this partial area under the ROC curve can be evaluated by the partial integral below rather than the whole area as described by the marginal distribution of z_2 described earlier in the thesis.

Now

$$A_c = \int_{-\infty}^{\phi-1(c)} \int_{x2=-\infty}^{a/\sqrt{1+b**2}} \phi(z_1, z_2; \rho)dz_1dz_2,$$

where ρ corresponds to $-b/\sqrt{(1+b^2)}$ and 'a' and 'b' correspond to the parameters estimated according to the discussion on page 40-44 of this thesis. 'c' corresponds to the particular value of the false positive rate that is subject to this evaluation. However it is not clear how any particular limits of integration can be selected 'objectively', as opposed to perhaps selecting a portion of two ROC plots where a particular cut off for the false positive rate appears, subjectively, to give the 'best' separation for different diagnostic modalities.

Chapter 4: External responsiveness: Correlation and Regression

Correlation

Measurement of the correlation between the change in a QoL measure and the change in an established measure of health status, over a period of treatment known to improve the disease process, is an intuitively straight forward way to assess external responsiveness. Husted and colleagues (32, p 463) state:

If r_{xy} [the correlation coefficient between a new QoL instrument, X, and a traditional clinical outcome, Y] *approaches +1 then X is thought to capture the information on Y (i.e. it responds to changes in Y).*

The formula given by Husted and colleagues (32) to calculate the correlation coefficient appears to be in error. It is given as:

$$r_{xy} = \Sigma(D_{xi} - D_{xm})(D_{yi} - D_{ym})/I,$$

where D_{xi} is the difference in the new QoL instrument score for the i^{th} subject measured on two occasions, D_{xm} is the mean change in the new QoL instrument score for the I subjects, D_{yi} is the difference in the traditional clinical measure for the i^{th} subject measured on two occasions, and D_{ym} is the mean change in traditional clinical measure for the I subjects. This represents a covariance. This is equivalent to a correlation only if the QoL instrument scores are measured on the same scale. The scale of QoL instruments is not mentioned in the paper (ibid.).

Given this the denominator for this equation, if the Pearson product moment correlation is to be calculated, should be:

$$[\Sigma(D_{xi} - D_{xm})^2\Sigma(D_{yi} - D_{ym})^2]^{1/2},$$

which acts as a scaling factor to allow comparison of QoL instruments on different scales. In addition the review by Husted and colleagues does not comment on the comparison of correlation coefficients when a number of different QoL instruments are compared, the role of sample size, or the recognition that sample correlation coefficients are estimates of population parameters, that is they have standard errors associated with them. Neither does the review discuss whether the Pearson product moment correlation coefficient or Spearman's rank correlation coefficient should be used. As will be illustrated papers which use a correlation coefficient to compare responsiveness do not comment or account for these issues and at most, if a number of new QoL instruments are to be compared with each other, rank the correlation coefficients and state that the instrument with the highest correlation with the traditional measure has the highest responsiveness.

Husted and colleagues (32) cite three examples of this procedure carried out as part of the examination of responsiveness of QoL instruments.

- Deyo and colleagues (46) in a study of QoL instruments in low back pain used 5 different external criteria to rate improvement. A Spearman's rank correlation coefficient was used to compare the change in each of four QoL measurements with changes in each of these 5 measures, measured twice, with a three week interval between measurements. The results were presented in a 20 cell table listing the individual correlation statistics (which ranged from 0.02 to 0.41). Accompanying 'P values' were assigned to the correlation measurements comparing each correlation coefficient with zero. The QoL instrument which had the highest correlation with each of the external criteria was nominated the most responsive.
- Fitzpatrick and colleagues (39) in a study of QoL instruments for rheumatoid arthritis used 4 different external criteria to rate improvement. They compared the correlation, although didn't state whether they used the Spearman or Pearson statistic, between the change in four QoL measurements, measured 3 times over 6 months, and the change in external criteria. Further complicating this analysis it was the dimension scores of individual instruments (mobility, activities of daily living, household activities, pain, emotions, and social) that were subjected to correlation analysis. Not all the instruments had all the dimensions. The table of the results of the correlation had 152 pairs of comparisons. 'P values' were

presented for these correlation coefficients comparing each to zero. The range of the correlation coefficients was between –0.42 and +0.5. The authors could find no consistent pattern amongst the various correlation coefficients to suggest one instrument was more responsive than another.

- Wright and colleagues (40) in a study of QoL instruments before and after a hip replacement operation for osteoarthritis used one external criterion for change and the Spearman's rank correlation coefficient and compared 5 different instruments, although dimension scores for two of the five instruments were used in the analysis. Fourteen correlation coefficients were presented ranging from 0.08 to 0.53.

Two further examples of the use of correlation to assess responsiveness include studies by Meenan and colleagues (74) and Stucki and colleagues (62).

- In the Meenan study of rheumatoid arthritis (74) 3 external criteria were examined for 3 dimensions of a single QoL instrument. A Pearson's correlation coefficient was calculated for each of the 9 sets of data. The correlation coefficients varied between 0.03 and 0.52. Significance was assigned to an individual correlation coefficient when it was significantly different from zero.
- In the Stucki study of osteoarthritis of the spine (62) a single external criterion was used for improvement, a change in satisfaction score, and 4 different QoL instruments were compared. A Pearson's correlation coefficient was calculated which ranged between 0.38 and 0.72. Significance was assigned to an individual correlation coefficient when it was significantly different from zero.

Issues with correlation coefficients and responsiveness

1. Confidence intervals should be presented for correlation coefficients

The 'Fisher's Z' transformation can be used to develop confidence intervals for the Pearson correlation coefficient (75). This is based on the assumption that the paired data are from a bivariate normal distribution and defines:

$V = \frac{1}{2} \ln [(1+R)/(1-R)]$ and $m = \frac{1}{2} \ln [(1+\rho)/(1-\rho)]$,

where R is the sample correlation coefficient and ρ is the population correlation coefficient. For large sample size V has an approximate distribution $N(m,1/(n-3))$.

For example using the coefficients presented in the paper by Stucki (62), with a sample size of 130, the confidence intervals are shown in Table 10:

Table 10: Confidence intervals for correlation coefficients

New QoL measure	**Pearson's Correlation Coefficient with external criterion**	**95% Confidence interval**
Physical function scale	-0.72	-0.79 to -0.63
Symptom severity scale	-0.68	-0.76 to -0.58
Roland scale	-0.53	-0.64 to -0.39
Sickness Impact Profile	-0.38	-0.52 to -0.22

The simple ranking based on the point estimate of the correlation coefficients hides the considerable overlap in the confidence intervals for the coefficients.

2. What is being compared with what?

The majority of the examples discussed compared individual correlation coefficients with zero. As both the QoL instrument and the external criteria are supposed to reflect QoL it is hardly surprising that the correlation coefficients describing their relationship are different from zero. What is more important in defining a responsive instrument is comparing correlation coefficients with each other. Statistical tests are available for comparing correlation coefficients where a number of predictor variables are compared with a single response variable using the same set of subjects. In short these are tests for comparison of correlated correlation coefficients. A number of these tests have been reviewed by May and Hittner (76). They include the Hotelling t statistic, Williams t statistic, Olkins Z statistic and a Z statistic attributed to Meng and colleagues (77).

The latter paper includes tests of equality of two correlated correlation coefficients, equality of a set of I correlated correlation coefficients between a set of I predictor variables and the same response variable, and of contrasts amongst correlated correlation coefficients.

1. Z test for 2 correlated correlation coefficients

Y is the response variable
X_1 and X_2 are predictor variables
r_1 is the correlation between Y and X_1
r_2 is the correlation between Y and X_2
r_x is the correlation between X_1 and X_2
z_{r1} is the Fisher's Z transformation of r_1
z_{r2} is the Fisher's Z transformation of r_2
N is the number of subjects
r_m^2 is the arithmetic mean of r_1^2 and r_2^2
$f = (1-r_x)/(1-r_m^2)$ and is set to 1 if its value is less than 1
$h = (1-fr_m^2)/(1-r_m^2)$

Then $Z = (z_{r1} - z_{r2}) \times \{(N-3)/(2(1 - r_x)h)\}^{1/2}$ is distributed as the standard normal.

Chi squared test for equality of the set of correlations of a number of predictors with a single response variable.

Y is the response variable

X_i, i = 1 to I, are the predictor variables

r_i is the correlation between Y and the i^{th} predictor variable

z_{ri} is the Fisher's Z transformation of r_i

h is defined as above except r_x is replaced by r_s, the median inter-correlation between the I predictor variables and r_m is the arithmetic mean of all the r_i^2.

I is the number of correlations

z_m is the mean of z_{ri}

Then $C = ((N - 3)(\Sigma_i(z_{ri} - z_m)^2)/((1 - r_s)h)$ is distributed as a χ^2 variable with (I - 1) degrees of freedom.

2. *Z test for contrasts between correlation coefficients*

Let λ_i is the contrast weight assigned to each of the z_{ri}.

Then $Z = \Sigma_i[\lambda_i z_{ri} \{(N-3)/((\Sigma_i \lambda_i^2)(1 - r_x)h)\}^{1/2}]$ is distributed as the standard normal

None of the papers discussed as examples of the use of correlation coefficients and responsiveness of QoL instruments give the inter-correlations between predictors, therefore these test scores cannot be calculated for the coefficients presented.

3. Normality assumptions should be met if the Pearson statistic is used

The normality assumption is not discussed in any of the papers discussed quoted. This is particularly important where Likert scales are used but not added, or where data are categorical.

4. The external criteria may be faulty

If the external criteria used are imperfect reflections of QoL then low correlation between the QoL instrument being examined and any particular external criterion may be because the external criterion is at fault. A common external criterion is the score on a Likert scale, or weighted sum of a combination of Likert scales, that is generated just for a particular study, and often given a name such as a 'satisfaction scale' (62) or a 'global rating scale' (40). The measurement properties of these external criteria are usually not stated.

5. The new QoL instrument and the external criterion may be related through an unmeasured third variable.

The inference drawn when the change in a QoL instrument score is correlated with the change in the external criterion is that there is a causal relationship. For example the pain from inflamed joints in a person with arthritis are thought to cause a reduction in QoL. The more a change in a measure of pain is correlated with a change in a QoL instrument score the more this suggests that the change in QoL is 'responding' to the change in external criterion. The issue that arises, particularly in observational studies where subjects are measured before and after an intervention, is whether the external criterion truly is associated in a causal way with QoL instrument score.

In the setting of an assessment of joint pain and quality of life a group of people may be subjected to intensive study involving a large number of QoL instruments. The change in the score on QoL instruments between two occasions may be due simply to researchers demonstrating interest in the subjects and enhancing the subject's feelings of self worth. The change in pain related to the joints might also be related to this extra attention, or a number of other factors, such as a greater interest in the joint problem and more efforts to get it treated. Both joint pain and the QoL score improve, and will be correlated, but one is not causing the other, in fact the change is mediated by a third variable, in this case the attention paid to the subjects, i.e. a Hawthorne effect. Simple correlation analysis will not detect this. Amongst the five examples of the use of correlation for assessment of responsiveness three studies were observational only

(39,40,62), one was based on placebo controlled trial data (74) and in one study it was not clear whether it was an observational study or part of a placebo controlled trial (46).

6. The external criterion and the new QoL instrument may not be related in a linear way.

An example of this is where there is a curvi-linear but essentially monotonic relationship between an external criterion and a QoL instrument, such that as the score on the external criterion increases the score on the QoL instrument increases but at a slower rate. A Pearson's correlation coefficient, in this situation, could be low, where as the Spearman rank correlation could be high. More usually the Pearson coefficient will be larger than the Spearman coefficient, especially for non-normal data. If plots of the relationship between external criteria and QoL instruments are not presented (the case in all the papers described above) this possibility cannot be examined.

7. Outliers and clustering may be present

Two situations where a correlation coefficient, particularly the Pearson's coefficient, may be misleading are if outliers or clustering are present. In the figures 10 and 11, Y might be a new QoL instrument score and X the score on an external criterion.

The hypothetical data illustrated in figure 10 may well produce a 'significant' correlation coefficient even though clearly there is little relationship between the two variables if the outlier is removed. Without plots of the data analysed in the papers discussed above it is difficult to know if the, usually, small correlation coefficients, could be describing this sort of relationship between QoL instrument scores and external criteria. Figure 11 illustrates that the relationship between Y and X conditional on whether X is at the extremes of the range of X, is different from the marginal distribution over the whole range of X. In fact the relationship between Y and X in each cluster, i.e. conditional on the particular range of X for each cluster, could be in the opposite direction to the marginal relationship between Y and X. Furthermore if the range of the external criteria for the QoL instrument is restricted in any particular study, for example only people with poor or excellent QoL, or low or high levels of the

external criteria, are subject to study, then the sample estimate of the correlation may be quite different to other studies performed in different populations.

Figure 10: Example of outlier

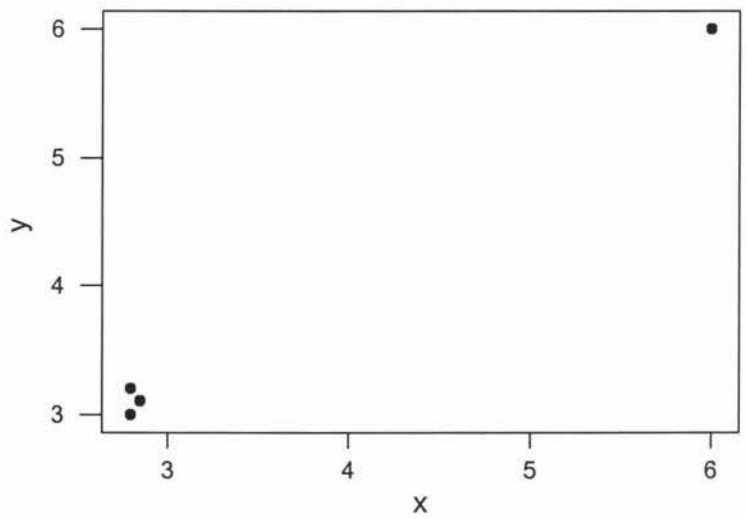
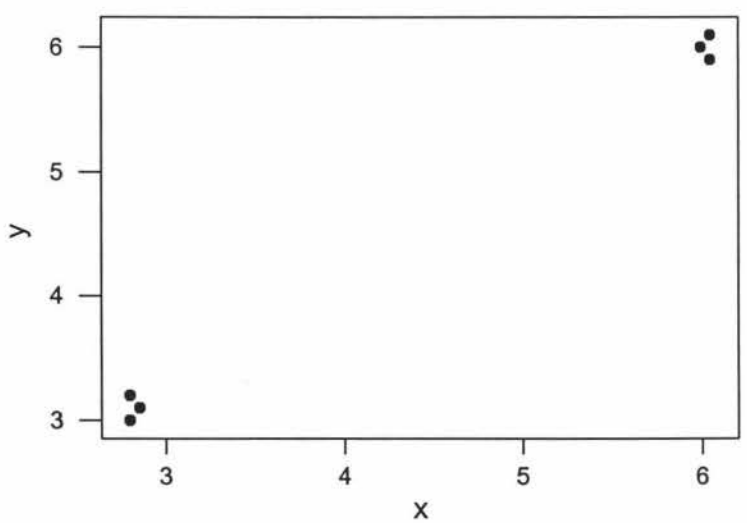


Figure 11: Example of clustering



8. Using dimensions of QoL instruments

If individual dimensions of a QoL instrument are compared to an external criterion that truly reflects the whole construct of QoL, then it is expected that the dimension will not be as correlated with the external criterion, as the whole instrument score. This is in part a consequence of the whole instrument containing more items than are present in any particular QoL dimension.

9. Multiple statistical testing

The study of Fitzpatrick and colleagues (39) examined 152 pairs of comparisons. They clearly considered a 'P value' of less than 0.05 to indicate that a particular correlation coefficient was different from zero. However the type I error rate for the set of correlation coefficients will be far in excess of this when the tests are considered in concert, since this becomes a multiple comparison problem. Studies examining the relationship between different instruments, or different dimensions of different instruments, against multiple external criteria may be susceptible to inflation of type I error rates. Just ranking correlation coefficients without regard to their standard errors is a problem for similar reasons.

10. Correlation does not mean agreement

In an important paper in the medical literature Bland and Altman distinguish between agreement and correlation (78). Clinical measurements, of which QoL measurements are examples, may be correlated, but not agree particularly well. For example the linear relationship detected by correlation measures any linear relationship even if the two measurements do not agree. An example is given in the paper of two different ways of measuring the same physiologic variable, expiration rate, which are correlated. When the differences between the measurements on the same subjects are analysed however the two measures do not agree very well. Consider the following synthetic data for QoL measured by two instruments on the same numeric scale of zero to 100, where the 'data' for QoL_1 are a random sample of size 10 from a uniform distribution and QoL_2 is

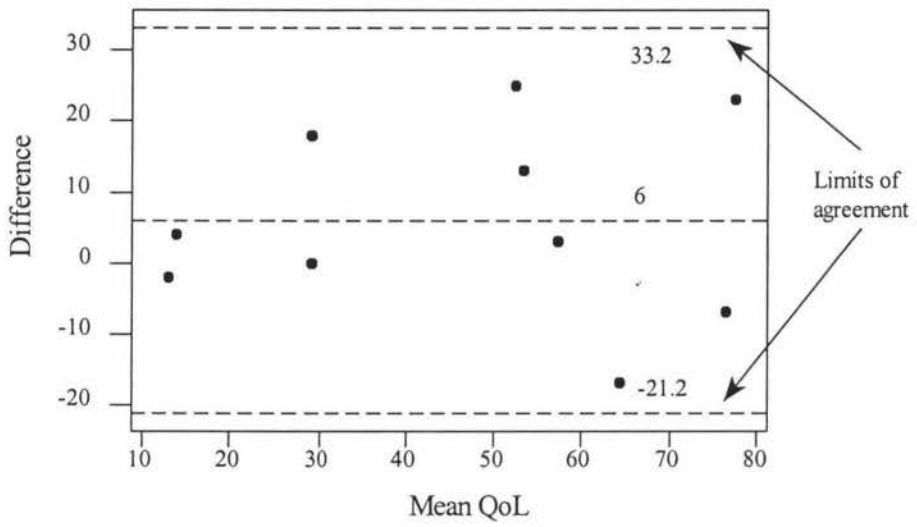
generated by $QoL_2 = a + b(QoL_1)$, where 'a' was a random sample of size 10 from N~(5,4) and 'b' a random sample of size 10 from N~(0.7, 0.04).

Table 11: Data to illustrate difference between correlation and agreement

QoL instrument 1	**QoL instrument 2**	**Difference QoL_1 - QoL_2**
73	80	-7
59	56	3
29	29	0
12	14	-2
16	12	4
65	40	25
56	73	-17
60	47	13
38	20	18
89	66	23

The correlation between the two measures of QoL is very good, the correlation coefficient is 0.85, because of the close linear relationship between the two instruments. Agreement between the two measures is however better assessed by examining the differences. This is sometimes assessed by the technique of 'limits of agreement', which refers to the mean difference plus or minus 2 times the standard deviation of the differences. It is illustrated with the so called Bland-Altman plot, of differences versus the mean of the measurement for the two techniques. For the data in the table above the mean difference was 6 units, with a standard deviation of 13.6 units, and the Bland-Altman plot is shown the figure 12.

Figure 12: Limits of agreement plot



In this example correlation was high but agreement is not very good as the limits of agreement constitute about 50% of the range of the instruments. This concept is most applicable to comparing a standard instrument for measuring QoL and a new instrument. Correlation may be high but agreement may not. Analysing agreement in this situation may be quite difficult as QoL instruments measure the abstract concept of QoL, but do so on different numeric scales.

Regression models

As a result of some of these concerns about the use of correlation coefficients, in particular if a non-linear association between a predictor and QoL was to be modelled and the effect of restricting the range of the predictor variables on the correlation, Husted and colleagues (32) suggest linear regression as a way of modelling the way a score on a new QoL instrument relates to external criteria, in order to determine responsiveness. The models they suggest and use in practice (79,80) are examples of multiple linear regression.

In the first published paper (79) subjects with psoriatic arthritis were examined four years apart. A QoL instrument was administered and two external criteria for change, the number of actively inflamed joints and the number of deformed joints were measured. Over the observation period the number of inflamed joints decreased slightly from a mean of 5.8 (standard deviation 6.8) to a mean of 5.4 (standard deviation 7.4). A decrease in the number of inflamed joints should be associated with an improvement in QoL. The number of deformed joints increased from a mean of 4.3 (standard deviation 7.9) to a mean of 9.1 (standard deviation 12.3). An increase in the number of deformed joints should be associated with decreased QoL. The regression had, as response variables, the change in the external criteria, i.e. the change in number of inflamed joints or deformed joints. The explanatory variables could include all three dimensions of each of two different QoL instruments relating to physical function, psychological status and pain, as well as the initial value of the external criteria. The dimensions of each of the two QoL instruments could be standardised, by the design of the instrument, to a score of between 0 and 10, with higher scores reflecting lower quality of life. Separate multiple regressions were performed for each of the external criteria, although the authors do not state why this was done. To complicate the analysis the particular QoL instrument had changed between the two measurement times so that the authors used the three separate dimensions of the original QoL instrument for the initial measurement, and the same three dimensions of the modified, albeit related, second QoL instrument as the second measurement. The models presented included, as individual explanatory variables, the scores on the dimensions of the particular QoL instrument, even if the individual parameter estimates for the regression coefficients were not statistically significant. No attempt was made to remove explanatory variables

from any particular model if they did not add explanatory power to a model. A model was stated to be a good one if its value for R squared was large. This implied, in the authors view, the values of QoL measured on two occasions some time apart predict the change in joint inflammation and deformity.

The validity of this approach can be questioned.

- The QoL score was used to predict the change in the number of inflamed joints or number of deformed joints, or state of the disease. The question of precedence is not addressed, that is in the model of how disease affects quality of life, the change in joint status should predict a change in QoL. In addition the analysis was carried out with the QoL score predicting the count of the number of inflamed joints or number of deformed joints and this required a Poisson (or similar type of) regression because the dependent variable is a count and is not continuous.
- The authors do not state why they analysed each of the disease variables separately, however this approach seems fundamentally flawed. If a disease causes an effect on QoL, and if it is postulated that the disease can manifest in a number of different ways, such as joint inflammation and joint deformity, then the response variables should be included as a multivariate response variable to examine if the mean vector variance is explained by the explanatory variables. Conducting two separate univariate analyses does not take into account the covariance between the response variables.
- No attempt was made in the paper to select models which could explain the variance in the response, the disease variables, based on a minimum number of explanatory variables, or to compare different models. No attempt in the paper is made to use estimates of bias of various models, e.g. Mallow's C_p. This led to the model, which was described as the best model, having 7 explanatory variables only two of which had parameter estimates that were significantly different from zero.
- The paper did not even address the same QoL instrument on the two occasions so that no conclusions regarding the responsiveness of a particular instrument could be made. In any case even if the same instrument had been used on the two occasions using only some of the dimensions of the QoL instruments rather begs the question

as to whether an instrument, overall, is responsive to disease changes that lead to QoL changes.

- One of the suggested disease variables did not change over the time period. Anticipation that QoL change could be detected if the disease has apparently not changed is a fundamental design flaw in this observational study.
- No information was presented to demonstrate that normality and other regression assumptions were met.

The second paper (80) examined the use of three QoL instruments in psoriatic arthritis. The instruments were applied to subjects on two occasions twelve to eighteen months apart but no specific intervention designed to influence the disease process was applied. The external criteria for change were counts of the numbers of diseased joints, both the number of inflamed joints and the number of deformed joints, and a short single question asking the patients to rate their general health on a 5 point ordinal scale: 1=much better than a year ago, 2=somewhat better than a year ago, 3=about the same, 4=somewhat worse than a year ago, 5=much worse than a year ago. The multiple linear regression analysis performed had the change in the dimension of a particular QoL instrument as the explanatory variable and the change in the external criteria as the response variable, although as a scale of this sort, i.e. Likert, certainly does not meet the usual multiple regression assumptions, this may not be sensible. The three instruments did not all have the same 4 dimensions of physical function, pain, psychological function, and social function. In order to compare different instruments analysis was also carried out by performing the same regression but standardising the scores of the explanatory and response variables, by dividing the 'change score' by the standard deviation of particular variable. Parameter estimates relating the change in QoL dimension score to the change in disease status score based on both the raw and standardised data were presented. Although multiple regression analyses were performed the response variables were each treated in a univariate way rather than as a multi-variate response. In addition the selection of the explanatory variables for the final model was only dependent on whether they were statistically significant in when they were the single explanatory variable. For example the physical function scores for each of the three instruments were left in a multivariable model that had the change in the 5 point ordinal scale as the response variable. An R squared is presented for this regression even though two of the explanatory variables did not have a parameter

estimate that was significantly different from zero, and the one dimension score that was significant was nominated the most responsive instrument.

Similar criticisms to the authors first paper could be made. The issue of precedence has not been addressed, the way in which univariate and multivariate analyses were carried is not explained or justified, models were not generated or compared sensibly, and normality assumptions and other regression assumptions were not commented on.

Thus while regression techniques may have something to offer in evaluating external responsiveness the recommendations in the review article (32) and the actual techniques suggested by Husted and colleagues (79,80) are significantly flawed.

Issues with regression models for responsiveness

1. Precedence and the model underlying QoL

In the two examples above (79,80) the external criteria were disease related variables. As such, based on conceptual models of QoL, they should predict QoL, so that disease related variables should be explanatory variables. In addition to the conceptual issue that deficits in QoL arise because of disease related impairments, there will be a difference in the two regression parameter estimates depending on which is the explanatory and which is the response variable.

Consider the simple linear regression model for N subjects examining the relationship between two types of measurement, y_i representing a change QoL measurement, and x_i representing a change in external criterion measurement. If x_i predicts y_i then the simple linear regression model, under a normality assumption, is:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$E(Y_i) = \beta_0 + \beta_1 x_i$$
$$Var(Yi) = \sigma^2$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

The estimates of β_0, b_0, and β_1, b_1, are given by:

$$b_0 = y_m - b_1 x_m$$
$$b_1 = \Sigma(x_i - x_m)(y_i - y_m) / \Sigma(x_i - x_m)^2,$$

where x_m is the sample mean of the x_i and y_m the sample mean of all the y_i.

Furthermore let:

$$SS_{XY} = \Sigma(x_i - x_m)(y_i - y_m)$$
$$SS_{XX} = \Sigma(x_i - x_m)^2$$
$$SS_{YY} = \Sigma(y_i - y_m)^2,$$

R^2 for this model is then defined as $[SS_{XY}]^2/SS_{XX}SS_{YY}$, the square of the correlation between X and Y.

Now if the 'roles' of the measurements are swapped so that now y_i predicts x_i then the regression estimates of the parameters in the new regression b_0^* of β_0^* and b_1^* of β_1^*, are given by:

$b_0^* = x_m - b_1 y_m$

$b_1^* = \Sigma(x_i - x_m)(y_i - y_m)/\Sigma(y_i - y_m)^2$

The relationship of the slopes of the simple linear regression is then the relationship between the values of SS_{XY}/SS_{XX} for the first regression to SS_{XY}/SS_{YY} in the second regression. One will not be simply the inverse of the other so that the magnitude of the relationship between the two measurements, based on the slope parameter, depends importantly on the precedence in the regression. R^2 for both models is identical and equals $[SS_{XY}]^2/SS_{XX}SS_{YY}$ as demonstrated below:

Model 1: x_i predicts y_i

R^2 = Regression Sum of Squares regression /Total Sum of Squares

Regression Sum of Squares $= \Sigma(\text{predicted } y_i - y_m)^2$

$= \Sigma(b_0 + b_1 x_i - y_m)^2$

$= \Sigma(y_m - b_1 x_m + b_1 x_i - y_m)^2$

$= b_1^2 \Sigma(x_i - x_m)^2$

$= [SS_{XY}]^2/SS_{XX}$

Total Sum of Squares $= SS_{YY}$

R^2 $= [SS_{XY}]^2/SS_{XX}SS_{YY}$

Model 2: y_i predicts x_i

R^2 = Regression Sum of Squares regression /Total Sum of Squares

Regression Sum of Squares $= \Sigma(\text{predicted } x_i - x_m)^2$

$= \Sigma(b_0^* + b_1^* y_i - x_m)^2$

$= \Sigma(x_m - b_1 y_m + b_1^* y_i - x_m)^2$

$= b_1^{*2}\Sigma(y_i - y_m)^2$

$= [SS_{XY}]^2/SS_{YY}$

Total Sum of Squares $= SS_{XX}$

R^2 $= [SS_{XY}]^2/SS_{XX}SS_{YY}$

The other invariant quantity, that does not depend on precedence for the simple linear regression, is the statistical significance of the slope parameter. If the first regression is considered then the test of whether the slope parameter is zero is given by the ratio of b_1 to its standard error. The standard error of b_1 is given by:

$$[((SS_{YY} - [SS_{XY}]^2/SS_{XX})/N\text{-}2)/(SS_{XX})]^{1/2}$$

With b_1 equal to SS_{XY}/SS_{XX} this simplifies to:

$$T = SS_{XY}/((N\text{-}2)\times(SS_{XX}SS_{YY} - [SS_{XY}]^2)^{1/2}$$

Which is compared to a t distribution with N-2 degrees of freedom.

If the roles of x_i and y_i are then interchanged then the T statistic for testing whether b_1^* is zero is identical.

