Quality of Life Outcomes in Clinical Trials and Health-Care Evaluation

A Practical Guide to Analysis and Interpretation

Stephen J. Walters

School of Health and Related Research, University of Sheffield, UK
Quality of Life Outcomes in Clinical Trials and Health-Care Evaluation
STATISTICS IN PRACTICE

Advisory Editor

Stephen Senn
University of Glasgow, UK

Founding Editor

Vic Barnett
Nottingham Trent University, UK

Statistics in Practice is an important international series of texts which provide detailed coverage of statistical concepts, methods and worked case studies in specific fields of investigation and study.

With sound motivation and many worked practical examples, the books show in down-to-earth terms how to select and use an appropriate range of statistical techniques in a particular practical field within each title’s special topic area.

The books provide statistical support for professionals and research workers across a range of employment fields and research environments. Subject areas covered include medicine and pharmaceutics; industry, finance and commerce; public services; the earth and environmental sciences, and so on.

The books also provide support to students studying statistical courses applied to the above areas. The demand for graduates to be equipped for the work environment has led to such courses becoming increasingly prevalent at universities and colleges.

It is our aim to present judiciously chosen and well-written workbooks to meet everyday practical needs. Feedback of views from readers will be most valuable to monitor the success of this aim.

A complete list of titles in this series appears at the end of the volume.
Quality of Life Outcomes in Clinical Trials and Health-Care Evaluation

A Practical Guide to Analysis and Interpretation

Stephen J. Walters

School of Health and Related Research, University of Sheffield, UK
Contents

Preface xi

1 Introduction 1
Summary 1
1.1 What is quality of life? 1
1.2 Terminology 2
1.3 History 2
1.4 Types of quality of life measures 4
1.5 Why measure quality of life? 10
1.6 Further reading 11

2 Measuring quality of life 13
Summary 13
2.1 Introduction 13
2.2 Principles of measurement scales 13
  2.2.1 Scales and items 13
  2.2.2 Constructs and latent variables 14
2.3 Indicator and causal variables 15
  2.3.1 Indicator variables 15
  2.3.2 Causal variables 15
  2.3.3 Why do we need to worry about the distinction between indicator and causal items? 16
  2.3.4 Single-item versus multi-item scales 16
2.4 The traditional psychometric model 16
  2.4.1 Psychometrics and QoL scales 17
2.5 Item response theory 17
  2.5.1 Traditional scales versus IRT 18
2.6 Clinimetric scales 18
2.7 Measuring quality of life: Indicator or causal items 19
2.8 Developing and testing questionnaires 19
  2.8.1 Specify the research question and define the target population 19
  2.8.2 Identify concepts 20
3 Choosing a quality of life measure for your study

Summary 31
3.1 Introduction 31
3.2 How to choose between instruments 31
3.3 Appropriateness 33
3.4 Acceptability 33
3.5 Feasibility 34
3.6 Validity 35
3.6.1 Tests for criterion validity 35
3.6.2 Tests for face and content validity 36
3.6.3 Tests for construct validity 36
3.7 Reliability 38
3.7.1 Repeatability reliability 40
3.7.2 Graphical methods for assessing reliability between two repeated measurements 40
3.7.3 Internal reliability or internal consistency reliability 42
3.8 Responsiveness 44
3.8.1 Floor and ceiling effects 44
3.9 Precision 49
3.10 Interpretability 51
3.11 Finding quality of life instruments 53