In the case where there are multiple predictors and the role of one of the predictors is swapped with the response variable it can also be demonstrated that the 'size' of the regression parameter estimate depends on the precedence.

The general linear model, with uncorrelated errors and no random components, is given by:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$

with $E(\mathbf{Y}) = \mathbf{X}\beta$ and $Var(\mathbf{Y}) = \sigma^2\mathbf{I}$, with Y multivariate normal.

Now $\mathbf{X}\beta$ can be partitioned such that:

$$\mathbf{Y} = \mathbf{X}_1\theta_1 + \mathbf{X}_2\theta_2 + \varepsilon$$

If the case is now considered where X_2 is an $n\times1$ vector and θ_2 is a scalar then the case where the column vector Y and X_2 are swapped and its effect on the least squares estimate of θ_2 can be considered.

The normal equations in the case of the partitioning can be written:

$$\begin{pmatrix} \mathbf{X}^T_1\mathbf{X}_1 & \mathbf{X}^T_1\mathbf{X}_2 \\ \mathbf{X}^T_2\mathbf{X}_1 & \mathbf{X}^T_2\mathbf{X}_2 \end{pmatrix} \times \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T_1\mathbf{Y} \\ \mathbf{X}^T_2\mathbf{Y} \end{pmatrix}$$

The solution of the equations gives $\hat{\theta}_2$.

The general method of inverting a partitioned matrix is (81):

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{F}\mathbf{E}^{-1}\mathbf{F}^T & -\mathbf{F}\mathbf{E}^{-1} \\ -\mathbf{E}^{-1}\mathbf{F}^T & \mathbf{E}^{-1} \end{pmatrix}$$

Where:

$\mathbf{E} = \mathbf{D} - \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}$ and $\mathbf{F} = \mathbf{A}^{-1}\mathbf{B}$

For the partitioned matrix relating to the least squares estimators of the parameters this suggests the solution can be derived:

$$\begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{D} \end{pmatrix}^{-1} \times \begin{pmatrix} \mathbf{X}^T_1\mathbf{Y} \\ \mathbf{X}^T_2\mathbf{Y} \end{pmatrix}$$

$$\begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{F}\mathbf{E}^{-1}\mathbf{F}^T & -\mathbf{F}\mathbf{E}^{-1} \\ -\mathbf{E}^{-1}\mathbf{F}^T & \mathbf{E}^{-1} \end{pmatrix} \times \begin{pmatrix} \mathbf{X}^T_1\mathbf{Y} \\ \mathbf{X}^T_2\mathbf{Y} \end{pmatrix}$$

So that:

$$\begin{aligned} \hat{\theta}_2 &= (-\mathbf{E}^{-1}\mathbf{F}^T\,\mathbf{X}^T_1\mathbf{Y} + \mathbf{E}^{-1}\,\mathbf{X}^T_2\mathbf{Y}) \\ &= \mathbf{E}^{-1}(\mathbf{X}^T_2 - \mathbf{F}^T\,\mathbf{X}^T_1)\mathbf{Y} \end{aligned}$$

Now let:

$$\mathbf{A} = \mathbf{X}^T_1\mathbf{X}_1\,,\ \mathbf{B} = \mathbf{X}^T_1\mathbf{X}_2\,,\ \mathbf{D} = \mathbf{X}^T_2\mathbf{X}_2$$

Then:

$$\begin{aligned} \mathbf{E} &= \mathbf{D} - \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B} \\ &= \mathbf{X}^T_2\mathbf{X}_2 - \mathbf{X}^T_2\mathbf{X}_1(\mathbf{X}^T_1\mathbf{X}_1)^{-1}\,\mathbf{X}^T_1\mathbf{X}_2 \\ &= \mathbf{X}^T_2(\mathbf{I} - \mathbf{X}_1(\mathbf{X}^T_1\mathbf{X}_1)^{-1}\,\mathbf{X}^T_1)\mathbf{X}_2 \\ &= \mathbf{X}^T_2(\mathbf{P}_1)\mathbf{X}_2 \end{aligned}$$

which is scalar where $\mathbf{P}_1 = (\mathbf{I}-\mathbf{H}_1)$ and $\mathbf{H}_1 = \mathbf{X}_1(\mathbf{X}^T_1\mathbf{X}_1)^{-1}\mathbf{X}^T_1$.

$$\begin{aligned} \mathbf{F} &= \mathbf{A}^{-1}\mathbf{B} \\ &= (\mathbf{X}^T_1\mathbf{X}_1)^{-1}\,\mathbf{X}^T_1\mathbf{X}_2 \\ \mathbf{F}^T &= \mathbf{X}^T_2\mathbf{X}_1\,(\mathbf{X}^T_1\mathbf{X}_1)^{-1} \end{aligned}$$

So that:

$$\begin{aligned} (\mathbf{X}^T_2 - \mathbf{F}^T\,\mathbf{X}^T_1) &= \mathbf{X}^T_2(\mathbf{I} - \mathbf{X}_1(\mathbf{X}^T_1\mathbf{X}_1)^{-1}\,\mathbf{X}^T_1) \\ &= \mathbf{X}^T_2(\mathbf{P}_1) \end{aligned}$$

So that:

$$\begin{aligned} \hat{\theta}_2 &= \mathbf{E}^{-1}(\mathbf{X}^T_2 - \mathbf{F}^T\,\mathbf{X}^T_1)\mathbf{Y} \\ &= [\mathbf{X}^T_2(\mathbf{P}_1)\mathbf{Y}]/\,[\mathbf{X}^T_2(\mathbf{P}_1)\mathbf{X}_2] \end{aligned}$$

Now if the roles of $\mathbf{Y}$ and $\mathbf{X}_2$ are swapped then the partitioned model can be written:

$$\mathbf{X}_2 = \mathbf{X}_1\theta_1 + \mathbf{Y}\theta^*_2 + \varepsilon^*$$

And the normal equations are:

$$\begin{pmatrix} \mathbf{X}^T_1\mathbf{X}_1 & \mathbf{X}^T_1\mathbf{Y} \\ \mathbf{Y}^T\mathbf{X}_1 & \mathbf{Y}^T\mathbf{Y} \end{pmatrix} \times \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}^*_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T_1\mathbf{X}_2 \\ \mathbf{Y}^T\mathbf{X}_2 \end{pmatrix}$$

The solution of the equations gives $\hat{\theta}^*_2$:

$$\hat{\theta}^*_2 = [\mathbf{X}^T{}_2\mathbf{P}_1\mathbf{Y}]/[\mathbf{Y}^T\mathbf{P}_1\mathbf{Y}]$$

The numerator for the two least squares estimates are equivalent as:

$$[\mathbf{X}^T{}_2\mathbf{P}_1\mathbf{Y}] = [\mathbf{X}^T{}_2\mathbf{P}_1\mathbf{Y}]^T = [\mathbf{Y}^T\mathbf{P}_1\mathbf{X}_2]$$

however the denominators are not as:

$$[\mathbf{Y}^T\mathbf{P}_1\mathbf{Y}] \neq [\mathbf{X}^T{}_2\mathbf{P}_1\mathbf{X}_2]$$

So that in the multi-variate case the size of the regression parameter depends on the precedence of predictor and explanatory variable also.

2. Interpreting estimates of parameters

The discussion regarding regression and estimation of regression coefficients, which is the underlying theme of the two papers cited as examples of this technique in the assessment of responsiveness, focuses, in part, on the absolute size of the estimated regression coefficient.

The review paper by Husted and colleagues (32, p 463) states:

Values for b [the regression coefficient estimate] *near zero suggest large changes observed in X may not be accompanied by changes in Y and large values of b mean the associated changes in Y will also be large.*

Later in the discussion the authors do acknowledge that this is scale dependent and the statistical significance of the estimate of the coefficient should also be examined.

In this context it is the ratio of the parameter estimate to its standard error that truly assesses responsiveness. This reflects the probability that QoL changes in relation to a unit change in the external criterion. In the setting of the general linear model, for each parameter estimate, b_i, which predicts a QoL instrument score, based on an external

criterion change, the standard error is equal to $[(\mathbf{X}'\mathbf{X})^{-1}{}_{ii}s^2]^{1/2}$, representing the i^{th} row and column of the matrix, $(\mathbf{X}'\mathbf{X})^{-1}$, and s^2 the estimate of the variance for each element of Y, these variances assumed to be independently and identically distributed. When comparing different QoL instruments, using the same external criterion as an explanatory variable, the ratio of the parameter estimate to its standard error can be described as the probability that a value as, or more extreme, as this can arise compared with the value from a t distribution with appropriate degrees of freedom. Sample size and the number of other parameter estimates that are made in the modelling process are important as they determine the degrees of freedom. The overall fit of the model is also important. If the other explanatory variables in the regression model are not accurately specified, that is if the model is over fitted or under fitted, then the variance structure of the model will be inaccurate. As a consequence statements about the size and variance of the external criterion parameter will also be inaccurate.

3. Is the scale on which QoL measured important?

QoL instruments are often measured on different numeric scales, for example 0-10 or 0-100. In comparing QoL instruments there is often a temptation to re-scale them so that changes, for example in regression models, can be directly compared. There are two problems with this. Firstly this sort of 're-scaling' assumes that the population minimum and maximum scores for each QoL instruments, i.e. 0 and 100, are then the same for each instrument, and that in any particular sample and sample size that the cumulative distribution function, and the actual distribution, of the minimum and maximum values will be the same. This may not be the case. Consider two QoL instruments, QoL_1 and QoL_2, that are scored on scales that range from 0-100 and 0-10 respectively. In order to transform a change in score from QoL_2 to QoL_1 so that the 'distances' are equivalent, what is important is not the nominal score range of 100 units and 10 units respectively, but the 'actual' range for the population rather than the sample, of each instrument in relation to the population minimum and maximum for each instrument. This treats each of the QoL instruments as instruments which measure on an 'interval' rather than a 'ratio' measurement scale. For example the population minimum and maximum for QoL_1 may be 20 units to 80 units, a range of 60 units, and that for QoL_2 may be from 1 unit to 9 units, a range of 8 units. The appropriate linear transformation so that a 10% change in QoL_1 is equivalent to a 10% change in QoL_2 is

to scale the instruments so that 0.8 unit change in QoL_2 is equal to a 6 unit change in QoL_1, thus making QoL_2 range from 0 to 75, rather than the intuitively appealing procedure of scaling it to 0 to 100. The second and perhaps more fundamental problem is that in any given study the population minimum and maximum have to be estimated from the sample used in the responsiveness study. Often the samples used in responsiveness studies are not particularly representative of the overall population with a particular disease, for example they attend a referral centre clinic, and are seldom randomly selected from any particular population, for example they may be an opportunity sample of those attending such a clinic in consecutive order. In any case the estimate of the range of QoL then in turn depends on both the sample size and the underlying probability distribution function for scores on a QoL instrument.

Gibbons (82, p 36) gives a general expression for the probability density function for the range, R, based on a sample size, n, with u =sample maximum – sample minimum and v = sample maximum, and $f_X(x)$ the probability distribution function the random variable X:

$$f_R(u) = \int n(n-1)(F_X(v)-F_X(v-u)^{n-2})f_X(v-u)f_X(v)dv$$

and further states (ibid. p 37) tables are available to evaluate the cumulative distribution function of this integral where f_X is the normal distribution, up to n of 20, and that an asymptotic distribution with f_X the normal distribution, also available.

What then are the implications of this discussion for using regression techniques to estimate responsiveness? Firstly the apparently simple approach of transforming QoL instruments measured on different scales will not actually lead to equivalence of interval distances as these depend on the population ranges for the instruments, which are usually unknown. Secondly in using the data to estimate the population ranges this is dependent both on the sample size and on assumptions about the probability distribution of the QoL instrument. Sample size may vary within studies and across studies. Both of these factors must lead to considerable caution in interpretation of the size of parameter estimates in regression carried out on inappropriately 're-scaled' QoL instruments. Finally the actual probability density function for QoL instruments may

not have a normal distribution and may in fact vary in different populations. How then to proceed to calculate an estimate of the population range is unclear.

In light of the discussion in section 2 above the issue of scaling is peripheral to the actual interpretation of responsiveness. In addition if the size of the regression parameters are to be used then it seems more appealing to present them in terms of the actual QoL instrument as it will be used in research.

4. The change in QoL should be in a predictable direction

For the examples used by Husted and colleagues the natural history of the disease was used to guarantee that change in QoL took place. This assumption and its consequences could be challenged in a number of ways. The first relates to the reason for wishing an instrument to be responsive, that is so that change in a clinical trial or other evaluative setting can be captured. Any setting that wishes to establish responsiveness should have some assurance that change will, in fact, occur and that this change should be measurable with a QoL instrument. This is not guaranteed in a natural history observational setting and any change in QoL that does occur may not be particularly related to changes in the disease, and may be confounded by accommodation to the altered life style induced by the disease. Secondly the external criteria, if more than one criterion is used, should change in the same direction that, at face value, should be associated with a consistent direction of change in QoL. The use of regression techniques should occur in the setting of an intervention or time period where change in a positive or negative direction will occur. A QoL instrument which had a score which changed even if the change in an external criterion was negligible would be a 'noisy' instrument, rather than an instrument that could give a strong 'signal' that change had occurred. The papers by Husted and colleagues refer to change measured over only one time interval, and then, in essence, analyse the differences between these two times. If there were a number of measurements made of an external criterion and a QoL instrument over time then it is possible to capture richer co-variance structures in the regression process by developing appropriate repeated measures models.

5. Model building and QoL

The very general model relating a QoL instrument to an external criterion is:

$$\delta_{QoL} = \alpha \times \delta_{External\ criterion}$$

Where δ_{QoL} refers to the change in the QoL instrument score, $\delta_{External\ criterion}$ refers to the change in external criterion. As discussed a larger value of the ratio of the estimate of 'α' in relation to the standard error of the estimate means that an instrument is more responsive. Now if δ_{QoL} depends on factors other than the change in the external criterion, that is that the model is under specified in some way, this will lead to inflation of the estimate of the mean square error of the model, and inflation of the estimate of the standard error of the parameter 'α'. It may be that different QoL instruments are some how sensitive to different co-variates in terms of their change in response to an external criterion. One instrument may have a gender bias, for example it penalises the ability to apply make up or bear children, where another instrument has no items relating to these activities. Thus in order to determine the sensitivity of an instrument to a change in an external criterion the best possible model should be constructed that incorporates measures of the explanatory power of the model. This will minimise the mean square error of the model, and hence decrease the standard error of the parameter estimate. Such measures could include R squared. However in addition the model should be parsimonious enough so that R squared is not spuriously inflated by merely including more explanatory variables. The model should not give rise to biased estimates. Measures such as Mallow's C_p, and variance inflation factors may be useful. The approach taken by Husted and colleagues in their second paper (80) of including change scores of all three of the QoL instruments they examined in the same multiple regression, in the knowledge that these were highly correlated, to predict change in disease status and then reporting the R squared for this analysis seems nonsensical.

Polynomial extensions of this model could also be examined, for example:

$$\delta_{QoL} = \alpha \times [\delta_{External\ criterion}]^b$$

Where if the model holds the power function could be assessed by taking the logarithm of both sides:

$$\log(\delta_{QoL}) = \log(\alpha) \times b\log(\delta_{External\ criterion})$$

and treating α as a nuisance parameter.

6. Instruments which are not designed to be reduced to a single score

Some QoL instruments are not designed to be reduced to a single score. It might be questionable whether instruments of this type should be subjected to analysis of responsiveness at all. Responsive instruments are of most utility in comparative trials. Multi-dimensional instruments are difficult to use as outcome measurements in randomised trials because of potential problems with multiple comparisons or, if the instrument is used as a multi-variate outcome measurement, interpreting differences between treatments in terms of the vector of the response. If the outcome measurement is the 'overall' QoL, that is treating QoL as a unitary, albeit abstract, construct suggests that responsiveness should be measured on this unitary construct, that is that some sort of scaling, perhaps of the simple sort suggested by Cox and colleagues (5) should be used regardless of how the instrument was constructed. The comparison of different regression coefficients for different dimensions of a QoL instrument in respect of a single external criterion for change is likely to lead to confusion rather than clarity as to whether the instrument as a whole is responsive.

7. Simple linear regression may be insufficient

QoL may not be related in a linear fashion to external criterion. For example a change in one inflamed joint for some one who has one inflamed joint already may lead to a marked change in QoL when a change in one inflamed joint for some one who already had 10 inflamed joints may make no overall difference to QoL. In addition simple external criteria may not be sufficient markers of disease severity. For the joint example if a simple joint count is used the effect on QoL when two finger joints in the non dominant hand are involved may be quite different from the effect when two knee

joints are involved. Non-linear association between external criteria and the QoL may exist and these relationships may not be captured in without, for example examination of models that include polynomial terms.

Conclusion: Regression

Intuition suggests that regression modelling may offer insight into the responsiveness of a new QoL instrument by assessing how much information derived from an external criterion for change is 'captured' by the new instrument, and by assessing the variance of this relationship. The use of regression modelling raises specific issues which should be addressed so that the technique can be usefully applied to the assessment of responsiveness. These issues include the accurate specification of variance-covariance structures, the completeness and correctness of proposed models, and testing assumptions that underlie the analysis.

Chapter 5: Mixed linear models for external responsiveness

Sophisticated modelling strategies for the relationship between QoL instrument scores and external criteria may be redundant. A simple modelling process may be all that is necessary to reach a decision on which QoL instrument may be more responsive because it is uncertain as to which, if any, strategy is necessary to reduce a QoL instrument to a single score (5). It is also uncertain which external criterion should be used to test any particular QoL instrument. However as external responsiveness, using regression techniques, is related to the ratio of the estimate of a regression parameter to its variance, then an ideal modelling process should give rise to estimates of these two parameters that are in some sense the best estimates. In particular a more sophisticated modelling process to determine external responsiveness of QoL instruments should take into account that repeated measures on the same subjects are often used to determine responsiveness. Repeated measures on the same subjects are often correlated and this affects the estimation of parameters and, to an even larger extent, the variance of parameters. The modelling process should also be able to account for multiple predictors and functions of the predictors, for example models which are polynomial in the predictors. Some of the issues discussed in the previous chapter, such as precedence, ensuring the change in QoL will be in the same direction for predictors, and having sufficient explanatory variables, are conceptual and structural issues for any particular study rather than statistical issues.

A mixed linear model seems a good candidate for modelling a variety of sources of variation and for repeated measures designs. A variety of synonyms are used for mixed linear models such as 'two stage random effects models', 'multi-level linear model', 'hierarchical linear model', and 'random regression coefficients' models (83-87).

The most 'usual' linear model, of which simple linear regression is an example, has the form:

$$\mathbf{Y} = \mathbf{X}\alpha + \varepsilon$$

Where **Y** is an n×1 response vector, **X** is an n×p design matrix, α is p×1 vector of fixed parameters, and ε is an n×1 vector of error terms associated with each element of the response vector. The expected value of **Y** is $\mathbf{X}\alpha$ and the variance-covariance matrix of **Y**, under the simplifying assumption of independent and identically distributed variance terms, is $\sigma^2\mathbf{I}$, where **I** is the n×n identity matrix and σ^2 is the variance of an individual element of the response vector. The simple linear model can be extended such that the variance-covariance matrix of **Y** is an n×n matrix of known constants. For the purposes of the following discussion **Y** will also be assumed to have a multivariate normal distribution.

The mixed linear model is an extension of the simple model but allows for a second element in the relationship between the response vector and the explanatory vectors and associated error terms (85-87). This model can be expressed as:

$$\mathbf{Y} = \mathbf{X}\alpha + \mathbf{Z}\beta + \varepsilon$$

Where **Y**, **X**, α and ε are defined as for the linear model but the term $\mathbf{Z}\beta$ allows modelling of other sources of variation in the response vector, because the additional parameters in the vector β are defined to be random rather than fixed.

Z is an n×q design matrix for explanatory variables. The variance-covariance of the vector $\mathbf{X}\alpha$ is zero but the variance-covariance matrix of the vector $\mathbf{Z}\beta$ is non-zero. β is the q×1 vector of parameters corresponding to these random effects.

In this expanded model the variance-covariance matrix of **Y**, **V**, has the more complex form:

$$\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$$

G is a q×q matrix which in general has a block diagonal structure corresponding the variance-covariance structure for the random effects parameters, the vector β, for each individual element of the response vector **Y**. **R** is an n×n matrix with a block diagonal structure that contains the 'residual' variance after the variance related to the random

effects is taken into account. Under assumptions of multivariate normality β is assumed to be distributed as N(**0**, **G**) and ε as N(**0**, **R**). Often **R** is $\sigma^2\mathbf{I}$ as for the simpler fixed parameter case which does not involve β.

The factorisation of **V** described above is a notational convenience for estimation of the individual parameters of the vectors α and β that can be carried out using, for example, the 'PROC MIXED' procedure available in SAS.

This thesis will use two techniques for estimation of the parameters and their variance. The first is that used by the SAS procedure 'PROC MIXED' based on maximum likelihood estimation, and the second based on Bayes techniques as implemented in the software 'WinBUGS' or 'MlWin'.

Maximum likelihood estimation

The default estimation procedure for parameters used in SAS is 'residual maximum likelihood' also known as 'restricted maximum likelihood', abbreviated REML. Rather than maximising the logarithm of the likelihood function for the multivariate normal distribution:

$$L = k - \tfrac{1}{2}[\log|\mathbf{V}| + (\mathbf{Y} - \mathbf{X}\alpha)'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\alpha)]$$

REML maximises a likelihood function based on the residuals: $(\mathbf{Y} - \mathbf{X}\hat{\mathbf{a}})$ where $\hat{\mathbf{a}}$ are the estimates of α. The estimate of α is found by solving the equation $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ and this leads to a log likelihood that is based only on **V** and the design matrix for the fixed effects **X**. In practice **V** is dependent on β which is unknown, and must be estimated, and on **Z**, so that estimating α, β and **V** becomes an iterative process. At each iteration α, and then β, and then **V**, are estimated. Restricted likelihoods can be maximised by numerical methods, such as the Newton-Raphson method.

The variance of the estimate of α is $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$ although this is based on the assumption that **V** is known. Brown (85) states that this results in downward bias in the variance of

the estimate of α although this bias can be, but is not always, small. The bias is most likely to be relevant when the variance parameters are imprecise, the ratio of the variance parameters to the residual variance is small or when there is a large degree of imbalance in the data. Brown goes onto suggest (ibid.) that bias may be 5% or more if the number of random effect categories relating to a variance parameter is less than about ten and the ratio of the variance parameter to the residual variance is less than one.

In addition tests for fixed effects may be based on statistics that are only approximately from an F distribution, and the degrees of freedom for the denominator of the F statistic may be difficult to calculate (87).

Measures of model fit

Overall measures of model fit (85-87) that are based on the likelihood include comparison of the log likelihood of two nested models, that is models which fit the same fixed effects and whose variance-covariance structures are also nested, by this meaning that one covariance structure is a restricted form of another covariance structure. An example of this is that a compound symmetry structure is a restricted version of a 'Toeplitz' structure. In the latter time points the same distance apart have the same correlation, where in the former the correlation between all different time points is the same. Akaike's Information Criterion (AIC) can be used to make direct comparisons between models with the same fixed effects. This is given by:

$$AIC = \log (L) - q$$

Where L is the likelihood and q the number of covariance parameters.

Schwarz's information criterion (SIC) takes into account the number of fixed effects, p, the number of observations, N, and the number of covariance parameters, q. For REML estimation this is given by:

$$SIC = \log (L) - (q \log(N - p))/2$$

A larger value of either of these criteria means a better overall fit. In effect the number of parameters fitted is a penalty, so that models with more parameters are more heavily penalised relative to the log likelihood which increases as the number of parameters fitted increases.

Littell and colleagues (87) suggest plots of correlation between repeated measures provide a visual check of how a chosen covariance structure compares with an 'unstructured' model and to compare between different models. Brown (85) suggests however that with few repeated measures reasonably robust estimates of fixed effects can be made using a simple structure for the covariance, such as a compound symmetry model.

Analysis of residuals may be difficult for covariance structure models, which are assumed to have zero mean but a multivariate normal covariance matrix. However residual plots, for example residuals plotted against predicted values, may be useful to detect outliers, and then temporary removal of outliers from the data set can be used to check if parameter estimates change a lot, a form of sensitivity analysis.

Bayesian techniques

Bayesian techniques are an alternative way of estimating the parameters of the mixed linear model (88-92).

Bayesian techniques use the general concept that:

$$p(\theta|\mathbf{y}) = L(\theta|\mathbf{y})p(\theta)/K$$

Where K is $\int L(\theta|\mathbf{y})p(\theta)\, d\theta$

$p(\theta|\mathbf{y})$ represents the probability density function of a vector of parameters, θ, given a vector of data, $\mathbf{y}$. It is referred to as the posterior density of θ, in the sense that it is calculated after specification of the functions on the right hand side of the equation and

after the data is incorporated into the Bayesian model. These functions are $p(\theta)$ and $L(\theta|\mathbf{y})$. $p(\theta)$ is the probability density function of the parameters based on some sort of prior knowledge of these parameters, both of the form of the probability density function, for example the normal distribution, and in terms of the parameters of the probability density function, such as values of the mean and variance of that normal distribution. It is often just known as the 'prior'. The function $L(\theta|\mathbf{y})$ is the usual likelihood function describing the likelihood of the parametric specification of θ given a vector of observations. The constant 'K' is necessary so that the probability density function $p(\theta|\mathbf{y})$ integrates to one. The prior distribution can be specified in a way that reflects 'ignorance' about the values of the parameters, for example specifying a normal distribution with a mean of zero and a very large variance. The prior distribution can also be specified in a way such that its form is 'reproduced' in the form of the posterior distribution, a property known as conjugacy. For the binomial parameter specifying a beta distribution as the prior will lead to a posterior that is also a beta distribution, and the normal and gamma distributions are jointly conjugate for the normal distribution. The evaluation of the integral to give a value for 'K' can use a Markov Chain Monte Carlo method that uses a Gibbs sampler(93), leading to a full specification of the posterior distribution of the vector of parameters, θ. In an analogy to forming expected values and confidence intervals using 'frequentist' methods, i.e. based on maximum likelihood and asymptotic theory, values of parameters of the posterior distribution can be calculated, for example the mean or median, together with 'credible intervals' calculated from the probability distributions for these parameters.

Monte Carlo integration is based on the idea that the expected value of a function can be approximated by taking a very large number of random samples from that function and dividing by the number of samples taken. A Markov chain is a sequence, or chain, of random variables such that the value of one random variable depends at most on the value of the random variable that immediately precedes it in the chain. The function that describes the relationship between a pair of adjacent members of the chain is called the 'transition kernel' of the chain. A property of Markov chains is that, subject to regularity conditions, after a sufficiently long sequence, say T, the distribution of the T^{th} variable X_T will be from a probability density function that does not depend on where the chain started. This probability density function is the 'stationary' or 'invariant'

distribution of the Markov chain. Thus with time the successive values of X resemble dependent samples from this stationary distribution. Gibbs sampling is a particular example of a Markov Chain Monte Carlo method. It works by sampling each parameter from a vector of parameters one at a time conditioned on the previous values for the other parameters. This process is iterative so that once a full cycle of sampling has taken place the process repeats. Repeating this process many times, for example thousands or tens of thousands of times, usually will lead to the parameters converging to their full conditional distribution, despite the dependencies in the process. Starting values of the parameters must be specified, as well as their prior distribution. Typically the values for the parameters found by the first 1000 or so iterations are unstable and are typically discarded. Gibbs sampling can be carried out, for example, using the software package 'WinBUGS'.

Measures of model fit

This can be carried out at both an individual observation level and at an overall model level (91,94-96). The individual observation level checks involve comparison of an observed data point, y_i, with a predictive distribution $p(Y_i)$. This can include residuals; $y_i - E(Y_i)$, standardised residuals; $((y_i - E(Y_i)/\sqrt{(V(Y_i))}$, the chance of a more extreme observation, the chance of a more 'surprising' observation, and the predictive ordinate of the observation $p(y_i)$.

The source of the predictive distribution can be from another data set, if one is available, or from within the source data set, where the predictive distribution is based on all the data except for the particular data point under consideration, that is cross validation.

Comparisons between two or more candidate models can be carried out based on the sum of squares or sum of absolute values of the cross validatory measure described above. In addition a 'pseudo-Bayes' factor, which is related to the sum of the log of the predictive ordinates for the observations, can be used to compare models.

These model checking procedures can be implemented from within 'WinBUGS' or by processing output from the program.

Application to external responsiveness of QoL instruments

Consider the situation where a QoL instrument is administered to a single group of subjects on more than one occasion. Between two of the measurements using the QoL instrument an intervention is applied to the subjects that is likely, based on previous experience with subjects in a similar situation, to improve the disease and the QoL of the subjects. Co-variates are measured on the subjects including a criterion that is anticipated to change as the disease changes.

A model for this situation could have the form specified earlier in the chapter where for the vector of QoL measurements, **Y**:

$$\mathbf{Y} = \mathbf{X}\alpha + \mathbf{Z}\beta + \varepsilon$$

$$\mathbf{Y} \sim \text{MVN}(\mathbf{X}\alpha, \mathbf{V})$$

Y represents the vector of QoL measurements for I subjects measured on J_i occasions, α represents a vector of fixed effect parameters that could include an overall mean as well as parameters describing the relationship of an external criterion to the QoL measurement, β a vector of random effect parameters, and ε, the vector of error terms. **V** has a block diagonal structure by individual subject since in general successive measurements taken on an individual are correlated, relative to the overall mean across subjects and due to the random differences between individuals, and measurements on different individuals are uncorrelated.

Brown (85) and Littell (87) recommend that in most circumstances the repeated measures variance covariance structure is most easily modelled by using the variance of ε, the **R** matrix, and setting the variance of β, the **G** matrix, to zero. The **R** matrix has a block diagonal structure with each block representing an individual, and the covariance between measurements of different individuals is set to zero.

This is equivalent to ignoring random effects. Conditions under which this gives equivalent estimates of the fixed effect parameters and their variance are however not sufficiently general to make this a viable solution in all cases (97).

The general structure of the i^{th} block which corresponds to the i^{th} individual, $\mathbf{R_i}$, is given below for the example of three measurements per subject:

$$\begin{bmatrix} \sigma^2_1 & \theta_{12} & \theta_{13} \\ \theta_{12} & \sigma^2_2 & \theta_{23} \\ \theta_{13} & \theta_{23} & \sigma^2_3 \end{bmatrix}$$

In this matrix the diagonal elements represent the variance of the measurement at an particular time point and the off diagonal elements in this symmetric matrix represent the covariance of measurements made at different times on the same subjects.

Using a univariate notation this modelling process corresponds to:

$$Y_{ij} = \mu + \alpha x_{ij} + d_i + \varepsilon_{ij}$$

Where Y_{ij} is a QoL measurement made on the i^{th} subject on the j^{th} occasion, μ represents the overall mean, α represents the parameter relating Y_{ij} to the value of the external criterion for the i^{th} subject on the j^{th} occasion, x_{ij}. d_i is a normally distributed variable with mean zero and variance σ^2_d representing a random subject effect, and ε_{ij} is a normally distributed variable with mean zero and variance σ^2_ε representing the residual error. All terms are independent of each other. The correspondence with the terms in the sub-matrix $\mathbf{R_i}$, above, are that:

$$\sigma^2_i = \sigma^2_d + \sigma^2_\varepsilon \text{ and}$$
$$\theta_{ij} = \sigma^2_d + \text{cov}(\varepsilon_{ij}, \varepsilon_{ik})$$
$$= \sigma^2_d + \rho\sigma^2_\varepsilon$$

The ratio of the estimate of the parameter relating the external criterion to the QoL measurement to its variance using either maximum likelihood or Bayes techniques, gives a measure of the responsiveness of the instrument.

A further fixed effect could be introduced that is an indicator of whether measurements occurred before or after an intervention.

$$Y_{ij} = \mu + \alpha x_{ij} + \beta I_j + d_i + \varepsilon_{ij}$$

With the notation described above but the fixed effect indicator variable I_j equal 1 if the ij^{th} measurement occurs after the intervention and equal to zero if before.

This could be used to assess the completeness of the model. For example if after taking into account the effect of the external criterion predictor a second model that also includes an intervention effect does not improve the model fit, this suggests that the QoL instrument is responsive to the external criterion and captures all the information about the change in status of the subjects that the external criterion does. If both effects are needed for a good model this suggests the QoL instrument is responsive but that the external criterion, on its own, is insufficient to explain the change in QoL, as measured by the external criterion, in relation to the intervention. If neither parameter is important in a model this suggests the instrument is not responsive to the particular external criterion, assuming the external criterion captures the change in the disease process due to the intervention, and that the intervention is successful in changing the external criterion and at changing QoL.

Further predictor variables could be added to the model to test for gender or age effects. Other disease related variables, such as the use of medication, marital status, job status, education, ethnicity, and duration of disease could also be added, as could functions of predictor variables, such as polynomials. If such a study was carried out at a number of different centres, or with interventions that were homogenous with respect to expected outcome but heterogeneous with respect to actual implementation, then these could be modelled by including random effect parameters.

The ideal model should be parsimonious with respect to the number of parameters specified yet complete enough to be good model for the data. In terms of modelling the random effect parameters, for the purposes of this discussion in the **R** matrix, a number of co-variance structures exist. These range from a simple structure, where all off diagonal values are set to zero, which corresponds to the assumption that repeated measurements on the same subjects are not correlated, to a completely unstructured model where all of the variance-covariance parameters are different. Between these two extremes are structures such as the compound symmetric structure, where all the off diagonal elements are identical, representing constant covariance between different measurements on the same individuals. A large number of ways for relating the difference in covariance parameters to the difference in time between the measurements exist, such as banded (Teoplitz) or auto-regressive structures, although these are usually more relevant to data analysis situations with a larger number of repeated measures than specified in the above discussion, since the extent of the decay of correlation between measures on an individual are likely to depend on the increase in time interval between measurements.

An example of the use of a linear mixed model in the analysis of a QoL instrument, although not in assessing external responsiveness, is given by Beacon & Thompson (98). This study used the software 'MLn' to study the change in a QoL instrument score of a group of subjects measured repeatedly over an eight week period following randomisation to one of two sorts of cancer treatment. A variety of models including baseline co-variate and random effects, and different specifications of co-variance structures were analysed, using as one criterion for goodness of fit, the log likelihood. Other ways of assessing the goodness of fit of the model available within SAS, such as Akaike's information criterion and Schwarz's information criterion, were not reported, although some information regarding normality assumptions was presented. The authors extended the repeated measures to also look at a two level vector of different dimensions of the particular QoL instrument analysed.

Chapter 6: The study

The data set for this thesis is based on the study 'Measuring change in Quality of Life: Three measures compared' performed at the Rehabilitation Research and Teaching Unit of the Department of Medicine, Wellington School of Medicine and Health Sciences. This study aims to compare three instruments which measure QoL in patients with rheumatoid arthritis in order to determine which instrument is most responsive.

Rheumatoid arthritis is a chronic inflammatory disorder of joints which leads, in the short term, to pain and loss of function of the joints and of the patients ability to do tasks, and in the long term to joint deformity and as a consequence the inability to use joints for tasks related to those joints (99). If the affected joints are in the lower limb, for example the knee and hip, then walking may be compromised. If the affected joints are in the hand, which is commonly affected in rheumatoid arthritis, then the ability to manipulate clothing, write and grasp objects may all be compromised. When the patient is very severely affected other organ systems may be affected, such as the lungs, which can lead to symptoms unrelated to the joints, such as shortness of breath. People with rheumatoid arthritis can also suffer from fatigue and malaise. There are a number of medical and surgical treatments for rheumatoid arthritis which can reduce pain and improve joint function. Some of these treatments are accompanied by the potential for adverse effects related solely to treatment. Rheumatoid arthritis tends to be a long term condition so that both the disease and factors related to treatment for the disease need to be taken into account in conducting trials of therapy. It is a disease where QoL measurements are thought to be very useful in controlled trials of therapy. However there is uncertainty about which QoL or health status instrument to use in evaluation of new therapies.

The QoL instruments

The three QoL instruments that are the subject of the study are the EuroQol quality of life scale (EuroQol), the World Health Organisation Quality Of Life-Abbreviated version (WHOQoL-Bref) and the Quality of Life Profile (QLP).

The EuroQol was described by the European Quality of Life Group in 1990 (100). It consists of 5 questions each with 3 possible responses, and a separate visual analogue scale. The question responses are each given a weight and the sum of the weights for the responses are subtracted from one to give an overall QoL score. The score on the visual analogue scale is not combined with the questionnaire score but is given as a separate rating.

The WHOQoL-Bref (101) consists of 26 items phrased as questions which are scored between 1 and 5. There are four dimensions covered by 24 of these questions and an additional two questions which relate to overall quality of life and general health. The scoring method recommended is to multiply the mean score for each of the dimensions by four.

The QLP was described by Raphael and colleagues (102,103). It is a relatively lengthy and complex instrument. It assesses 9 dimensions of QoL. These are labelled: My body and my health, my thoughts and feelings, my beliefs attitudes and values, where I live, the people around me, access to resources, practical things I do, things I do for enjoyment, and things I do to improve myself. Each of these 9 dimensions is assessed in 4 different ways. Firstly as to how important the dimension is and secondly how satisfied in the particular dimension the person is. For both these assessments there are a number of different items, from 9 to 13, with each item scored between 1 and 5, reflecting 'not at all important' to 'extremely important' and 'not at all satisfied' to 'extremely satisfied'. The other two ways of assessing the dimensions are firstly to ask how much control the subject has over each dimension, there are 9 items for this section, each item related to one of the dimensions, and each rated 1 to 5, 'almost no control' to 'almost total control'; and secondly asking how much opportunity there is to change one of the dimensions, again there are 9 items in this section, each item relating to one of the dimensions, and each rated 1 to 5, 'almost none' to 'great many'. The scoring system for the instrument can produce a single score.

All three instruments are completed by the subject ticking or writing in a manual that contains the instruments. The order in which the instruments were administered was randomly changed for each subject. All three instruments are in Appendix 1.

The external criteria used to analyse the external responsiveness

1. The Ritchie Articular Index score (104). This rates 25 joints by the amount of tenderness elicited by an examiner. Each joint is scored between 0 and 3, from not tender, tender, tender and winced, to tender and winced and withdrew. The index score is the sum of the scores for the 25 individual joint scores, and can range from zero to 75. A higher joint index score reflects increased disease activity. This index is completed by an examiner.
2. The Health Assessment Questionnaire (HAQ). This is an established instrument for the assessment of QoL. It has three dimensions. The first is physical ability and is assessed using 8 items each rated on 0 to 3 from 'Without any difficulty' to 'Unable to do' and the scores on the 8 items averaged to give a score from 0 to 3, a higher score reflecting worse QoL. The second is a general question on how the arthritis is affecting the person overall, the question 'How well you have been doing over the last 24 hours, rated on a visual analogue scale from 'Very well' to 'Not at all well'. The third is a question about pain, rated on a visual analogue scale for 'How much pain have you been having over the last 24 hours', rated from 'No pain' to 'Pain as bad as it could be' (37,39,105-110). The instrument is completed by the subject.
3. Simply asking the subject to rate on a 5 point ordinal scale whether they have feel worse or better since the previous assessment.
4. The ESR, a blood test that is a marker of inflammation.

The Ritchie Articular Index and the HAQ are in Appendix 1.

Structure of the study

Patients with rheumatoid arthritis who are treated in the tertiary referral centre at Hutt Hospital were approached for participation. Consenting subjects complete four interviews. An interview two weeks before admission for a period of intensive medical treatment for their rheumatoid arthritis, interview on the day of admission to the unit, an interview on discharge from the unit and an interview two weeks after discharge. At each interview the subjects complete all the instruments that are part of the study, as well as the external criterion QoL instrument, the HAQ, and have their joints assessed

using the articular index. For the second, third, and fourth assessments they are also asked if they have changed, using the simple 5 point ordinal scale. Although the study aims to recruit 100 subjects a smaller number, based on the number of subjects who have completed all the assessments, will be analysed in this thesis. A randomised clinical trial has demonstrated that a period of intensive inpatient therapy will improve rheumatoid arthritis (111). The measurement times in relation to admission of the subjects to the inpatient treatment program could vary considerably between subjects.

Chapter 7: Results

Simple summary statistics

Table 12 shows simple summary statistics for the QoL instruments in the study. Visit 1 is a pre-admission visit approximately two weeks prior to admission, visit 2 is the day of admission, visit 3 is the day of discharge and visit 4 is about 2 weeks after discharge.

Table 12: QoL instrument scores

Quality of life profile

Visit	Number of subjects	Mean	Standard deviation
1	62	0.46	0.72
2	62	0.50	0.73
3	56	0.83	0.88
4	59	0.77	0.88

EuroQol Visual analogue scale

Visit	Number of subjects	Mean	Standard deviation
1	67	49.3	16.7
2	63	53.9	16.7
3	64	69.4	15.9
4	56	69.9	18.1

EuroQol Rating scale

Visit	Number of subjects	Mean	Standard deviation
1	68	37.3	14.9
2	62	41.8	14.4
3	65	53.2	16.6
4	55	54.4	18.7

WHOQoL total

Visit	Number of subjects	Mean	Standard deviation
1	65	226	50.5
2	64	235.5	49
3	62	262.1	51.6
4	56	273.8	60.2

Table 13 shows simple summary statistics for the external criteria.

Table 13: External criteria scores

HAQ

Visit	Number of subjects	Mean	Standard deviation
1	70	1.86	0.65
2	69	1.70	0.67
3	69	1.31	0.67
4	60	1.30	0.73

Ritchie articular index

Visit	Number of subjects	Mean	Standard deviation
1	67	12.1	6.99
2	69	12.2	7.45
3	66	7.80	5.49
4	59	8.60	6.51

ESR

Visit	Number of subjects	Mean	Standard deviation
2	61	61.4	31.2
3	44	50.9	27.4

Note that a decrease in the HAQ, Ritchie articular index and ESR imply an improvement.

Measures of internal responsiveness

Definition and discussion of internal responsiveness is covered in chapter 2 of this thesis.

Table 14 shows the measures of internal responsiveness for each of the QoL instruments.

Table 14: Measures of internal responsiveness

	QLP	**EuroQol VAS**	**EuroQol Rating scale**	**WHOQoL Total**
RI_2 (Paired t test statistic)	3.74	5.37	5.34	6.47
RI_3[1] (Relative efficiency index)	1	2.1	2.0	3.0
RI_4 (Standardised effect size)	0.52	0.90	0.85	0.75
RI_5 (Standardised response mean)	0.51	0.73	0.73	0.81

[1]QLP used as denominator for calculation of RI_3

All the QoL measures are responsive, that is there is clear evidence of change across the period of the intervention. There seems to be little to chose between the EuroQol Visual analogue scale, the EuroQol rating scale and total of the WHOQoL scores. The

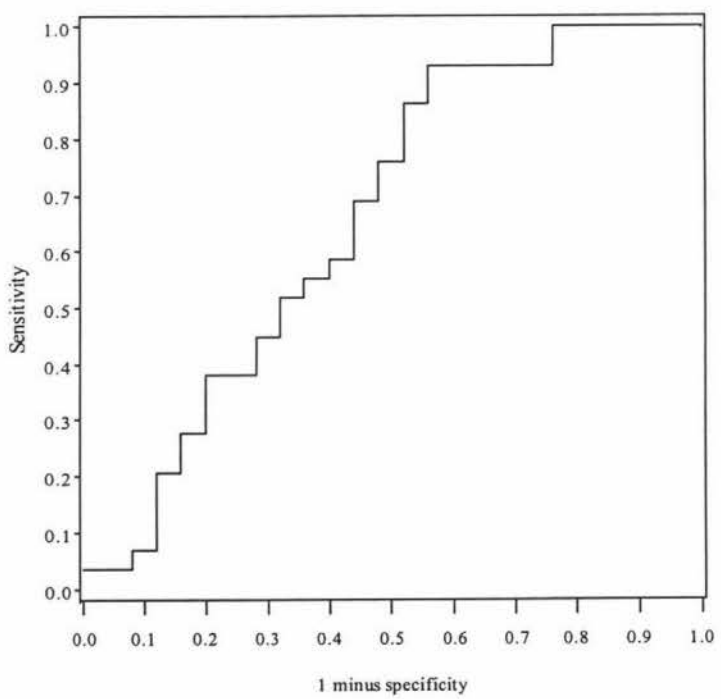
QLP may have inferior responsiveness compared to the other three measures based on these internal responsiveness statistics.

ROC curves

Definition and discussion of ROC curves is covered in chapter 3 of this thesis.

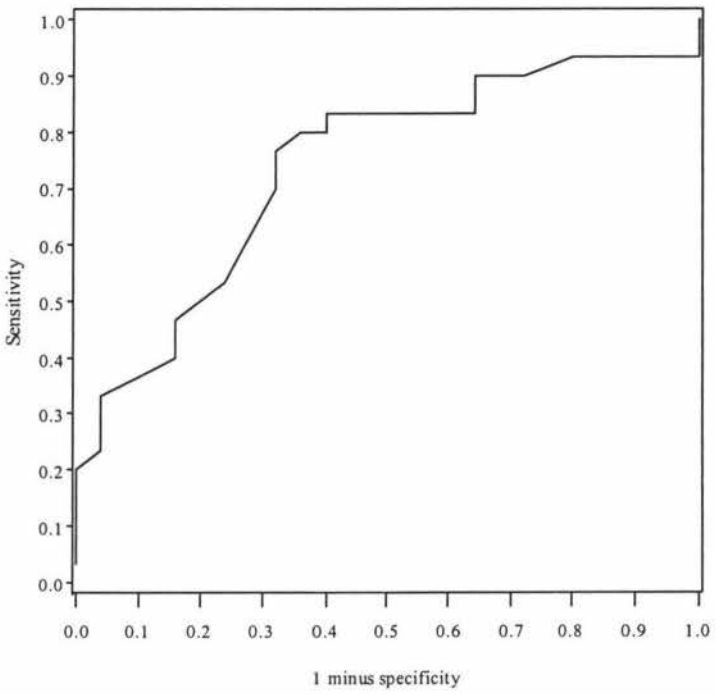
ROC curves were calculated using the difference in QoL instrument scores between visit 4 and visit 2. The external criterion for change was a marked improvement in quality of life on the 5 point Likert scale coded as a dichotomous variable.

Figure 13: QLP ROC curve



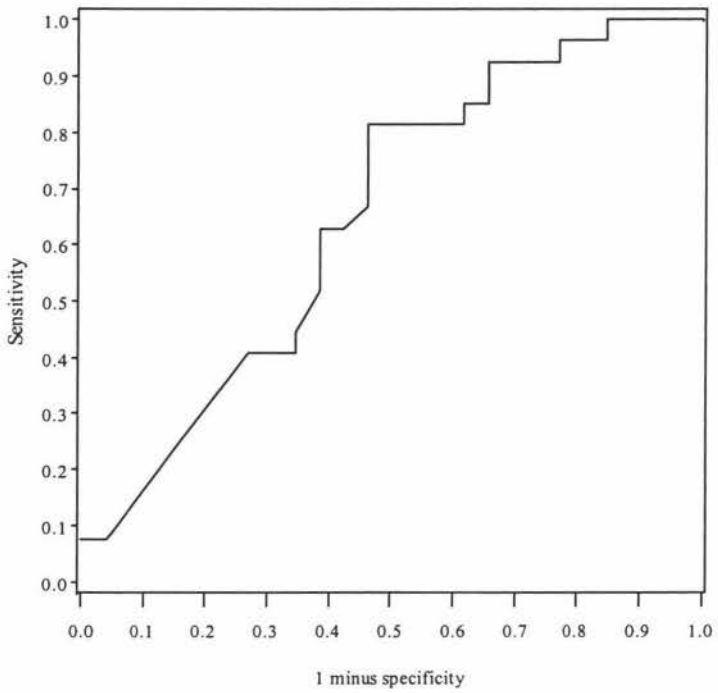
Area under the curve: 0.66

Figure 14: EuroQol Visual analogue scale ROC curve



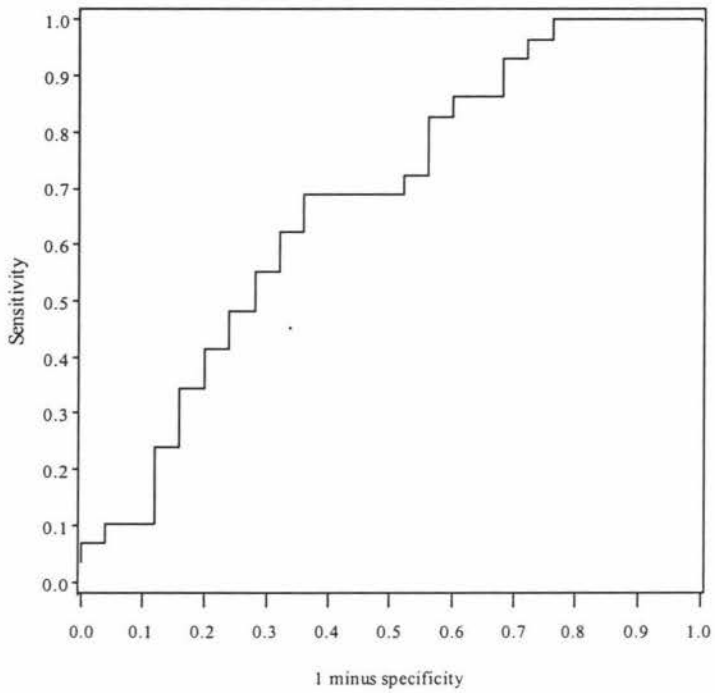
Area under the curve: 0.74

Figure 15: EuroQol Rating scale ROC curve



Area under the curve: 0.65

Figure 16: WHOQoL total score ROC curve



Area under the curve: 0.67

The AUC ROC are summarised in table 15.

Table 15: Summary for AUC ROC

QoL instrument	**AUC ROC**
QLP	0.66
EuroQol visual analogue scale	0.74
EuroQol Rating scale	0.65
WHOQoL simple sum of domain scores	0.67

An omnibus test for whether any of the ROC AUC were different from each other for QLP, the EuroQol visual analogue scale, EuroQol rating scale and the simple sum of the WHOQoL domains had a Chi-square statistic value of 1.36 on 3 df, P=0.71. This used the method suggested by Delong and colleagues (56) as implemented in a SAS program, SAS.ROC, available from the SAS web site (112).

For individual contrasts the estimates and 95% CI were:

QLP minus EuroQol visual analogue scale: -0.079 (–0.24 to 0.09)

QLP minus EuroQol rating scale 0.006 (-0.11 to 0.12)

QLP minus WHOQoL simple sum of domains: -0.017 (-0.18 to 0.15)

Based on this criterion for external responsiveness all the QoL measures had only moderate responsiveness and statistical testing was unable to distinguish between them. The point estimate for the ROC AUC was the least for the EuroQol Rating scale.

Correlation

Discussion of correlation in the assessment of responsiveness is covered in chapter 4 of this thesis.

Table 16 shows the Pearson's correlation coefficients and their 95% confidence intervals for the relationship between the QoL instruments and the external criteria for change.

Table 16: Correlation between the change in QoL and external criteria

	QLP total	**EuroQol VAS**	**EuroQol Rating scale**	**WHOQoL simple sum**
HAQ[1]	-0.37	-0.45	-0.50	-0.59
(95% CI)	(-0.58,-0.11)	(-0.64,-0.21)	(-0.68,-0.26)	(-0.74,-0.39)
Ritchie[1]	-0.31	-0.41	-0.35	-0.32
(95% CI)	(-0.53,-0.04)	(-0.62,-0.41)	(-0.57, -0.09)	(-0.54,-0.06)
ESR[2]	-0.23	-0.13	-0.34	-0.18
(95% CI)	(-0.53, 0.13)	(-0.46,0.22)	(-0.62,0.02)	(-0.50, 0.18)

[1] Change in scores Visit 4 minus Visit 2

[2] Change in scores for ESR Visit 3 minus Visit 2

Table 17 shows the omnibus tests for whether any of the correlation coefficients in the previous table, in each of the rows, are different from each other using the methods of Meng and colleagues (77) as outlined on page 58 of this thesis.

Table 17: Omnibus tests for differences in correlations

	C Statistic	Df	P value
HAQ	2.80	3	0.424
Ritchie	0.62	3	0.89
ESR	0.24	3	0.97

When correlation coefficients are used as a measure of responsiveness based on the external criteria statistical testing is unable to distinguish between the QoL instruments. The direction of the correlation is as expected for all the instruments, that is an increase in the score on each of the QoL instruments is associated with a decrease in the score on the external criteria. There is a modest, at most, relationship between the QoL instrument scores and the HAQ as the external criterion, a poor relationship between the QoL instrument scores and the Ritchie Articular index, and little relationship between the QoL instrument scores and the ESR.

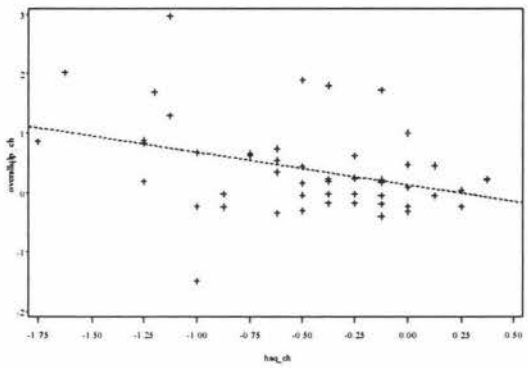
Regression analysis

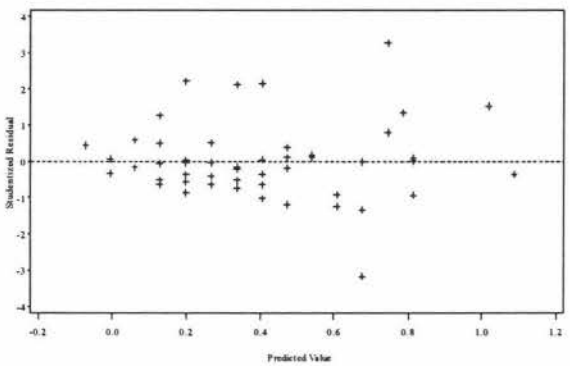
The theory and discussion of regression analysis for assessment of responsiveness is covered in Chapter 4 of this thesis.

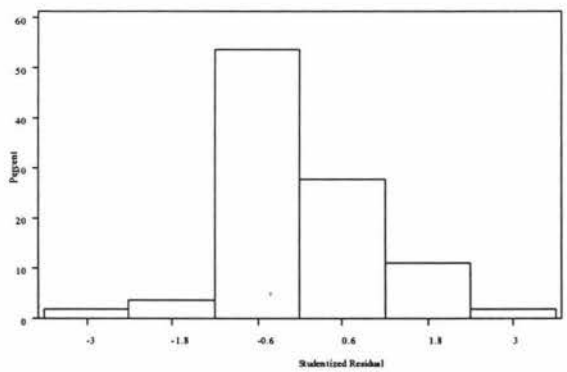
In the following simple regression analyses the change in the QoL instrument was regressed on the change in the external criteria. The change in the QoL instruments was the score on Visit 4 minus the score on Visit 2, and similarly for the external criteria for change, except for the ESR which was only measured at Visit 2 and Visit 3. The explanatory variable for all of the following plots is the change in the HAQ score. As will be clear from the summary of the regression analyses, the R squared value for the regression of the QoL instruments on the Ritchie articular index and the ESR all had very poor explanatory power, with none of the values for R squared greater than about 10%.

Plots of the regression variables with the regression line are displayed, as well as a plot of studentized residuals versus predicted values, and a histogram of the studentized residuals. Outliers and problems with residuals will be discussed for each QoL instrument in turn.

Figure 17: QLP versus HAQ







QLP versus HAQ discussion

Estimated model 1

QLP change = 0.13 – 0.55(HAQ change)

Table 18: QLP vs. HAQ model 1

Slope parameter	Standard error of slope parameter	T score	P(slope not equal to zero)	R squared for regression
-0.547	0.192	-2.85	0.0063	13.5 %

Outliers

Two subjects, subjects 1 and 28, had residuals with an absolute value greater than 3. Both these subjects had very large changes in QLP between the visits but changes in the HAQ that were disproportionately large or small. Analysis with these two subjects removed gave the following model.

Estimated model 2 (with outliers 1 and 28 removed)

QLP change = 0.14 – 0.52(HAQ change)

Table 19: QLP vs. HAQ model 2

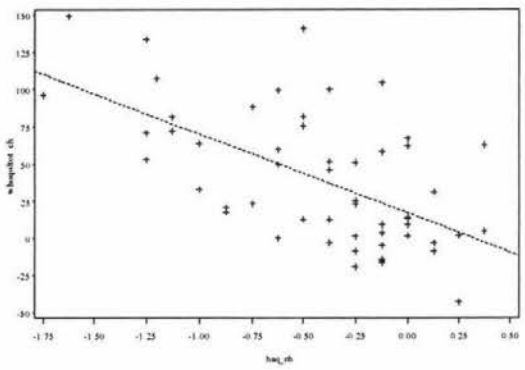
Slope parameter	Standard error of slope parameter	T score	P(slope not equal to zero)	R squared for regression
-0.521	0.159	-3.28	0.0019	18%

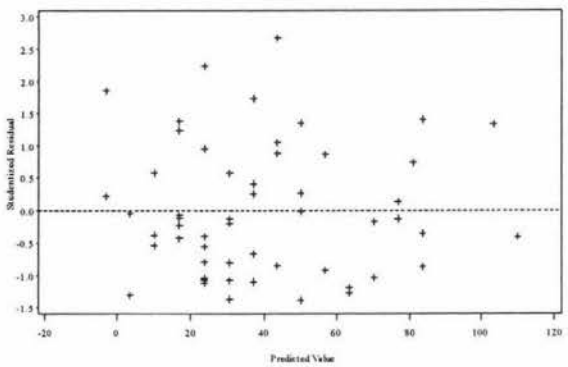
The size of the parameter estimate for slope did not change much for the model with outliers removed although the R squared improved slightly.

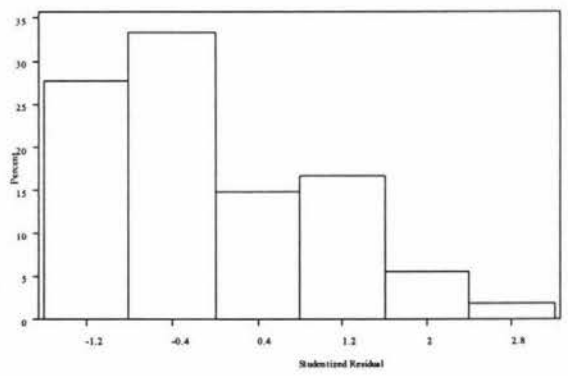
Residuals

For both models the distribution of the studentized residuals were fairly symmetrical, and there was no clear evidence of heteroscedascity on the residual versus predicted value plot.