4 Design and sample size issues: How many subjects do I need for my study? 55

Summary 55
4.1 Introduction 55
4.2 Significance tests, \( P \)-values and power 56
4.3 Sample sizes for comparison of two independent groups 58
4.3.1 Normally distributed continuous data – comparing two means 58
4.3.2 Transformations 61
4.3.3 Comparing two groups with continuous data using non-parametric methods 61
4.3.4 Dichotomous categorical data – comparing two proportions 63
4.3.5 Ordered categorical (ordinal) data 66
4.4 Choice of sample size method with quality of life outcomes 69
4.5 Paired data 70
4.5.1 Paired continuous data – comparison of means 70
4.5.2 Paired binary data – comparison of proportions 72
4.6 Equivalence/non-inferiority studies 73
4.6.1 Continuous data – comparing the equivalence of two means 74
4.6.2 Binary data – comparing the equivalence of two proportions 75
4.7 Unknown standard deviation and effect size 75
4.7.1 Tips on obtaining the standard deviation 76
4.8 Cluster randomized controlled trials 76
4.9 Non-response 77
4.10 Unequal groups 77
4.11 Multiple outcomes/endpoints 79
4.12 Three or more groups 80
4.13 What if we are doing a survey, not a clinical trial? 80
4.13.1 Sample sizes for surveys 80
4.13.2 Confidence intervals for estimating the mean QoL of a population 81
4.13.3 Confidence intervals for a proportion 82
4.14 Sample sizes for reliability and method comparison studies 85
4.15 Post-hoc sample size calculations 86
4.16 Conclusion: Usefulness of sample size calculations 86
4.17 Further reading 86

5 Reliability and method comparison studies for quality of life measurements 91
Summary 91
5.1 Introduction 91
5.2 Intra-class correlation coefficient 92
5.2.1 Inappropriate method 94
5.3 Agreement between individual items on a quality of life questionnaire 95
5.3.1 Binary data: Proportion of agreement 95
5.3.2 Binary data: Kappa 95
5.3.3 Ordered categorical data: Weighted kappa 96
5.4 Internal consistency and Cronbach’s alpha 98
5.5 Graphical methods for assessing reliability or agreement between two quality of life measures or assessments 99
5.6 Further reading 102
5.7 Technical details 102
5.7.1 Calculation of ICC 102
5.7.2 Calculation of kappa 103
5.7.3 Calculation of weighted kappa 104
5.7.4 Calculation of Cronbach’s alpha 104

6 Summarizing, tabulating and graphically displaying quality of life outcomes 109
Summary 109
6.1 Introduction 109
6.2 Graphs 110
6.2.1 Dot plots 112
6.2.2 Histograms 112
6.2.3 Box-and-whisker plot 114
6.2.4 Scatter plots 116
6.3 Describing and summarizing quality of life data 116
6.3.1 Measures of location 117
6.3.2 Measures of spread 119
6.4 Presenting quality of life data and results in tables and graphs 122
6.4.1 Tables for summarizing QoL outcomes 122
6.4.2 Tables for multiple outcome measures 124
6.4.3 Tables and graphs for comparing two groups 126
6.4.4 Profile graphs 129

7 Cross-sectional analysis of quality of life outcomes 133
Summary 133
7.1 Introduction 133
7.2 Hypothesis testing (using $P$-values) 134
7.3 Estimation (using confidence intervals) 137
7.4 Choosing the statistical method 138
7.5 Comparison of two independent groups 138
7.5.1 Independent samples $t$-test for continuous outcome data 140
7.5.2 Mann–Whitney $U$-test 144
7.6 Comparing more than two groups 146
7.6.1 One-way analysis of variance 147
7.6.2 The Kruskal–Wallis test 150
7.7 Two groups of paired observations 150
7.7.1 Paired $t$-test 153
7.7.2 Wilcoxon test 157
7.8 The relationship between two continuous variables 157
7.9 Correlation 160
7.10 Regression 165
7.11 Multiple regression 168
7.12 Regression or correlation? 171
7.13 Parametric versus non-parametric methods 171
7.14 Technical details: Checking the assumptions for a linear regression analysis 173

8 Randomized controlled trials 181
Summary 181
8.1 Introduction 181
8.2 Randomized controlled trials 182
8.3 Protocols 182
8.4 Pragmatic and explanatory trials 182
8.5 Intention-to-treat and per-protocol analyses 183
8.6 Patient flow diagram 186
8.7 Comparison of entry characteristics 186
8.8 Incomplete data 189
8.9 Main analysis 191
8.10 Interpretation of changes/differences in quality of life scores 196
8.11 Superiority and equivalence trials 197
8.12 Adjusting for other variables 199
8.13 Three methods of analysis for pre-test/post-test control group designs 202
8.14 Cross-over trials 203
8.15 Factorial trials 206
## 8.16 Cluster randomized controlled trials
8.17 Further reading