Figure 18: WHOQoL versus HAQ







WHOQoL versus HAQ discussion

Estimated model

WHOQoL total change = 17 – 53.3(HAQ change)

Table 20: WHOQoL vs. HAQ

Slope parameter	Standard error of slope parameter	T score	P(slope not equal to zero)	R squared for regression
-53.3	10.04	-5.30	<0.0001	35%

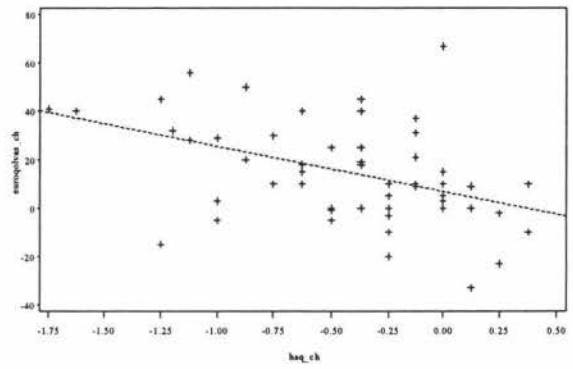
Outliers

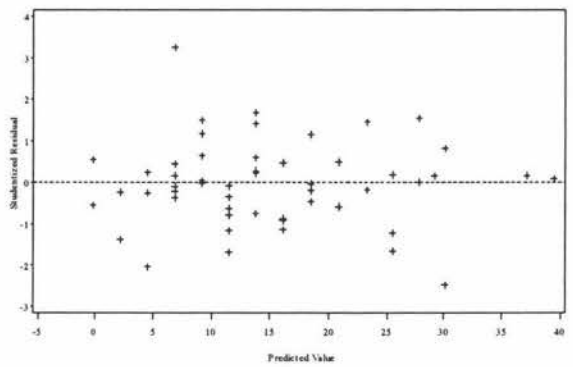
One subject, subject 59, had a residual with an absolute value greater than 2.5. This didn't seem particularly large and this subject was left in the analysis.

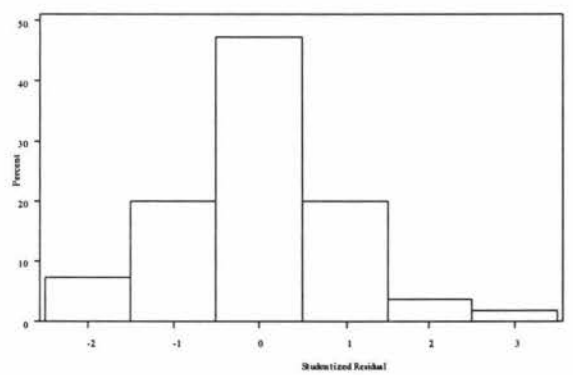
Residuals

Although there was little evidence of heteroscedascity of residuals on the residual versus predicted value plot the histogram of studentized residuals was not particularly symmetrical.

Figure 19: EuroQol visual analogue scale versus HAQ







EuroQol visual analogue scale (VAS) versus HAQ discussion

Estimated model 1

EuroQol visual analogue scale change = 6.86 – 18.6(HAQ change)

Table 21: EuroQol VAS vs. HAQ model 1

Slope parameter	Standard error of slope parameter	T score	P(slope not equal to zero)	R squared for regression
-18.6	5.1	-3.65	0.0006	20%

Outliers

One subject, subject 21, had a large studentized residual of 3.25. Analysis with this subject removed gave the following model.

Estimated model 2 (with outlier 21 removed)

EuroQol visual analogue scale change = 4.84 – 20.63(HAQ change)

Table 22: EuroQol VAS vs. HAQ model 2

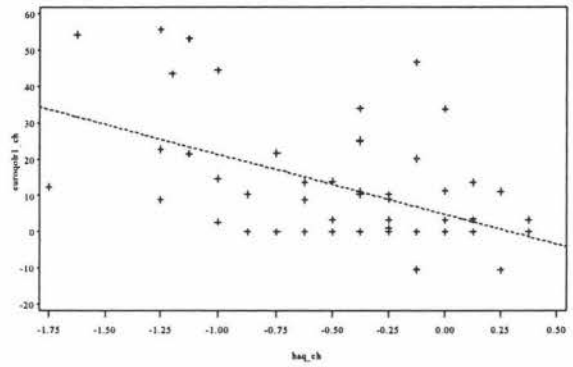
Slope parameter	Standard error of slope parameter	T score	P(slope not equal to zero)	R squared for regression
-20.63	4.64	-4.45	<0.0001	28%

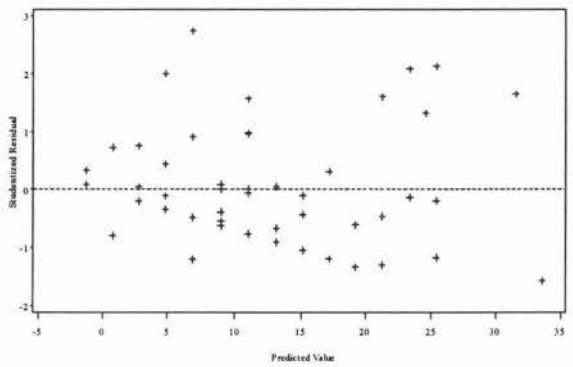
The size of the parameter estimate for slope changed by 10% and the model R squared improved by 8%.

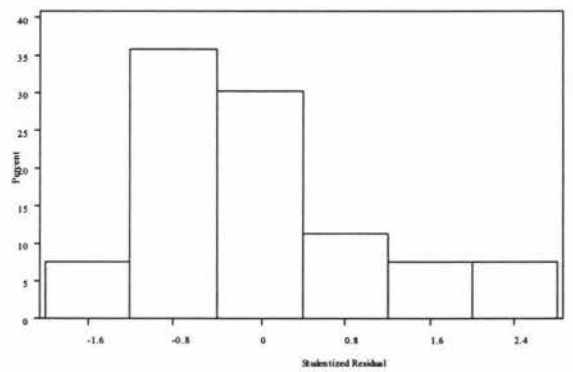
Residuals

For both models there was little evidence of heteroscedascity of residuals on the residual versus predicted value plot. The histogram of studentized residuals was symmetrical for both models.

Figure 20: EuroQol Rating Scale versus HAQ







EuroQol rating scale versus HAQ discussion

Estimated model

EuroQol rating scale change = 4.85 – 16.4(HAQ change)

Table 23: EuroQol rating vs. HAQ

Slope parameter	Standard error of slope parameter	T score	P(slope not equal to zero)	R squared for regression
-16.43	4.02	-4.08	0.0002	25%

Outliers

One subject, subject 15, had a studentized residual of greater than 2.5, this didn't seem particularly large and the subject was left in the analysis.

Residuals

There was little evidence of heteroscedascity of residuals on the residual versus predicted value plot. The histogram of studentized residuals showed some asymmetry.

Summary of regression statistics

Table 24: Summary of regression statistics

External criterion: HAQ

	Slope parameter	Standard error of slope parameter	T score	P(slope not equal to zero)	R squared for regression
QLP	-0.55	0.19	-2.85	0.0063	13.5%
WHOQoL sum of domains	-53.3	10.0	-5.30	<0.0001	35%
EuroQol VAS	-18.6	5.1	-3.65	0.0006	20%
EuroQol Rating	-16.4	4.0	-4.08	0.0002	25%

External criterion: Ritchie Articular Index

	Slope parameter	Standard error of slope parameter	T score	P(slope not equal to zero)	R squared for regression
QLP	-0.038	0.017	-2.27	0.0278	9.3%
WHOQoL sum of domains	-2.41	0.99	-2.42	0.0191	10.5%
EuroQol VAS	-1.45	0.45	-3.18	0.0025	17%
EuroQol Rating	-1.02	0.39	-2.64	0.0112	12%

External criterion: ESR

	Slope parameter	Standard error of slope parameter	T score	P(slope not equal to zero)	R squared for regression
QLP	-0.0077	0.0061	-1.26	0.22	5%
WHOQoL sum of domains	-0.38	0.38	-1.01	0.32	3%
EuroQol VAS	-0.14	0.19	-0.74	0.47	2%
EuroQol Rating	-0.28	0.14	-1.95	0.06	11%

Discussion of simple regression

The simple regression analyses show that there is a modest association, at best, between the QoL instruments and the external criteria. The relationship seems strongest where the HAQ is the external criterion however this should be considered with caution. Firstly the data is likely to give rise to very wide prediction intervals for individual subjects, as opposed to prediction intervals for the mean relationship. Secondly there is some evidence from the raw data plots that a small number of subjects, those who have both large changes in the QoL instrument scores and the external criterion located in the top left of the graphs, may be strongly influencing the slope parameter. This is particularly evident in the plots for the QLP (about 5 subjects) and EuroQol rating scale (about 5 subjects). This may well reflect a small sub-group of subjects who gained the most benefit from the intervention because they had the most potential to benefit. The poor relationship for the other subjects may reflect that these QoL instruments and the external criteria are relatively blunt instruments for detection of change, that is that they are poorly responsive. An additional concern may be that a boundary (ceiling or floor) effect may be present, that is some subjects were so mildly or severely affected that there is little potential for change and hence for change to be detected. For at least one of the instruments, the WHOQoL, the studentized residuals suggested that normality assumptions may have been violated however simple transformations, square root and logarithm, of the WHOQoL change score, while correcting the distribution of the residuals created heteroscedascity, apparent in the plots of residuals versus predicted values (analysis not shown).

Finally the relationship between the QoL instruments and the Ritchie Articular index and the ESR seems very poor. This may reflect a problem in the external criteria, that is that they are poorly responsive to change.

Mixed models: Maximum likelihood techniques

The theory and discussion of the use of mixed linear models for analysing responsiveness is covered in chapter 5 of this thesis.

Covariance pattern models with fixed effects

The first model fit using maximum likelihood techniques is:

$Y_{ij} = \mu + \alpha x_{ij} + \varepsilon_{ij}$

Where:

Y_{ij} is the measurement of the QoL instrument on the i^{th} subject on the j^{th} measurement occasion.

μ is the overall mean, or intercept, and is treated as fixed.

x_{ij} is the measurement of the external criterion on the i^{th} subject on the j^{th} measurement occasion and α is the fixed slope parameter describing the relationship between the QoL instrument and the external criterion.

ε_{ij} are the error terms and for each subject have an expected value of zero and a variance covariance matrix of the patterns described below. The covariance between different subjects is zero.

As a particular detailed example table 25 and table 26 shows a range of covariance pattern models for the relationship between the QLP and, as the external criterion for change, the HAQ, as well as the parameter estimates for the relationship, their standard error and other details of the model fitting.

Table 25: Examples of covariance patterns for QLP versus HAQ

Covariance pattern	**Number of covariance parameters**	**Estimated covariance parameters**
Unstructured	10	0.45
		0.45 0.53
		0.43 0.47 0.74
		0.39 0.37 0.59 0.68
Heterogeneous Toeplitz	7	0.43
		0.39 0.53
		0.41 0.54 0.80
		0.39 0.42 0.61 0.68
Heterogeneous compound symmetry	5	0.44
		0.37 0.55
		0.43 0.48 0.73
		0.43 0.48 0.56 0.73
Toeplitz	4	0.60
		0.48 0.60
		0.41 0.48 0.60
		0.44 0.41 0.48 0.60
Compound symmetry	2	0.61
		0.46 0.61
		0.46 0.46 0.61
		0.46 0.46 0.46 0.61

Table 26: Example of slope parameter estimates and model fitting for QLP versus HAQ

Covariance pattern	Parameter estimate	Standard error of parameter estimate	AIC	SBC	-2 log Likelihood
Unstructured	-0.529	0.0621	-180	-191	340
Heterogeneous Toeplitz	-0.535	0.0623	-182	-190	351
Heterogeneous compound symmetry	-0.575	0.0612	-195	-200	381
Toeplitz	-0.548	0.0634	-186	-191	365
Compound symmetry	-0.596	0.0621	-198	-200	393

Overall there seems to be little change in the size of the parameter estimates for any of the covariance patterns nominated, varying relatively only by 12% between the smallest and largest estimate. The estimate of the variance of the parameter estimates does not change much for the different covariance patterns.

The heterogeneous Toeplitz model had the largest value of SBC and the second largest value of AIC. It seems, based on the covariance pattern in the unstructured model , that it is important to incorporate into the covariance pattern the difference in covariance between the two sets of visits before and after the intervention. For the other QoL instruments and the HAQ the heterogeneous Toeplitz pattern also seemed to give one of the better models. For the EuroQol VAS, and the EuroQol rating scale the Toeplitz model may have been better than the heterogeneous Toeplitz with larger values of the AIC and SBC.

Similarly to the effects on parameter estimates illustrated for the QLP and the HAQ there is little change in the measures of model fit across the different covariance patterns.

Table 27 summarises the model fitting information for these models.

Table 27: Model fitting for QoL instruments versus HAQ

WHOQoL sum of domain scores

Covariance pattern	Slope parameter estimate	Standard error of parameter estimate	AIC	SBC	-2 log Likelihood
Unstructured	-38.1	4.43	-1228	-1239	2436
Heterogeneous Toeplitz	-38.5	4.39	-1227	-1235	2441
Heterogeneous compound symmetry	-46.8	4.11	-1234	-1239	2458
Toeplitz	-41.2	4.39	-1228	-1232	2448
Compound symmetry	-49	4.09	-1233	-1235	2462

EuroQol visual analogue scale

Covariance pattern	Slope parameter estimate	Standard error of parameter estimate	AIC	SBC	-2 log Likelihood
Unstructured	-15.6	1.63	-1026	-1037	2032
Heterogeneous Toeplitz	-16.3	1.66	-1027	-1035	2041
Heterogeneous compound symmetry	-18.7	1.61	-1035	-1040	2060
Toeplitz	-15.8	1.67	-1025	-1030	2043
Compound symmetry	-18.2	1.63	-1031	-1036	2063

EuroQol rating scale

Covariance pattern	Parameter estimate	Standard error of parameter estimate	AIC	SBC	-2 log Likelihood
Unstructured	-14.6	1.42	-990	-1002	1961
Heterogeneous Toeplitz	-15.8	4.45	-996	-1004	1979
Heterogeneous compound symmetry	-16.6	1.39	-998	-1004	1987
Toeplitz	-15.8	1.50	-996	-1001	1985
Compound symmetry	-16.8	1.43	-999	-1001	1994

Addition of a dummy variable for an intervention effect

As discussed earlier in this thesis a dummy variable, labelled 'Intervention', is now added to represent the effect of the intervention. A value of 1 for the dummy variable represents after the intervention and 0 before the intervention. The heterogeneous Toeplitz covariance pattern, or the equivalent unstructured pattern where the ESR is the external criterion for change, is used and for both the dummy variable and parameter estimating the relationship between the QoL instrument and the external criteria. Satterthwaites approximation is used for the degrees of freedom and P values are now are presented.

The model fitted is:

$$Y_{ij} = \mu + dI_k + \alpha x_{ij} + \varepsilon_{ij}$$

Where I_k is the dummy variable that takes a value of 1 when the measurement occurs after the intervention and 0 when it occurs before the intervention. The estimate of the value of d then captures the change in the QoL instrument that is not associated with the change in the external criterion. The other terms are as described in the last section.

Tables 28 to 30 shows the parameter estimates, standard errors, degrees of freedom and P values for the hypothesis that the parameter estimates are zero. No adjustments for multiple comparisons are made.

Table 28: QoL instrument versus HAQ and dummy variable: fixed effects model

Instrument	Parameter	Parameter estimate	Standard error	DF	T value	P value
QLP	Intervention	- 0.12	0.069	163	-1.81	0.073
	Slope	- 0.47	0.071	202	-6.6	<0.0001
WHOQoL sum of domains	Intervention	- 19.2	4.34	166	-4.43	<0.0001
	Slope	- 29.7	4.9	229	-6.06	<0.0001
EuroQol VAS	Intervention	- 11.4	1.79	232	-6.34	<0.0001
	Slope	- 12.6	1.74	150	-7.25	<0.0001
EuroQol Rating scale	Intervention	- 7.1	1.62	161	-4.37	<0.0001
	Slope	- 13.1	1.55	141	-8.42	<0.0001

Table 29: QoL instrument versus Ritchie with dummy variable: fixed effects model

Instrument	Parameter	Parameter estimate	Standard error	DF	T value	P value
QLP	Intervention	- 0.31	0.073	144	-4.26	<0.0001
	Slope	- 0.008	0.007	184	-1.12	0.27
WHOQol sum of domains	Intervention	- 28.59	4.41	158	-6.49	<0.0001
	Slope	- 0.62	0.45	189	-1.38	0.17
EuroQol VAS	Intervention	- 13.39	1.98	163	-6.74	<0.0001
	Slope	- 0.80	0.17	172	-4.59	<0.0001
EuroQol Rating scale	Intervention	- 9.94	1.82	155	-5.45	<0.0001
	Slope	- 0.76	0.17	162	-4.56	<0.0001

Table 30: QoL instrument versus ESR with dummy variable: fixed effects model

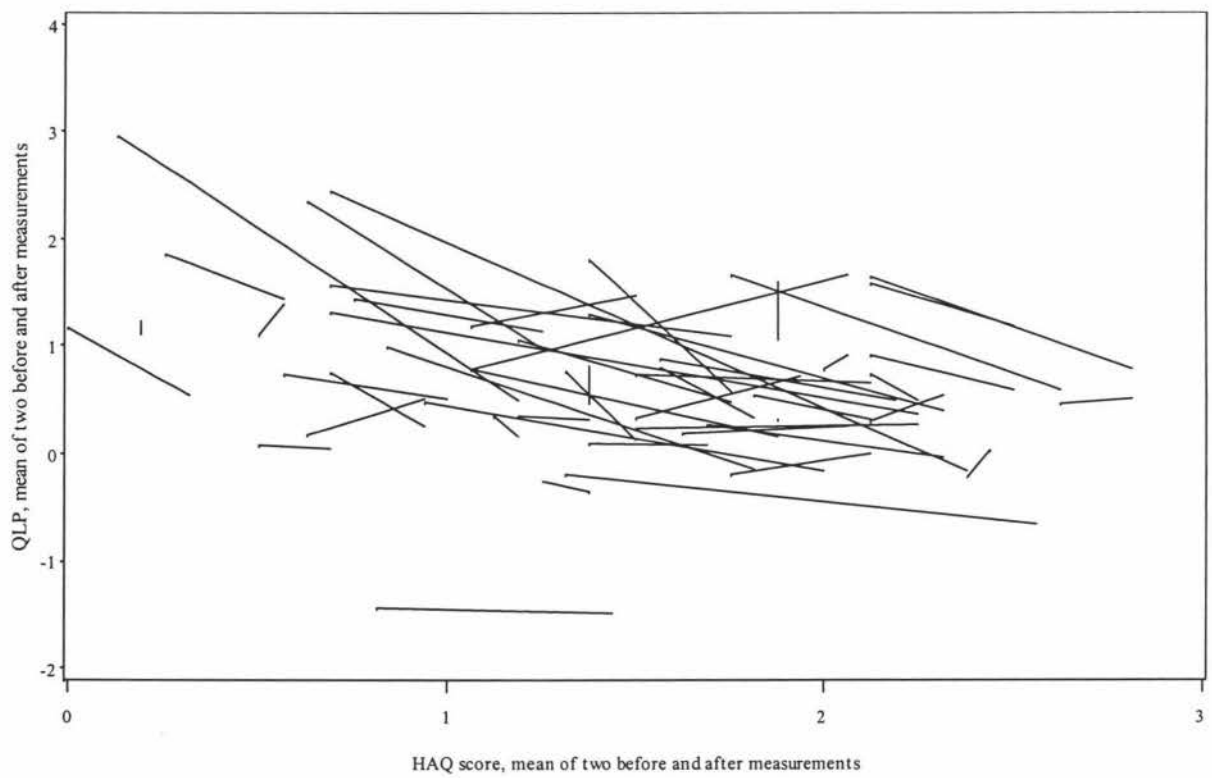
Instrument	Parameter	Parameter estimate	Standard error	DF	T value	P value
QLP	Intervention	-0.38	0.091	52.1	-4.17	<0.0001
	Slope	-0.0034	0.0025	79.5	-1.37	0.18
WHOQoL sum of domains	Intervention	-27.19	5.86	53.4	-4.64	<0.0001
	Slope	-0.33	0.167	94.5	-1.99	0.0495
EuroQol VAS	Intervention	-18.2	2.88	51.8	-6.32	<0.0001
	Slope	-0.046	0.061	76.4	-0.75	0.46
EuroQol Rating scale	Intervention	-11.3	2.87	52.4	-3.94	0.0002
	Slope	-0.1125	0.57	76.9	-1.97	0.053

Based on this analysis and with the HAQ as the external criterion the results suggest that apart from the QLP the other QoL instruments capture change in the subjects in addition to change in the external criterion. The QLP variation appears to be related more with the change in the HAQ. With the Ritchie Articular Index as the external criterion the analysis suggests that the QLP and WHOQoL are changing after the intervention but that this change is not in relation to the change in the Ritchie score. With the ESR as the external criterion there is little evidence that the change in the QoL instrument scores is related at all to the ESR change.

Random coefficients models

As an illustration as to why random coefficient models might be appropriate figure 21 shows a plot of the mean of the two measurements before the intervention and the mean of the two measurements after the intervention, for the QLP versus the HAQ.

Figure 21: QLP versus HAQ, mean of two measurements



The figure suggests that although the general trend is for QLP to decrease as the HAQ increases there is a considerable variation in the slopes and the intercepts for this relationship. This can be modelled by including random effects for the slope and intercept term for the relationship between the QoL instrument and the external criteria.

The model fit is:

$$Y_{ij} = \mu + s_i + \alpha x_{ij} + (s_i * \alpha)x_{ij} + \varepsilon_{ij}$$
$$= \mu_i + \alpha_i x_{ij} + \varepsilon_{ij}$$

Where:

Y_{ij} is the measurement of the QoL instrument on the i^{th} subject on the j^{th} measurement occasion.

μ is the overall mean, or intercept, and is treated as fixed.

s_i is a random subject effect and together with the interaction effect with the slope parameter has an expected value of zero and a variance covariance matrix that is multi-variate normal, of dimension two.

μ_i represents a random intercept term for the i^{th} subject

x_{ij} is the measurement of the external criterion on the i^{th} subject on the j^{th} measurement occasion and α is the fixed slope parameter describing the relationship between the QoL instrument and the external criterion.

α_i represents a random slope term for the i^{th} subject

ε_{ij} are the error terms and for each subject have an expected value of zero and a variance covariance matrix of the patterns shown previously. The covariance between different subjects is zero.

Tables 31 to 33 show model fitting information comparing random coefficient models with fixed effect models. When modelling the relationship between the QoL instruments and the HAQ as the external criterion the heterogeneous Toeplitz covariance pattern could only be fit, because of numerical problems, for the QLP. For

the WHOQoL a compound heterogeneous symmetry covariance pattern was used and for the two EuroQol instruments a compound symmetry pattern. When modelling the relationship between the QoL instruments and the Ritchie Articular Index the heterogeneous Toeplitz covariance pattern was fit for the QLP and the WHOQoL and a compound symmetry for the two EuroQol instruments. For the final external criterion, the ESR, as only two measurements of the ESR were made, a random coefficients model, without specifying any structure for the R matrix, was directly compared with unstructured covariance pattern model.

Table 31: Random coefficients versus fixed effects models: HAQ

QoL Instrument		**Number of Covariance parameters**	**AIC**	**SBC**	**-2 log Likelihood**
QLP	Fixed effects	7	-182	-190	351
	Random coefficient	10	-175	-185	332
WHOQoL[1]	Fixed effects	5	-1234	-1239	2458
	Random coefficient	8	-1230	-1239	2444
EuroQol VAS[2]	Fixed effects	2	-1031	-1036	2063
	Random coefficient	5	-1028	-1033	2046
EuroQol Rating scale[2]	Fixed effects	2	-999	-1001	1994
	Random coefficient	5	-995	-1001	1981

[1]Using heterogeneous compound symmetry covariance structure for R matrix

[2]Using compound symmetry covariance structure for R matrix

Note that the value of a Chi square statistic of 11.34 on 3 degrees of freedom has a probability of 0.01.

Table 32: Random coefficients versus fixed effects models: Ritchie Articular Index

QoL Instrument		Number of Covariance parameters	AIC	SBC	-2 log Likelihood
QLP	Fixed effects	7	-208	-215	400
	Random coefficient	10	-208	-216	403
WHOQoL[1]	Fixed effects	5	-1231	-1237	2452
	Random coefficient	8	-1216	-1225	2417
EuroQol VAS[2]	Fixed effects	2	-1032	-1034	2060
	Random coefficient	5	-1034	-1039	2057
EuroQol Rating scale[2]	Fixed effects	2	-998	-1001	1994
	Random coefficient	5	-995	-1001	1981

[1]Using heterogeneous compound symmetry covariance structure for R matrix

[2]Using compound symmetry covariance structure for R matrix

Table 33: Random coefficients versus fixed effects models: ESR

QoL Instrument		Number of Covariance parameters	AIC	SBC	-2 log Likelihood
QLP	Fixed effects	3	-105	-109	205
	Random coefficient	4	-104	-108	199
WHOQoL[1]	Fixed effects	3	-514	-517	1021
	Random coefficient	4	Unable to fit	Unable to fit	Unable to fit
EuroQol VAS	Fixed effects	3	-424	-427	842
	Random coefficient	4	-424	-428	839
EuroQol Rating scale[1]	Fixed effects	3	-417	-420	828
	Random coefficient	4	Unable to fit	Unable to fit	Unable to fit

[1]Unable to fit this model using random coefficients as algorithm would not converge

It seems likely that attempting to fit a random coefficients model where the ESR was the external criterion for change failed because of the poor relationship between the QoL instruments and this particular external criterion.

Where the HAQ was the external criterion for change the random coefficient model seemed to give a better fit than the fixed effect model. This was not so prominent for the Ritchie Articular Index as the external criterion. For example the fit for the QLP and EuroQol VAS was not much better. For the ESR as the external criterion the random coefficients model gave much the same fit as the covariance pattern model although there were convergence problems with the WHOQoL and the EuroQol rating scale. It seems likely that attempting to fit a random coefficients model where the ESR was the external criterion for change failed because of the poor relationship between the QoL instruments and this particular external criterion. The failure of improvement of

model fit where the Ritchie Articular Index was the external criterion for change may have been due to a similar problem.

The effects of using a random coefficients model on the parameter estimates are summarised in tables 34 to 36.

Table 34: Parameter estimates for fixed and random coefficients models for the HAQ

QoL Instrument	Model	Parameter estimate	Standard error	DF	T statistic	P value
QLP	Fixed	-0.54	0.062	205	-8.58	<0.0001
	Random	-0.53	0.08	51.4	-6.66	<0.0001
WHOQoL[1]	Fixed	-46.8	4.11	238	-11.37	<0.0001
	Random	-44.2	5.12	38.6	-8.63	<0.0001
EuroQol VAS[2]	Fixed	-18.2	1.64	199	-11.13	<0.0001
	Random	-18.6	1.99	45	-9.35	<0.0001
EuroQol Rating scale[2]	Fixed	-16.8	1.43	198	-11.78	<0.0001
	Random	-15.0	2.05	48.8	-7.34	<0.0001

[1]Using heterogeneous compound symmetry covariance structure for R matrix

[2]Using compound symmetry covariance structure for R matrix

Table 35: Parameter estimates for fixed and random coefficient models: Ritchie Articular Index

QoL Instrument	Model	Parameter estimate	Standard error	DF	T statistic	P value
QLP[1]	Fixed	-0.025	0.006	197	-3.83	0.0002
	Random	-0.011	0.006	146	-1.73	0.087
WHOQoL[2]	Fixed	-3.0	0.48	234	-6.26	<0.0001
	Random	-3.2	0.57	11.7	-5.71	0.0001
EuroQol VAS[2]	Fixed	-1.34	0.19	197	-7.20	<0.0001
	Random	-1.52	0.21	13.9	-7.20	<0.0001
EuroQol Rating scale[2]	Fixed	-1.24	0.17	195	-7.47	<0.0001
	Random	-1.19	.20	21.8	-5.94	<0.0001

[1]Using heterogeneous compound symmetry covariance structure for R matrix

[2]Using compound symmetry covariance structure for R matrix

Table 36: Parameter estimates for fixed and random coefficient models for the ESR

QoL Instrument	Model	Parameter estimate	Standard error	DF	T statistic	P value
QLP[1]	Fixed	-0.007	0.002	89.8	-2.94	0.004
	Random	-0.007	0.002	37.7	-2.74	0.009
EuroQol VAS[1]	Fixed	-0.11	0.065	73.8	-1.67	0.009
	Random	-0.13	0.075	6.14	-1.69	0.14

[1]Using compound symmetry covariance structure for R matrix

Despite a suggestion from the model fitting information, particularly where the HAQ is the external criterion for change, that a random coefficients model gave a better fit to the data, there is little change in the estimates of the slope parameters or their standard errors by using the more complex model. This is likely to represent the poor overall relationship between the QoL instruments and the external criteria for change. Modelling the variation in the data by transferring the magnitude of the error terms for subjects and residual error does not lead to a great change in the 't' statistics and their 'P' values for the slope parameter estimates.

Addition of a dummy variable for an intervention effect

Finally a dummy variable was fit to take into account effects of the intervention not captured by the relationship between the QoL instrument and the external criteria.

The model fit was:

$$Y_{ij} = \mu + s_i + dI_k + \alpha x_{ij} + (s_i * \alpha)x_{ij} + \varepsilon_{ij}$$
$$= \mu_i + dI_k + \alpha_i x_{ij} + \varepsilon_{ij}$$

Where I_k is the dummy variable that takes a value of 1 when the measurement occurs after the intervention and 0 when it occurs before the intervention. The estimate of the value of d then captures the change in the QoL instrument that is not associated with the change in the external criterion. The other terms are as described in the last section.

Tables 37 to 39 show the effects on the slope parameter estimates and the log likelihoods of adding a dummy variable for the intervention for the relationship between the QoL instrument and the external criteria.

Table 37: Random coefficients models with a dummy variable for the intervention: HAQ

QoL Instrument[1]		Slope parameter (se)	T statistic (DF)	P	-2 log Likelihood
QLP	No Dummy	-0.54 (0.08)	-6.59 (56.4)	<0.0001	347.6
	Dummy	-0.46 (0.09)	-4.9 (81.6)	<0.0001	348
WHOQoL[2]	No Dummy	-44.2 (5.12)	-8.63(38.6)	<0.0001	2443.2
	Dummy	-31.2 (5.76)	-5.41 (57.6)	<0.0001	2412.7
EuroQol VAS[2]	No Dummy	-18.8 (2.0)	-9.35 (45)	<0.0001	2046
	Dummy	-13.3 (2.1)	-6.38 (52.3)	<0.0001	2005
EuroQol Rating scale[2]	No Dummy	-15.03 (2.1)	-7.34 (48.8)	<0.0001	1976.7
	Dummy	-11.7 (2.1)	-5.13 (54.7)	<0.0001	1949.9

[1]Using heterogeneous compound symmetry covariance structure for R matrix

[2]Parameter for dummy variable statistically significant although not shown

Table 38: Random coefficients models with a dummy variable for the intervention: Ritchie Articular Index

QoL Instrument[1]		Slope parameter (se)	T statistic (DF)	P	-2 log Likelihood
QLP[2]	No Dummy	-0.11 (0.007)	-1.73 (46.1)	0.087	414.9
	Dummy	0.003 (0.007)	0.45 (41.5)	0.65	395.2
WHOQoL[2]	No Dummy	-2.12 (.49)	-4.3 (30.6)	0.0002	2417
	Dummy	-1.19 (0.564)	-2.11 (25.3)	0.045	2375.5
EuroQol VAS[2]	No Dummy	-1.52 (0.21)	-7.15 (32.3)	<0.0001	2057.3
	Dummy	-0.89 (0.21)	-4.15 (37.3)	0.0002	1996.5

[1]Using compound symmetry covariance structure for R matrix

[2]Parameter for dummy variable statistically significant although not shown

There were convergence problems for the EuroQol rating scale.

Table 39: Random coefficients models with a dummy variable for the intervention: ESR

QoL Instrument[1]		Slope parameter (se)	T statistic (DF)	P	-2 log Likelihood
QLP[2]	No Dummy	-0.007 (0.002)	-2.74 (37.7)	0.009	199.1
	Dummy	-0.003 (0.002)	-1.18 (47.5)	0.24	182.3
EuroQol VAS[2]	No Dummy	-0.13 (0.073)	-1.69 (36.1)	0.10	838.9
	Dummy	-0.07 (0.69)	-0.98 (35.9)	0.33	803.4

[1]Using compound symmetry covariance structure for R matrix

[2]Parameter for dummy variable statistically significant although not shown

There were convergence problems for the WHOQoL and the EuroQol rating scale.

The convergence problems are again consistent with the poor overall relationship between QoL instruments and the external criteria. Of the QoL instruments the QLP does not seem as responsive to the HAQ as the other instruments. There is still, however, only a modest difference in the slope parameter estimates, and their standard error estimates, with and without the dummy variable, consistent with the weak to moderate relationship between all the QoL instruments and the external criteria for change.

Analysis of residuals

This is discussed for the final model incorporating an intervention effect as well as random coefficients for intercept and slope parameters.

HAQ

Residuals showed no subjects that consistently gave large residual values. For the QLP and WHOQoL the residuals were distributed symmetrically although for the EuroQol VAS and EuroQol rating scale there was a skewed distribution. There were no problems suggested by the residual versus predicted plots.

Ritchie Articular Index

Residuals showed no subjects that consistently gave large residuals. For all the QoL instruments the residuals were distributed symmetrically and there were no problems suggested by the residual versus predicted plots.

ESR

Again residuals showed no subjects that consistently gave large residuals and the residuals were symmetrically distributed. For the QLP there was evidence of heteroscedascity on the residual plot with the residuals declining with increasing predicted values.

Analysis of random effects coefficients

Plots of the estimates of the random effect estimates for each subject for the intercept versus slope did not show strong evidence of outlying values. The bivariate normal assumption may not be met particularly well, based on the plot of random coefficients. This may be consistent with a ceiling or floor effect for the instruments, that people at the top or bottom of the range may not have a capacity to change very much.

As particular examples figures 22 to 24 show examples of residual plots for the QLP versus the HAQ.

Figure 22: Residuals versus predicted values for final model of QLP versus HAQ

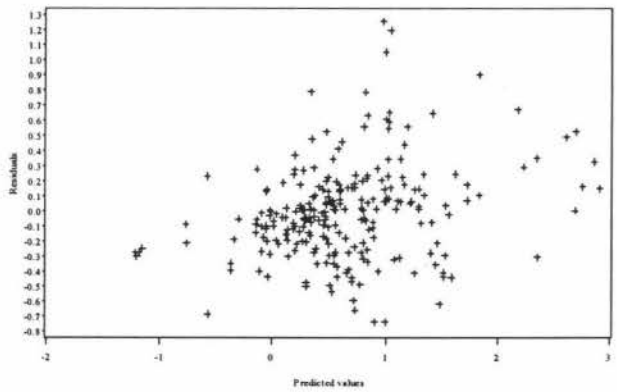


Figure 23: Histogram of residuals for final model of QLP versus HAQ

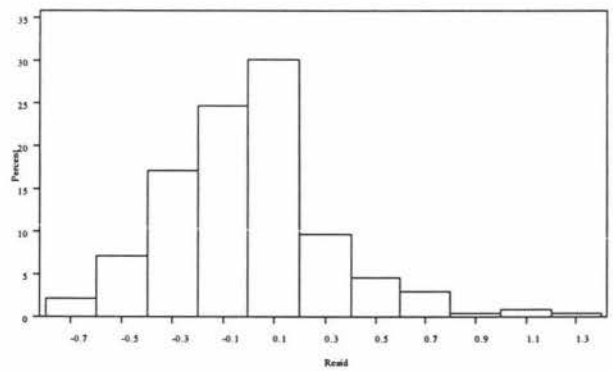
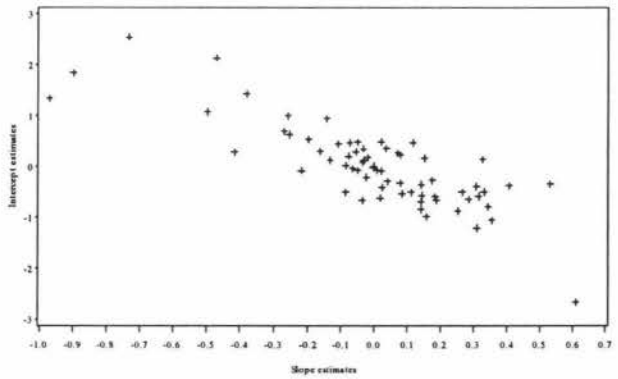


Figure 24: Random coefficient estimates for final model of QLP versus HAQ



Mixed models: Bayesian techniques

A mixed linear model was fit using the program 'WinBUGS' adapting an example from Volume 2 of the examples '7 Jaw: repeated measures analysis of variance' (113).

The measurements of the subjects on the 4 occasions were each treated as independently distributed with a constant variance but the slopes and intercepts, for the relationship between the QoL instruments and an external criterion for change, were treated as random effects. Only the HAQ was used as the external criterion for change because of the poor relationship between the QoL instruments and the other external criteria. For each of the models the slope estimates together with their standard errors were based on 2000 iterations after a 1000 iteration burn in. Where the response variable was the WHOQoL and the EuroQol rating scale the 'over relaxation' option in the sampling scheme for 'WinBUGS' was chosen to give better convergence to a solution. For completeness the equivalent random effects parameter estimate are shown from maximum likelihood estimation. For the QLP and WHOQoL a heterogeneous Toeplitz variance covariance matrix was specified for the R matrix and for the two EuroQol instruments a compound symmetry variance covariance matrix for the R matrix was specified.

The results are presented in table 40.

Table 40: Bayesian estimates of slope parameters

QoL Instrument	**Parameter estimate (se) Bayesian**	**Parameter estimate (se) Maximum likelihood**	**MC error**
QLP	-0.604 (0.105)	-0.53 (0.08)	0.0035
WHOQoL	-55.2 (3.02)	-56.8 (11.83)	0.41
EuroQol VAS	-19.2 (2.18)	-18.6 (1.99)	0.18
EuroQol Rating scale	-17.8 (1.7)	-15.0 (2.05)	0.12

There seems to be little substantive change in either the parameter estimates for the two estimation methods. For the WHOQoL instrument the standard error of the slope estimate based on the Bayesian analysis was substantially less than for the maximum likelihood estimation. The Bayesian analysis suggests that WHOQoL instrument was more responsive than the maximum likelihood analysis.

Chapter 8: Discussion

Quality of Life

Quality of life (QoL) is a concept that has great intuitive appeal in the assessment of people with chronic health conditions. It captures the distinction between merely staying alive and having a life that is in some sense worth living. The assessment of interventions that attempt to improve chronic health conditions should identify changes that are 'worthwhile'. The use of QoL to measure the change induced by interventions attempts to balance the severity of a condition and the extent to which it affects the life a person leads, against the positive and negative effects of interventions designed to influence the chronic health condition.

Movement from an intuition of QoL to a more detailed construction of the concept is difficult. Some of these difficulties are fundamentally linked to the aspects of measurement and analysis. For example is QoL fundamentally a unitary concept or is it intrinsically a multi-dimensional concept? If it is a unitary concept then in what sense can the concept be captured in a single number and how can this single number be generated? If it is an intrinsically multi-dimensional concept then how can the dimensions be conceptualised and measured in a way that reduces the correlation between numerical measurements of the different dimensions?

There is a strong subjective element to QoL. Individual variation in the way people view themselves and society, as well as the different ways individuals respond to chronic illness, and the heterogeneity in the manifestations of any one health condition and the differences between different health conditions, compounds the potential difficulties in conceptualising and operationalising QoL measurement.

Comparison of different instruments which are conceptualised and operationalised in different ways offers additional challenges. Repeated measures designs are often used to track changes in individual responses to health care interventions and the necessity to

account for the correlation of measurements made on individual subjects also increases the complexity of using QoL instruments in clinical trials.

Responsiveness is one aspect of the development of an instrument that attempts to measure QoL. Responsiveness is evaluated by comparing the change in an instrument score with a measure of background variation. A more responsive instrument changes more, in relation to background variation, after an intervention. Two broad approaches to evaluating responsiveness have been discussed in this thesis, internal and external responsiveness.

Internal responsiveness

Internal responsiveness considers an instrument, and the change in score of the instrument, in relation only to whether a measurement was made before or after an intervention that 'should have' changed QoL. The suggested measurements of internal responsiveness, outlined in Chapter 2 of this thesis, are all variations on a theme of standardised mean response. 'Standardised' is used in the sense of using a measure of background variation to compare a mean response of a group of subjects who experience an intervention.

The concept of internal responsiveness can be extended by the suggestion that if a clinically meaningful change in QoL can be specified the relationship between this and background variation can be used to rank instruments. An appropriate instrument can then be used that minimises the number of subjects recruited into a clinical trial in order to detect a difference in QoL.

Internal responsiveness measurements all suffer from difficulties in choosing their numerator and denominator, i.e. which mean change should be standardised and which measure of background variation should be used. Inappropriate choices, particularly in the setting of repeated measures designs, can lead to different conclusions. When internal responsiveness is measured by the techniques suggested in the literature simple ranking of the internal responsiveness measurements is unlikely to be helpful.

The measurements of internal responsiveness are based on samples, but are derived from complex combinations of means and standard deviations, so that the actual statistical properties of internal responsiveness measurements are unclear. Any form of ranking that does not take into account the sampling process and then looks simply at the point estimates of differences in measurements may lead to incorrect conclusions.

The term 'sample' used in the preceding discussion is used advisedly. Groups of subjects used for the measurement of new QoL instruments are not usually random samples of some larger population. They are usually highly selected convenience samples that are likely to be subject to all sorts of selection bias. When measurements of mean change and its variation are dependent on a biased sample, generalising results of internal responsiveness may be difficult.

In the particular case of the data set used for this thesis, although the QLP was rated by all the measurements as the worst performing instrument, there was no consistent rank order for the other instruments. There was no way of comparing the different internal responsiveness measurements to determine if the differences were in some sense 'significant'.

Receiver operating characteristic curves

This method of evaluating responsiveness is the first of the methods of assessing external responsiveness. Internal responsiveness uses only the fact that an intervention has been applied to subjects that should have affected QoL, to measure change against background variability. External responsiveness attempts to relate the change in QoL to a criterion for change that is different from the QoL measurement. For receiver operating characteristic curves this change is a categorical statement that change in a subject has or has not occurred. Rather than assessing uncertainty by some measure of variation such as a standard deviation, receiver operating characteristic curves associate the probability of correctly allocating subjects based on the change in a QoL score to whether change has occurred based on the external criterion. Serious underlying conceptual problems with this method of assessment are, firstly, it explicitly treats QoL as a unitary concept. Subjects are allocated explicitly into a changed and unchanged

group. Secondly the allocation of subjects by the external criteria is often based on methods that have not been as rigorously developed as the instruments the external criteria are to test. For example a complex instrument with evidence of validity and reliability may be tested by a simple question as to whether QoL has improved or not. Finally reducing QoL to a dichotomous variable, for the purposes of classification, loses the potential richness in QoL as a spectrum of experience.

Trying to compare conclusions reached about responsiveness based on the same instruments but different external criteria for change, in different studies, by a summary statistic such as the area under the curve of the receiver operating characteristic curve (AUC ROC), is also fraught with difficulties.

The same difficulty in generalising from a particular convenience sample of subjects to a wider or different group of subjects discussed for the internal responsiveness statistics applies to receiver operating characteristic curves.

This thesis has illustrated that it is important to not only consider the point estimates of a summary statistic but also the proper construction of analysis of the differences in these summary statistics when measured on the same subjects. For this thesis despite a range of point estimates of the AUC ROC of 0.65 to 0.74, for the different instruments, none could be distinguished from each other on statistical testing. The instruments had moderate external responsiveness based on AUC ROC.

Correlation coefficients

QoL may form a spectrum of experience. Analytical methods based on continuous variables, such as correlation and regression approaches, might offer a useful way of evaluating whether new instruments 'capture' the same change as external criteria.

A simple way of doing this is by the use of correlation coefficients. However use of correlation coefficients implicitly assumes a particular model, in particular a simple linear regression. Use of correlation coefficients without examining the underlying, implicit, basis for their use can lead to incorrect conclusions about responsiveness.

Correlation coefficients presented with confidence intervals and an omnibus test for differences between correlated correlation coefficients highlight that differences in point estimates based on relatively small samples can lead to misleading conclusions regarding differences in responsiveness. This is shown in the data set for this thesis where the magnitudes for the correlation coefficients for the QoL instruments versus the HAQ as the external criterion ranged from –0.37 for the QLP to –0.59 for the WHOQoL. This 'obvious' difference which might suggest that the WHOQoL was far more responsive than the QLP, if the simple rank order of the point estimates was used to make this judgement, is put in its correct context when the confidence intervals for each point estimate are taken into account and in addition when the correlation between the estimates for the correlation coefficients, which were in fact measured on the same subjects, is also taken into account.

Simple linear regression

The poor performance and dubious utility of the correlation coefficient to assess responsiveness is emphasised when attempts are made to fit a linear model to the relationship between the QoL instruments and the external criteria. The simple linear regression plots of the data points and simple model fitting information show that there is considerable variation in the data not explained by the relationship between the QoL instruments and the external criteria.

The simple linear regression of the differences in the QoL instruments versus the differences in the external criteria suggest that this is in part related to the unreliable measurement of both these sets of variables and then using differences in these unreliable variables as the basis for inference. In addition it seems likely, particularly for the Ritchie Articular Index and the ESR, that these are poorly performing external criteria for change in this setting, further complicating efforts to relate these external criteria to a change in QoL.

An advantage of the use of simple linear regression models is, providing the same external criteria for change are used as the predictor variables, it may be easier to compare different studies assessing responsiveness of QoL instruments. This is by the

compare different studies assessing responsiveness of QoL instruments. This is by the ratio of the parameter estimating the slope relationship to its standard error, and adjusting for sample size through the medium of the statistical significance of this slope relationship.

Mixed linear models

If the information gathered from studies of change in QoL instrument scores compared to external criteria is more complex, for example more than two measurements are made on individual subjects, then more detailed modelling of the covariance structure of the relationship between the QoL instruments and the external criteria for change could more accurately estimate the standard error of slope parameter estimates. Individual variation related to the subjective nature of QoL, and the heterogeneity in disease and response to disease and treatment, could also be explicitly modelled and the ratio of a parameter estimate to its standard error can be better evaluated.

In practice the actual covariance pattern used when maximum likelihood techniques were applied in the data set for this thesis made little difference to the probability statements about the relationship between the QoL instruments and the external criteria. The relationships between the estimates of slopes and slope standard errors, across all the QoL instruments tested, were much the same, based on the 'P' value assessing whether the slope value was different from zero. This was the case even if account was taken of correlation between measurements made on the same individuals, the repeated measures design. It seems likely that this may in part be due to the poor overall relationship between the QoL instruments and the external criteria. Complex co-variance structures could simply not be fit to some of the relationships between QoL instruments and the external criteria. This was because the algorithms for fitting the maximum likelihood models would not converge, reflecting that there was in fact very little that could be meaningfully extracted from the data set. This was particularly the case where the Ritchie Articular Index and the ESR were the external criteria. For the HAQ as the external criterion all the QoL instruments had evidence of responsiveness but all seemed similarly responsive.

The random intercepts and slopes models seemed to fit the data better than fixed effect models. They did not provide greater insight into the responsiveness of the instruments.

Use of Markov Chain Monte Carlo methodology to fit models in a Bayesian framework was relatively straightforward but again offered little additional insight into the relationships modelled.

Conclusions

All the QoL instruments in this study had some evidence for responsiveness based on the measures of internal responsiveness and all of the measures of external responsiveness. The ROC analysis suggests that responsiveness of the instruments was at best modest and this is consistent with the correlation analysis. The QoL instruments had a disappointingly poor relationship with the external criteria when analysed by fitting linear models of varying degrees of sophistication. While in part this may well reflect that no adequate 'gold standard' for change was used in this study, it may also reflect that even apparently well developed QoL instruments are blunt instruments when used on individual subjects. QoL has to change a large amount, when measured with one of these instruments, to be detected above the background noise of variation in how people view themselves and their lives.

If responsiveness is used as a criterion to decide which QoL instrument might be used in a clinical trial then it is difficult to recommend one of these instruments above another based on the numerical and statistical analysis. Other ways of deciding which instrument to use may be more useful, such as ease of use, both by researchers and subjects, ability to reduce non-response, and how easily the results might be transmitted in analysis and publication of the results of randomised trials. In this setting the EuroQol instrument, both the visual analogue scale and the rating scale, have the virtue of considerable simplicity of administration. The results are also relatively easy to explain.

Appendix 1

The EuroQol

The WHOQoL-Bref

The Quality of Life Profile

The Ritchie Articular Index

The Health Assessment Questionnaire

The EuroQol

By placing a tick in one box in each group below, please indicate which statements best describe your own health state today.

Mobility

I have no problems walking about ☐

I have some problems walking about ☐

I am confined to bed ☐

Self care

I have no problems with self care ☐

I have some problems with self care ☐

I am unable to wash and dress myself ☐

Usual activities

I have no problems performing my usual activities (e.g. work, study, housework, family or leisure activities) ☐

I have some problems with performing my usual activities ☐

I am unable to perform my usual activities ☐

Pain/discomfort

I have no pain or discomfort ☐

I have moderate pain or discomfort ☐

I have extreme pain or discomfort ☐

Anxiety/depression

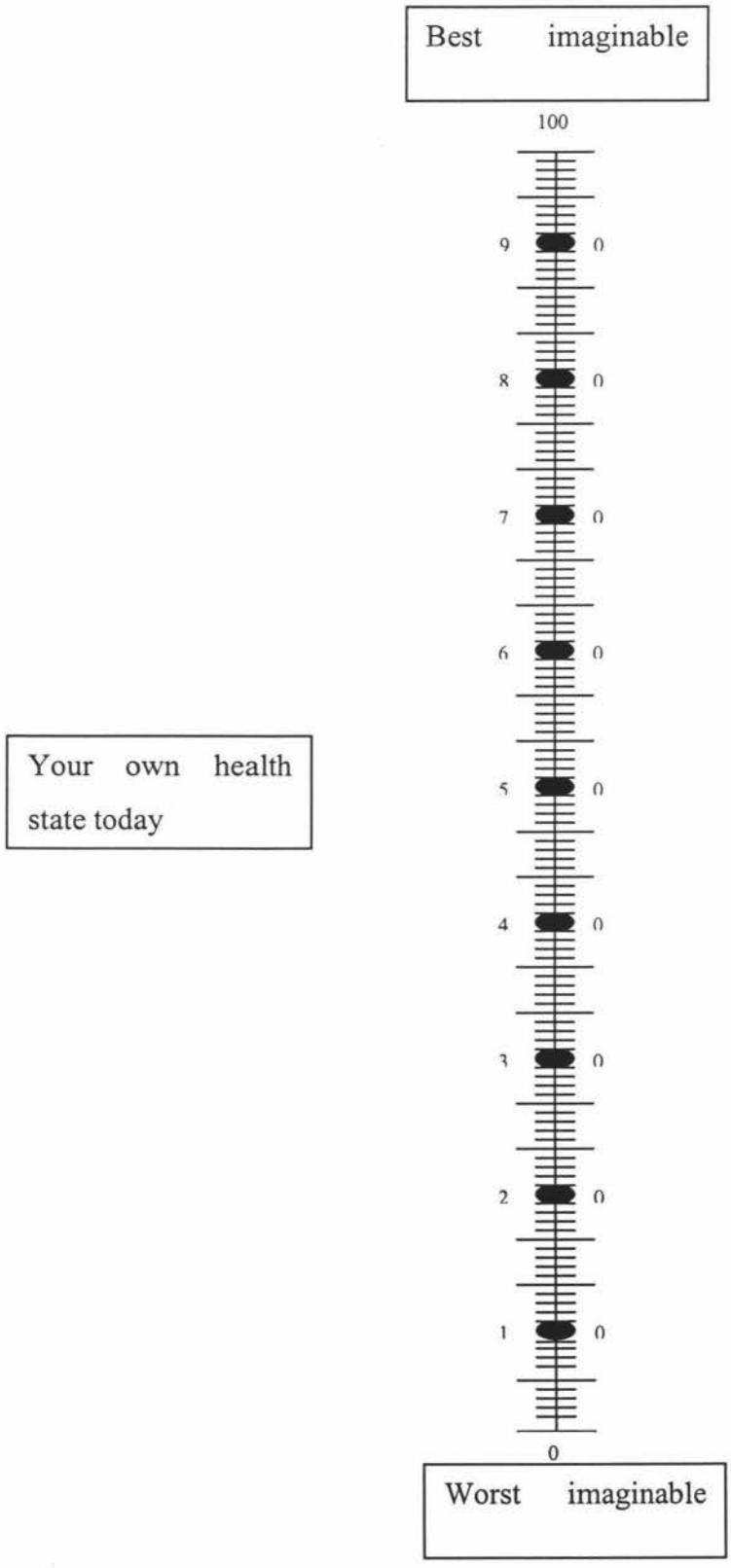
I am not anxious or depressed ☐

I am moderately anxious or depressed ☐

I am extremely anxious or depressed ☐

To help people say how good or bad a health state is, we have drawn a scale (rather like a Thermometer) on which the best state you can imagine is marked by 100 and the worst state you can imagine is marked by 0.

We would like you to indicate on this scale how good or bad is your health today, in your opinion. Please do this by drawing a line from the box below to whichever point on the scale indicates how good or bad your current health state is.



The WHOQoL-Bref

Instructions

This assessment asks how you feel about your quality of life, health, or other areas of your life. **Please answer all the questions.** If you are unsure which response to give to a question, **please choose the one** that appears most appropriate. This can often be your first response.

Please keep in mind your standards, hopes, pleasures and concerns. We ask that you think about your life **in the last two weeks.** For example, thinking about the last two weeks, a question might ask:

	Not at all	Not Much	Moderately	A great deal	Completely
Do you get the kind of support from others that you need?	1	2	3	4	5

You should circle the number that best fits how much support you get from others over the last two weeks. So you would circle the number 4 if you got a great deal of support from others as follows.

	Not at all	Not Much	Moderately	A great deal	Completely
Do you get the kind of support from others that you need:	1	2	3	(4)	5

You would circle number 1 if you did not get any of the support that you needed from others in the last two weeks.

Section 1: Please read each question, assess your feelings, and circle the number on the scale for each question that gives the best answer for you

	Very poor	Poor	Neither poor nor good	Good	Very good
1. How would you rate your quality of life?	1	2	3	4	5

	Very dissatisfied	Dissatisfied	Neither satisfied nor dissatisfied	Satisfied	Very satisfied
2. How satisfied are you with your health?	1	2	3	4	5

Section 2: The following questions ask *how much* you have experienced certain things in the last two weeks.

	Not at all	A little	A moderate amount	Very much	An extreme amount
3. To what extent do you feel that physical pain prevents you from doing what you need to do?	1	2	3	4	5
4. How much do you need any medical treatment to function in your daily life?	1	2	3	4	5
5. How much do you enjoy life?	1	2	3	4	5
6. To what extent do you feel your life to be meaningful	1	2	3	4	5

	Not at all	A little	A moderate amount	Very much	An extreme amount
7. How well are you able to concentrate?	1	2	3	4	5
8. How safe do you feel in your daily life?	1	2	3	4	5
9. How healthy is your physical environment?	1	2	3	4	5

Section 3: The following questions ask *how completely* you experience or were able to do certain things in the last two weeks

	Not at all	A little	Moderately	Mostly	Completely
10. Do you have enough energy for everyday life?	1	2	3	4	5
11. Are you able to accept your bodily appearance?	1	2	3	4	5
12. Have you enough money to meet your needs?	1	2	3	4	5

	Not at all	A little	Moderately	Mostly	Completely
13. How available to you is the information that you need in your day-to-day life?	1	2	3	4	5
14. To what extent do you have the opportunity for leisure activities	1	2	3	4	5

	Very poor	Poor	Neither poor nor good	Good	Very good
15. How well are you able to get around?	1	2	3	4	5

Section 4: The following questions ask you to say how *good or satisfied* you have felt about various aspects of your life over the last two weeks

	Very dissatisfied	Dissatisfied	Neither satisfied nor dissatisfied	Satisfied	Very satisfied
16. How satisfied are you with your sleep?	1	2	3	4	5
17. How satisfied are you with your ability to perform your daily living activities?	1	2	3	4	5
18. How satisfied are you with your capacity to work?	1	2	3	4	5
19. How satisfied are you with yourself?	1	2	3	4	5
20. How satisfied are you with your personal relationships?	1	2	3	4	5
21. How satisfied are you with your sex life?	1	2	3	4	5
22. How satisfied are you with the support you get from your friends?	1	2	3	4	5

	Very dissatisfied	Dissatisfied	Neither satisfied nor dissatisfied	Satisfied	Very satisfied
23. How satisfied are you with the conditions of your living place?	1	2	3	4	5
24. How satisfied are your with your access to health services?	1	2	3	4	5
25. How satisfied are your with your transport?	1	2	3	4	5

Section 5: The following question refers to *how often* you have felt or experienced certain things in the last two weeks

	Never	Seldom	Quite often	Very often	Always
26. How often do you have negative feelings such as blue mood, despair, anxiety, and depression?	1	2	3	4	5

Did some one help you fill out this form? _____

How long did it take you to fill this form out? _____

Do you have any comments about the assessment?

_____

_____

_____

_____

_____

Thank you for your help

Quality of life profile: Physical and sensory disabilities version

What is Quality of Life?

Quality of Life, in simple terms, means:

"How good is your life for you?"

The answer to this question is a measure of a person's **Quality of Life.**

To answer the question "How good is your life for you?' you are asked to focus on yourself and to rate some aspects of your life. These are all rated on a simple scale of 1-5. The aspects of your life are divided evenly into 9 areas—areas we think are part of the lives of all people.

The nine areas that are part of the lives of all people are:

1. My body and my health
2. My thoughts and feelings
3. My beliefs and values

4. Where I live and spend my time
5. The people around me
6. My access to things in my community

7. The practical things I do
8. The things I do for fun and enjoyment
9. The things I do to cope and change

First, you will rate the aspects two times, using two different questions: How important to me is...? And How satisfied am I with...? Then, you will indicate how much control and possibility for improvement exist in the 9 areas of your life. This sounds like a lot, but you will find that you can rate them rather quickly.