### 9 Exploring and modelling longitudinal quality of life data

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summary</td>
<td>217</td>
</tr>
<tr>
<td>9.1 Introduction</td>
<td>217</td>
</tr>
<tr>
<td>9.2 Summarizing, tabulating and graphically displaying repeated QoL assessments</td>
<td>218</td>
</tr>
<tr>
<td>9.3 Time-by-time analysis</td>
<td>222</td>
</tr>
<tr>
<td>9.4 Response feature analysis – the use of summary measures</td>
<td>223</td>
</tr>
<tr>
<td>9.4.1 Area under the curve</td>
<td>223</td>
</tr>
<tr>
<td>9.4.2 Acupuncture study – analysis of covariance</td>
<td>227</td>
</tr>
<tr>
<td>9.5 Modelling of longitudinal data</td>
<td>229</td>
</tr>
<tr>
<td>9.5.1 Autocorrelation</td>
<td>231</td>
</tr>
<tr>
<td>9.5.2 Repeated measures analysis of variance</td>
<td>232</td>
</tr>
<tr>
<td>9.5.3 Marginal general linear models – generalized estimating equations</td>
<td>232</td>
</tr>
<tr>
<td>9.5.4 Random effects models</td>
<td>237</td>
</tr>
<tr>
<td>9.5.5 Random effects versus marginal modelling</td>
<td>239</td>
</tr>
<tr>
<td>9.5.6 Use of marginal and random effects models to analyse data from a cluster RCT</td>
<td>241</td>
</tr>
<tr>
<td>9.6 Conclusions</td>
<td>243</td>
</tr>
</tbody>
</table>

### 10 Advanced methods for analysing quality of life outcomes

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summary</td>
<td>249</td>
</tr>
<tr>
<td>10.1 Introduction</td>
<td>249</td>
</tr>
<tr>
<td>10.2 Bootstrap methods</td>
<td>251</td>
</tr>
<tr>
<td>10.3 Bootstrap methods for confidence interval estimation</td>
<td>251</td>
</tr>
<tr>
<td>10.4 Ordinal regression</td>
<td>255</td>
</tr>
<tr>
<td>10.5 Comparing two independent groups: Ordinal quality of life measures (with less than 7 categories)</td>
<td>257</td>
</tr>
<tr>
<td>10.6 Proportional odds or cumulative logit model</td>
<td>258</td>
</tr>
<tr>
<td>10.7 Continuation ratio model</td>
<td>259</td>
</tr>
<tr>
<td>10.8 Stereotype logistic model</td>
<td>260</td>
</tr>
<tr>
<td>10.9 Conclusions and further reading</td>
<td>264</td>
</tr>
</tbody>
</table>

### 11 Economic evaluations

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summary</td>
<td>265</td>
</tr>
<tr>
<td>11.1 Introduction</td>
<td>265</td>
</tr>
<tr>
<td>11.2 Economic evaluations</td>
<td>266</td>
</tr>
<tr>
<td>11.3 Utilities and QALYs</td>
<td>266</td>
</tr>
<tr>
<td>11.4 Economic evaluations alongside a controlled trial</td>
<td>267</td>
</tr>
<tr>
<td>11.5 Cost-effectiveness analysis</td>
<td>267</td>
</tr>
<tr>
<td>11.6 Cost–effectiveness ratios</td>
<td>268</td>
</tr>
<tr>
<td>11.7 Cost–utility analysis and cost–utility ratios</td>
<td>269</td>
</tr>
<tr>
<td>11.8 Incremental cost per QALY</td>
<td>270</td>
</tr>
</tbody>
</table>
Preface

Quality of life (QoL) outcomes or person/patient reported outcome measures (PROMs) are now frequently being used in randomized controlled trials (RCTs) and observational studies. This book aims to be a practical guide to the design, analysis and interpretation of studies that use such outcomes. Since there are numerous QoL instruments now available, it emphasizes that, for busy and time-constrained researchers, it is easier to use an ‘off-the-shelf’ QoL instrument than to design your own. This book gives practical guidance on how to choose between the various instruments.