Please give the ratings that best match your views.

I. Importance

Instructions:

The first question to ask yourself is:

How *important* is this to me in my life?

Another way to think about this question is:

How much do I care about this?

2. Rate each of the items from 1 to 5, using the rating scale on each of the following pages.

Rate an item 5 if it is extremely important to the way you lead your life, and very strongly directs your thinking or your activities; this item is a dominant and driving force in your life.

Rate an item 4 if it is very important to you, and it is very relevant to the way you lead your life.

Rate an item 3 if it is important to you, but you do not think about it especially.

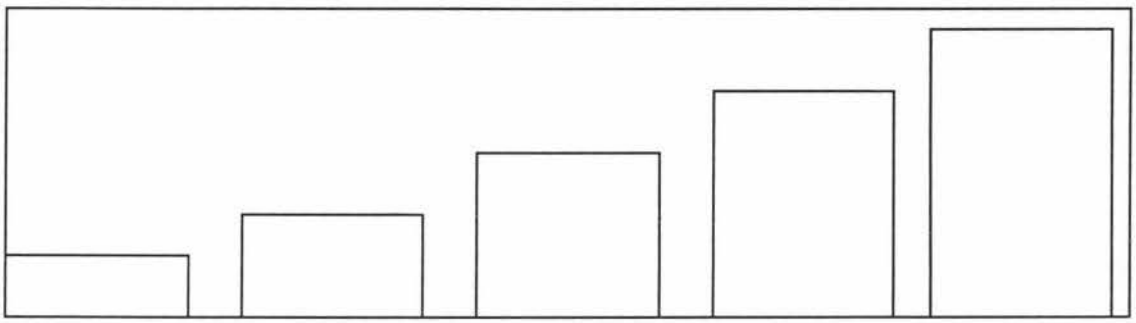
Rate an item 2 if it is slightly important or irrelevant to the way you lead your life.

Rate an item 1 if it is unimportant or irrelevant to the way you lead your life.

Note: Answer each question in terms of your life as it is right now.

You may score **N/A** for Not Applicable, or **DK** for Don't Know.

Rating scale: IMPORTANCE



NOT AT ALL IMPORTANT	*NOT VERY IMPORTANT*	*IMPORTANT*	*VERY IMPORTANT*	*EXTREMELY IMPORTANT*
1	**2**	**3**	**4**	**5**

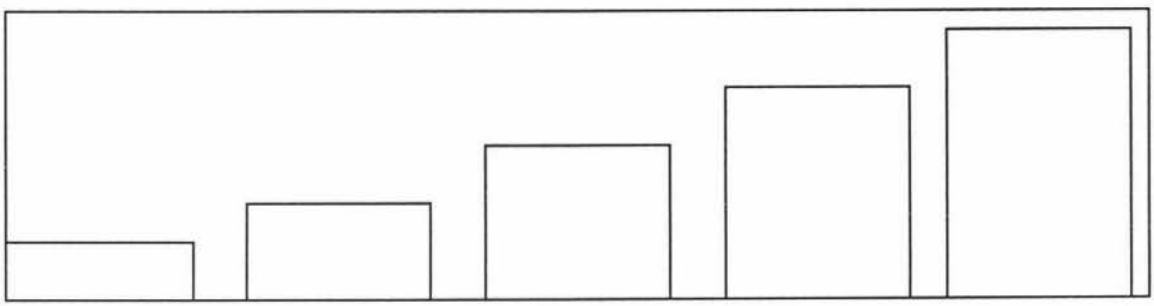
DON'T KNOW: **DK** NOT APPLICABLE: **NA**

My body and my health

How important to me is -- ?

1. Being physically able to get around my home _____
2. Being physically able to get around my neighbourhood _____
3. Being physically able to use public transportation _____
4. Being physically active and keeping fit _____
5. Getting enough sleep and rest _____
6. Good nutrition and eating the right foods _____
7. Having enough energy to do the things I want to _____
8. Maintaining my personal hygiene and caring for myself, by MYSELF _____
9. Maintaining my personal hygiene and caring for myself, WITH THE ASSISTANCE OF OTHERS _____
10. My personal appearance _____
11. How I am able to manage the pain that I have _____
12. My overall physical health _____

Rating scale: IMPORTANCE



NOT AT ALL IMPORTANT	*NOT VERY IMPORTANT*	*IMPORTANT*	*VERY IMPORTANT*	*EXTREMELY IMPORTANT*
1	**2**	**3**	**4**	**5**

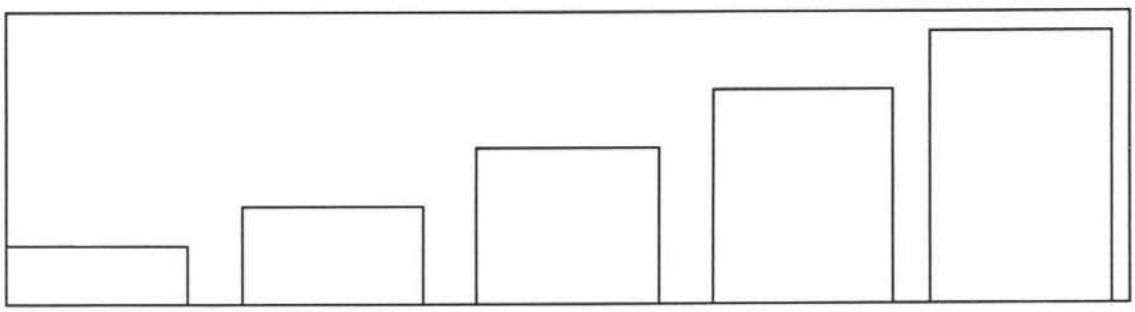
DON'T KNOW: **DK** NOT APPLICABLE: **NA**

My thoughts and feelings

How important to me is -- ?

1. Accepting the way I am _____
2. Making my own decisions _____
3. Acting independently, on my own _____
4. Being able to remember things _____
5. My level of stress _____
6. The mood I am usually in _____
7. Coping with what life brings _____
8. Feeling good about myself _____
9. Having a positive attitude towards life _____
10. Having a sense of humour _____
11. My mental health _____

Rating scale: IMPORTANCE



NOT AT ALL IMPORTANT	*NOT VERY IMPORTANT*	*IMPORTANT*	*VERY IMPORTANT*	*EXTREMELY IMPORTANT*
1	**2**	**3**	**4**	**5**

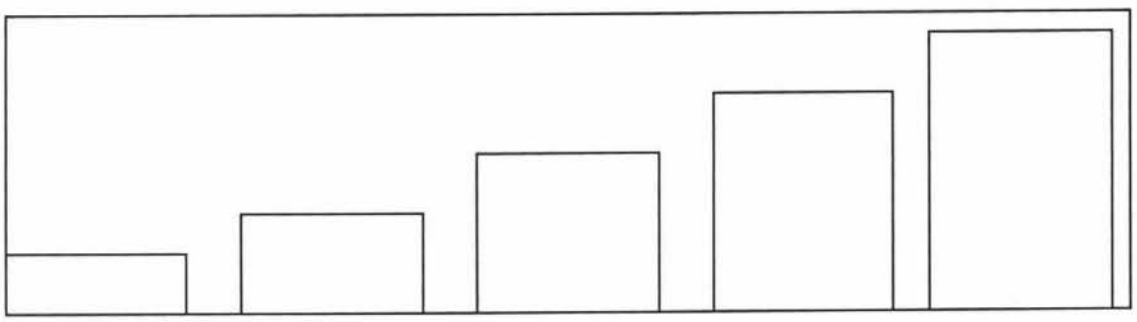
DON'T KNOW: **DK** NOT APPLICABLE: **NA**

My beliefs, attitudes and values

How important to me is -- ?

1. Being caring towards other people _____
2. Celebrating birthdays or special events _____
3. Feeling peaceful within myself _____
4. Feeling that my life has purpose _____
5. Sharing love with other people _____
6. Having my own ideas of right and wrong _____
7. Having religious or spiritual beliefs _____
8. Having things to look forward to _____
9. Participating in religious or spiritual activities _____

Rating scale: IMPORTANCE



NOT AT ALL IMPORTANT	*NOT VERY IMPORTANT*	*IMPORTANT*	*VERY IMPORTANT*	*EXTREMELY IMPORTANT*
1	**2**	**3**	**4**	**5**

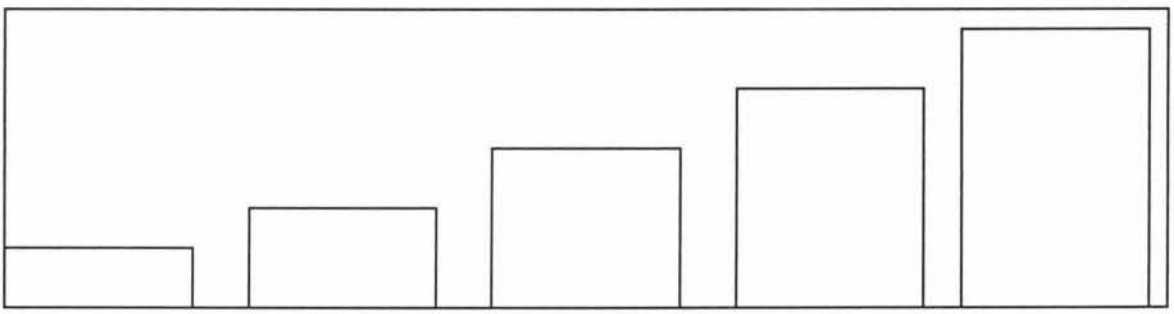
DON'T KNOW: **DK** NOT APPLICABLE: **NA**

Where I live

How important to me is -- ?

1. Where I live _____
2. Living in a safe place _____
3. Having a space for privacy _____
4. Having my own personal things _____
5. Living in a comfortable place _____
6. Living in a place with enough space _____
7. Living in a place that is physically accessible to me _____
8. What part of New Zealand I live in _____
9. Living near my family or friends _____
10. What neighbourhood I live in _____
11. Living in a safe neighbourhood _____

Rating scale: IMPORTANCE



NOT AT ALL IMPORTANT **1** | *NOT VERY IMPORTANT* **2** | *IMPORTANT* **3** | *VERY IMPORTANT* **4** | *EXTREMELY IMPORTANT* **5**

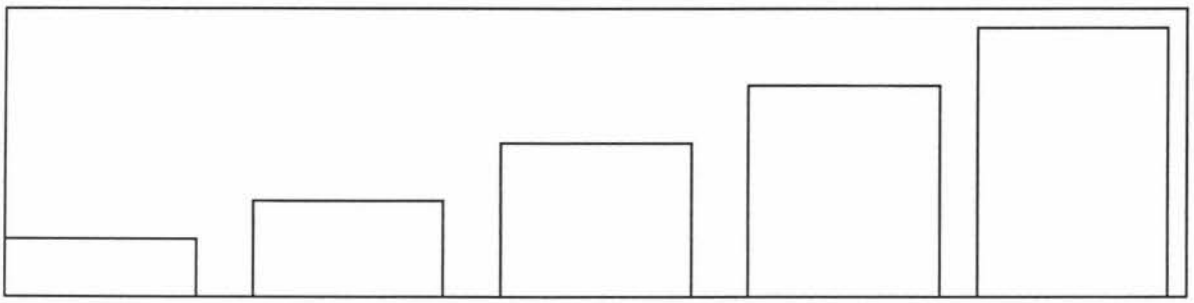
DON'T KNOW: **DK** NOT APPLICABLE: **NA**

The people around me

How important to me is -- ?

1. Having a spouse, partner or special person _____
2. Having friends _____
3. Being close to some members of my family _____
4. Having acquaintances _____
5. Having neighbours I can turn to _____
6. Being able to count on family members for help _____
7. Having people nearby who I can communicate with _____
8. Meeting in social/cultural/interest/faith groups _____
9. The degree to which I depend on people in my family _____
10. Having social events to attend _____
11. Being accepted by the people I see regularly (at work, school, etc.) _____
12. Sexual intimacy _____
13. Being respected by people around me _____

Rating scale: IMPORTANCE



NOT AT ALL IMPORTANT	*NOT VERY IMPORTANT*	*IMPORTANT*	*VERY IMPORTANT*	*EXTREMELY IMPORTANT*
1	**2**	**3**	**4**	**5**

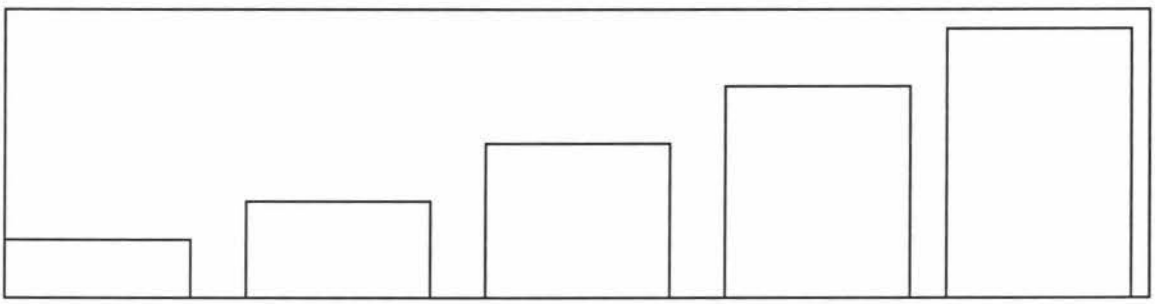
DON'T KNOW: **DK** NOT APPLICABLE: **NA**

My access to resources

How important to me is -- ?

1. Being able to get health services (from doctors, therapists, nurses, dentists etc.) _____
2. Being able to get social services (vocational services, social worker, etc.) _____
3. Being able to get other special services (attendant care etc.) _____
4. Being able to live in affordable housing _____
5. Going to places in my neighbourhood (stores, etc.) _____
6. Feeling the government understands my needs _____
7. Having access to meaningful work _____
8. Having courses, classes, or programs that I can take _____
9. Having enough money to live comfortably _____
10. Having events in my community to go to (movies, concerts, etc.) _____
11. Having programs and services in a language or form I understand _____
12. Having transportation that allows me to get where I want to be _____
13. Having adaptive equipment or resources (wheelchair, Braille formats, telephone adaptations, etc.) _____

Rating scale: IMPORTANCE



NOT AT ALL IMPORTANT	*NOT VERY IMPORTANT*	*IMPORTANT*	*VERY IMPORTANT*	*EXTREMELY IMPORTANT*
1	**2**	**3**	**4**	**5**

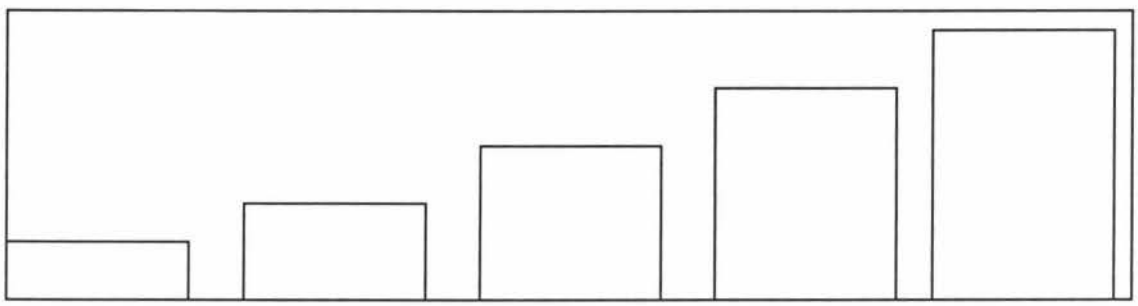
DON'T KNOW: **DK** NOT APPLICABLE: **NA**

The practical things I do

How important to me is -- ?

1. The everyday things I do for a spouse or other adult (laundry, cleaning, etc.) _____
2. Looking after a pet _____
3. Doing volunteer work through an organisation _____
4. Doing work around my home (cooking, repairs, etc.) _____
5. Doing work I get paid for _____
6. Going to appointments (doctor, dentist, therapist, etc.) _____
7. Looking after my children or other children _____
8. Shopping for myself or others _____
9. Helping family, friends, or neighbours in practical ways _____
10. Doing school work or course work _____

Rating scale: IMPORTANCE



NOT AT ALL IMPORTANT	*NOT VERY IMPORTANT*	*IMPORTANT*	*VERY IMPORTANT*	*EXTREMELY IMPORTANT*
1	**2**	**3**	**4**	**5**

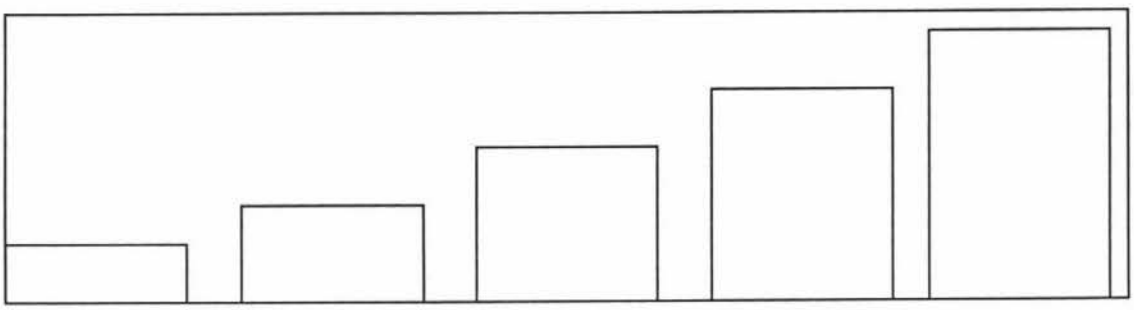
DON'T KNOW: **DK** NOT APPLICABLE: **NA**

The things I do for enjoyment

How important to me is -- ?

1. Having vacation and holiday activities _____
2. Getting out with others (shopping, lunch, etc.) _____
3. Going to community events like fairs and sales _____
4. Going to movies or shows _____
5. Doing hobbies (painting, gardening, knitting, etc.) _____
6. Doing indoor activities (TV, reading, etc.) _____
7. Doing outdoor activities (walks, driving, etc.) _____
8. Participating in holiday activities (Christmas, Waitangi Day, Queens Birthday) _____
9. Participating in organised recreation activities (cards, sports, bingo, etc.) _____
10. Visiting and socialising with friends and neighbours _____
11. Visiting and socialising with people in my family _____
12. Taking breaks from my usual routines _____

Rating scale: IMPORTANCE



NOT AT ALL IMPORTANT	*NOT VERY IMPORTANT*	*IMPORTANT*	*VERY IMPORTANT*	*EXTREMELY IMPORTANT*
1	**2**	**3**	**4**	**5**

DON'T KNOW: **DK** NOT APPLICABLE: **NA**

The things I do to improve myself

How important to me is -- ?

1. Adjusting to changes in my personal life _____
2. Creating new challenges and/or projects in my life _____
3. Improving or maintaining my skills (mental, manual, communication, etc.) _____
4. Improving or maintaining my mental health _____
5. Improving or maintaining my physical health _____
6. Learning about new things _____
7. Learning to get along better with others _____
8. Solving my own problems _____
9. Trying things I haven't tried before _____
10. Sharing ideas with other people _____
11. Working towards goals I set for myself _____

II. Satisfaction

Instructions

The second question to ask yourself is:

How *satisfied* am I with this part of my life?

Another way to think about this question is:

How happy am I with this aspect of my life?

Rate each of the items from 1 to 5, using the rating scale I am giving you now. Rate items 5 if you are *extremely satisfied* with these parts of your life; rate items 4 if you feel *very satisfied*. Rate items 3 if you think you are feeling *satisfied* with these parts of your life; rate items 2 if you are *not very satisfied*; rate items 1 if your are not at all *satisfied* with these parts of your life.

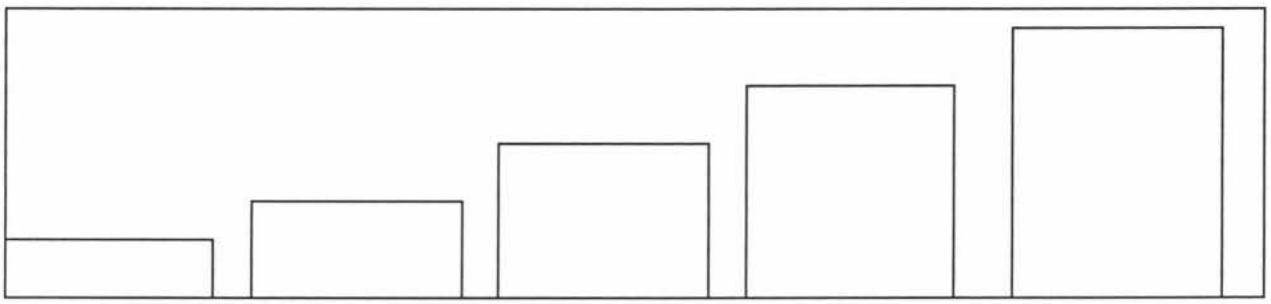
Answer each question in terms of your life as it is right now.
Answer each question whether or not you can actually participate in or do the activities described.

If you feel that the question does not apply to you, you would give a rating of "N/A" (Not Applicable).

I you cannot answer a question because you are very unsure, give a rating of "DK" (Don't Know).

SATISFACTION

Rating scale:



NOT AT ALL SATISFIED	*NOT VERY SATISFIED*	*SATISFIED*	*VERY SATISFIED*	*EXTREMELY SATISFIED*
1	**2**	**3**	**4**	**5**

DON'T KNOW: **DK**

NOT APPLICABLE: **NA**

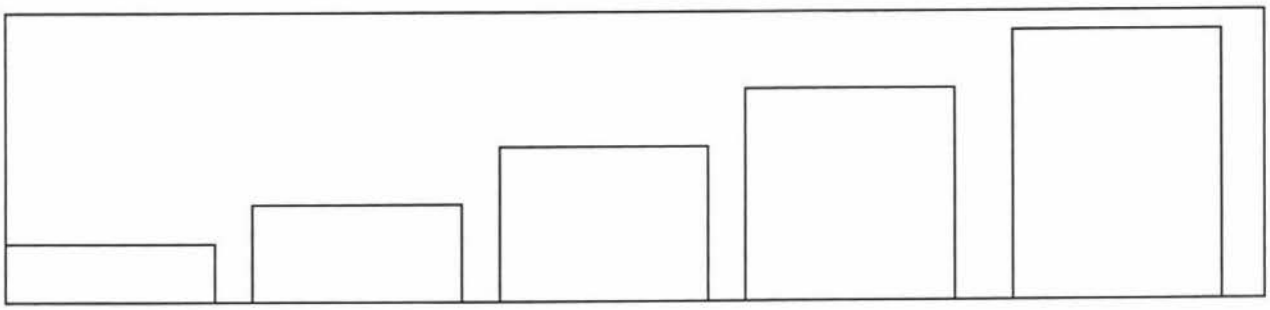
My body and my health

How satisfied am I with -- ?

1. My physical ability to get around my home _____
2. My physical ability to get around my neighbourhood _____
3. My physical ability to use public transportation _____
4. How I keep physically active and keeping fit _____
5. The sleep and rest I get _____
6. My nutrition and the food I eat _____
7. The energy I have to do the things I want to _____
8. How I maintain my personal hygiene and caring for myself, by MYSELF _____
9. How I maintain my personal hygiene and caring for myself, WITH THE ASSISTANCE OF OTHERS _____
10. My personal appearance _____
11. How I am able to manage the pain that I have _____
12. My overall physical health _____

SATISFACTION

Rating scale:



NOT AT ALL SATISFIED	*NOT VERY SATISFIED*	*SATISFIED*	*VERY SATISFIED*	*EXTREMELY SATISFIED*
1	**2**	**3**	**4**	**5**

DON'T KNOW: **DK**

NOT APPLICABLE: **NA**

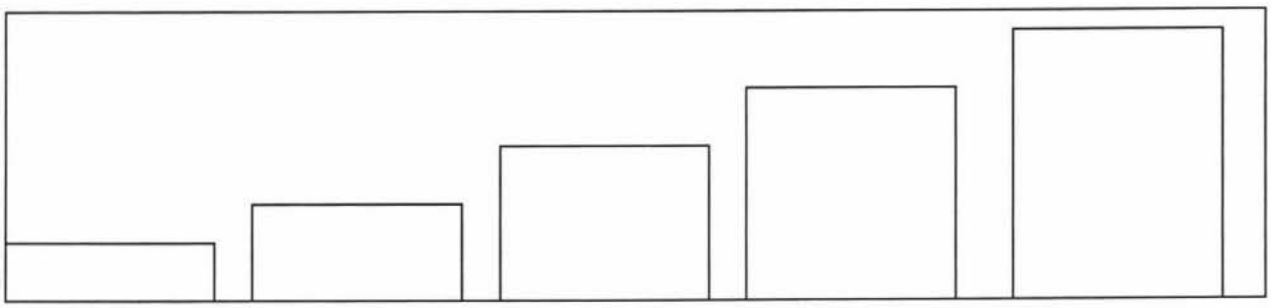
My emotional state

How satisfied am I with -- ?

1. How much I accept myself _____
2. How much I make my own decisions _____
3. How much I act independently, on my own _____
4. My ability to remember things _____
5. How free I am of stress _____
6. The mood I am usually in _____
7. How I cope with what life brings _____
8. How I feel about myself _____
9. My attitude towards life _____
10. My sense of humour _____
11. My mental health _____

SATISFACTION

Rating scale:



NOT AT ALL SATISFIED	*NOT VERY SATISFIED*	*SATISFIED*	*VERY SATISFIED*	*EXTREMELY SATISFIED*
1	**2**	**3**	**4**	**5**

DON'T KNOW: **DK**

NOT APPLICABLE: **NA**

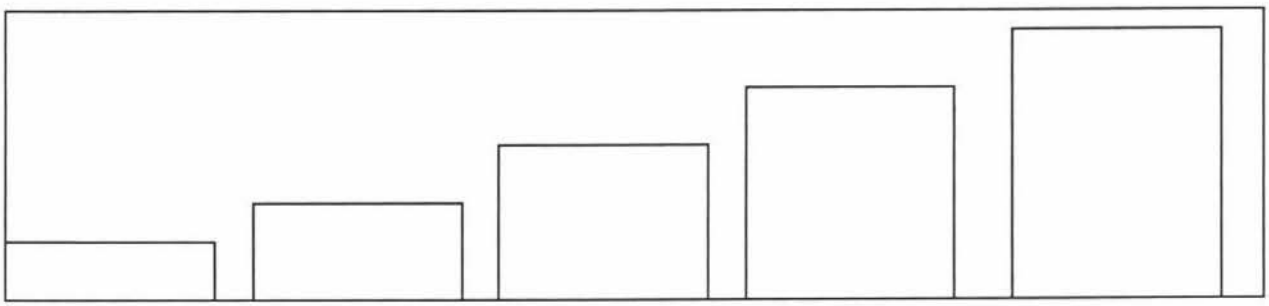
My beliefs, attitudes and values

How important to me is -- ?

1. How caring I am towards other people _____
2. How I celebrate birthdays or special events _____
3. How peaceful I feel within myself _____
4. How much I feel that my life has purpose _____
5. How much I share love with other people _____
6. My own ideas of right and wrong _____
7. My religious or spiritual beliefs _____
8. How much I have things to look forward to _____
9. My participation in religious or spiritual activities _____

SATISFACTION

Rating scale:



NOT AT ALL SATISFIED **1**

NOT VERY SATISFIED **2**

SATISFIED **3**

VERY SATISFIED **4**

EXTREMELY SATISFIED **5**

DON'T KNOW: **DK**

NOT APPLICABLE: **NA**

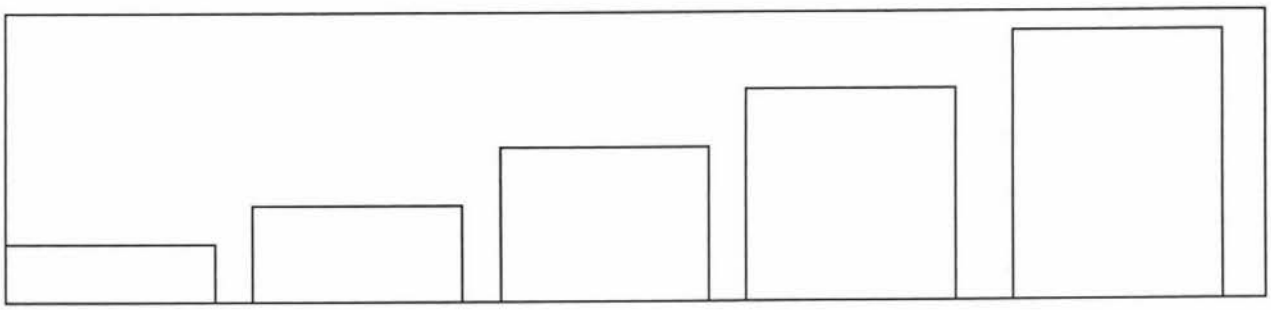
Where I live

How satisfied am I with -- ?

1. Where I live _____
2. The safety of my place _____
3. The space I have for privacy _____
4. The personal things I have _____
5. How comfortable my living place is _____
6. How much space I have in the place where I live _____
7. How physically accessible my place is _____
8. The part of New Zealand I live in _____
9. How near I live to my family or friends _____
10. What neighbourhood I live in _____
11. How safe my neighbourhood is _____

SATISFACTION

Rating scale:



NOT AT ALL SATISFIED	*NOT VERY SATISFIED*	*SATISFIED*	*VERY SATISFIED*	*EXTREMELY SATISFIED*
1	**2**	**3**	**4**	**5**

DON'T KNOW: **DK**

NOT APPLICABLE: **NA**

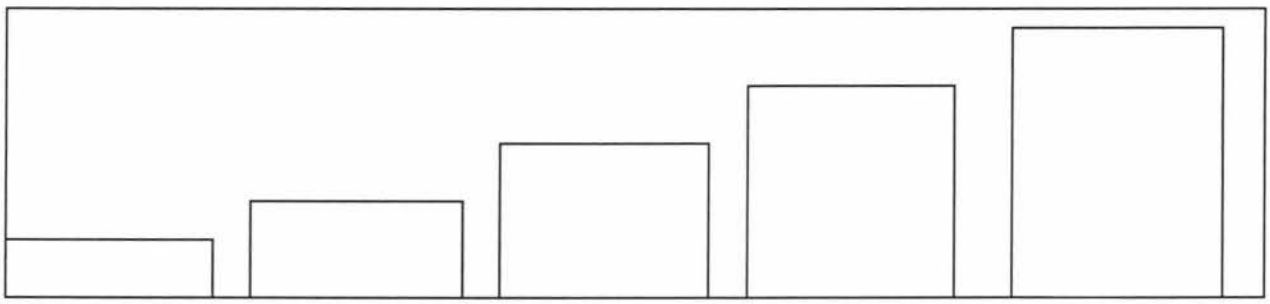
The people around me

How satisfied am I with -- ?

1. My spouse, partner or special person _____
2. My friends _____
3. How close I am to some members of my family _____
4. My acquaintances _____
5. How much I have neighbours I can turn to _____
6. How much I can count on family members for help _____
7. My access to people nearby who I can communicate with _____
8. My involvement in social/cultural/interest/faith groups _____
9. The degree to which I depend on people in my family _____
10. The social events I attend _____
11. How accepted I am by the people I see regularly (at work, school, etc.) _____
12. How much sexual intimacy I have _____
13. How respected I am by people around me _____

SATISFACTION

Rating scale:



NOT AT ALL SATISFIED	*NOT VERY SATISFIED*	*SATISFIED*	*VERY SATISFIED*	*EXTREMELY SATISFIED*
1	**2**	**3**	**4**	**5**

DON'T KNOW: **DK**

NOT APPLICABLE: **NA**

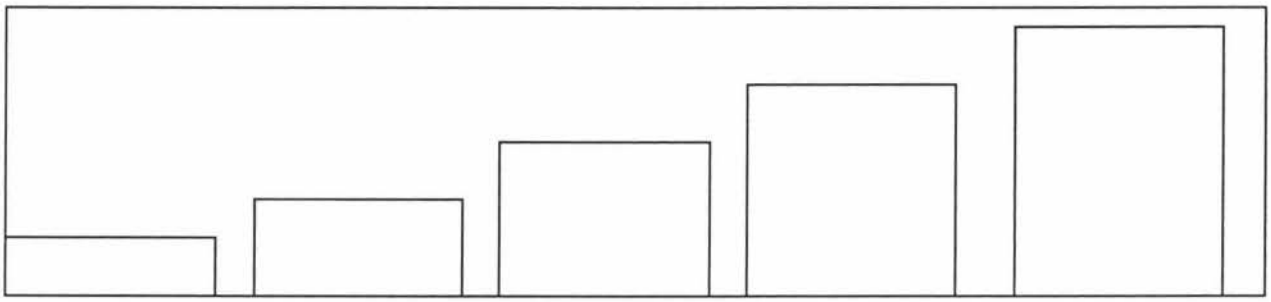
My access to community resources

How satisfied am I with -- ?

1. How able I am to get health services (from doctors, therapists, nurses, dentists etc.) _____
2. How able I am to get social services (vocational services, social worker, etc.) _____
3. How able I am to get other special services (attendant care etc.) _____
4. How able I am to live in affordable housing _____
5. How much I go to neighbourhood places (stores, etc.) _____
6. How much the government understands my needs _____
7. My access to meaningful work _____
8. The courses, classes, or programs that I can take _____
9. The amount of money I have _____
10. Events in my community to go to (movies, concerts, etc.) _____
11. Programs and services in a language or form I understand _____
12. The transportation available to get where I want to be _____
13. The adaptive equipment or resources I have (wheelchair, Braille formats, telephone adaptations, etc.) _____

SATISFACTION

Rating scale:



NOT AT ALL SATISFIED	*NOT VERY SATISFIED*	*SATISFIED*	*VERY SATISFIED*	*EXTREMELY SATISFIED*
1	**2**	**3**	**4**	**5**

DON'T KNOW: **DK**

NOT APPLICABLE: **NA**

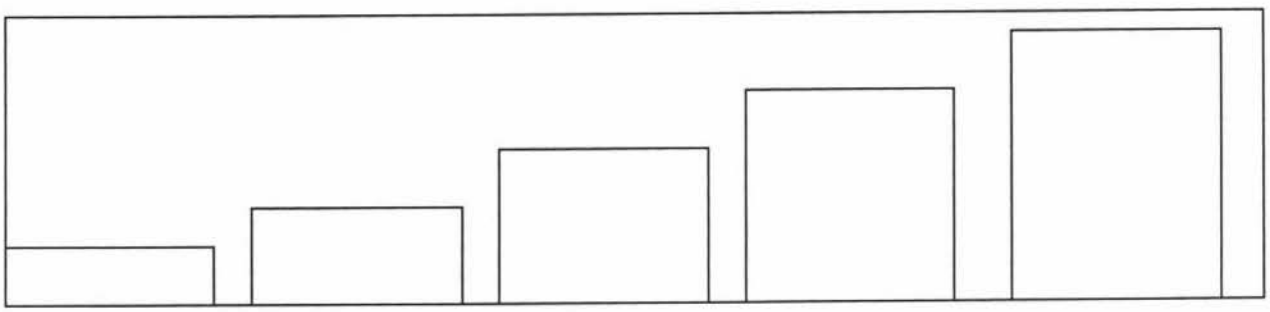
The practical things I do

How satisfied am I with -- ?

1. The everyday things I do for a spouse or other adult (laundry, cleaning, etc.) _____
2. Looking after a pet _____
3. The volunteer work I do _____
4. The work I do around my home (cooking, repairs, etc.) _____
5. The work I do that I get paid for _____
6. The appointments I have (doctor, dentist, therapist, etc.) _____
7. The looking after I do for my children or other children _____
8. The shopping I do for myself or others _____
9. The help I give to family, friends, or neighbours in practical ways _____
10. Doing school work or course work _____

SATISFACTION

Rating scale:



NOT AT ALL SATISFIED	*NOT VERY SATISFIED*	*SATISFIED*	*VERY SATISFIED*	*EXTREMELY SATISFIED*
1	**2**	**3**	**4**	**5**

DON'T KNOW: **DK**

NOT APPLICABLE: **NA**

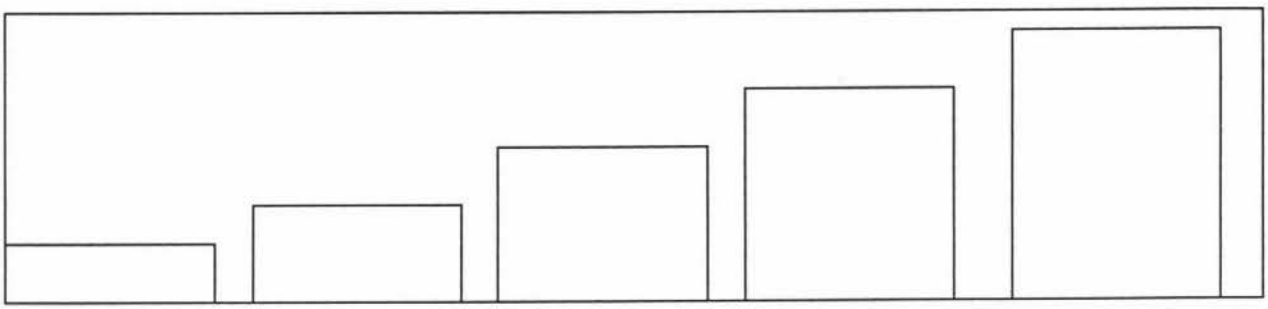
The things I do for enjoyment

How satisfied am I with -- ?

1. My vacation and holiday activities _____
2. How much I get out with others (shopping, lunch, etc.) _____
3. The community events I go to (fairs and sales, etc.) _____
4. The movies or shows I go to _____
5. My hobbies (painting, gardening, knitting, etc.) _____
6. My indoor activities (TV, reading, etc.) _____
7. My outdoor activities (walks, driving, etc.) _____
8. My holiday activities (Christmas, Waitangi Day, Queens Birthday) _____
9. My organised recreation activities (cards, sports, bingo, etc.) _____
10. My visiting and socialising with friends and neighbours _____
11. My visiting and socialising with people in my family _____
12. The breaks I take from my usual routines _____

SATISFACTION

Rating scale:



NOT AT ALL SATISFIED	*NOT VERY SATISFIED*	*SATISFIED*	*VERY SATISFIED*	*EXTREMELY SATISFIED*
1	**2**	**3**	**4**	**5**

DON'T KNOW: **DK**

NOT APPLICABLE: **NA**

The things I do to improve myself

How satisfied am I with -- ?

1. How I am adjusting to changes in my personal life _____
2. How I am creating new challenges and/or projects in my life _____
3. How I am improving or maintaining my skills (mental, manual, communication, etc.) _____
4. How I am improving or maintaining my mental health _____
5. How I am improving or maintaining my physical health _____
6. How I am learning about new things _____
7. How I am learning to get along better with others _____
8. How I am solving my own problems _____
9. How I am trying things I haven't tried before _____
10. How I share ideas with other people _____
11. How I am working towards my own goals _____

III. Control

Instructions:

The third question to ask yourself is:

How much *control* do I have over this part of my life?

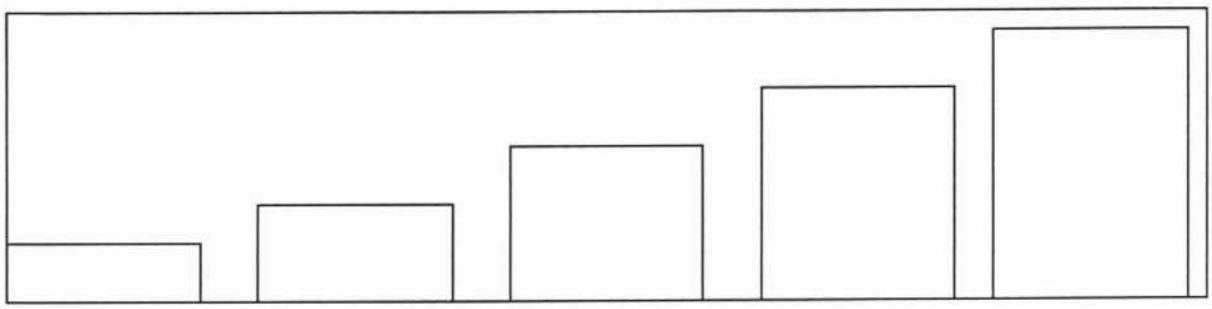
Another way to think about the question is:

How much am I in charge of this aspect of my life?

Rate each of the items from 1 to 5, using the rating scale I am giving to you now. Rate items 5 if you have *almost total control* in this area of your life; rate items 4 if you have *much control* in this area of your life. Rate items 3 if you think you have about the *same amount of control* as most people in this aspect of your life; rate items 2 if you *do not have much control*; rate items 1 if you *almost no control* in this part of your life.

CONTROL

Rating Scale



ALMOST NO CONTROL	*NOT MUCH CONTROL*	*SOME CONTROL*	*MUCH CONTROL*	*ALMOST TOTAL CONTROL*
1	**2**	**3**	**4**	**5**

DON'T KNOW: **DK**

NOT APPLICABLE: **NA**

Control

How much control do I have over -- ?

1. My physical health and self care _____
2. My thoughts and feelings _____
3. The spiritual part of my life _____
4. Where I am living or will be living _____
5. Who I spend my time with _____
6. Being able to use what my community has to offer (transportation, services, resources, etc.) _____
7. The everyday things I can do in my life _____
8. The things I can do for fun and enjoyment _____
9. The things I can do to improve myself _____

IV. Potential opportunities

Instructions:

The last question you ask yourself is:

Are there many *opportunities* for me to improve or change this part of my life?

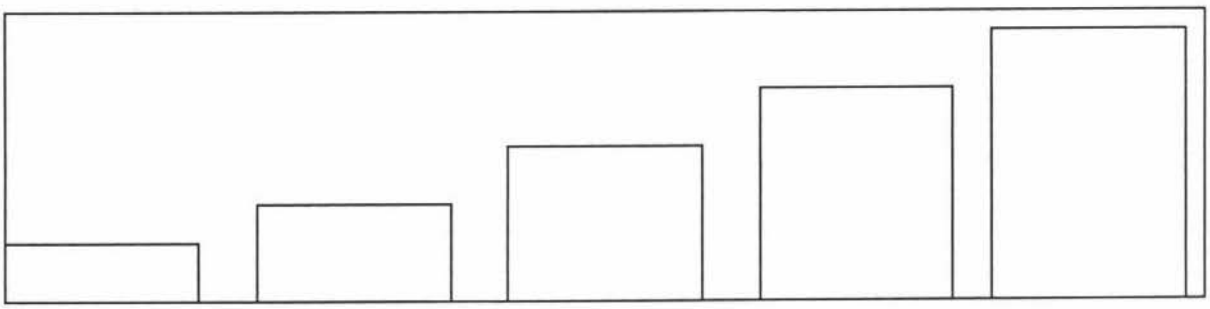
Another way to think about the question is:

Are there alternative choices available to me about this aspect of my life:

Rate each of the items from 1 to 5 using the rating scale I am giving you now. Rate items 5 if you have a *great many opportunities* in this part of your life; rate items 4 if you have *many opportunities* in this part of your life. Rate items 3 if you think you have about the *same number of opportunities* as most people in this part of your life; rate items 2 if you have *a few opportunities*; rate items 1 if you have *almost no opportunities* in this part of your life.

OPPORTUNITIES

Rating Scale



ALMOST NONE	*A FEW*	*SOME*	*MANY*	*GREAT MANY*
1	**2**	**3**	**4**	**5**

DON'T KNOW: **DK**

NOT APPLICABLE: **NA**

Opportunities

Are there opportunities for me to -- ?

1. Improve or maintain my physical health and self care _____
2. Improve or maintain how I think and feel about things _____
3. Improve or maintain the spiritual part of my life _____
4. Live in a comfortable or pleasing place _____
5. Spend time with different people _____
6. Use more of what my community has to offer (transportation, services, resources, etc.) _____
7. Do different daily activities than I do now _____
8. Do different things for fun and relaxation than I do now _____
9. Learn and do new things _____

The Ritchie Articular Index

Joints examined	**Not tender (0)**	**Tender (+1)**	**Tender and winced (+2)**	**Tender, winced and withdrew (+3)**	**Joint score**
Temporo-mandibular					
Cervical spine					
Acromio-clavicular					
Shoulder Left					
Shoulder Right					
Elbow Left					
Elbow Right					
Wrist Left Wrist Right					
Metacarpophalangeal Left					
Metacarpophalangeal Right					
Proximal interphalangeal Left					
Proximal interphalangeal Right					
Hip Left					
Hip Right					
Knee Left					
Knee Right					
Ankle Left					
Ankle Right					
Talocalcaneal Left					
Talocalcaneal Right					
Midtarsal Left					
Midtarsal Right					
Metatarsal Left					
Metatarsal Right					
Total					

The Health Assessment Questionnaire

In this section we are interested in learning how your illness affects your ability to function in daily life. Please tick the **one response** which best describes your usual abilities OVER THE PAST WEEK.

	Without ANY difficulty	With SOME difficulty	With MUCH difficulty	UNABLE to do
1. DRESSING AND GROOMING				
Are you able to:				
a. Dress yourself, including tying shoelaces and doing buttons?	☐	☐	☐	☐
b. Shampoo your hair?	☐	☐	☐	☐
2. RISING				
Are you able to:				
a. Get in and out of bed?	☐	☐	☐	☐
b. Stand up from a straight chair without using your arms for support?	☐	☐	☐	☐
3. EATING				
Are you able to:				
a. Cut your meat	☐	☐	☐	☐
b. Lift a full cup or glass to your mouth	☐	☐	☐	☐
c. Open a new carton of juice, milk or soap powder?	☐	☐	☐	☐
4. WALKING				
Are you able to:				
a. Walk outdoors on flat ground?	☐	☐	☐	☐
b. Climb up five steps?	☐	☐	☐	☐

PLEASE TICK ANY AIDS OR DEVICES THAT YOU USUALLY USE FOR ANY OF THESE ACTIVITIES:

Cane ☐ Walking frame ☐ Crutches ☐ Wheelchair ☐

Devices used for dressing (button hook, shoe horn etc) ☐

Built up or special utensils ☐ Special or built up chair ☐ Anything else ☐

Please tick the **one response** which best describes your usual abilities OVER THE PAST WEEK.

	Without ANY difficulty	With SOME difficulty	With MUCH difficulty	UNABLE to do
5. HYGIENE				
Are you able to:				
a. Wash and dry your entire body	☐	☐	☐	☐
b. Take a bath	☐	☐	☐	☐
c. Get on and off the toilet	☐	☐	☐	☐
6. REACH				
Are you able to:				
a. Reach and get down a 3lb object (e.g. a bag of potatoes) from just above your head?	☐	☐	☐	☐
b. Bend down and pick clothing from the floor?	☐	☐	☐	☐
7. GRIP				
Are you able to:				
a. Open car doors?	☐	☐	☐	☐
b. Open jars which have previously been opened?	☐	☐	☐	☐
8. ACTIVITY				
Are you able to:				
a. Get in and out of the car?	☐	☐	☐	☐
b. Run errands and shop?	☐	☐	☐	☐
c. Do chores such as vacuuming or sweeping the drive?	☐	☐	☐	☐

Please tick any AIDS OR DEVICES that you usually use for any of these activities:

Raised toilet seat	☐	Bath seat	☐	Bath rail	☐
Jar opener	☐	Long handled appliances for reach			☐

Long handled appliances in the bathroom ☐ Other ☐

Please tick any categories for which you usually need HELP FROM ANOTHER PERSON:

Dressing and grooming ☐ Eating ☐ Arising ☐ Walking ☐

Errands and chores ☐ Reach ☐ Hygiene ☐

Gripping and opening ☐

1. Considering all the ways your arthritis is affecting you, mark X on the line below to indicate how well you have been doing over the last 24 hours?

Very well _____ Not well at all

2. How much pain have you had over the last 24 hours?

No pain _____ Pain as bad as it could be

Appendix 2: Data sets

1. Data set 1: QoL instrument scores and external criteria scores by subject and visit
2. Data set 2: Change in QoL and external criteria scores

Data set 1: QoL and external criteria scores by subject and visit

Data set description:

Fields are space delimited

Field label	**Description**
Subject	Individual subject label
Visit	Data collection point
1	Pre-admission
2	Admission
3	Discharge
4	Follow up
Qlp	Quality of life profile score
EuroQol_rate	EuroQol rating scale score
EuroQol_vas	EuroQol visual analogue scale score
Whoqol	WHOQoL-Bref score
Haq	Health assessment questionnaire score
Ritchie	Ritchie articular index score
ESR	ESR measurement

Subject Visit Qlp EuroQol_rate EuroQol_vas Whoqol Haq Ritchie ESR
1 1 0.75 35 20 289.6 1.25 7
1 2 0.25 46.8 70 321.0 1.125 2 90
1 3 2.69 100 98 367.6 0.25 0 40
1 4 3.22 100 98 393.3 0 3
2 1 1.59 56.7 72 308.3 0.375 13
2 2 0.86 63.6 70 308.3 0 9 80
2 3 1.15 67 80 312.5 0 7 35
2 4 1.09 67 60 313.1 0.375 8
3 1 0.87 33.1 65 260.6 1.375 14
3 2 0.06 42 65 246.3 1.375 16
3 3 0.57 53.3 75 275.9 1.375 22
3 4 1.06 53.3 80 260.4 1.375 8
4 1 1.31 56.7 65 318.3 0.75 7
4 2 1.58 100 75 330.8 0.375 3 30
4 3 1.90 100 85 347.2 0.25 0 15
4 4 1.80 100 96 389.1 0.25 1
5 1 0.45 33.9 70 282.1 0.875 12
5 2 0.58 44.3 70 289.3 1 12 35
5 3 0.13 44.3 79 290.8 0.875 7
5 4 0.24 44.3 80 289.3 0.375 2
6 1 -.10 33.1 70 175.0 2.375 22
6 2 -.36 44.3 70 217.9 2.375 16 70
6 3 -.05 44.3 68 249.0 2.375 15
6 4 0.10 44.3 37 249.0 2.5 20
7 1 0.98 35 233.3 2 23
7 2 0.47 37 238.5 1.875 24
7 3 0.53 44.3 70 234.8 1.125 10 65
7 4 0.15 40 240.0 1.875 27
8 1 0.57 33.9 48 248.2 2.75 21
8 2 0.99 22.6 50 274.3 2.875 36 70
8 3 1.68 44.3 60 332.9 2.125 40
8 4 1.62 44.3 80 297.6 2.125 15
9 1 0.37 30.4 37 241.1 2.25 9
9 2 0.18 33.1 29 237.5 2.25 9 73
9 3 0.46 54.6 70 255.4 1.125 5 50
9 4 0.01 44.3 47 250.0 1.875 8
10 1 44.3 55 241.1 1.25 10
10 2 -.51 44.3 55 257.1 1.25 17 65
10 3 0.00 56.7 100 312.2 0.5 2 20
10 4 0.37 100 100 390.6 0 5
11 1 33.1 60 213.4 1.5 10
11 2 0.25 33.1 50 225.4 1 16 105
11 3 1.24 100 80 299.6 0 0 75
11 4 2.05 67 90 277.2 0.625 1
12 1 0.35 44.3 57 218.0 1.875 14
12 2 0.26 44.3 37 255.8 1.875 14 105
12 3 0.28 44.3 46 268.2 1.875 10 60
12 4 0.35 44.3 37 268.6 1.875 9

13 1 1.13 44.3 25 309.5 1.75 12
13 2 1.04 44.3 25 301.8 1.75 14 123
13 3 1.25 54.6 60 326.3 0.875 5
13 4 1.87 67 70 372.8 0.5 2
14 1 -1.5 22.6 30 85.71 1.375 15
14 2 -1.5 33.1 20 88.84 1.5 18 45
14 3 -1.4 74 20 100.0 1 15
14 4 -1.5 33.1 70 109.5 0.625 21
15 1 0.85 53.3 65 274.9 2 3
15 2 0.96 53.3 65 286.8 2.125 3 62
15 3 0.81 100 85 278.3 2 2 45
15 4 0.77 100 74 269.9 2 0
16 1 2.52 53.3 50 300.7 1.25 5
16 2 2.70 63.5 50 289.1 0.75 7 60
16 3 3.06 63.5 50 0.625 6
16 4
17 1 0.08 44.3 70 214.1 0.75 5
17 2 0.01 75 186.2 0.625 6 120
17 3 0.07 67 90 212.9 0.375 4 110
17 4 0.10 56.7 80 248.1 0.625 6
18 1 22.6 25 147.6 2.75 25
18 2
18 3
18 4
19 1 0.43 11.3 35 222.2 1.75 19
19 2 0.52 220.1 1.75 21 60
19 3 1.16 42 85 275.3 1.125 13 32
19 4 0.96 42 85 302.1 1.25 6
20 1 0.15 19.1 50 144.6 2 4
20 2 -.16 22.6 50 133.6 2.25 6 90
20 3 -.18 7.9 50 196.1 1.75 5 70
20 4 -.20 50 209.4 1.75 7
21 1 0.25 55.7 30 136.0 1.25 9
21 2 0.27 66.2 33 208.0 0.625 6 15
21 3 0.76 66.2 40 218.6 0.75 3 10
21 4 0.74 100 100 275.0 0.625 3
22 1 0.67 33.9 70 292.6 1.875 17
22 2 0.50 44.3 70 274.4 1.625 22 118
22 3 1.37 44.3 80 302.8 1.25 14 90
22 4 2.22 44.3 80 379.0 1.5 13
23 1 -.05 29.6 30 187.1 2.375 11
23 2 0.38 33.1 40 171.3 1.375 11 38
23 3 0.65 44.3 80 243.8 1.375 9 42
23 4 0.92 42 80 270.7 0.75 7
24 1 33.1 50 208.5 1.625 15
24 2 33.1 50 199.1 2 17 15
24 3 44.3 50 204.5 1.625 13
24 4 22.6 60 193.9 1.875 20

25 1 1.45 53.3 75 280.5 0.625 6
25 2 1.34 50 80 286.8 0.5 1
25 3 1.08 50 86 307.0 0.5 2 40
25 4 1.11 53.3 90 295.5 0.5 4
26 1 0.33 44.3 42 241.5 2.25 13
26 2 0.65 44.3 35 249.3 2.25 9
26 3 0.88 44.3 30 249.3 2.125 13 60
26 4 0.60 44.3 66 234.1 2.125 7
27 1 -.01 7.9 20 178.1 2.5 26
27 2 -.06 19.1 17 170.5 2.125 28 73
27 3 0.42 54.6 64 247.0 1.625 18 40
27 4 0.13 44.3 62 270.5 1.75 23
28 1 1.10 22.6 55 2.375 1
28 2 2.24 42 64 231.4 1.75 6
28 3 0.81 42 60 258.6 1.375 5
28 4 0.75 56.7 67 0.75 5
29 1 0.48 44.3 70 247.5 2.125 34
29 2 0.04 44.3 71 220.2 2 23 50
29 3 0.18 44.3 70 1.75 18
29 4 0.20 44.3 70 232.9 1.5 21
30 1 -.35 18.1 40 192.9 2.5 4
30 2 0.02 9.2 30 174.7 2.25 13 45
30 3 2.85 63.5 80 211.2 0.75 1
30 4 2.04 63.5 70 324.1 0.625 1
31 1 0.76 22.6 237.8 2.75 8
31 2 0.28 33.1 288.5 2.875 4 120
31 3 0.71 53.3 308.3 2.5 4 100
31 4 0.22 53.3 274.3 2.75 3
32 1 0.10 30.4 50 243.2 2.25 14
32 2 0.55 40.9 60 2 19 20
32 3 0.82 40.9 55 221.1 2.125 13 10
32 4 0.25 44.3 55 258.5 1.5 15
33 1
33 2
33 3
33 4
34 1 -.20 36.5 57 191.8 1.875 11
34 2 -.08 22.8 52 200.4 1.75 15 20
34 3 0.39 66.2 85 258.0 1.125 9 10
34 4 1.61 66.2 84 307.7 0.55 4
35 1 -.16 22.6 40 240.3 1.375 9
35 2 0.47 63.5 70 315.0 1 6 10
35 3 0.27 67 69 311.5 1.125 6
35 4 0.43 67 79 311.8 1.125 3
36 1 0.07 33.1 36 174.4 1.625 11
36 2 0.17 33.1 60 190.0 1.375 10 70
36 3 1.12 33.1 55 178.6 1.5 14 88
36 4 0.41 33.1 40 170.6 1.125 15

37 1 0.75 22.6 70 248.7 1.875 17
37 2 1.35 44.3 70 264.1 1.875 24 85
37 3 1.59 44.3 65 277.2 2.125 13 45
37 4 1.60 54.6 60 255.4 1.625 12
38 1 -.97 33.1 40 133.8 2.75 8
38 2 -.34 54.6 75 163.4 2.375 6 45
38 3 -.24 44.3 50 206.8 1.5 4 20
38 4 -.16 63.5 60 216.4 1.125 5
39 1 49 80 258.2 2.5 10
39 2 0.71 29.6 80 245.1 2.25 8 115
39 3 0.79 70 259.2 0.75 3 70
39 4 0.47 74 75 278.1 1.25 4
40 1 0.44 42 50 198.7 0.375 2
40 2 0.64 55.7 75 218.3 0.25 1
40 3 1.46 67 90 269.2 0 1 25
40 4 0.88 56.7 80 269.5 0 0
41 1 0.23 30.4 40 232.3 2.25 37
41 2 0.57 30.4 36 249.6 2.375 35 41
41 3 1.35 33.1 66 305.2 1.375 18 30
41 4 1.25 33.1 65 313.5 1.375 23
42 1 1.45 53.3 50 282.0 1.25 18
42 2 1.48 53.3 60 287.4 1.75 21 10
42 3 1.13 63.7 70 267.4 1.25 7
42 4 1.24 63.7 80 305.1 0.875 3
43 1 0.23 53.3 60 247.0 1.875 12
43 2 0.44 44.3 65 267.9 1.75 13 40
43 3 0.54 54.6 80 282.1 1.625 12 30
43 4 1.07 53.3 65 293.3 1.5 14
44 1 0.29 22.6 25 215.6 2.125 18
44 2 0.43 29.6 44 234.4 2.375 18 100
44 3 1.34 56.7 79 296.9 0.75 13 50
44 4 1.29 42 85 330.8 0.625 14
45 1 -.01 33.1 30 197.9 2 8
45 2 -.30 33.1 34 197.8 2 9 40
45 3 -.01 67 80 299.4 1 10 10
45 4 0.99 54.6 90 279.5 0.875 3
46 1 0.79 44.3 70 280.7 0.875 7
46 2 0.72 53.3 296.6 0.75 12 80
46 3 1.09 53.3 90 312.5 0.625 72
46 4
47 1 0.29 33.1 40 233.5 2.625 15
47 2 33.1 65 2.5 16 50
47 3 33.1 75 2.25 15 50
47 4 0.20 58.1 84 315.5 2.125 8
48 1 33.1 40 259.1 2.125 16
48 2 33.1 40 261.9 1.375 11 20
48 3 -.06 52.3 60 280.4 1.5 13 25
48 4