QoL outcomes tend to generate data with discrete, bounded and skewed distributions. Hence, many investigators are concerned about the appropriateness of using standard statistical methods to analyse QoL data and want guidance on what methods to use. This book provides such practical guidance, based on the author’s extensive experience. Other texts, on the analysis of QoL outcomes, concentrate mainly on clinical trials and ignore other frequently used study designs such as cross-sectional surveys and non-randomized health-care evaluations. Again this book rectifies this and provides practical guidance on the analysis of QoL outcomes from such observational designs. It presents simple conventional methods to tackle these problems (such as linear regression), before addressing more advanced approaches, including ordinal regression and computer-intensive methods (such as the bootstrap).

The book is illustrated throughout with real-life case studies and worked examples from RCTs and other observational studies, taken from the author’s own experience of designing and analysing studies with QoL outcomes. Each analysis technique is carefully explained and the mathematics, as far as possible, is kept to a minimum. Hopefully, it is written in a style suitable for statisticians and clinicians alike!

The practical guidance provided by this book will be of use to professionals working in and/or managing clinical trials, in academic, government and industrial settings, particularly medical statisticians, clinicians and trial co-ordinators. Its practical approach will appeal to applied statisticians and biomedical researchers, in particular those in the biopharmaceutical industry, medical and public health organizations. Graduate students of medical statistics will also find much of benefit, as will graduate students of the medical and health sciences who have to analyse QoL data for their dissertations and projects.

Most of the book is written at an intermediate level for readers who are going to collect and analyse their own QoL data. It is expected that readers will be familiar with basic statistical concepts such as hypothesis testing (P-values), confidence intervals, simple
statistical tests (e.g. the $t$-test and chi-square test) and simple linear regression. The more advanced topics, in the later chapters, such as marginal generalized linear models for longitudinal data, will require a more thorough statistical knowledge, but are explained in as simple a way as possible with examples.

Stephen J. Walters
Sheffield, UK
1

Introduction

Summary

Quality of life (QoL) is a complex concept with multiple dimensions. This book will assume a wide definition for this concept. It will describe the design, assessment, analysis and interpretation of single- and multi-item, subjective measurement scales. These measurement scales all have the common feature of using a standardized approach to assessing a person’s perception of their own health by using numerical scoring systems, and may include one or several dimensions of QoL. This chapter will provide a brief history of QoL assessment; describe the different types of QoL assessment tools available and give reasons why it is important to measure QoL.

1.1 What is quality of life?

Quality of life (QoL) is a complex concept with multiple aspects. These aspects (usually referred to as domains or dimensions) can include: cognitive functioning; emotional functioning; psychological well-being; general health; physical functioning; physical symptoms and toxicity; role functioning; sexual functioning; social well-being and functioning; and spiritual/existential issues (see Figure 1.1). This book will assume a wide definition for this concept. It will describe the design, assessment, analysis and interpretation of single- and multi-item, subjective measurement scales. This broad definition will include scales or instruments that ask general questions, such as ‘In general, how would you rate your health now?’, and more specific questions on particular symptoms and side effects, such as ‘During the past week have you felt nauseated?’ These measurement scales all have the common feature of using a standardized approach to assessing a person’s perception of their own health by using numerical scoring systems, and may include one or several dimensions of QoL.
1.2 Terminology

Researchers have used a variety of names to describe QoL measurement scales. Some prefer to use the term *health-related quality of life* (HRQoL or HRQL), to stress that we are only concerned with health aspects. Others have used the terms *health status* and *self-reported health*. The United States Food and Drug Administration (FDA) has adopted the term *patient-reported outcome* (PRO) in its guidance to the pharmaceutical industry for supporting labelling claims for medical product development (FDA, 2006). However, not all people who complete such outcomes are ill and patients, and hence PRO could legitimately stand for *person-reported outcome*. Mostly, we shall assume that the QoL instrument or outcome is self-reported, by the person whose experience we are interested in, but it could be completed by another person or proxy. The term *health outcome assessment* has been put forward as an alternative which avoids specifying the respondent. This book will follow convention and use the now well-established term *quality of life*.

1.3 History

The World Health Organisation (WHO, 1948) declared health to be ‘A state of complete physical and mental social well-being, and not merely the absence of disease and infirmity’. This definition was one of the first to emphasize other facets of health, such as physical, mental and social, in connection with disease and infirmity.

The Karnofsky Performance Scale (Karnofsky and Burchenal, 1949) was one of the first instruments to undertake a wider assessment of patients’ functional impairment apart from clinical and physiological examination. It involves health-care staff assessing patients, using a simple single-item 11-point scale ranging from 0 for ‘dead’ to 100.
Table 1.1  The Karnofsky Performance Scale.

<table>
<thead>
<tr>
<th>Description</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal; no complaints; no evidence of disease</td>
<td>100</td>
</tr>
<tr>
<td>Able to carry on normal activity; minor signs and symptoms of disease</td>
<td>90</td>
</tr>
<tr>
<td>Normal activity with effort; some signs and symptoms of disease</td>
<td>80</td>
</tr>
<tr>
<td>Cares for self; unable to carry on normal activity or do work</td>
<td>70</td>
</tr>
<tr>
<td>Requires occasional assistance, but is able to care for most personal needs</td>
<td>60</td>
</tr>
<tr>
<td>Requires considerable assistance and frequent medical care</td>
<td>50</td>
</tr>
<tr>
<td>Disabled; requires special care and assistance</td>
<td>40</td>
</tr>
<tr>
<td>Severely disabled; hospitalization indicated although death not imminent</td>
<td>30</td>
</tr>
<tr>
<td>Very sick; hospitalization necessary; requires active support treatment</td>
<td>20</td>
</tr>
<tr>
<td>Moribund; fatal processes progressing rapidly</td>
<td>10</td>
</tr>
<tr>
<td>Dead</td>
<td>0</td>
</tr>
</tbody>
</table>

for ‘Normal’ (see Table 1.1). It can be used to compare effectiveness of different therapies and to assess the prognosis in individual patients.

This led to the development of the next generation of questionnaires which focused on broader aspects of QoL, such as emotional well-being, social functioning, impact of illness, perceived distress and life satisfaction. These included the Nottingham Health Profile (NHP, Hunt et al., 1980, 1981) and the Sickness Impact Profile (SIP, Deyo et al., 1982). Again, I shall describe the NHP and SIP as QoL scales although their developers neither designed them nor claimed them as QoL scales.

Newer instruments such as the Medical Outcomes Study (MOS) Short Form (SF)-36 (Ware and Sherbourne, 1992) now place more emphasis on the subjective aspects of QoL, such as emotional, role, social and cognitive functioning. The SF-36 is the most commonly used QoL measure in the world today. It contains 36 questions measuring health across eight dimensions: Physical Functioning (PF); Role-Physical (role limitations due to physical health, RP); Social Functioning (SF); Vitality (VT); Bodily Pain (BP); Mental Health (MH); Role-Emotional (role limitations due to emotional problems, RE); and General Health (GH).

Quality of life was introduced by the MEDLINE (Medical Literature Analysis and Retrieval System Online) international literature database of life sciences and biomedical information as a heading in 1975, and accepted as a concept by Index Medicus in 1977. Since then there has been a rapid expansion of interest in the topic, with an exponential increase in the number of citations of QoL in the medical literature (see Figure 1.2).

In 1991, the first edition of a new international, multidisciplinary journal devoted to the rapid communication of original research, theoretical articles and methodological reports related to the field of QoL in all the health sciences was published, entitled Quality of Life Research. The February 2004 issue was largely devoted to the publication of abstracts from the first meeting of the International Society for Quality of Life Research (ISOQOL), held in Brussels. ISOQOL’s mission is the scientific study of QoL relevant to health and health care. The Society promotes the rigorous investigation of health-related QoL measurement from conceptualization to application and practice. ISOQOL fosters the worldwide exchange of information through scientific publications, international conferences, educational outreach, and collaborative support for QoL initiatives.
1.4 Types of quality of life measures

The SF-36 is an example of a QoL instrument that is intended for general use, irrespective of the illness or condition of the patient. Such instruments are often termed generic measures and may often be applicable to healthy people too and hence used in population surveys. Figure 1.3 shows the distribution of the eight main dimensions of the SF-36 from a general population survey of United Kingdom residents (Brazier et al., 1992). The SF-36 dimensions are scored on a 0 to 100 (‘good health’) scale. Figure 1.3 shows that the SF-36 outcome, in common with many other QoL scales, generates data with a discrete, bounded and skewed distribution. Figure 1.4 shows how physical functioning in the general population (Walters et al., 2001a) declines rapidly with increasing age.

The SF-36 is also an example of a profile QoL measure since it generates eight separate scores for each dimension of health (Figure 1.3). Other generic profile instruments include the SIP and NHP (see Section 1.3). Conversely, some other QoL measures generate a single summary score or single index, which combines the different dimensions of health into a single number. An example of a single index QoL outcome is the EuroQol or EQ-5D as it is now named (EuroQol Group, 1990).

Generic instruments are intended to cover a wide range of conditions and have the advantage that the scores from patients with various diseases may be compared against each other and against the general population. For example, Figure 1.5 compares the mean SF-36 dimension scores of a group of patients six months after acute myocardial infarction (AMI) with an age and sex matched general population sample (Lacey and Walters, 2003). The AMI sample has lower QoL on all eight dimensions of the SF-36 than the general population sample. On the other hand, generic instruments may fail to focus on the issues of particular concern to patients with disease, and may often lack the sensitivity to detect differences that arise as a consequence of treatments that are compared in clinical trials. This has led to the development of condition- or disease-specific questionnaires. Disease-specific QoL measurement scales are comprehensively reviewed by Bowling (2001, 2004). Examples of disease-specific QoL questionnaires include the
Figure 1.3  Distribution of the eight SF-36 dimensions from a general population survey ($n = 1372$); a score of 100 indicates ‘good health’ (data from Brazier et al., 1992).
Figure 1.3  (Continued)
Figure 1.3 (Continued)
Figure 1.3 (Continued)
Figure 1.4  Mean SF-36 Physical Functioning age profile by sex (data from Walters et al., 2001a).

Figure 1.5  Profile of mean SF-36 scores for an acute myocardial infarction sample (six weeks after infarction) compared with an age and sex matched general population sample (data from Lacey and Walters, 2003).

cancer-specific 30-item European Organisation for Research and Treatment of Cancer (EORTC) QLC-30 questionnaire (Aaronson et al., 1993) and the cancer-specific 30-item Rotterdam Symptom Checklist (RSCL, de Haes et al., 1990).

The instruments described above claim to measure general QoL, and usually include at least one question about overall QoL or health. Sometimes investigators may wish to explore particular aspects or concepts in greater depth. There are also instruments
for specific aspects of QoL. These specific aspects may include anxiety and depression, physical functioning, pain and fatigue. Examples of instruments which evaluate specific aspects of QoL are: the Hospital Anxiety and Depression Scale (HADS, Zigmond and Snaith, 1983) and the Beck Depression Inventory (BDI, Beck et al., 1961) instruments for measuring anxiety and depression; the McGill Pain Questionnaire (MPQ, Melzack, 1975) for the measurement of pain; the Multidimensional Fatigue Inventory (MFI, Smets et al., 1995) for assessing fatigue and the Barthel Index (Mahoney and Barthel, 1965) for assessing disability and functioning.

1.5 Why measure quality of life?

There are several reasons why we should measure quality of life in both a research setting and in routine clinical practice. The use of QoL assessment in routine clinical practice may make communication with patients easier and help find out information about the range of problems that affect patients. Medicine and health care have traditionally tended to focus on symptom relief as the main outcome measure. QoL assessment may help improve symptom relief, care or rehabilitation for an individual patient. Using QoL instruments may reveal other issues that are equally or more important to patients than just symptom relief. The patient’s self-assessment of their own QoL may differ substantially from the judgement of other health-care staff. Individual patient preferences may also differ from those of other patients. Therefore it is important to measure QoL from the patient’s perspective, using a self-completed questionnaire to establish their views and preferences. Cured patients and long-term survivors may have ongoing problems long after their treatment is successfully completed. These ongoing problems may be overlooked, so again it important to measure QoL long term and to look for late problems of psychosocial adaptation.

QoL assessments may be included in research studies such as randomized controlled trials (RCTs). The main reason is to compare the study treatments with respect to those aspects of QoL that may be affected by the treatment. These treatment comparisons will include both the positive benefits from trials that are expected to improve QoL, and any negative changes, from toxicity and side effects of treatment.

QoL can be a predictor of treatment success, and hence pre-treatment assessment of QoL may have prognostic value. Fayers and Machin (2007) suggest that the direction of the association between QoL scores and treatment outcome is not clear. Do QoL scores reflect an early perception by the patient of the disease progression? Alternatively, does QoL status in some way influence the course of the disease? Whatever the nature of the association, it is important to assess QoL and use it when making medical decisions for individual patients.

QoL assessment can also be used to make decisions on treatments at a population level, rather than an individual patient level. QoL outcomes can be used in economic evaluations alongside clinical trials to assess the clinical and cost-effectiveness of new health technologies.

There is an ongoing thoughtful discussion about the meaning of QoL, and about what should be measured. In the face of this debate, it is still important to measure quality of life as well as clinical and process-based outcomes. This is because ‘All of the these [QoL] concepts reflect issues that are of fundamental importance to patients’ well-being. They are all worth investigating and quantifying’ (Fayers and Machin, 2007).
1.6 Further reading

Measuring quality of life

Summary

This chapter describes the principles of measurement scales and introduces the methods for developing and validating new questionnaires. Psychometric methods lead to scales that are based upon items reflecting a patient’s level of QoL. The clinimetric approach makes use of composite scales that may include symptoms and side effects. The remainder of the chapter provides an overview of the stages of developing and validating new questionnaires and the principles that are involved.

2.1 Introduction

Questionnaires for assessing QoL usually contain multiple questions or items, although rarely a few may attempt to rely upon a single global question to assess overall QoL. For example, ‘Overall, what has your quality of life been like over the last week?’ Some QoL questionnaires are designed so that all items are combined together to produce an overall score. Most instruments attempt to group the items into separate ‘scales’ corresponding to different dimensions of QoL. This chapter explores the relationship between items and scales and introduces the concepts underlying QoL scales and their measurement.

2.2 Principles of measurement scales

2.2.1 Scales and items

Most QoL instruments consist of many questions or items. These items are usually combined to generate a dimension or domain score. Figure 2.1 shows this process graphically. Some of these items may aim to measure a relatively simple aspect of QoL, such as physical symptoms like nausea, vomiting or constipation. For these relatively simple aspects of QoL a single question or item may be sufficient to measure the underlying dimension.
For example, the EORTC QLQ-C30 questionnaire (Aaronson et al., 1993), measures the symptom of constipation with the single question, ‘During the past week, have you been constipated?’ The question has four possible response options: not at all; a little; quite a bit; very much.

The more complex psychological dimensions of QoL such as anxiety and depression are usually more vaguely defined in a subject’s understanding of QoL. These dimensions are typically measured by the use of several questions in multi-item scales. For example, the Hospital Anxiety and Depression scale (HADS) consists of 14 items, with seven items on the ‘anxiety’ aspect and the other seven items assessing ‘depression’ (Zigmond and Snaith, 1983).

### 2.2.2 Constructs and latent variables

Fayers and Machin (2007) describe QoL as a complex construct that cannot be adequately measured by a single global question. They suggest that QoL has a number of dimensions (see Figure 1.1), each of which should be thought of as an underlying ‘construct’. These constructs are represented or measured by ‘latent variables’, which we measure by asking the subject one or, more typically, a number of separate questions. For this reason QoL instruments commonly contain multiple questions to assess the underlying latent variables.

#### Example: Hospital Anxiety and Depression Scale (Zigmond and Snaith, 1983)

The HADS questionnaire (see Appendix A) is a QoL instrument with a simple theoretical structure (see Figure 2.2). It assumes that there are two different and distinct constructs of ‘anxiety’ and ‘depression’, which are meaningful to patients and can be quantified. It is assumed that anxiety and depression cannot be adequately measured by a single question, such as ‘How anxious are you today?’ (not at all, a little, quite a bit, very much), and that multiple questions must be employed. The HADS consists of 14 items, with seven questions relating to anxiety and seven questions relating to depression.
2.3 Indicator and causal variables

2.3.1 Indicator variables

Most items in personality tests, intelligence tests, educational attainment tests and other psychometric assessments reflect a level of ability or a state of mind. Such items do not alter or influence the latent construct that they measure. These items are indicator variables (Fayers and Machin, 2007). In common with most questionnaires that assess psychological aspects of QoL, the HADS items (see Appendix A) are mainly indicator variables. For example, ‘During the past week, I feel tense or “wound up”’. The question has four possible response options: most of the time; a lot of the time; from time to time; not at all.

2.3.2 Causal variables

The symptoms (such as nausea, vomiting and constipation) assessed in QoL scales, such as the EORTC QLQ-C30 (Aaronson et al., 1993) may cause a change in QoL. A patient who gets serious symptoms is likely to have their QoL affected by those symptoms. The reason for including symptoms in QoL instruments is principally that symptoms are believed to affect QoL. However, having a poor QoL does not imply that the patient has specific symptoms (such as nausea, vomiting and constipation). Typically, a single causal item may be enough to change the latent QoL variable. It is unnecessary, and usually rare, for each patient to suffer from all the symptoms in order to have a poor QoL. One serious symptom, such as extreme nausea, may be enough to reduce overall QoL.

Fayers and Machin (2007) caution that the above distinction between indicator and causal variables is not entirely clear-cut. Variables may frequently be partly indicator...
Box 2.1 Identifying causal items (Fayers and Machin, 2007)

- “Thought test”
  - Consider a typical patient from the target population. For an item called, say, item X:
    - a) If the level of item X changes, is the patient’s quality of life likely to change?
    - b) If the patient’s quality of life improves (or deteriorates), do we expect this to be reflected by a change in item X?
  - If the answer to (a) is “yes” and (b) is “no”, the item is likely to have a causal component.

and partly causal. For example, a patient may experience symptoms such as nausea and vomiting, become anxious and depressed, and then perceive and report the symptoms as being worse than they are. An initial causal variable has acquired indicator properties. So how can we identify causal variables? Fayers and Machin (2007) describe the thought test for identifying causal variables (see Box 2.1).

2.3.3 Why do we need to worry about the distinction between indicator and causal items?

Indicator variables assume that the observed responses to the items depend solely upon the level of the underlying latent variable. That is, if QoL is ‘good’, then this should be reflected in good or high levels of response on the various items. Furthermore, if the observed values of the items are correlated, then these correlations arise solely because of the effect of the latent variable. Causal variables are not correlated with each other through the different levels of QoL. They do not have correlations that arise through their parallel nature. Their correlations arise through an underlying variable – such as treatment, or stage or extent of disease. Thus causal variables may exhibit seemingly strange correlations that are nothing to do with changes in QoL. Causal items do not reflect QoL, they affect it. Therefore indicator and causal items behave in fundamentally different ways and this will have a considerable impact upon the design of QoL scales.

2.3.4 Single-item versus multi-item scales

Multi-item scales are commonly used to assess specific aspects of QoL. Responses from multiple items usually have several advantages over a score estimated from the responses to a single item in terms of reliability, precision, validity and scope (see Box 2.2).

2.4 The traditional psychometric model

The most common psychometric model is the parallel tests model. In this model each measurement item is a ‘test’ or question that reflects the level of the underlying construct or latent variable. Each item is distinct from the others, but is similar and comparable in all important respects. They differ only as a consequence of random error. These items