49 1 1.19 44.3 90 2.75 13
49 2 1.17 44.3 70 240.6 2.25 11 100
49 3 1.75 44.3 80 291.7 2.375 10 100
49 4 1.40 44.3 70 237.5 1.875 7
50 1 -.23 44.3 50 211.2 1.25 3
50 2 -.28 33.1 42 189.6 1.25 8 80
50 3 -.47 44.3 60 178.9 1.25 98
50 4 -.24 44.3 40 191.4 1.5
51 1 -.17 33.1 50 190.0 1.5 18
51 2 -.19 38.5 60 225.7 1.625 10 75
51 3 0.34 66.2 80 260.1 0.375 2 62
51 4
52 1 0.55 33.1 70 253.1 1 7
52 2 0.46 53.3 76 264.9 1 11 69
52 3 0.27 42 84 241.7 0.75 10 100
52 4 1.20 67 94 314.6 0.375 8
53 1 1.07 53.3 65 257.3 1.375 15
53 2 1.19 53.3 70 267.7 1.125 17 42
53 3 1.33 53.3 71 260.6 1 17 45
53 4 1.54 53.3 85 327.7 0.5 9
54 1 0.51 40.9 30 210.1 2.5 13
54 2 0.67 40.9 23 200.7 2.5 13 110
54 3 1.56 44.3 50 294.8 1.875 12
54 4 0.27 40.9 44 204.2 2.375 15
55 1 0.58 44.3 40 199.7 2.625 18
55 2 0.61 44.3 40 179.6 2.625 16 65
55 3 2.07 44.3 80 287.8 1.625 8 55
55 4 1.27 44.3 50 268.0 1.875 14
56 1 -.85 24.1 56 150.7 2.875 14
56 2 10.4 50 167.4 2.625 13 50
56 3 33.1 80 157.7 2.624 12 32
56 4 -1.3 24.1 50 158.5 2.75 13
57 1 0.58 44.3 60 267.9 2.375 12
57 2 0.72 54.6 49 288.7 1.875 17 40
57 3 0.58 54.6 60 289.6 1.25 9
57 4 0.89 44.3 86 297.9 1.75 9
58 1 33.1 30 151.3 2.375 10
58 2 -.53 33.1 60 177.8 2 6
58 3 -.71 33.1 50 132.7 2.25 0
58 4 -.76 22.6 37 134.8 2.25 7
59 1 1.10 42 60 227.5 1.375 5
59 2 0.85 53.3 70 230.7 1.125 3 65
59 3 1.95 67 90 325.6 0.625 3
59 4 2.74 67.2 95 371.6 0.625 2
60 1 0.27 22.6 40 200.0 2.5
60 2 0.73 44.3 45 210.4 1.875 9 51
60 3 1.04 53.3 65 304.2 1.625 2
60 4 0.71 54.6 70 256.5 1.5

61 1 0.81 44.3 30 250.4 2.125 8
61 2 0.29 44.3 30 227.8 2.5 1 100
61 3 0.49 44.3 70 243.8 2 2
61 4 0.11 44.3 40 251.0 2.25 8
62 1 0.05 100 63 286.8 1.875 10
62 2 0.11 53.3 64 249.0 1.5 9 30
62 3 0.09 56.7 67 238.8 1.5 0
62 4 0.09 56.7 61 250.0 1.25 8
63 1 0.29 53.3 60 220.8 1.625 17
63 2 0.35 42 60 222.5 1.125 15 5
63 3 0.13 42 50 208.5 0.875 6
63 4 0.57 42 70 285.3 1.5 8
64 1 -.15 8.9 30 228.3 2.125
64 2 0.08 29.6 45 216.2 2.125 14 50
64 3 0.17 44.3 85 222.8 0.75 3
64 4
65 1 51.1 70 2.625 6
65 2 -.03 44.3 60 225.9 2.125 1 71
65 3 0.56 54.6 60 211.9 1.625 2 65
65 4 2.5 5
66 1 -.19 0 30 212.1 3 11
66 2 0.15 21.7 30 166.4 3 14 71
66 3 0.29 44.3 70 261.5 2.75 6 42
66 4 6
67 1 2.92 44.3 301.9 0.875
67 2 3.10 44.3 313.8 1 13 36
67 3 3.19 44.3 328.9 0.875 4
67 4
68 1 53.3 45 0.999 6
68 2 0.84 60 240.8 1.875 8
68 3 56.7 60 1.75 8
68 4 1.25 2
69 1 43.3 30 133.5 1.75 3
69 2 50 1.5 5 55
69 3 2 4 65
69 4
70 1 0.21 33.1 40 168.6 2.25 7
70 2 1.75 9 25
70 3 1.25 10
70 4
71 1 1.875 7
71 2 1.875 4 51
71 3 1.75 5
71 4 1.25 7

Data set 2: Change in QoL and external criteria scores

Data set description:

Fields are space delimited

Field label	**Description**
Subject	Individual subject label
Qlp_ch	Quality of life profile score change, visit 4 minus visit 2
EuroQolvas_ch	EuroQol visual analogue scale score change, visit 4 minus visit 2
EuroQolrate_ch	EuroQol rating scale score change, visit 4 minus visit 2
Whoqol_ch	WHOQoL-Bref score change, visit 4 minus visit 2
Qollikert_ch	Code for QoL change based on a Likert scale rating change in QoL.
1	Significant improvement in QoL
0	No significant improvement in QoL

subject qlp_ch eurovas_ch euroqolrate_ch whoqol_ch qollikert_ch
1 2.97 28 53.2 72.32 1
2 0.23 -10 3.4 4.762 0
3 1.00 15 11.3 14.14 0
4 0.22 21 0 58.33 1
5 -.35 10 0 0.000 1
6 0.46 -33 0 31.10 0
7 -.31 3 1.488 0
8 0.62 30 21.7 23.36 0
9 -.17 18 11.2 12.50 0
10 0.88 45 55.7 133.5 1
11 1.80 40 33.9 51.79 1
12 0.09 0 0 12.80 1
13 0.83 45 22.7 70.98 1
14 -.03 50 0 20.68 1
15 -.19 9 46.7 -16.8 1
16 0
17 0.09 5 61.90 0
18 0
19 0.44 81.99 1
20 -.04 0 75.74 0
21 0.47 67 33.8 66.96 0
22 1.73 10 0 104.6 1
23 0.54 40 8.9 99.40 0
24 10 -10.5 -5.21 0
25 -.23 10 3.3 8.780 0
26 -.05 31 0 -15.2 0
27 0.19 45 25.2 100.0 1
28 -1.5 3 14.7 0
29 0.16 -1 0 12.65 0
30 2.02 40 54.3 149.4 1
31 -.05 20.2 -14.3 1
32 -.30 -5 3.4 1
33 0
34 1.69 32 43.4 107.3 1
35 -.05 9 3.5 -3.27 0
36 0.24 -20 0 -19.4 0
37 0.25 -10 10.3 -8.78 0
38 0.19 -15 8.9 52.98 0
39 -.23 -5 44.4 33.04 0
40 0.24 5 1 51.19 1
41 0.68 29 2.7 63.96 1
42 -.24 20 10.4 17.71 0
43 0.62 0 9 25.45 1
44 0.86 41 12.4 96.43 1
45 1.29 56 21.5 81.70 0
46 0
47 19 25 1
48 0
49 0.23 0 0 -3.13 0

50 0.04 -2 11.2 1.786 0
51 0
52 0.74 18 13.7 49.70 1
53 0.35 15 0 59.97 0
54 -.39 21 0 3.423 1
55 0.66 10 0 88.39 0
56 0 13.7 -8.93 1
57 0.17 37 -10.3 9.226 1
58 -.23 -23 -10.5 -43.0 0
59 1.90 25 13.9 140.9 1
60 -.02 25 10.3 46.13 1
61 -.18 10 0 23.21 0
62 -.02 -3 3.4 1.042 0
63 0.22 10 0 62.80 0
64 0
65 0
66 0
67 0
68 0
69 0
70 0
71 0

Appendix 3: Sample analysis programmes

1. SAS program for mixed linear model
2. SAS program for ROC curves
3. WinBUGS program for mixed linear model

SAS program for mixed linear model

```
1. proc mixed data=raqol.mixed;
2. class visit subject bfafter;
3. model qlp=haq bfafter/solution ddfm=satterth outp=raqol.resid;
4. id subject visit;
5. random int haq/subject=subject type=un solution;
6. repeated visit/subject=subject type=toeph;
7. ods output solutionr=raqol.temp;
8. run;
```

Description of programming statements.

1. Invokes the SAS procedure 'Mixed' and nominates the particular data set, in this case the first data set described in Appendix 2.
2. Declares the categorical variables, in this case 'visit' for the data collection point, 'subject' for the individual subject label, and 'bfafter' a dummy variable for whether a particular data collection point was before (visits 1 and 2) or after (visits 3 and 4) the inpatient intervention.
3. This describes the model to be fitted with 'qlp', the quality of life profile score, and the response variable, 'haq', the health assessment questionnaire score, and 'bfafter' as the explanatory variables. The options were 'solution' to output the numerical estimates of the effects of the explanatory variables, 'ddfm=satterth' to set Satterthwaites approximation as the method of calculating the degrees of freedom for statistical testing, and 'outp=raqol.resid' to output a data set containing the residuals for the model.
4. This includes the specified variables in the residuals data set.
5. This declares that the slope and intercept term for the explanatory variable, 'haq', is a random effect and distributed as multivariate normal. The option 'solution' gives an estimate of the overall slope and intercept term.

6. This declares the variance covariance structure of the random effects, in this case heterogeneous Toeplitz with the categorical variables 'subject' and 'visit' determining the block diagonal structure.
7. This outputs a data set containing predicted values for the slopes and intercept terms for the individual subjects
8. Runs the programming statements within SAS.

Output from this set of programming statements is described in Table 37. Residual plots based on the output data sets from the programming statements are shown in Figures 23 to 25.

SAS program for ROC curves

```
1. proc logistic data=raqol.rocdata;
2. model qollikert_ch=qlp_ch/outroc=raqol.roc;
3. run;
4. proc gplot data=raqol.roc;
5. axis1 length=25 label=(angle=90 'Sensitivity' h=1.5) minor=none;
6. axis2 length=80 label=('1 minus specificity' h=1.5) minor=none;
7. symbol1 i=join;
8. plot _sensit_*_1mspec_/
9. haxis=axis2
10. vaxis=axis1;
11. run;
12. quit;
```

Description of programming statements.

1. Invokes the SAS procedure 'Logistic' and nominates the particular data set, in this case the second data set described in Appendix 2.
2. Fits a logistic regression with the 'qollikert_ch', a variable which takes a value of 1 if the QoL changes based on the score on a simple Likert scale asking subjects to rate QoL, and 0 otherwise, as the response variable. The explanatory variable is the change in the particular QoL instrument, in this case the Quality of life profile, score. The 'outroc' option generates a data set from which a ROC curve can be generated.
3. Runs this first set of programming statements. The 'c' statistic in the output is the area under the curve for the receiver operating characteristic curve (AUC ROC).

4 to 12. Plots the sensitivity versus 1- specificity based on the output data set from the 'Logistic' procedure.

Output from these programming statements is described in Figure 13.

WinBUGS program for mixed linear model

```
1. model qlpvshaq;
2. const N = 50,
3. T= 4;
4. {for (i in 1:N) {
5. beta[i,1:2] ~ dmnorm(mu.beta[],R[,])
6. for (j in 1:T) {
7. Y[i , j] ~ dnorm(mu[i , j],tau.c)
8. mu[i, j] <- beta[i, 1] + beta[i, 2]*x[i, j]}}
9. mu.beta[1:2]~dmnorm(mean[],prec[,])
10. tau.c~dgamma(1.0E-3,1.0E-3);
11. sigma<-1.0/sqrt(tau.c);
12. R[1:2,1:2]~dwish(Omega[,], 2)}
```

Description of programming statements.

1. Declares that the following programming statements describe the model.
2. Declares constants for the model, the total number, 'N', of subjects with complete data.
3. Declares a further constant, the number, 'T', visits for the subjects.
4. Commences a labelling loop that for each individual subject, with subscript 'i'.
5. Declares that the variable beta[i,1] and beta[i,2], which represent the slope and intercept terms for the individual subjects for the relationship between the explanatory variable, x[i,j] and the response variable, Y[i,j], are distributed as multivariate normal, 'dmnorm', with an expected value vector of mu.beta and variance-covariance matrix R.
6. Commences a labelling loop for the data on each individual subject, with subscript 'i', and visit 'j'.
7. Declares a distribution for the individual measurements, in this case, univariate normal with a different expected value for each subject, mu[i,j] but a common precision, tau.c.

8. Describes the relationship between the expected value of the individual measurements on each subject at each visit of the QLP, Y[i,j], in terms of the random slope and intercept terms, variables beta[i,1] and beta[i,2], and the explanatory variable the external criterion measurement, in this case the HAQ, x[i,j].
9. Sets a prior distribution for mu.beta.
10. Sets a prior distribution for tau.c.
11. Generates a term, the inverse of the precision tau.c, that represents the variance for mu[i,j].
12. Sets a prior distribution for the R variance covariance matrix.

Output from these programming statements is described in table 40.

References

1. Gill TM, Feinstein AR. A critical appraisal of the quality of quality of life measurements. JAMA 1994;272:619-626

2. Schumaker M, Olschewski M, Schlugen G. Assessment of quality of life in clinical trials. Statist Med 1991;10:1915-1930

3. Testa MA, Nackley JF. Methods for quality of life studies. Annu Rev Public Health 1994;15:535-59

4. Wood-Dauphinee S. Assessing quality of life in clinical research: From where have we come and where are we going. J Clin Epidemiol 1999;52:355-363

5. Cox DR, Fitzpatrick R, Fletcher AE, Gore SM, Speigelhalter DJ, Jones DR. Quality of life assessment: Can we keep it simple? J R Statist Soc A 1992;155:353-393

6. Fitzpatrick R, Fletcher A, Gore S, Jones D, Spiegelhalter D, Cox D. Quality of life measures in health care I: Applications and issues in assessment. BMJ 1992;305:1074-

7. McDowell I, Newell C. Measuring Health: A guide to rating scales and questionnaires, second edition. Oxford University Press, New York, 1996

8. Fayers PM, Machin D. Quality of life: Assessment, analysis and interpretation. J Wiley and Sons, Chichester, 2000.

9. Coste J, Fermanian J, Venot A. Methodological and statistical problems in the construction of composite measurement scales: A survey of six medical and epidemiological journals. Statist Med 1995;14:331-345

10. Fletcher A. Quality of life measurements in the evaluation of treatment: Proposed guidelines. Br J Clin Pharmacol 1995;39:217-222

11. Aaronson NK. Quality of life assessment in clinical trials: Methodologic issues. Cont Clin Trial 1989:10:195S-208S.

12. Kirshner B, Guyatt G. A methodologic framework for assessing health indices. J Chron Dis 1985;38:27-36

13. Guyatt G, Kirshner B, Jaeschke R. Measuring health status: What are the necessary measurement properties. J Clin Epidemiol 1992;45:1341-1345

14. Guyatt GH, Feeny DH, Patrick DL. Measuring health related quality of life. Ann Int Med 1993;118:622-629

15. Fletcher A, Gore S, Jones D, Fitzpatrick R, Spiegelhalter D, Cox D. Quality of life measures in health care II: Design, analysis, and interpretation. BMJ 1992;305:1145-8

16. World Health Organisation 2001 web site: www.who.int/aboutwho/en/definition.html

17. World Health Organisation. Towards a common language for functioning and disablement: ICIDH-2 International classification of impairments, activities, and participation. World Health Organisation. Geneva, 1998 (available at WHO 2001 web site www.who.int/icidh)

18. Coste J, Walter E, Venot A. A new approach to selection and weighting of items in evaluative composite measurement scales. Statist Med 1995;14:2565-2580

19. Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. Biometrics 1987;43:487-498

20. Cook RJ, Farewell VT. Multiplicity considerations in the design and analysis of clinical trials. J R Statist Soc A 1996;159(Part 1)93-110

21. Fairclough DL. Summary measures and statistics for comparison of quality of life in a clinical trial of cancer therapy. Statist Med 1997;16:1197-1209

22. Sankoh AJ, Huque MF, Dubey SD. Some comments on frequently used multiple endpoint adjustment methods in clinical trials. Statist Med 1997;16:2529-2542

23. Gray SM, Brookmeyer R. Estimating a treatment effect from multidimensional longitudinal data. Biometrics 1998;54:976-988

24. Allison PJ, Locker D, Feine JS. Quality of life: A dynamic construct. Soc Sci Med 1997;45:221-230

25. Hobart JC, Lamping DL, Thompson AJ. Evaluating neurological outcome measures: the bare essentials. J Neurol Neurosurg Psych 1996;60:127-30

26. Agresti A. An introduction to categorical data analysis. Wiley New York, 1996.

27. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures: Statistics and strategies for evaluation. Cont Clin Trial 1991;12:142S-158S

28. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psychol Rep 1966;19:3-11

29. Shrout PE, Fleiss JL. Intra-class correlations: Uses in assessing rater reliability. Psychol Bull 1979;86:420-428

30. Fleiss JL. The design and analysis of clinical experiments. J Wiley and Sons, New York, 1986

31. Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: An illustration of appropriate statistical analyses. Clin Rehabil 1998;17:187-199

32. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: A critical review. J Clin Epidemiol 2000;53:459-468

33. Guyatt G, Walter S, Norman G. Measuring change over time: Assessing the usefulness of evaluative instruments. J Chron Dis 1987;40:171-178

34. Guyatt GH, Deyo RA, Charlson M, Levine MN, Mitchell A. Responsiveness and validity in health status measurement: A clarification. J Clin Epidemiol 1989;42:403-408

35. Guyatt GH, Eagle DJ, Sackett B, Willan A, Griffith L, McIlory W, Patterson CJ, Turpie I. Measuring quality of life in the frail elderly. J Clin Epidemiol 1993;46:1433-1444

36. Tuley MR, Mulrow CD, McMahan CA. Estimating and testing an index of responsiveness and the relationship of the index to power. J Clin Epidemiol 1991;44:417-421.

37. Liang MH, Larson MG, Cullen KE, Schwartz JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arth Rheum 1985;28:542-547

38. Beaton DE, Hogg- Johnson S, Bombardier C. Evaluating changes in health status: Reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. J Clin Epidemiol 1997;50:79-93

39. Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A, Mowat A. A comparison of sensitivity to change of several health status instruments in rheumatoid arthritis. J Rheumatol 1993;20:429-436

40. Wright JG, Young NL. A comparison of different indices of responsiveness. J Clin Epidemiol 1997;50:239-246

41. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: The lesson of Cronbach. J Clin Epidemiol 1997;50:869-879

42. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: Ascertaining the minimal clinically important difference. Cont Clin Trial 1989;10:407-415

43. Cohen J. Statistical power analysis for the behavioural sciences 2nd edition. Lawrence Erlbaum Associates, Hillsdale New Jersey, 1988.

44. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care 1989;27:S178-S189

45. Murawski MM, Mederhoff PA. On the generalizability of statistical expressions of health related quality of life instrument responsiveness: A data synthesis. Quality of Life Research 1998;7:11-22

46. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: An analogy to diagnostic test performance. J Chron Dis 1986;39:897-906

47. Erdreich LS, Lee ET. Use of relative operating characteristic analysis in epidemiology. Am J Epidemiol 1981;114:649-662

48. Beck JR, Shultz EK. The use of relative operating characteristic (ROC) curves in test performance evaluation. Arch Pathol Lab Med 1986;110:13-20

49. Swets JA. Measuring the accuracy of diagnostic systems. Science 1988;240:1285-1293

50. Begg CB. Advances in statistical methodology for diagnostic medicine in the 1980's. Statist Med 1991;10:1887-1895

51. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: A fundamental tool in clinical medicine. Clin Chem 1993;39:561-577

52. Campbell G. Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. Statist Med 1994;13:499-508

53. Pepe MS. Receiver operating characteristic methodology. J Am Stat Ass 2000;95:308-311

54. Hanley JA, McNeill BJ. The meaning and use of the area under the receiver operating characteristic curve. Radiology 1982;143:29-36

55. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J Math Psychol 1975;12:387-415

56. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A non-parametric approach. Biometrics 1988;44:837-845

57. Dorfman DD. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals. J Math Psychol 1969;6:487-496

58. Metz CE, Kronman HB. Statistical significance tests for binormal ROC curves. J Math Psychol 1980;22:218-243

59. Metz CE, Herman BA, Shen JH. Maximum likelihood estimation of receiver operating characteristic curves from continuously-distributed data. Statist Med 1998;17:1033-1053

60. Metz CE, Pan X. Proper binormal ROC curves: Theory and maximum likelihood estimation. J Math Psychol 1999;43:1-33

61. Thompson ML, Zucchini W. On the statistical analysis of ROC curves. Statist Med 1989;8:1277-1290

62. Stucki G, Liang MH, Fossel AH, Katz JN. Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. J Clin Epidemiol 1995;48:1369-78

63. Beurskens AJHM, de Wet HCW, Koke AJA. Responsiveness of functional status in low back pain: A comparison of different instruments. Pain 1996;65:71-76

64. Van Bennekom CAM, Jelles F, Lankhorst GJ, Bouter LM. Responsiveness of the rehabilitation activities profile and the Barthel index. J Clin Epidemiol 1996;49:39-44

65. Bronfort G, Bouter LM. Responsiveness of general health status in chronic low back pain: A comparison of the COOP charts and the SF-36. Pain 1999;83:201-209

66. Vliet Vlieland TPM, Zwinderman AH, Breedveld FC, Hazes JMW. Measurement of morning stiffness in rheumatoid arthritis clinical trials. J Clin Epidemiol 1997;50:757-763

67. Coffine M, Sukhatme S. Receiver operating characteristic studies and measurement errors. Biometrics 1997;53:823-837

68. Faraggi D. The effect of random measurement error on receiver operating characteristic (ROC) curves. Statist Med 2000;19:61-70

69. Reiser B. Measuring the effectiveness of diagnostic markers in the presence of measurement error through the use of ROC curves. Statist Med 2000;19:2115-2129

70. Goddard MJ, Hinberg I. Receiver operator characteristic (ROC) curves and non-normal data: An empirical study. Statist Med 1990;9:325-337

71. Le CT, Lindgren BR. Construction and comparison of two receiver operating characteristic curves derived from the same samples. Biom J 1995;37:869-877

72. Lee WC, Hsiao CK. Alternative summary indices for the receiver operating characteristic curve. Epidemiology 1996;7:605-611

73. Lee WC. Probabilistic analysis of global performances of diagnostic tests: Interpreting the Lorenz curve based summary measures. Statist Med 1999;18:455-471

74. Meenan RF, Anderson JJ, Kazis Le, Egger MJ, Altz-Smith M, Samuelson CO et al. Outcome assessment in clinical trials: Evidence for the sensitivity of a health status measure. Arth Rheum 1984;27:1344-1352

75. Bain LJ, Engelhardt M. Introduction to probability and mathematical statistics, 2nd edition. PWS-Kent 1991, Boston.

76. May K, Hittner JB. A note on statistics for comparing dependent correlations. Psychol Rep 1997;80:475-80

77. Meng XL, Rosenthal R, Rubin DB. Comparing correlated correlation coefficients. Psychol Bull 1992;111:172-175

78. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;i:307-310

79. Husted JA, Gladman DD, Long JA, Farewell VT. Relationship of the arthritis impact measurement scales to changes in articular status and functional performance in patients with psoriatic arthritis. J Rheumatol 1996;23:1932-1937

80. Husted JA, Gladman DD, Cook RJ, Farewell VT. Responsiveness of health status instruments to changes in articular status and perceived health in patients with psoriatic arthritis. J Rheumatol 1998;25:2146-2155

81. Hocking RR. Methods and applications of linear models: Regression and analysis of variance. J Wiley and Sons, New York, 1996

82. Gibbons JD, Chakraborti S. Nonparametric statistical inference. 3rd edition. Marcel Decker, New York. 1992

83. Cnaan A, Laird NM, Slasor P. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. Statist Med 1997;16:2349-2380

84. Verbeke G, Molenberghs G. Linear mixed models in practice: A SAS orientated approach. NY, USA, Springer-Verlag, 1997.

85. Brown H, Prescott R. Applied mixed linear models in medicine. John Wiley and Sons 1999. Chichester.

86. Khattree R, Naik DN. Applied multivariate statistics with SAS software, 2nd edition, Cary,NC:SAS Institute Inc., 1999.

87. Littell RC, Pendergast J, Natarajan R. Modelling covariance structure in the analysis of repeated measures data. Statist Med 2000;19:1793-1819

88. Gilks WR, Clayton DG, Spiegelhalter DJ, Best NG, McNeil AJ, Sharples LD, Kirby AJ. Modelling complexity: Applications of Gibbs sampling in medicine. J R Statist Soc B 1993;55:39-52

89. Smith AFM, Roberts GO. Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo methods. J R Statist Soc B 1993;55:3-23

90. Best NG, Spiegelhalter DJ, Thomas A, Brayne CEG. Bayesian analysis of realistically complex models. J R Statist Soc A;1996:159 Part 2:323-342

91. Carlin BP. Hierarchical longitudinal modelling. In Markov Chain Monte Carlo in practice. Gilks WR, Richardson S, Spiegelhalter DJ eds. Chapman Hall 1996. London. pp 305-319

92. Clayton DG. Generalized linear mixed models. In Markov Chain Monte Carlo in practice. Gilks WR, Richardson S, Spiegelhalter DJ eds. Chapman Hall 1996. London. pp 275-301

93. Casella G, George EI. Explaining the Gibbs sampler. American Statistician 1992;46:167-174

94. Rafferty AE. Bayesian model selection in social research. Sociological Methodology 1995;25:111-163

95. Gelfland AE. Model determination using sampling based methods. In Markov Chain Monte Carlo in practice. Gilks WR, Richardson S, Spiegelhalter DJ eds. Chapman Hall 1996. London. pp 145-161

96. Spiegelhalter D, Thomas A, Best N, Gilks W. BUGS 0.5 Bayesian inference using Gibbs sampling manual (version ii). 1996 (available at the BUGS web site www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml)

97. Puntanen S, Styan GPH. The equality of the ordinary least squares estimator and the best linear unbiased estimator. American Statistician 1989;43:153-61

98. Beacon HJ, Thompson SG. Multi-level models for repeated measurement data: Application to quality of life data in clinical trials. Statist Med 1996;15:2717-2732

99. Semble EL. Rheumatoid arthritis: New approaches for its evaluation and management. Arch Phys Med Rehabil 1995;76:190-201

100. EuroQol group. EuroQol: a new facility for the measurement of health related quality of life. Health Policy 1990;16:199-208

101. Harper A, Power M. Development of the World Health Organisation WHOQOL-BREF quality of life assessment. Psychological Medicine 1998;28:551-558

102. Raphael D, Brown I, Renwick R, Cava M, Weir N, Heathcote K. The quality of life of seniors living in the community: A conceptualization with implications for public health practice. Can J Pub Health 1995;86:228-233

103. Raphael D, Rukholm E, Brown I, Hill-Bailey P, Donato E. The quality of life profile-Adolescent version: Background, description, and initial validation. J Adolescent Health 1996;19:366-375

104. Ritchie DM, Boyle JA, McInnes JM, Jasani MK, Dalakos TG, Grieveson P, Buchanan WW. Clinical studies with an articular index for the assessment of joint tenderness in patients with rheumatoid arthritis. Quart J Med 1968;37:393-406

105. Fries JF, Spitz P, Kraines G, Holman HR. Measurement of patient outcome in arthritis. Arth Rheum 1980;23:137-145

106. Pincus T, Summey JA, Soraci SA, Wallston KA, Hummon NP. Assessment of patient satisfaction in activities of daily living using a modified Stanford health assessment questionnaire. Arth Rheum 1983;26:1346-1353

107. Bombardier C, Raboud J. A comparison of health related quality-of-life measures for rheumatoid arthritis research. Con Clin Trials 1991;12:243S-256S

108. Wells GA, Tugwell P, Kraag GR, Baker PRA, Groh J, Reidlmeier DA. Minimum important difference between patients with rheumatoid arthritis: The patients perspective. J Rheumatol 1993;20:557-60

109. Buchbinder R, Bombardier C, Yeung M, Tugwell P. Which outcome measurements should be used in rheumatoid arthritis trials. Arth Rheum 1995;38:1568-1580

110. Kosinski M. Zhao SZ, Dedhiya S, Osterhaus JT, Ware JE. Determining the minimally important changes in generic and disease-specific health-related quality of life questionnaires in clinical trials of rheumatoid arthritis. Arth Rheum 2000;43:1478-1487

111. Vliet Vlieland TPM, Zwinderman AH, Vanendbroucke JP, Breedveld FC, Hazes JMW. A randomized clinical trial of inpatient multidisciplinary treatment versus routine outpatient care in active rheumatoid arthritis. Br J Rheumatol 1996;35:475-482

112. SAS Institute 2001 web site www.SAS.com

113. WinBUGS 2001 web site: http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml