The Valuation of Health Outcomes Data from Clinical Trials for Use in Economic Evaluation

Volume 2

A thesis presented for the degree of PhD at the School of Health and Related Research,
the University of Sheffield
by

Isabel Margaret Falcon Towers, BSc (York), MMedSc (Birmingham)

October 2005

Funded by the Medical Research Council

Chapter 7

Study 3: Preferences over treatments for varicose veins – a test of linearity over time and probability

7.1 Introduction

The aim of this study was to construct health states describing varicose veins based on patient focus group discussions. These states were then used in the construction of profiles incorporating pre-treatment, treatment, and post-treatment states. Some of the profiles also incorporated the risks of the treatment procedures, including risk of recurrence. The conventional QALY algorithm and a holistic method were used to value these profiles. The results from the two valuation methods were compared to determine the extent to which estimates of disutility associated with varicose veins depend upon the methods used to calculate values for HRQoL. The effect of adding risk to the profiles and choices over treatment process were also examined.

7.2 Background

7.2.1 Varicose veins and its treatment

Varicose veins most commonly occur in the legs, and are the result of failures in the closing of venous valves (Tibbs, 1992). If a valve ceases to close properly, blood may flow back down when the leg muscle is relaxed. If the amount of blood involved is significant, the walls of the section of vein below the faulty valve are stretched so that they bulge out. This prevents the valve below from closing properly, and so on down the vein. Smaller veins, such as those near the skin surface, may become so stretched by the extra volume of blood that they increase in length. They are then forced to fold over themselves, forming the characteristic appearances of varicose veins.

Varicose veins can obviously lead to cosmetic defects. They may also manifest physical symptoms. These include feelings of heaviness or tension in the legs, aching, sensation of swelling, restless legs, cramps, itching, tingling, burning sensation, fatigue, pruritis, throbbing, bleeding, and ulceration (Bradbury *et al* 1999, Tibbs 1992, Weiss 1999, Wyatt 1999).

Over 50,000 operations to treat varicose veins are currently carried out each year in England and Wales, with an annual cost to the NHS of around £400-600 million (Bradbury *et al*, 1999). There is debate among clinicians about which patients should be treated, and the appropriate type of treatment.

Varicose veins have low priority in the NHS, reflecting the view that they are cosmetic defects rather than being an actual issue of health and quality of life. Although some health authorities fund treatment for varicose veins, such funding is generally limited, and other health authorities are not prepared to allocate any of their health care budgets to the treatment of varicose veins.

EQ-5D data was examined for the first 79 patients to be entered into a clinical trial of the different treatments and their effects on varicose veins (Michaels *et al*, 1999). The York MVH TTO algorithm was used to convert the raw EQ-5D data into utility scores before treatment and one month after treatment (MVH Group, 1995). The results are reported in Table 7.A.1 in Appendix 3. The mean utility before treatment was 0.779, and the mean utility one month after treatment was 0.786 ($p > 0.10$). This result indicates that the treatment used had a negligible effect on HRQoL within one month of treatment. However, amongst varicose veins patients there is a strong drive towards receiving treatment. This begs the question of whether the generic measure used to obtain this result was insensitive to the benefits of treatment. One month is a relatively short period of follow-up considering that it may take at least that long to completely get over the effects of surgery. It is possible that the effects of treatment on HRQoL could change over a longer period of follow-up than one month. Unfortunately, the EQ-5D data for treatment and one-month of follow-up for the first 79 patients in the trial was all that was available at this point in time. It would have been informative to make the comparison between EQ-5D data at time of treatment and six months or a year after treatment.

Although there are condition-specific measures of HRQoL for several other illnesses, to the knowledge of this author there are no formal preference-based condition-specific measures of HRQoL for varicose veins. One of the aims of this study was therefore to construct varicose veins specific vignettes. These would then be used to determine whether the symptoms of varicose veins had a greater impact on HRQoL than is generally recognised.

7.2.2 *Rationale of this study*

This study was undertaken alongside a research project examining the cost-effectiveness of the main treatments for varicose veins using the results of a clinical trial in a decision analysis model (Michaels et al, 1999). The data for modelling in this trial will be obtained through a combination of systematic literature review and the collection of retrospective and prospective data on patients undergoing treatment for varicose veins. This will include randomised controlled studies in three sub-groups of patients in whom conservative treatment, sclerotherapy and surgery will be compared. The model will allow an assessment of the incremental cost effectiveness of each treatment modality in subgroups of patients based upon their symptomatic, investigative and demographic features. Patient and societal priorities for treatment will be assessed using a number of stated preference techniques including the traditional QALY model.

This present study constructs condition-specific health states and profiles based on discussions with patients. It examines how benefits from varicose veins treatment should be measured, asking the question of whether generic QALY instruments such as EQ-5D measure disutility and benefits associated with varicose veins sufficiently.

This study also compares different methods of valuing the benefits of alternative treatments. The main comparison is between the QALY algorithm applied to health profiles against valuations of these profiles by a holistic method.

Varicose veins are a particularly good condition upon which to base this study. The treatment of varicose veins involves changes in HRQoL with no changes in lifetime duration. Although commonly described as a mild clinical condition, work by Garratt *et al* (1993) has shown that, in terms of physical and emotional functioning, fatigue, and pain, varicose vein sufferers score significantly lower than the general population on the SF-36 scale. Different treatments have differing degrees of effectiveness in improving HRQoL.

Treatments for varicose veins may be daunting, and can involve risks and other unpleasant consequences from the process of treatment. This study therefore presents the opportunity to look at how processes of treatment affect valuations. This study attempts to value treatments for varicose veins. It will test the QALY assumptions that health is the only attribute that matters. Process utility can be thought of as a non-health outcome because it simply relates to processes of care such as added comfort or reassurance, which some would argue do not relate to health itself. On the other hand, process utility can be thought of as a health outcome because different processes of care may have different effects in terms of levels of pain, anxiety, or other health factors (Donaldson and Shackley, 1997; Ratcliffe and Buxton, 1999).

This study also explored the effects of *ex ante* risks on valuations of profiles. The risks attached to entering or leaving a health state are often ignored by the QALY model. This study looked at the risks of mortality and recurrence after treatment.

The Michaels *et al* clinical trial uses conventional preference-based measures for QALYs, such as EQ-5D and SF-6D. This study aims to address the concern that these methods fall short of an adequate sensitivity to the benefits important to patients by constructing condition-specific measures.

7.3 Methods overview

A research protocol was written and submitted to the North Sheffield Local Research Ethics Committee. The protocol outlined the details of the study, methods used, and the intended recruitment of varicose veins patients to participate in focus groups and for the elicitation of health state and health profile valuations. The meeting of the ethics committee was attended by the author, and ethical approval was obtained for this study.

Health states were constructed based largely on patient focus groups, in which patients discussed aspects of varicose veins that they found important. Care was taken to use patient phraseology in describing symptoms wherever possible. These states were used in the formation of health profiles, which consisted of a pre-treatment "waiting list" state, followed by either sclerotherapy or surgery, followed by a post-treatment state which would last the rest of their life.

A questionnaire was designed using TTO to evaluate the states and profiles. Varicose veins patients were recruited to complete the questionnaire. Values for the profiles were constructed using discrete health state valuations multiplied by duration as per usual with the QALY method.

The profiles for this study were valued holistically using a single-stage generalised TTO (Reid, 1998). QALY and holistic values were compared for each profile. This study was to be used as a pilot study for the Michaels *et al* clinical trial, and this comparison was requested along with some research into which method more accurately reflected individual preferences.

7.4 Development of health state and profile descriptions

Varicose veins patients attending hospital tend to be very well-informed about their condition, and often have a clear opinion on which treatment they think they should receive before they actually see the consultant. It was felt to be important that patients views be incorporated into the profile descriptions. However, it was also felt that medical professionals and previous evidence could provide an important insight into the problems associated with varicose veins.

7.4.1 Questionnaire to health professionals

A short questionnaire was designed and sent to 25 health professionals, including consultants and nurses (see Appendix 3). It asked for their comments on issues relating to symptoms, treatment processes, outcomes and prognosis for varicose veins. Only four (16%) were returned. The response rate was disappointing, and the extent to which these four are representative of health professionals in general is unclear.

The comments are outlined in Table 7.A.2 in Appendix 3. Pain and discomfort were mentioned by one respondent as important symptoms. The possibility of ulceration was mentioned by three respondents. "Cosmetic" factors were mentioned by two respondents, with bulging appearance and unsightliness being cited by one of them. Another respondent mentioned heaviness of legs and swelling.

The responses to the treatment item of the questionnaire were variable. Surgery was mentioned by three respondents, with reference to unspecified "risks" by two. Compression stockings or hosiery were mentioned by two respondents. Two respondents implied a good prognosis from either surgery or compression methods. Only one respondent mentioned injection treatment, and nothing much was said about it. This questionnaire had little writing on it, giving the impression that it had been completed in a hurry without much thought. One respondent thought that, although there would be a good response to compression treatment or surgery, there were likely to be long-term problems such as varicose recurrence and ulcers.

One respondent did not use the questionnaire to describe varicose veins, but rather used it to air his/her concern that patients may not have enough understanding of possible complications of surgery, the probabilities that the treatment will work, and the recurrence rates after treatment. S/he was also concerned that symptoms might be exaggerated by patients in order to obtain treatment; thus if they know the NHS will not fund treatment for cosmetic reasons, the patients might pretend the veins are affecting their health more than they are.

Although these four respondents may not be representative of health professionals, it was still possible to use the information they provided qualitatively to guide the construction of the health states and profiles.

7.4.2 Patient focus groups

Methods

Two focus groups of patients were held in order to gain insight into the aspects of varicose veins that they found important. The groups were selected to obtain different perspectives of treatment:

1) Patients who had received treatment within the last 5 years. A mix of patients was selected so that the group included some patients who had not had surgery within the last year (to avoid responses being unduly weighted by the recent unpleasantness associated with surgery), and some patients who have had surgery within the last 12 months (these patients may recall relevant aspects of surgery which are forgotten as time goes by).

2) Patients who had not yet received treatment (they could discuss their expectations - what they hoped to achieve from treatment).

Patient details were obtained from lists held at the Sheffield Vascular Institute. Prospective patients were contacted by telephone in order to determine whether they would be interested in attending a focus group. Patients expressing an interest were then sent an invitation pack containing a letter signed by Mr Michaels, a covering letter from the author explaining the nature of the focus groups, and a pre-focus group questionnaire. This questionnaire asked them what topics they would like to discuss in the focus groups. This contact was carried out by a member of staff at the Sheffield Vascular Institute in order to preserve patient confidentiality. Patients who were willing to participate returned an attendance confirmation along with the questionnaire by post to the author, and their responses were written up into a list of topics for discussion. The day before the interviews were due to be held, patients were telephoned again (by the author) to remind them of the group and to give them the opportunity to say whether they were still able to attend.

The focus group discussions were tape recorded. There was a facilitator (the author) and a moderator present at both the focus groups. The facilitator conducted the group, while the moderator made relevant notes to accompany the transcripts. The focus groups were semi-structured. An interview plan was drawn up prior to the group sessions (see Tables 7.1 and 7.2). These interview plans drew from ideas suggested by Morgan (1998).

After the groups, the facilitator transcribed the tapes verbatim.

Results

Six patients were invited to each focus group, and each group had an attendance of four. There were three women and one man in each group. The transcripts can be provided on request.

These focus groups provided a great wealth of information about varicose veins symptoms and the effects they have on quality of life. The transcripts were examined for key categories of symptoms and other factors that were important to patients.

Patients in both groups were asked to define "good health". It was noticeable that the post-treatment group were able to instantly answer the question without referring to their varicose veins, whereas the pre-treatment group referred to their veins a great deal while answering the question. However, in the end they all agreed closely on a definition of good health. They defined good health as

> "being able to do the things that you would like to do ... when you want to do them, or for how long you want to do them within reason ... within the bounds of work ... or other commitments ... not being in pain ... not being short of breath, and not having to think about can I do this without being absolutely shattered".
>
> "To be able to get on with what you want to do, and be able to do it when you *want* to, not oh I'm not up to it today, I'll have to do it when I feel better. To be able to just do it *then*, because you want to be able to do it."

When asked how varicose veins affected their lives, it became clear that the condition had a negative affect on patients' lives. Itching and irritation was mentioned by several with comments such as "It drives me mad! I don't know what to do about that." Itching and irritation were cited as the worst thing about having varicose veins by several of the pre-treatment group. Although itching was mentioned by the post-treatment group, none of the post-treatment group cited it as the worst factor of varicose veins.

There was concern that varicose veins might ulcerate, especially if they were scratched in response to itching. Some patients also worried about the possibility of getting a deep vein thrombosis (DVT), especially as there had been a lot of coverage about travel-related DVTs in the media.

Self-consciousness of appearance seemed to be a major factor. Some patients said that they never wore shorts, or felt very self-conscious wearing shorts because of their veins, even when very hot.

Swelling was also an important factor, with many comments about the discomfort caused by swelling. Some people said that their ankles could get so swollen that it was painful to wear elasticated socks or stockings. Some patients commented that they had to plan their days around their varicose veins, largely because of the problems with swelling and having to plan what to wear because of this. The need to keep moving around to avoid cramps and aches was also mentioned. Cramps and aches were cited as the worst thing about having varicose veins by some of the post-treatment group and one of the pre-treatment group.

The necessity of keeping weight down was mentioned repeatedly in the pre-treatment group.

7.4.4 Designing the health states

It was a difficult decision to make regarding whether to follow the "clinical" argument or the "functional" argument (see Chapter 4). In the end a "middle of the road" approach was taken. The advantages of objective descriptions such as of appearance is that they would be equal for anyone visualising a particular state. However, the more subjective emotional or pain aspects of a particular state would be likely to differ widely between individuals. Thus appearance was described in terms used by members of the focus groups, but without reference to self-consciousness. Anything more subjective, such as worry over getting leg ulcers or planning life round the symptoms was described in terms of "you may worry..." or "you may find that...". This allowed for the uncertainty which is inevitable in medical decision making.

DVT was not included in the descriptions, because there is no available evidence that varicose veins sufferers are more concerned than the general population. Worry about getting DVT and ulcers was mentioned in both focus groups. However, as far as DVT is concerned, there has been much media publicity about whether certain aspects of long flight journeys put passengers at risk of DVT (CNN, 2001; BBC, 1999. 2000). There is no evidence that people with varicose veins are more concerned about this than members of the general population who do not have varicose veins. It is uncertain whether varicose veins sufferers are actually at greater risk of developing DVTs than members of the population as a whole (Campbell and Ridler. 1995; Oger *et al*. 2002). but there is some research to suggest they may be (Kakkar *et al*. 1970; Crandon *et al*, 1980; Heit *et al*, 2000).

Another point for discussion was the way in which the relevant factors were described in the health state. Thus if patients report that pain is a significant factor. should the description be of the type:

> "You may experience pain in your legs."

or:

> "You may experience mild/moderate/severe pain in your legs."

Since pain was the most significant factor relating to the trial varicose veins patients when compared to the Sheffield general population, consideration was given to whether an SF-36 description of pain should be used. There are two pain questions in the SF-36:

- How much bodily pain have you had during the past 4 weeks?

  None/Very mild/Mild/Moderate/Severe/Very severe

- During the past 4 weeks, how much did pain interfere with your normal work (including work both outside the home and housework)?

  Not at all/A little bit/ Moderately/Quite a bit/Extremely

It was decided that the descriptions in this survey should be condition-specific unlike the SF-36. As previously stated, one of the aims of this study was to construct condition-specific health states and profiles in order to address the concern that generic measures may not have sufficient sensitivity to detect the effects on HRQoL of some of the potential benefits of treatment that may be of importance to varicose veins patients. The wording was designed using phrases and words used by the patients in the focus groups in order to make the states familiar and recognisable to patients.

Symptoms mentioned as important in the postal survey of health professionals were also mentioned by patients in the patient focus groups. Health states for mild, moderate and severe varicose veins were designed based upon these sources of information. These are shown in Figure 7.1.

Consideration was given to the presentation of the health states (see Chapter 4). The fact that the health state descriptions relied upon the phraseology used by actual patients meant that each state was described by a set of bulky text. In an attempt to make the descriptions as accessible as possible, the descriptive words were highlighted in bold print. In order to further clarify the health state descriptions, each set of symptoms was separated and negative descriptions (*i.e.* the presence of symptoms) were appointed a sad face icon, whereas positive descriptions (*i.e.* the absence of symptoms) were left without an icon. The intention was to allow respondents to easily read the text and differentiate between the different health state descriptions in terms of the presence or absence of symptoms.

The process descriptions for the treatments sclerotherapy and surgery were constructed based upon information obtained from patient information leaflets (Campbell and Bickerton, 1999; Royal Devon and Exeter Healthcare NHS Trust, 1996). These are shown in Figure 7.2.

7.4.5 Designing the health profiles

For the profiles, there was debate about whether the pre-treatment state is required. The original draft of each profile consisted of:

pre-treatment state → treatment state → end state

However, it could be argued that it should not be necessary to include the pre-treatment state. Hence a profile would consist of:

treatment state → end state

This argument follows the line of reason that it should be the treatment being valued. However, the argument for including the pre-treatment state is that there is a tendency for people to prefer sequences that get better with time to sequences that get worse with time (Loewenstein and Prelec, 1993). If no pre-treatment state is included, respondents would be likely to value their own current health state as the pre-treatment state. The work of Loewenstein and Prelec (1993) suggests that if the patient's current health is near to full health they would place a lower value on the scenario than if their current health was very bad. This would be because the healthy patient would be looking at a declining sequence, whereas the unhealthy patient would be looking at an improving sequence. It was therefore decided that health profiles should consist of a pre-treatment health state followed by a treatment process followed by the final health state in order to ensure all respondents were valuing the same sequence.

There were three health states and two treatment processes under consideration, making a total of 18 possible profiles which could be created by combining pre-treatment states, treatment processes and end outcomes. In order to avoid cognitive overload it was decided that five profiles should be chosen to be included in the valuation questionnaire. Selection would be on the basis of what would be most interesting for this thesis in terms of comparing the QALY methodology with a holistic method of valuation, and what would be most interesting in terms of the clinical trial. As regards the thesis, it would be interesting to consider some profiles for which no divergences between the two methods would be expected, and some profiles for which divergences would be expected. The matrix of possible profiles is shown in Table 7.3. The conservative treatment was not included in this study, thus eliminating six of the profiles. A decision was made to limit the study to those going from moderate to mild or from moderate to moderate for both surgery and sclerotherapy. The profile severe → surgery → mild was also included for the sake of informing the clinical trial. These limitations were due to limited time and resources. This resulted in a total of five profiles (the shaded cells in Table 7.3).

In order to assess the degree to which uncertainty in the outcomes of treatment affects valuations or decision making, two questions included the risk of the process. These were incorporated in order to determine the extent to which *ex ante* risks affected patients' valuations of profiles. Surgery was estimated to have a 1 in 10,000 risk of mortality, compared to a zero risk of mortality for sclerotherapy. To aid respondents in considering the degree of risk, an everyday frame of reference was used. The risk of death was presented in the context of road traffic accidents, thus "You have a 1 in 10,000 chance of dying under the anaesthetic. (This is similar to the risk of having a fatal road accident in a year.)". This comparison was derived from data taken from mortality statistics in England and Wales in 1981 (Bandolier, 1996; Cooper, 1985) and the International Road Traffic and Accident Database (2002).

It was estimated that 20% of varicose veins would recur after surgery enough for re-treatment to be requested over the next 20-30 years. The recurrence rate for sclerotherapy was estimated to be 75-80% over the next 20-30 years (Michaels, 2001).

In terms of the presentation of the health profiles for valuation, there was the need to fit each profile description onto one side of A4 paper so that it could be valued in the questionnaire (see Appendix 3). This was managed by labelling the pre-treatment state (*e.g.* Moderate), detailing the process, and just using the label of the end outcome (*e.g.* Mild). The respondent would be told that the pre-treatment and outcome states were those that they had already ranked and valued, and they were encouraged to get the health state cards out again and refresh their memories of these health states.

Both the health states and health profile descriptions benefited during the draft processes from the advice of health professionals (Michaels, 2001; Rigby, 2001).

7.5 Valuation techniques

The self-completed version of the TTO suggested by Gudex (1994) was used for valuations of health states and health profiles (see Chapter 4).

The questionnaire incorporated valuations of the treatment processes of surgery and sclerotherapy. Respondents were asked to value these processes on their own in order to enter their values into the QALY algorithm, in addition to valuing them as a profile which included a pre- and a post-treatment state. There has been previous debate about how temporary health states should be valued (Badia and Herdman, 2001; Cook *et al*, 1994). Drummond *et al* (1997) suggested a method using TTO, whereby two temporary states are compared. Thus state H_i lasts for time t followed by full health, and state H_j lasts time x followed by full health, where $H_i \succ H_j$ and $t > x$. Time x is varied until the individual is indifferent between H_i for time t and H_j for time x. In order to obtain valuations on a scale of death to full health, H_j must also be valued as a short-term chronic state against death. Cook *et al* (1994) were concerned that the second stage of valuing H_j as a short-term chronic state against death would provide biased data because of death being so prominent. However, when they compared valuations of a state over 12 weeks, 12 months, and 12 years, they found reassuringly little differences in the valuations, thereby upholding the assumption of constant proportional trade-off.

A valuation procedure was drafted using this method to value surgery and sclerotherapy. The TTO questions involved were complex, and there was concern that this would have proved too much cognitively for most respondents. They already had to grasp the regular type of TTO question, in which they had to make a choice between a health state/profile for t years and full health for x years, where $x < t$. In the end it was decided that this additional complexity would create too much of a cognitive overload for respondents.

In 1995 the MVH Group (Measurement and Valuation of Health) at York published a report on valuation tariffs for EQ-5D health states. They had modelled tariffs from TTO and VAS valuations of a sample of EQ-5D states. One of their interests was in differences between VAS and TTO scores for the same health states. As a general observation, they found that VAS scores tended to be lower than TTO scores for states at the more mild end of the spectrum, and higher than TTO scores for states at the more severe end of the spectrum. The MVH Group modelled the relationship between the TTO and the VAS over health states derived from the EQ-5D, and found the quadratic equation in (7.1) to describe the relationship. In equation (7.1), VAS_i is the VAS score for the health state, TTO_i is the predicted TTO score for that state, and a_0, a_1, and a_2 are coefficients.

$$TTO_i = a_0 + a_1 * VAS_i + a_2 * VAS_i^2 \qquad (7.1)$$

This transformation was considered as an alternative procedure to the method for valuing short-term health states described above. The use of a VAS would be a simple way for respondents to value the treatment processes associated with surgery and sclerotherapy.

One consideration was whether the MVH model for transforming VAS values of EQ-5D health states to TTO would be comparable to transforming VAS values of short-term treatment processes to TTO. The MVH valuations were of health states of 10-year durations. Although the findings of Cook *et al* (1994) (described above) are reassuring in this aspect, the MVH study (1995) found that the value attached to a health state was significantly higher for a one-month duration than a 10-year duration, and the relationship existed (though not at a significant level) between a one-month duration and a one-year duration. The MVH Group dealt with this problem by providing different coefficients for their EQ-5D tariffs for different durations. However, the coefficients used for transforming VAS valuations of EQ-5D states to TTO remained constant regardless of the duration of the health state. The MVH Group did not have the data to say whether this was valid or not.

The MVH Group compared the mean TTO values for their selection of EQ-5D health states with the predicted mean values from the transformed VAS model. The difference in absolute values ranged from 0.002 to 0.27, with a mean absolute difference of 0.065. Placed in the context of a study such as this present one, an error of 0.065 seems relatively large. This study is powered to detect a difference in utility of 0.05, and this is considered to be the MEID. An estimated error in values for treatment processes of 0.065 is greater than the MEID, and this is of concern. However, this must be balanced against the great importance of using a method for obtaining values that the respondents find possible to use.

Respondents used a VAS scale to value the treatment processes of surgery and sclerotherapy. The valuation thus obtained was then converted to TTO using the MVH transformation involving coefficients based on means (MVH Group, 1995). Equation (7.2) shows the coefficient values used in this study, where $VAS_{process}$ is the VAS score for the description of the treatment process, and $TTO_{process}$ is the TTO score for that treatment process as predicted by the MVH model. Coefficients a_0, a_1, and a_2 have been derived from the MVH study and are –0.445, 2.112, and –0.580 respectively.

$$TTO_{process} = -0.445 + 2.112 * VAS_{process} + -0.580 * VAS_{process}^2 \qquad (7.2)$$

The relationship between the VAS scores and the predicted or transformed TTO scores is shown graphically in Figure 7.3 for all values of VAS between 0 and 1. It can be clearly seen that, for low values of VAS the transformed TTO values are lower, and for high values of VAS the transformed TTO values are slightly higher. The magnitude of the difference is greater at the lower end of the scale, where values of TTO are actually negative for VAS values of 0.2 or less. There is a slight problem with the upper end of the scale, because for VAS values of greater than 0.9 the transformed TTO values are greater than 1, which seems rather meaningless. However, the majority of VAS values would produce reasonable TTO scores.

In summary, the MVH method of transforming VAS scores to TTO does have some problems at the extremes of the scale, but it has the advantage of ease of completion for respondents and avoidance of cognitively demanding methods in an already difficult survey. The results section will determine the extent to which the extremes of the scale affect the scores.

7.6 The pilot

This formulation of the TTO valuation procedure has already been applied (Gudex, 1994), so it was not necessary to pilot the TTO method. However, the states and profiles were complex, and it was important to determine whether respondents were able to understand them, and also whether patients found them to be realistic.

Patients were recruited for the pilot study from lists held at the Sheffield Vascular Institute. Their consultant explained the study to them and handed them a pre-prepared information sheet and letter about the study. The information sheet explained the purpose of the study and the procedures involved. The letter stated explicitly that the study had the full approval of their consultant. It also stressed that they were under no obligation to take part in the study, and that their care would not be affected in any way should they choose to decline. The letter and information sheet are contained in Appendix 3. Patients who were willing to participate in the pilot study returned an attendance confirmation and consent form by post to the author.

Four patients agreed to take part in the pilot study. Of these, three attended. They were one man and two women, aged 49, 42 and 70 respectively. The author was present during the completion of the questionnaire in order to be on hand to discuss any aspects of the questionnaire and provide explanations as necessary. It was made clear to the attendees that this was a pilot, and they were specifically requested to criticise the process in order to aid in the finalisation of the questionnaire. They completed the questionnaire in 35 minutes with relative ease. The pilot took place in a seminar room of the Medical Education Centre, the Northern General Hospital, Sheffield.

During the piloting process it was decided that it was not logical to rank the descriptions of the process of surgery and sclerotherapy with the varicose veins health states. Patients were asked to imagine being in those states for the rest of their lives without change, whereas the processes are by nature transitory. The two process descriptions were not ranked in a separate exercise from their valuation on the same rating scale.

After the pilot, numbering was added to the table of current symptoms for ease of data entry.

The questionnaire was originally designed to be completed by a pre-treatment group of patients. However, the man had received surgery two weeks prior to the pilot, but since being invited. The question on what sort of treatment they wanted therefore was not appropriate in his case. It was considered likely that this event would occur throughout the surveying, because even though patients might not have received treatment when they received their invitation it was possible that they would have by the time they attended the interview. The question was altered accordingly. They were asked whether, if they had not already received treatment, they had particular treatment preferences.

The relative ease with which the patients completed the pilot questionnaire indicated that it was perfectly comprehensible. It was not therefore deemed necessary to pilot it on a larger group of patients.

7.7 The final questionnaire

The final version of the questionnaire is shown in Appendix 3, and consisted of:

- Background characteristics (age, sex, occupation, highest level of education, health rating, duration of varicose veins, whether received treatment, and opinions on what treatment should be received.
- Current symptoms
- Ranking exercise for health states of full health, current health, mild, moderate, severe states of varicose veins, and immediate death
- A practice TTO question for valuing a hypothetical health state
- TTO valuations of severe, moderate, and mild varicose veins followed by a valuation of current health
- VAS ratings of surgery and sclerotherapy
- A practice TTO valuation of a hypothetical profile
- Five TTO valuations of profiles without risks described
- Two TTO valuations of profiles containing descriptions of risks

- A section for patients' comments

It was possible to divide the patients into four groups according to treatment:

- Had received treatment

- Had not received treatment and would prefer surgery

- Had not received treatment and would prefer sclerotherapy

- Had not received treatment and was neutral

Since varicose veins treatment is very much patient driven, and patients are often very well-informed about treatments and may have an opinion of what treatment they want when referred, it was thought likely that preferences for particular treatments might influence their valuations. By creating these groups it would be possible to categorise people and examine the effect of prior preferences on valuations. Of course, with a sample size of 56 or somewhat higher, grouping members of the sample in this way would create four subgroups which would be too small in sample size to perform meaningful statistical tests.

The current symptom table used to ascertain current health was designed to be similar to the descriptions of symptoms used in the health states. It would then be possible to classify people roughly into mild, moderate and severe. Of course, this could not be exact, because there were more than three possible combinations of symptoms. But at least it would provide some insight into the extent to which patients value their own state differently to the same state described hypothetically. The same caveat regarding the small sizes of these subsamples applies as described in the previous paragraph.

7.8 Recruitment

Two batches of invitation letters were sent to varicose veins patients who were on waiting lists or referred pending lists (see Appendix 3). These letters were sent by staff at the Vascular Institute on the author's behalf. Each batch contained 100 invitations, making a total of 200. There was a low response rate of 41 (20.5%). In an effort to increase the sample size, patients were also recruited from a Barnsley weekend clinic. The consultant or the nurse asked the patient if they were willing to participate, and those that were completed the questionnaire after seeing their consultant. A total of 26 patients were recruited by this method. An E-mail was sent to staff and students at the

School of Health and Related Research, the University of Sheffield, asking for people with varicose veins to volunteer to take part in the study. Two respondents were recruited as a result.

Patients were interviewed in groups of one to five by a trained and experienced interviewer. She explained the TTO procedure, giving patients the opportunity to ask questions. She remained in the room in case they needed further explanations.

7.9 Analysis plan

7.9.1 Analysis of background characteristics

A descriptive analysis was carried out of respondents' background characteristics, in terms of age, sex, occupation, highest level of education, number of years suffered from varicose veins, and whether they had received previous treatment for their veins.

7.9.2 Determining treatment choices and underlying reasons

It was thought that, since varicose veins patients sometimes have prior opinions about different treatments, it could be useful to determine the relationship between their treatment preferences and their health profile valuations. Respondents were first asked whether they had received treatment for their varicose veins. Those who had not recently received treatment were then asked if they had a preference over treatment options, and if so to state it. They were also asked to state the reasons for their preferences. Stated treatment preferences were compared with VAS ratings of sclerotherapy and surgery to determine whether the preferred treatment was rated higher by VAS.

7.9.3 Classification of current symptoms

Statistics were drawn up describing the current symptoms of the sample as described by the classification tickbox system near the beginning of the questionnaire.

7.9.4 Statistics for general health

The questionnaire contained the SF-36 general health question. The responses to this were described in terms of proportions of respondents reporting their health as excellent, very good, good, fair, and poor. The general health ratings from this sample were compared to those of a large sample of general practice patients from Sheffield.

7.9.5 *Data completion and exclusion criteria*

The data were checked for completion. One of the aims of this study was to compare valuation data for health profiles from two different valuation methods: the QALY and a holistic method of valuation. In order to calculate the QALYs for health profiles, the valuation data for the constituent health states must be available. In order to compare QALY values with values from using the holistic method, the data from holistic valuations of the health profiles must be available. Respondents were therefore excluded from the analysis if they had missing valuation data for the constituent health states, missing valuation data for the treatment processes, or missing valuation data for the health profiles. Any respondents who had valuation data that was irremediably unclear was also excluded at this stage.

7.9.6 *Logical consistency and convergent validity of health states*

A test of logical consistency was carried out to determine if respondents ranked health states logically. There was only one logical ranking order for the health states, and this was full health $\succ$ mild varicose veins $\succ$ moderate varicose veins $\succ$ severe varicose veins. The number of respondents who did not fulfil logical consistency of ranking was determined.

In addition to the check of logical consistency of health state ranking, a test of convergent validity for health states was carried out. The ordinal ranking of health states was compared to the implied ranking provided by the TTO valuations of the health states. Respondents were considered to show convergent validity if their health state ranking as implied by their TTO ratings were the same as their original rankings (strong convergency), or if states ranked immediately above or below a given state were rated equal to that state in the TTO valuations (weak convergency). Respondents were considered non-convergent if their ranking orders of health states were not in the same order as their TTO valuations of these states or equal to the states ranked immediately above or below, or if states were not given equal values by TTO when they were ranked equally.

7.9.7 *Valuations of treatment process*

The treatment processes of sclerotherapy and surgery were rated against a VAS ranging from 0 to 100. The results were then divided by 100 to place them on a scale of 0 to 1.

These values were then transformed to TTO values using the method described in Section 7.5. Both non-transformed and transformed data were examined.

7.9.8 Valuations of health states

Health state values were calculated using the equation *x/t* where *x* is the number of years in full health stated to be equivalent to *t* years in the health state, where *t* = 20. How the value of 20 years for t was arrived at was discussed in detail in Chapter 4. In brief, it was considered to be a realistic life expectancy for the respondents of this study. The paired t-test and Wilcoxon-sign test were used to determine whether the health states differed significantly from each other or from full health.

Respondents were classified, using their classification of their current health, as closely as possible into the states of mild, moderate, and severe varicose veins. An analysis was then conducted with the aim of determining whether peoples' values of hypothetical health states was related to their current health at the time of the valuation exercise.

7.9.9 The QALY algorithm for health profiles

QALY values were calculated for the health profiles. The preliminary "waiting list" state lasted 6 months, and was followed by the treatment process. Duration of surgery was taken as 6 weeks, as this included recovery time. Duration for sclerotherapy was set as 1 week, because the recovery time is generally much shorter than for surgery. Treatment was followed by 19.38 years in the post-treatment state for surgery, and 19.48 years in the post-treatment health state for sclerotherapy. Equations (7.3) and (7.4) were used to calculate QALY values for the profiles using transformed VAS scores for treatment processes.

$$U \text{ (surgery profile)} = U \text{ (health state) } 0.5 + U \text{ (surgery) } 0.12 + U \text{ (health state) } 19.38 \tag{7.3}$$

$$U \text{ (sclerotherapy profile)} = U \text{ (health state) } 0.5 + U \text{ (sclerotherapy) } 0.02 + U \text{ (health state) } 19.48 \tag{7.4}$$

For the profiles containing risk, patients were told that there was a 1-*p* chance that their varicose veins would return to the original health state within the remainder of their lifetime. However, they did not know at what point the recurrence would take place. It could be at any point in time. Therefore this risk is ambiguous. For the sake of calculating values, this study assumed for the main analysis that recurrence would occur halfway through the remaining life expectancy and a time preference rate of zero was assumed (alternative assumptions are explored below). Equations (7.5) and (7.6) were used to obtain the QALY valuations for these risk profiles.

Surgery:
$$0.5U_{moderate}+0.12U_{surgery}+\underbrace{[0.20\{(U_{moderate}*19.38*0.5)+(U_{mild}*19.38*0.5)-\underbrace{(1/10000*19.38)}_{\text{(risk of death)}}\}]}_{\text{(risk of recurrence)}}+\underbrace{[0.80(U_{mild}*19.38)-(1/10000*19.38)]}_{\text{(chance of non-recurrence)}} \tag{7.5}$$

Sclerotherapy:
$$0.5U_{moderate}+0.02U_{sclerotherapy}+\underbrace{[0.75\{(U_{moderate}*19.48*0.5)+(U_{mild}*19.48*0.5)\}]}_{\text{(risk of recurrence)}}+\underbrace{[0.25(U_{mild}*19.48)]}_{\text{(chance of non-recurrence)}} \tag{7.6}$$

7.9.10 Holistic valuation of health profiles

In valuing profiles holistically, respondents gave an indifference value x for each profile, such that x years in full health was equivalent to t years with the health profile. These values of x were taken as the "healthy years equivalent" for each profile.

In order to obtain a value on a scale of 0 to 1 for the sake of comparison with other studies, health profile values were also calculated using the equation x/t where x is the number of years in full health stated to be equivalent to t years in the health profile, and $t = 20$.

7.9.11 Statistical comparisons for health profiles

QALY and holistic values for profiles were compared using the paired t-test and the Wilcoxon-sign test. The null hypothesis was that there were no significant differences between the QALY and holistic methods of valuing health profiles.

The t-test and Wilcoxon-sign test were also used to determine whether the values of health profiles differed significantly from each other for each valuation method. Thus for each of the QALY and holistic results, each health profile value was compared to that of every other health profile.

7.9.12 Incorporating risk into health profiles

The QALY values for the health profiles containing risk were calculated as stated in equations 7.5 and 7.6. The risks were inherent in the profile descriptions, and were therefore valued directly for the holistic valuation method. The values for the equivalent profiles containing no descriptions of risk were compared with the values of the profiles containing risk for each of the QALY and holistic valuations methods.

7.9.13 Logical consistency for health profiles

Some of the health profiles have a logical ranking order, as listed below:

Mod-Scl-Mild $\succ$ Mod-Scl-Mod

Mod-Sur-Mild $\succ$ Mod-Sur-Mod

Mod-Sur-Mild $\succ$ Sev-Sur-Mild

Mod-Scl-Mild $\succ$ Mod-Scl-Mild (risk)

Mod-Sur-Mild $\succ$ Mod-Sur-Mild (risk)

The degree to which respondents followed these logical rankings was examined, both for the QALY method and the holistic method of valuation. If the ranking order was the same as that listed above for any of the pairwise comparisons, the respondent was said to be strongly consistent for that comparison. If the values given to the two profiles were equal, that respondent was said to be weakly consistent. If the preferences were in the reverse order, that respondent was said to be non-consistent. The degree of strong, weak, and non-consistency was reported for each of the above pairwise comparisons.

7.9.14 Unwillingness to trade

A sub-analysis was conducted, excluding those respondents who were unwilling to trade. Unwillingness to trade would raise the mean values of the profiles for the sample. This analysis allowed the extent of this effect to be seen.

7.9.15 Sensitivity analysis

Time of recurrence

As previously stated, the risk of recurrence was explicit in two of the profiles, but the time of recurrence was unknown. The QALY algorithm used in this study made the assumption that recurrence would occur halfway through the remainder of the individual's life. A sensitivity analysis was performed to explore the possible range of values for these risky health profiles if recurrence occurred soon after treatment or close to the end of the person's life. At one extreme the recurrence was assumed to take place approximately 1 year after treatment (0.05 of life expectancy), and at the other extreme recurrence was assumed to take place approximately 1 year before the end of life (0.95 of life expectancy). These are demonstrated for both surgery and sclerotherapy in Equations (7.7), (7.8), (7.9), and (7.10).

Surgery: $0.5U_{\text{moderate}}+0.12U_{\text{surgery}}+[0.20\{(U_{\text{moderate}}*19.38*0.05)+(U_{\text{mild}}*19.38*0.95)-(1/10000*19.38)\}]+[0.80(U_{\text{mild}}*19.38)-(1/10000*19.38)]$ (7.7)

Surgery: $0.5U_{\text{moderate}}+0.12U_{\text{surgery}}+[0.20\{(U_{\text{moderate}}*19.38*0.95)+(U_{\text{mild}}*19.38*0.05)-(1/10000*19.38)\}]+[0.80(U_{\text{mild}}*19.38)-(1/10000*19.38)]$ (7.8)

Sclerotherapy:
$0.5U_{\text{moderate}}+0.02U_{\text{sclerotherapy}}+[0.75\{(U_{\text{moderate}}*19.48*0.05)+(U_{\text{mild}}*19.48*0.95)\}]+[0.25(U_{\text{mild}}*19.48)]$ (7.9)

Sclerotherapy:
$0.5U_{\text{moderate}}+0.02U_{\text{sclerotherapy}}+[0.75\{(U_{\text{moderate}}*19.48*0.95)+(U_{\text{mild}}*19.48*0.05)\}]+[0.25(U_{\text{mild}}*19.48)]$ (7.10)

The paired t-test and Wilcoxon-sign test were used to determine whether there were significant differences between the holistic valuations of the risky profiles and the QALY valuations with the different times of recurrence. These statistical tests were also used to examine the differences between the different QALY values for these profiles obtained by varying the time of recurrence.

Sensitivity to process

A sensitivity analysis was carried out to determine the extent to which treatment process (as valued in this study) could affect QALY valuations of profiles. The value of surgery was adjusted to zero for the entire sample while the value of sclerotherapy was adjusted to 1. Then they were adjusted the opposite way, so that the value of surgery was 1 and the value of sclerotherapy was zero.

The purpose of this sensitivity analysis was to gauge the range of effects that might be expected from the values of treatment process. This was because the duration of the treatment phase was not included in the profile descriptions for holistic valuations (see Section 7.11.1).

7.9.16 Time discounting

Time preferences were not measured for this varicose veins sample, but it was considered desirable to use time preference estimates to determine the possible effects of time preferences on the QALY valuations of the profiles. The literature on population discount rates reveals several different discount rates. The UK government recommends a discount rate of 0.035 (HM Treasury, 2003). Cairns and van der Pol (2000) conducted a review of previous empirical time preference literature in health (see also Chapter 3). Each study was different in terms of sample size and delay periods considered, and time preference rates differed accordingly. The lowest mean and median time preference rates from these studies were –0.029 and 0.000 respectively (Dolan and Gudex, 1995), and the highest mean and median time preference rates found were 1.240 and 1.000 respectively (Chapman and Elstein, 1995).

These extreme values were used to determine the range of possible effects of discounting QALY valuations for time (see equation 7.11). The QALY valuations were also adjusted by 0.073, which was the mean time preference rate for own health from the open-ended method used by Cairns and van der Pol (2000). QALY valuations were also discounted by 0.035, which is the discount rate recommended by the UK government.

For each health profile, each year was adjusted for the discount rate (r), as shown in equation 7.11. This was done for all the health profiles, for each discount rate. The results were reported and compared with holistic valuations, for which it was assumed that time discounting was inherent to the holistic TTO process.

$$\{1/(1+r)^0 * U_1\} + \{1/(1+r)^1 * U_2\} + \ldots + \{1/(1+r)^{n-1} * U_n\} \qquad (7.11)$$

7.9.17 Analysis by treatment groups

Responses were analysed according to whether they had already received treatment for their veins. Where respondents had merely answered "yes" to the question of whether they had received treatment, it was assumed that they had received surgery or sclerotherapy recently. Respondents who had answered "yes" were put into the "treated" group. Those who had either answered "no", or that they had received surgery or sclerotherapy several years previously, or had received other treatments, were put into the "untreated" group. QALY and holistic health profile values were compared between these two groups, using the paired t-test and the Wilcoxon-sign test, to determine whether recent treatment with surgery or sclerotherapy affected valuations by the two methods. Results from the two groups were compared with each other, and with the results from the whole sample.

7.9.18 Patients' comments

Patients' written comments were examined to determine whether they could throw any light on the results. Comments were classified according to the nature of the comment. Some respondents made verbal comments, which the interviewer noted down. These verbal comments are also reported in this section, as are some of the interviewer's personal observations of some respondents where relevant.

7.10 Results

7.10.1 Background characteristics

There were 67 varicose veins patients in the sample. This was 33.5% of the 200 patients invited by letter, and less than this percentage of the entire number of patients approached if patients at the clinic are included. The number of patients approached at the clinic in Barnsley is unknown, because this was done by their consultants and the information on numbers was not made available. In terms of age, the present sample was similar to the IBS samples in Chapters 5 and 6, with a mean age of 48.4 years (SD 13.2). The median for the sample was 50.0 years, with an interquartile range of 36.9 to 58.3 years. There was a range of 23 to 78 years of age. There was one missing value for age (Table 7.A.3).

The mean number of years from which the sample had suffered from varicose veins was 13.3 years (SD 10.8). The median number of years suffered from varicose veins was 10.0, with an interquartile range of 4.0 to 22.5 years. The range for this variable was 1 to 40 years. There were seven missing values (Table 7.A.3).

The sample contained 13 (19.4%) men and 54 (80.6%) women. Occupational status is outlined in Table 7.A.4. There were two missing values for occupation. A total of 40

(59.7%) respondents were in paid employment. and 27 (40.3%) were not in paid employment.

A total of 50 (74.6%) had obtained educational qualifications below the level of higher education, and 17 (25.4%) had taken courses in higher education (Table 7.A.5).

7.10.2 Treatment choices

Respondents were asked if they had received treatment for their varicose veins. Twenty-two (32.8%) replied "yes", 41 (61.2%) replied "no", one (1.5%) replied that they had been injected 20 years previously. one (1.5%) stated that they had been given anti-inflammatory drugs for phlebitis, and two (3%) replied that they had received treatment some years ago.

If respondents had not yet received treatment for their veins, they were asked if they had already decided which treatment they wanted. If they had decided, they were asked to state what treatment they thought they should receive, and the reasons for their choice. A total of 46 (68.7%) had either been treated already, or had no particular opinions about what treatment they should receive. However, 21 (31.3%) respondents had already decided what treatment they thought they should have. The responses to this question are presented in Table 7.A.6. and summarised in Table 7.4. Thirteen of the respondents gave an opinion that they would like to receive surgery at the beginning of the questionnaire. Six of these rated sclerotherapy higher than surgery on the VAS which was later in the questionnaire. Six rated surgery higher, and one rated surgery and sclerotherapy equally on the VAS. It is possible that these changes of mind were due to the descriptions of sclerotherapy and surgery presented later in the questionnaire being different from what respondents had imagined at the beginning of the questionnaire. It is also possible that they had not previously considered sclerotherapy.

Pain. discomfort or aching was cited as a reason for wanting treatment by nine respondents. The overall choice seemed to be for surgery. Four respondents cited unsightliness as a reason for wishing their veins to be treated. Two respondents said that they would choose surgery because it was what the clinic had advised. Four respondents cited alternative types of treatment such as heat treatment or homeopathy. A reason for this choice was wishing to avoid anaesthetic.

7.10.3 Current symptoms

Table 7.5 describes the current symptoms of the sample. The symptoms were widely prevalent amongst the sample, with all categories of symptoms described in the health states present to a greater or lesser extent in the sample. The highest reported symptom was that the veins were noticeable (98.5%). Most of the sample (92.5%) considered their varicose veins to look unsightly. Various degrees of swelling were commonly reported, with 86.6% saying that their veins sometimes became swollen and 56.7% of the sample saying that their veins often became very swollen. Aching and pain were common symptoms, with 94.1% reporting that their legs often or sometimes ached or felt painful. Cramp was another common symptom (76.2%), as was itching and irritation of legs (82.1%). A significant proportion of the sample were worried about the possibility of getting an ulcer (48.8%). However, only 16.4% found themselves planning their lives around their symptoms.

7.10.4 General health

Eight (11.9%) respondents rated their general health as excellent. Thirty-one (46.3%) rated it as very good. Nineteen (28.4%) rated it as good. Seven (10.4%) rated it as fair. Two (3.0%) rated general health as poor. This was compared to the results of a survey of 1582 patients randomly selected from two GP lists in Sheffield (data obtained from the study by Brazier *et al*, 1992). In the Sheffield population, 10.4% rated their health as excellent, 37.0% rated it as very good, 34.1% rated it as good, 14.5% rated their health as fair, and 2.7% rated their health as poor. Thus in terms of general health, this varicose veins sample was similar to the Sheffield general population. The mean TTO valuation of current health for this sample was 0.862.

Surprisingly, a higher proportion of this varicose veins sample rated their health as excellent or very good than the general practice sample obtained by Brazier *et al* (1992). Characteristics of the general practice sample were compared to the 1988 General Household Survey, and were found not to differ in terms of use of health services. This would indicate that the Sheffield general practice sample was reasonably representative of the general population of Sheffield in terms of general health. A possible explanation for the apparent better health in these varicose veins patients is that this result was a fluke of the relatively small sample size.

7.10.5 Data completion

A total of eight (11.9%) respondents had incomplete data for the valuation questions. These were excluded from further analysis (Table 7.A.7). After the eight were excluded, a total of 59 remained in the analysis.

7.10.6 Logical consistency and convergent validity for states

Five respondents ranked the health states in an illogical order (see Section 7.9.6).

Implied TTO values for health states were compared with the original ranking of the states in order to check for convergent validity for the 59 respondents. A total of 51 (86.4%) gave responses that were either strongly or weakly convergent. Of these, 41 (80.4%) were only weakly convergent, valuing two or more of the three health states equally by the TTO method, although they were ranked unequal during the ranking procedure.

The ranking exercise was also a warm-up exercise, which gave respondents the opportunity to become familiar with the health state descriptions in addition to the concept of valuing different states of health. This level of total non-convergence is comparable with that displayed in the other studies of this thesis, and is therefore not unexpected. However, the degree of weak convergency is very high at 80.4%. Many of the sample were unwilling to trade (see Section 7.10.12), and this could explain them valuing all states the same at 19 years (the point of indifference between 18 and a maximum of 20 years).

7.10.7 Values of process

Table 7.6 shows the results of valuing the treatments of sclerotherapy and surgery. Both non-transformed and transformed TTO values are shown (see Section 7.5). Sclerotherapy has higher mean and median values than surgery for both transformed and non-transformed sets of data.

The mean VAS score for sclerotherapy is 0.679 (median 0.7) (Table 7.6). Under the MVH transformation to TTO, the mean value of sclerptherapy is transformed to 0.696 (median 0.749). Thus the transformation has the effect of raising the average values for sclerotherapy. The minimum VAS score for sclerotherapy is 0.2, and the maximum is 1. The minimum and maximum values under the MVH TTO transformation are –0.046 and 1.087 respectively. As discussed in Section 7.5 and demonstrated in Figure 7.3, for low VAS values the MVH transformation gives lower TTO scores, and for high VAS scores it gives higher values for TTO. This explains the transformed values of less than zero and greater than one seen in Table 7.6.

The mean VAS score for surgery is 0.605 (median 0.6) (Table 7.6). This is transformed to 0.599 (median 0.613) under the MVH transformation. Thus the MVH transformation has the effect of lowering the mean value for surgery while raising the median value. The non-transformed VAS minimum score is 0.1, maximum 1. Under the MVH transformation the minimum score is –0.24, maximum 1.087. This lower minimum may be the reason for the lowered transformed mean.

As already mentioned in Section 7.5, values over 1 (full health) seem meaningless. There was no impression that any members of the sample thought either treatment process was worth than death, leaving the impression that the negative values under the MVH transformation are also unreliable. As a reminder, VAS values below approximately 0.25 are transformed to negative TTO values, and VAS values greater than approximately 0.91 are transformed to TTO values greater than 1 (Figure 7.3).

There was one (1.7%) respondent who gave a VAS score for sclerotherapy of less than 0.25, and there were seven (11.9%) respondents who gave VAS scores for sclerotherapy of greater than 0.91. There were three (5.1%) respondents who gave VAS values for surgery of less than 0.25, and one (1.7%) respondent who gave a VAS value of greater than 0.91 for surgery. Thus, for the valuation of sclerotherapy, the values of eight (13.6%) respondents were questionable under the MVH transformation. For the valuation of surgery, the values of four (6.8%) respondents were questionable under the MVH transformation.

The absolute differences range from 0.009 to 0.246 for sclerotherapy, and 0.009 to 0.34 for surgery. The mean difference between the VAS and transformed values for sclerotherapy is 0.017 (median 0.049), and for surgery the mean difference is 0.006 (median 0.013). These are small differences, which are below the MEID. It is therefore a reasonable assumption that the error margin for the MVH transformation is small in this study.

7.10.8 Health state valuations

Table 7.7 shows the statistics for the health state valuations. The mean valuations follow a logical order, with mild varicose veins rated highest (mean 0.88) and severe varicose veins rated lowest (mean 0.79). The mean value for current health is 0.86, and is directly between mild and moderate varicose veins. The differences between the valuations of the states are small, with a range of mean values from mild to severe of 0.09.

It can be seen from Table 7.7 that the median values for current health, mild and moderate varicose veins are all equal at 0.95. The variability in the valuation for mild varicose veins was so small that the IQR was zero. For current health the IQR covered a range of 0.10, and for moderate varicose veins the IQR covered a range of 0.20. The median value for severe varicose veins was lower at 0.85, and the IQR covered a broader range of 0.30. This suggests that there was more variability in responses as the severity of the health state increased.

Valuations of the different health states were compared in order to determine whether there were any significant differences between them (Table 7.8). The only tests for which the differences were not significant were between the mild state and current health, and current health and moderate varicose veins. The value of 1 was attributed to each member of the sample as the value for full health. It is interesting to note that the mean value for all the health states of mild, moderate, severe, and current health are significantly lower than the value of 1 for full health according to both the paired t-test and the Wilcoxon-sign test ($p < 0.001$). This is particularly interesting with respect to their valuations of their current health, because the majority of the sample rated their general health as excellent, very good, or good (see Section 7.10.4). This indicates a discrepancy between the SF-36 question on current health and the TTO method.

An attempt was made to classify respondents into mild, moderate or severe varicose veins states according to the current symptoms they expressed (see Section 7.10.3). It was only possible to perform a rough classification, because respondents rarely completely fitted exactly into the descriptive system for symptoms in each state description. Two (3.4%) out of the 59 respondents were classified as having mild varicose veins, 32 (54.2%) as having moderate varicose veins, and 25 (42.4%) as having severe varicose veins. Mean and median health state values for the states of mild, moderate, and severe varicose veins, and also current health were calculated for each of these groupings (Table 7.9). The size of each of these sub-samples is relatively small, and therefore the results need to be viewed with caution. This caveat applies especially to the valuations for people classified as mild, because there were only two such respondents.

Valuations for the mild, moderate and severe health states within the groups classified as severe and moderate are typically very close to the values for the whole sample (Tables 7.7 and 7.9). However, some differences should be noted. One of these differences is in the valuations of current health across the three classification groups. The severe group valued current health at 0.842, which is the whole sample value of moderate varicose veins. The moderate class valued current health at 0.871, which is close to the whole sample value of mild varicose veins. The two members of the sample who were classified as having mild varicose veins gave very high valuations for current health (mean 0.975). They also gave a particularly high mean value to the mild state (0.95), but their mean valuations for the states of moderate and severe varicose veins are similar to those of the entire sample (0.85 and 0.8 respectively). As stated above, it was only possible to roughly classify the respondents, and it may be that the group classified as severe were closer to moderate, and the group classified as moderate closer to mild than seemed clear from the symptoms stated.

7.10.9 Health profile valuations

The values of the profiles were calculated, and the results are set out in Table 7.10 and Figure 7.4. For five of the seven profiles the mean values are higher for the QALY method. These are the profiles which consist of a sequence that improves over time. In the context of this particular study these are the profiles describing a treatment and subsequent improvement in health. The two profiles for which mean holistic values are higher than QALY values are Moderate → Sclerotherapy → Moderate and Moderate → Surgery → Moderate. These are the profiles for which there is no improvement subsequent to treatment, and thus no improving sequence.

The values in Table 7.10 are presented in years in full health equivalent to 20 years in the lesser profile. This is a meaningful way of looking at the results in the context of this study. However, in order to view the results in a broader context of effects on a HRQoL scale which may be compared with other studies, these values were transformed to a scale of 0 to 1 by dividing each value by 20. The results are shown in Table 7.11. The mean differences between QALY and holistic valuations range from -0.018 to 0.034. As described in Chapter 4, this study was powered to find a minimal economically important difference (MEID) of 0.05. The study did not have the statistical power to detect significant differences of less than 0.05.

There were no significant differences between the means of QALY and holistic valuations when tested by the paired *t*-test. However, the Wilcoxon-sign test found significant differences for three of the profiles (Table 7.10). One possible reason for this difference between statistical tests could be that the *t*-test tests for differences between means, whereas the Wilcoxon-sign tests for differences in distributions. Histograms for each variable are plotted in Figure 7.A.1 (Appendix 3), but there are no outstanding differences between the plots for the different methods. The median and IQRs are examined below to explore the root of the differences between the results of the t-test and the Wilcoxon-sign test.

The median values for all the holistic valuations are 19 years (Table 7.10). The median values for the valuations by the QALY method range from 18.91 to 18.99 years. The IQRs for most of the holistic valuations of the health profiles are 2 years, with the exception of the profile Moderate → Surgery → Mild containing risk which has an IQR of 4 years (Table 7.10). In contrast, the IQRs for most of the QALY valuations are much lower, indicating a narrower distribution of values for the QALY method than the holistic method. The IQRs for the QALY valuations of the successful treatment outcomes are between 0.1 and 0.5 years (Table 7.10). Interestingly, for the QALY valuation of the profile Moderate → Sclerotherapy → Moderate the IQR is 4 years, and for the profile Moderate → Surgery → Moderate the IQR is 3.95 years. The other QALY valuation with a large IQR is for the profile Moderate → Sclerotherapy → Mild containing risk, which has an IQR of 1.56 years (Table 7.10). In summary, the holistic method has consistently wide IQRs for all the profile valuations. The QALY method has very low IQRs where the outcome of treatment is successful and where there is no *ex ante* risk described. However, where the outcome of treatment is unsuccessful and for the sclerotherapy profile containing a description of *ex ante* risk the IQRs are much wider. For the risky sclerotherapy profile the IQR is comparable to those of the holistic valuations. However, for the profiles containing unsuccessful treatment the IQRs are twice as wide for the QALY method than for the holistic method. These differences in the distributions could explain the different findings of the two statistical tests.

The relatively wide IQRs around the holistic median values for the health profiles indicate a greater degree of variability in respondents' values for the health profiles. This explanation is straightforward. Consideration should be given to the differences in IQRs for the QALY valuations of the different profiles. As already explained, the QALY values are derived from the values given to the composite health states. The

SDs for the health state valuations in Table 7.7 are comparable to those of the QALY valuations of the health profiles in Table 7.11. The IQRs for the health states vary between 0.0 (mild) and 0.3 (severe). The SDs for the non-transformed valuations of treatment processes (Table 7.6) are comparable to those of the health states (Table 7.7). However, the SDs for the transformed values for treatment processes are greater, as are the IQRs. The IQR is greater for the moderate state (0.2) than the mild state (0.0). Therefore there is the potential for greater variability in values for a profile beginning and ending in the moderate state than one beginning in moderate and ending in mild. This is what is found in Tables 7.10 and 7.11.

Both the paired t-test and the Wilcoxon-sign test are based on the difference between the QALY and holistic values given by each member of the sample. The t value is given by dividing the mean of this difference by the standard deviation of the difference. Thus the t-test is a parametric method relying on standard deviations and means. Although the distributions of QALY and holistic values are heavily skewed, the sample size is greater than 30 and therefore the paired t-test should be expected to provide valid results. The Wilcoxon-sign test relies on the sum of the signed ranks of the differences between QALY and holistic valuations (W). If the null hypothesis is that there is a mean difference of zero between QALY and holistic valuations, the sampling distribution of W values approximates to a normal distribution around a mean of zero for samples of greater than 10 (Lowry, 1999-2005). By dealing with signed ranks, the Wilcoxon-sign tells which is greater of QALY and holistic values, and gives some indication of degree of difference. However, it is crude in terms of exact values. It seems likely that in some cases the two tests would give different results, due to the different methods used. In fact, there is no reason to think that two different tests should give the same answer when applied to the same data, because they make different assumptions and use different aspects of the observed data (Altman, 1991). However, one would expect similar results to be achieved by two methods if they were both valid. When the sample size is over 30, even though non-normal in distribution, a parametric method has greater power and is useful because it provides confidence intervals for the difference between QALY and holistic values.

In summary, the two tests give different results because they are measuring different things. According to the Wilcoxon-sign, a difference may be significant because many of the respondents in the sample give a higher value for the QALY measurement than the holistic measurement. However, according to the paired t-test the difference between the results from the two methods is non-significant because of the width of the confidence intervals. There is a 95% probability that the population mean difference between the results from the two methods would be between some negative and some positive value (Table 7.10), in other words the 95% confidence interval crosses zero. Thus there is no evidence that one method would consistently give higher values than the other method. Whereas the Wilcoxon-sign looks at how many respondents have values of QALYs greater than those from the holistic method (or vice versa) but does not take the magnitude of the difference into account, the t-test estimates the standard error and mean difference between the two methods for the population from which the sample is drawn, and adds 95% confidence intervals to the mean difference between the two methods for the sample. In samples with more than 30 respondents, the t-test has more power than the Wilcoxon-sign.

Of interest is the difference between successful treatment and non-successful treatment in terms of HRQoL. Table 7.10 shows the values in terms of years in full health equivalent to 20 years incorporating the health profile being valued. The difference between mean QALY valuations of the profiles Moderate → Sclerotherapy → Mild and Moderate → Sclerotherapy → Moderate is 0.792, and that between mean QALY valuations of the profiles Moderate → Surgery → Mild and Moderate → Surgery → Moderate is 0.788 years. The corresponding differences between mean holistic valuations are both 0.23 years. Table 7.11 shows the values transformed onto a scale of 0 to 1 (death to full health). Here the difference between the mean QALY valuations of the two sclerotherapy profiles is 0.040 and between the surgery profiles is 0.039, and the differences between the mean holistic valuations are 0.012. It is striking that the differences between the valuations for treatment with successful outcome and unsuccessful outcome are virtually equal for both the treatment processes of surgery and sclerotherapy, and this is the case for both valuation methods. Whereas the differences for both the holistic method and the QALY method of valuing the profiles are less than the MEID of 0.05 (see Table 7.11), those for the QALY method are close to 0.05.

The t-test and Wilcoxon-sign test were used to determine if there were significant differences between valuations within each method. The valuation of each health profile was tested against every other profile valuation, and this test was performed for both valuation methods. There were a total of 21 combinations of health profiles to be tested. The most highly significant finding for the holistic valuations was a difference between the profile Moderate-Surgery-Mild (risk) and the profile without risk (p =

0.054). The profile value was reduced by 0.57 when risk was added. This finding was with the Wilcoxon-sign test. No significant differences were found by the *t*-test for the holistic method.

The story was different for the QALY method. By the Wilcoxon-sign test, there were only seven out of the 21 combinations that were not significantly different (Table 7.12). All other combinations were significantly different from each other ($p < 0.005$). For the t-test, $p < 0.05$ for all combinations of QALY valuations.

Thus, when the profiles were valued by the holistic method, there were virtually no differences between valuations for each profile. However, when the QALY algorithm was applied to the profiles there were significant differences between valuations for most of the profiles.

7.10.10 The effect of incorporating risk

The incorporation of the risks of treatment to the profiles affected both QALY and holistic valuations in different ways. The holistic mean value for the riskless sclerotherapy profile was 17.37 compared to a mean value of 17.24 when risk was included. However, there was a more marked effect on the difference between the holistic value of the surgery riskless profile and the equivalent profile with risk incorporated. The riskless surgery profile was given a mean value of 17.37 compared to a mean value of 16.80 when risks were included (Table 7.10). As stated above, the difference between the risky and riskless profiles for surgery were found to be significant by the Wilcoxon-sign test.

The inclusion of risks had an important effect on QALY valuations. For the riskless profiles there was a mean preference for sclerotherapy over surgery by the QALY method (17.59 compared to 17.56 respectively) (Table 7.10). However, when risks were incorporated into the profiles, there was a mean preference for surgery over sclerotherapy (17.47 versus 17.29 respectively). Both these differences were significant (Table 7.12).

These findings clearly demonstrate that the way in which risk is incorporated into profile valuations has important implications to the valuation of health profiles, and this will be discussed further in the Discussion section.

7.10.11 Logical consistency

The results were found to follow the logical consistency conditions outlined in Section 7.9.13 for both mean QALYs and mean holistic valuations. For QALYs the differences were significant, although for holistic results they were not.

Responses were also tested at an individual level for logical consistency in the ordering of health profiles. The results are encouraging, with high proportions being either strongly or weakly consistent for each method of valuation (see Table 7.13). The QALY method always resulted in a higher level of strong consistency. It also gave greater weaker consistency, but when the weakly consistent results were included the differences between the two methods were less marked. When the weakly consistent were included 88.1 – 93.2% of responses were consistent for the holistic method compared to 93.2 – 94.9% for the QALY method. Non-consistent responses were non-consistent by between 0.003 and 0.4 for the QALY method, and between 0.1 and 0.45 for the holistic method.

The changes between profile values tend to be more marked for the QALY valuations than the holistic valuations, as is demonstrated in Figure 7.4. The differences in utility between profiles ending in the mild state and those ending in the moderate state (*i.e.* treatment unsuccessful) were more drastic for the QALY method.

Table 7.13 shows that strong consistency is achieved by 30.5 – 32.2% of respondents by the QALY method for most of the comparisons. However, it is interesting to note that 94.9% of respondents achieved strong consistency by the QALY method for the comparison between Moderate → Surgery → Mild and the version of this profile incorporating risk. There was no equivalent increase in consistency for by the holistic method for this profile, or for the equivalent comparison for sclerotherapy. Both the risky profiles incorporate risks of recurrence, which are far greater for sclerotherapy. However, the surgery profile incorporates a small risk of death.

It is perhaps a little damning for both valuation methods that, for the most part, a relatively low proportion of respondents reached high levels of strong consistency. The QALY scored better, but still achieved strong consistency in less than a third of responses for most comparisons. There were high levels of weak consistency (*i.e.* high proportions of respondents rated the pairs equally) for both methods. This suggests the possibility that the scales used in this study were not sensitive enough to detect the small differences in HRQoL involved. This will be discussed further in the Discussion section of this chapter.

7.10.12 Respondents' unwillingness to trade

A total of 21 (35.6%%) respondents were unwilling to trade any of their life expectancy for improvements in health as represented by the health states and profiles. This was defined by consistently giving a value of over 18 years for all the TTO valuations with which they were presented. These respondents would choose Choice B (full health) if they would not have to trade any of their life. However, the next choice involved a trade-off of two years, and these respondents chose Choice A (the health profile being valued) rather than making this trade-off. They therefore gave an indifference point of 19 years, and this value was used in the above analysis.

This sub-section examines the effect of excluding those respondents who were unwilling to trade on the sample statistics. The results for the 38 respondents who were willing to trade are set out in Table 7.14. Readers should bear in mind the caveat that the sample is relatively small at only 38 individuals, and the resulting statistics should be viewed with appropriate caution.

The first obvious point to make is that the mean values for the health profiles from both valuation methods are lower as a result of excluding the 21 individuals who gave an indifference value of 19 (Table 7.14). This difference from the original mean values ranges from 0.78 to 1.22. There are no significant differences in the results from the two valuation methods, and this time both the paired t-test and the Wilcoxon-sign are in agreement. The difference between the mean values from the QALY and holistic methods are greater (compare Tables 7.10 and 7.14). The confidence intervals, SDs, and (in most cases) the IQRs are wider for this smaller sample. The ranges are similar between this sample and the whole sample.

The results of this analysis demonstrate the possible effects of unwillingness to trade on economic evaluations. The author suggests that the level of unwillingness to trade in this sample was high at over one-third (although this level may have been due to the insensitivity of the methods – see Discussion). The differences in mean values between the whole sample and the sub-sample of individuals who were willing to trade are relatively large, and are greater than the suggested MEID of 0.05.

7.10.13 Sensitivity analysis

Sensitivity to time of recurrence

A sensitivity analysis was applied to the two profiles incorporating risk, and it was found that the mean values of the QALY profiles were higher the later the recurrence was expected to occur (Table 7.15). This is what would be expected, because HRQoL would be better for longer.

Differences between QALY valuations and holistic valuations were not significant no matter when the recurrence was expected. However, the QALY values for recurrence approximately 1 year after treatment, halfway through the remaining life expectancy, and approximately one year before death were significantly different according to the t-test ($p < 0.01$). This finding was verified by the Wilcoxon-sign test, which found significant differences for all values except between Moderate-Surgery-Mild (risk) with recurrence near the beginning and recurrence halfway.

Mean QALY values are higher than mean holistic values for the profile Moderate → Surgery → Mild (risk) whether the recurrence occurs at one year, middle of the period, or at the end of life. However, for the Moderate → Sclerotherapy → Mild (risk) profile the mean QALY valuation is lower than the mean holistic valuation if the recurrence takes place at one year after treatment, but higher if it takes place in the middle of the time period or towards the end of life. As stated in the previous paragraph, however, these differences between the results from the two methods are non-significant.

These results show that the time of recurrence is important to the QALY method, and its timing can have an effect on the results of economic evaluations. However, readers should be reminded that the timing of recurrence was not stated; it was merely stated that the veins would recur some time in the next 19.5 years after treatment. The QALY results were calculated by incorporating probabilities of recurrence with the timing occurring as shown in Table 7.15. The timing of recurrence is totally ambiguous for the holistic method. Whatever consideration respondents gave to the matter is unknown. Perhaps this is fair, because it is not known when recurrence would occur.

Sensitivity to treatment process

Table 7.16 shows the results of adjusting values for sclerotherapy and surgery to 1 and zero alternately on QALY valuations of health profiles. The results show the maximum and minimum values for each profile as relates to the treatment process content. For the profiles involving sclerotherapy, the mean profile values when sclerotherapy is given the value of 1 are 0.02 greater than the values when sclerotherapy is given a value of 0. For the profiles containing surgery, this difference is 0.12. Thus the mean effect of sclerotherapy on QALY valuations will be within 0.02, and that of surgery will be within 0.12.

The range for sclerotherapy is below the MEID of 0.05, and therefore this study could lack sensitivity to detect important differences in HRQoL resulting from different treatment processes.

7.10.14 *Effects of discounting QALY valuations for time*

Table 7.17 demonstrates the effect of discounting the QALY valuations of the profiles using the four discount rates mentioned in Section 7.9.16. The results show that the choice of discount rate has a profound effect on values for the health profiles. The greater the discount rate, the lower the values of the profiles. When QALY values were discounted by the rate of 0.035 recommended by the UK government, the results were reduced from the range of 16.77 – 17.59 to the range of 12.33 – 12.93. All the positive discount rates demonstrated cause discounted QALY valuations to be lower than holistic valuations even when the undiscounted QALY valuations were higher. It is clear from the values shown in Table 7.17 that the discount rate chosen could have drastic effects on cost-effectiveness analyses. All discounted values for QALY valuations were significantly different from holistic valuations of the same profile for both the paired t-test and the Wilcoxon-sign test ($p < 0.05$). There were no significant differences between the two methods by the t-test before discounting. It is assumed that discounting occurs inherently within the holistic valuations.

7.10.15 *Analysis by treatment group*

Nineteen (32.2%) of the 59 respondents included in the analysis answered "yes" to the question of whether they had already received treatment for their varicose veins. Forty (67.8%) had either not received treatment yet, or had received treatment years previously or received alternative treatment. Table 7.18 and Figure 7.5 show the profile values in these categories.

It is notable that the treated group gave average values below the non-treated group for most profiles. The mean health profile values for the treated group ranged from 15.20 – 16.54, compared to a range of 17.51 – 18.08 for the non-treated group. The holistic values are higher than the QALY values in the treated group. However, as there are only 19 in this group, the statistical tests may not be reliable. The variance is higher in this group, as indicated by the SDs and IQRs (Table 7.18).

The untreated group (n = 40) has higher mean values for some of the profiles than the mean values for the whole sample (compare Tables 7.10 and 7.18). However. the rankings of preferences are the same as in Table 7.10. If the mean QALY value is higher than the mean holistic value for a profile in Table 7.10, it is also higher in Table 7.18.

In the case of the treated group (n = 19), mean profile values are generally lower than for the entire sample. There are also reversals of preferences for five out of seven of the health profiles, such that the mean value of the holistic valuation is greater than that of the QALY, whereas the reverse was so for the entire sample (Tables 7.10 and 7.18).

It should be stressed that these results must be viewed with caution since the sub-samples of treated and non-treated respondents are small.

7.10.16 Patient comments

Patients were given the opportunity to write their comments on the back of the questionnaire. Patients' comments are reproduced in Table 7.A.9 (Appendix 3). A total of 28 (42%) out of the original 67 patients wrote comments. The comments are categorised with the number of respondents in each category as follows:

- General comments or suggestions on how they would have done the research (7)
- Initial difficulty in understanding the task (3)
- Overall difficulty with or criticism of the methods (5)
- Difficulty in imagining hypothetical states or profiles (2)
- Expressing interest in or general praise of the exercise (5)
- Expressing preference for surgery (6)
- Expressing preference for sclerotherapy (2)

Below are listed some comments or characteristics of individual respondents. which were not written in the comments section but were noted down by the author.

ID 15 kept saying that she was sticking to her opinion as she went through the valuation procedures. This may suggest that she was using an anchor point from which she was reluctant to budge. Her valuations for all the health states were 19, and also for one of the profiles. However, for the rest of the profiles her valuations were 17.

ID 18 left near the beginning of the interview, stating that she could not think about life in terms of TTO.

ID 20 had great difficulty understanding the TTO procedure. She had very severe veins that had been treated several times previously, but she said that the doctors were unwilling to treat them further because they were so severe. Many of the state/profile descriptions were difficult for her to imagine. For example, her veins had been severe for so long that it was difficult for her to imagine having mild veins. She also found it hard to imagine going through the treatment processes, because she knew she was not going to. She was 75 years old, and found it difficult to imagine living another 20 years.

ID 19 clearly had a preference for surgery over sclerotherapy. She was willing to trade off most of her life for the profiles containing sclerotherapy, because she did not want to take a treatment that would not work. It is unclear whether she realised that she was saying she would rather die a lot sooner than have sclerotherapy. It is also possible that she was deliberately answering strategically as she knew the questions were hypothetical.

ID 28 wanted to receive surgery, because she thought it was more likely to succeed. However, she rated both treatments equally on the visual analogue scale. She explained that she was rating each of the treatments at that score for different reasons. She felt that sclerotherapy would be less intrusive. It may be that she was bringing her own knowledge to bear on her initial opinion, but when she was asked to rate the two descriptions which did not contain information on chances of success, she ignored other information.

ID 37 was aged 78 years. He appeared to have difficulty understanding the questionnaire. In his valuations of the health states he seemed to be basing his answers in some way on the fact that he would not live for 20 years. This was indicated by his verbalisations as he completed the tasks.

The difficulty that some respondents had with the questionnaire will be discussed further in the Discussion section of this chapter.

7.11 Discussion

7.11.1 A critique of the strengths and weaknesses of the study

One of the problems confronted in this study was the question of how short-term treatment phases should be valued. The other health states and health profiles were valued by TTO. However, as already discussed, the TTO process for valuing short-term health states would have added complexity to an already complex questionnaire. The MVH transformation was used, whereby VAS values were given to the two treatment processes of sclerotherapy and surgery, and these values were transformed to TTO values. The concern was that the transformed values may be prone to errors, and may not be a true reflection of preferences. The MVH Group suggested that there might be an absolute mean error of 0.065 (MVH, 1995). This is relatively large, and is greater than the MEID of 0.05. However, the mean difference between VAS and transformed TTO values for sclerotherapy for this study was 0.017, and for surgery it was 0.006. There was a problem with scaling, in that some values were transformed to less than zero or greater than one. However, this was only the case for eight respondents in the case of sclerotherapy, and four respondents in the case of surgery. Overall, the transformation seemed reliable.

A possible criticism of this study is that the holistic valuations of the health profiles did not include duration of treatments (see the questionnaire in Appendix 3). The holistic profiles incorporated six months in the pre-treatment state and 19.5 years in the post-treatment state. However, for the QALY valuations the duration of the treatment was taken into account. Thus the QALY profiles incorporated six months in the pre-treatment state, 0.12 years for surgery or 0.02 years for sclerotherapy, and either 19.38 years (surgery) or 19.48 years (sclerotherapy) in the post-treatment state. The holistic profiles were therefore slightly inaccurate. The question is, to what extent would this have affected the valuations of these profiles. Since sclerotherapy took 0.001 of the 20-year profile, and surgery took 0.006 of the 20-year profile, these figures indicate the degree to which it might be expected that the holistic valuations might be inaccurate.

thus suggesting a reassuringly small scale of error. The sensitivity analysis carried out in Section 7.10.13 indicated that, for the QALY method, the maximum affect of including the processes sclerotherapy and surgery were 0.02 and 0.12 respectively. The purpose of valuing health profiles holistically is to determine whether the short-term treatment phases effect valuations in a way that is not predicted by the QALY algorithm. When valued holistically, such a profile might be given a greater, lesser. or equal value to the QALY valuation based on preferences over the treatment phase of the profile. It is possible, for example, that the short-term phase for sclerotherapy could be viewed with such horror by a respondent who was phobic of needles and imagined this would be a ghastly process that the entire profile could be given a very low value. This result would not be predicted by the QALY algorithm. However, since there is evidence that people tend to use heuristics to make judgements (Lloyd and Hutton, 2002), it would not particularly be the expectation that each respondent would make arithmetic calculations of the precise durations of each part of the health profile and use these to calculate a value for the profile when valuing it holistically. Indeed, the results from Chapters 5 and 6, tend to support this, showing that respondents did not value health profiles proportionately to the amount of time spent in each IBS health state when asked to value health profiles holistically. It is therefore considered that the error in not including the duration of the treatment phases explicitly in the holistic health profile valuations would not have led to significant errors in valuations. In other words, it is considered unlikely that, had the post-treatment stage for sclerotherapy been sited as 19 years and 6 months minus one week, the values given to the profiles would have been significantly different from those given under these circumstances.

Table 7.11 shows that the maximum mean difference between QALY and holistic profile valuations on a scale of 0-1 is 0.034. The sample size was chosen based on the power to detect a difference of 0.05 (which was considered a reasonable estimate of an MEID) at the 95% probability level. This study was therefore underpowered for the detection of differences of the lower magnitude that are reported in Table 7.11. In order to have the power to detect a difference in value of 0.034, a sample size of at least 119 would have been required. If the MEID is 0.05, differences of less than this value are not considered economically important, and therefore the sample size limitations are irrelevant in this respect. However, if the MEID of 0.05 was to be questioned, and it was desirable to explore lesser differences, the difficulties in recruitment during this study would come into play as relevant limiting factors. The difficulties in recruitment are discussed further in Chapter 9, and are demonstrated by the low response rate for this study reported in Section 7.10.1. There were time limits to the study, and under such circumstances it would have been implausible to extend the sample size. This demonstrates the possible effects of recruitment difficulties.

A strength of this study was the level of care taken to ensure that the descriptions used in the hypothetical health states and profiles had construct validity for respondents. In other words, care was taken to ensure that the health state and profile descriptions were realistic portrayals of life with varicose veins. The use of patient focus groups, the survey of health professionals, and literature sources in the construction of the health states and profiles were important factors in this. The success of the questionnaire pilot indicates a reasonable degree of construct validity.

There were possible framing issues with the way in which health profiles were presented. During the health state valuation process the descriptions of health states were presented in full in the TTO exercise. However, during the holistic valuations of the health profiles, the descriptions of the constituent health states were shortened to the title of the health state. A full description of the process of treatment was given. The pieces of paper containing the health state descriptions, which were used to rank the health states prior to their valuation, were available so that respondents were able to remind themselves of the health state descriptions during the valuations of the health profiles. However, it is possible that there may have been issues to do with the framing of the valuation questions. These issues may have involved, for example, respondents not referring back to the original descriptions of the health states while valuing the health profiles, but using their own mental descriptions of the health states of severe, moderate and mild varicose veins. However, the way in which the health profiles were presented led to the health profile valuation exercises being more accessible and simple in format. This is an important factor, because the amount of information the human mind can process at one time is limited (Lloyd and Hutton, 2002).

It would have been useful to rank the health profiles before the valuation procedure. This would have enabled a comparison of ranking between the direct ranking of the profiles and the ranking implied by the holistic valuations, and thus a test of convergent validity. This direct ranking exercise was excluded because of the relatively complex nature of the health profile descriptions.

At the beginning of the interview respondents were asked to complete a tickbox classification to describe their current varicose veins symptoms (see Table 7.5). This enabled categorisation of respondents' current health in terms of varicose veins symptoms. These data were used later to classify respondents as having mild, moderate, and severe varicose veins. The classification system was based on the descriptions of varicose veins used for the health states. Not all the possible combinations of symptoms and levels of symptoms were included in the table. It may seem on the face of it that valuable information was potentially lost due to respondents being potentially unable to classify themselves accurately. However, with the relatively small sample size in this study, the wider range of possible combinations of symptoms would have led to each member of the sample having a different combination of symptoms. It would have been difficult to classify in a meaningful way, and the small numbers of respondents in each classification group would have mean that statistical tests were unusable. It was not the aim of this study to develop a fully-fledged health state classification system, but rather to provide a simple way to roughly classify respondents.

A total of 10 (16.9%) respondents made written comments suggesting that they had some level of difficulty with the questionnaire. This ranged from initial difficulties in understanding to overall difficulty with the methods and difficulty in imagining the hypothetical tasks. Out of the entire sample of 67 who began the interviews, three people had difficulties with the process. One person left near the beginning of the interview on the grounds that she could not think about life in terms of trading off time. The other two respondents expressed verbally that they found the questionnaire difficult and confusing (see Section 7.10.16). This is a relatively high percentage of those who began or completed the interview who had some level of difficulty with the tasks. The TTO is a commonly used method of valuing health states, and if there is a percentage of the population who would have difficulty using this method, then their preferences may not be represented. However, this is a problem which is not only relevant to the present work, but to health economics evaluations in general. The other difficulties encountered may have been more specific to this study. For example, these respondents may have been referring to the TTO method, or they may have been referring to the cognitive demands of reading, understanding, and valuing the health states, treatment processes, and health profiles. The comments lacked enough specificity to be sure.

7.11.2 Categorisation of health states

There are a number of possible ways to classify varicose veins (Michaels *et al*, 2001). These include the following:

- Those used in this study, which were developed from focus group discussions. However, there was no access to patients with ulcers, because these patients were excluded from the clinical trial.

- The Michaels *et al* clinical trial, which used an anatomical classification based on whether the varicose veins are above or below the knee, size of varicose veins, and whether there is reflux up certain veins.

- The DEC report, which was developed along clinical manifestations. They have Asymptomatic, Mild, Moderate, and Severe health states.

The Mild, Moderate and Severe used in this study are linked most closely with the DEC Asymptomatic, Mild, and Moderate. The reason for this is that Asymptomatic are people who present at GPs but are too mild to be referred. The present study opted to include the whole range of varicose veins symptom levels rather than just those who have been referred to clinic. However, it is probable that the severe state was not fully incorporated because patients with ulcers were excluded from the trial. Indeed the patients who most closely fitted into the severe health state according to their self-reported current symptoms gave a TTO rating of their current health of almost equal to the average rating of the moderate state (Table 7.9). However, the numbers were too small to do statistical comparisons.

7.11.3 Are varicose veins just cosmetic?

As discussed in Section 7.2.1, the low priority given to the treatment of varicose veins by the NHS reflects the impression that they are merely cosmetic defects rather than health problems that affect HRQoL. If they cause negative effects on HRQoL, one might expect there to be some indications from health status measures and valuation tests.

Section 7.10.4 describes the health status of the sample using the general health item from the SF-36. Respondents rated their health as excellent, very good, good, fair, or poor. These results indicate that overall this sample enjoyed a healthy life. Only 13.4% of the sample rated their general health as fair or poor. The results from this sample were similar to the results from a large sample of 1582 patients randomly selected from two GP lists in Sheffield by Brazier *et al* (1992). As mentioned in Section 7.10.4, this GP list sample from Sheffield were comparable to the Sheffield general population in terms of use of health services, indicating that their general health was representative of the population. These findings indicate that varicose veins does not adversely affect general health as measured by this SF-36 item. This lends support to the view that varicose veins, in most cases, are mere cosmetic defects.

One of the concerns that this study aimed to address was the possibility that generic measures such as the SF-36 may not be sensitive to detect condition-specific effects on HRQoL. In particular, they may not be sufficiently sensitive to detect effects that may be relatively small, but nevertheless important. In order to address this issue, respondents were asked to rate their current health plus hypothetical health states describing mild, moderate, and severe varicose veins using the TTO. The results are described in Section 7.10.8, and show that the three hypothetical health states relating to varicose veins were all rated below full health, and the differences between each health state valuation and full health were highly significant ($p < 0.001$). The same was also the case for valuations of respondents' current health state. This raises to the surface a discrepancy between the results of the TTO valuation of current health and the results from the general health item of the SF-36. The mean value of current health for this sample was 0.86. The differences between the health states valued (including the three hypothetical states plus current health) and full health ranges between 0.12 for the mild state and 0.21 for the severe state (Table 7.8). These differences are greater than the commonly used MEID of 0.05.

The health state valuation data suggest that varicose veins can have an economically important adverse effect on HRQoL, which may not be detectable by the SF-36. This refutes the suggestion that they are merely cosmetic defects.

Members of the present sample had been referred to the hospital by their GPs, and the sample therefore did not include asymptomatic levels of varicose veins according to the DEC classification (see previous section). It is probable that the sample used in this study were not representative of the majority of people with varicose veins, because they were a select group who had varicosities of a severity great enough for their GPs to refer them to clinic. However, the hypothetical health state referred to as "mild" most closely fitted the DEC asymptomatic category, and this sample gave this state a mean value of 0.88 with a mean difference of 0.12 from full health (Tables 7.7 and 7.8).

7.11.4 Differences between QALY and holistic valuations of health profiles

According to the t-test there were no significant differences between the means for the profiles derived from the holistic and QALY methods of valuation (Table 7.10), and this result was not affected by when recurrence was assumed to occur for the QALY valuations of the two health profiles containing risk. This initial result would suggest that, in this instance, QALY and holistic results give the same results and therefore the QALY, as the simpler method to apply, would be the method of choice.

However, there are differences in valuations within the results from each valuation method. Whereas the results of the QALY valuations are, for the most part, significantly difference from each other across the different health profiles, those from the holistic valuations are not. In other words, if the QALY algorithm is applied each health profile is valued differently to the other profiles, and preferences can be inferred between the different health profiles. However, if the results of the holistic method are taken, the implication is that there are no preferences over the different health profiles. This latter finding is possibly the result of insensitivity in the method, and this is discussed further in the following section.

7.11.5 Insensitivity of the valuation methods

This study provides some evidence to suggest that the valuation methods used may have been insensitive to the levels of disutility described. As stated in a previous section, the TTO appeared to be more sensitive to the disutility of varicose veins health states than the general health item of the SF-36. However, the test of convergent validity for health states compared the implied ranking of the TTO valuations of health states with the original ranking of the health states (Section 7.10.6) and showed that 16.9% of respondents were strongly convergent, 69.5% of the sample were weakly convergent, and 13.6% were non-convergent. The large majority of the sample, therefore, valued health states equally when they were not originally ranked equally. Thus, in a choiceless situation one state was preferred, but no preference was expressed when a trade in life years was the mode of expressing preference.

A lack of sensitivity was also demonstrated in health profile valuations. Although valuations of health profiles follow a logical ordering by the holistic method, the differences between profile valuations were not significant. These non-significant findings suggest that the holistic method picks up no preferences over treatment, and also no preferences between end states, which is more remarkable. One would expect a profile ending in a mild state to be preferred to the equivalent profile ending in a moderate state.

These findings seem an unlikely reflection of reality. In ranking the states, respondents showed an ordinal effect, with only five ranking them in an illogical order. However, out of the 44 (74.6%) respondents who showed convergent validity for states, only 17 (38.6%) implied an ordinal ranking from their TTO valuations of the states. The remainder rated the states equally. This indicates that, for the health states, TTO was not sensitive enough to detect the ordinal effect. The health states were ranked in a sacrifice-free context, and there was no cost to showing a preference for one state over another. However, it may be that the patients used in the study sample did not wish to trade years of life for the health states described. This possibility was confirmed by some of the patients' comments outlined in Table 7.A.9:

> "...from a health point of view none of the scenarios presented was sufficiently bad to give up any years of life."
>
> "No matter what pain/swelling/unsightly veins my kids and family are worth fighting for."

Although the QALY means were not significantly different from the means of the holistic valuations, there were differences between the valuations of the different profiles for the QALY method. These findings seem to make more sense than the findings of the holistic method. For example, profiles Moderate-treatment-Mild are preferred to profiles Moderate-treatment-Moderate. According to the QALY method, sclerotherapy seemed to be slightly preferred to surgery when risk was not taken into account (Table 7.10). Moderate-Sclerotherapy-Mild was preferred to the same profile containing risk, and similarly for the Moderate-Surgery-Mild profiles. When risks were taken into account, surgery seemed to be preferred to sclerotherapy. This was probably due to the higher chance of recurrence with sclerotherapy (see below for discussion of risks).

Tests of logical consistency in the implied ranking of the valuations of the health profiles are reported in Table 7.13 and Section 7.10.11. The levels of strong consistency are relatively low for both methods, thought the QALY performs consistently better in this respect.

Despite the discomfort suffered by varicose veins patients, the findings of this study suggest that varicose veins may be too mild to warrant trading off life years. The TTO may therefore be too crude a tool to measure differences between health states that lie within the top 5% of the scale. Indeed a significant proportion of respondents (35.6%) showed unwillingness to trade. This unwillingness to trade in life years might not imply that the health states and profiles described had little disutility attached. It is possible that this sizeable group believed that the states and profiles would lead to a greater level of disutility than they indicated with their TTO scores, but the TTO method may not have been a suitable assessment technique for measuring the preferences of these respondents. As a comparison, for the study of IBS patients in Chapter 6 only two out of 49 (4.1%) of respondents were unwilling to gamble over the IBS health states and profiles.

One way of getting around this problem could be to use a process of chaining. Varicose veins health states could be valued against a worse state such as leg ulcers. The state of leg ulcers would then be valued against full health (for the TTO method). Problems associated with chaining have already been discussed in Chapter 4.

Another approach to exploring sensitivity to valuations of mild health states and profiles would be to offer respondents the option of trading in smaller chunks of time than were used in this study. This study used a scale of zero to 20 years, with intervals of two years. A follow-up study could elicit values for health states and profiles using a scale that allowed respondents to trade off in days, or even hours or minutes. The advantage with the scale in this study was that it could be printed off in a booklet questionnaire that each respondent could complete independently as long as the interviewer was on hand to offer explanations as required. However, there may be practical implications of increasing the sensitivity of the scale in the way just suggested. If this led to a more interviewer intensive valuation session, fewer respondents could be interviewed in one session and the study would become more expensive and time consuming. Another possibility is that, if the differences between the values of the profiles were so small as to be well below the MEID, it may not be worth knowing the exact differences. However, this would be an interesting way to further explore the issues of sensitivity in the holistic measurements.

The analysis of a sub-sample of patients entered into the Michaels *et al* clinical trial revealed no significant differences between EQ-5D scores before treatment and one month after treatment. One of the concerns this present study aimed to address was whether generic measures such the EQ-5D are sensitive enough to pick up disutility associated with comparatively mild conditions such as varicose veins. The patient focus groups, the health professional questionnaires, and the study by Garratt et al (1993) all suggest that varicose veins cause enough disutility to be important to patients. Yet the sub-sample from the Michaels *et al* clinical trial would suggest that treating the veins makes no difference to HRQoL, at least within a month of treatment. It would be interesting to determine whether ratings of health improved after this time, because it may be that patients were still suffering from post-treatment discomfort one month after treatment with either surgery or sclerotherapy.

This study used condition-specific health states based on the way patients themselves describe their varicose veins in order to overcome the insensitivity of generic measures. The results of the holistic valuations of profiles used in this study are insensitive to different levels of quality of life with varicose veins. However, this is not so with the QALY results, which follow a logical order. The holistic results were similar to the finding of Chapter 6, which found that respondents were unwilling to value IBS profiles below around 0.95.

If, as suggested above, the lack of sensitivity to the disutility described in the varicose veins health profiles was due to the unwillingness to trade off life years for such small improvements in health (because the condition is so mild), the question remains as to why the holistic method appears less sensitive than the QALY method. After all, they both use the TTO: the holistic method uses it directly to value an entire profile, and the QALY uses it to value the constituent health states.

It could be a similar phenomenon to that shown in Chapters 5 and 6, in which the holistic method used in those studies showed a lack of sensitivity to proportion of time spent in each health state for the profiles. The valuations of the constituent health states are simpler by nature. They involve imagining the rest of one's life in one health state, and giving it a TTO value. These values are then entered into the QALY algorithm for each health profile. The holistic method, however, involves giving a TTO value for the profile as a whole. It may be a case of using heuristics such as the general gist to aid decision-making. Since the valuation decision for the holistic method is more complex than that for constituent health states, it would be more prone to heuristics.

A willingness to pay (WTP) study might be more appropriate in this case, because of the unwillingness to trade years for such a mild condition. Respondents might find it more acceptable to value the profiles in terms of how much they would be willing to pay for improvements than giving up years of life for these improvements. Indeed the Michaels *et al* trial opted for a WTP method. However, the health states used were less subjective than the ones used in this study.

7.11.6 Importance of how risk is incorporated into health profiles

In five out of the seven health profiles valued in this study risk was not mentioned or taken into account. In two of the profiles, risks of mortality and recurrence were incorporated. The chance of recurrence is greater for sclerotherapy than surgery (75% versus 20% respectively), but there is approximately a 1/10,000 risk of death under the anesthetic for surgery. These risks are realistic for the scenarios concerned. The profiles which incorporated risk were Moderate → Sclerotherapy → Mild and Moderate → Surgery → Mild.

There was a key difference between the QALY and holistic method in the way that risk was incorporated in the health profile valuations. For the QALY valuations respondents valued the constituent health states and short-term treatment phases, and their values were entered into the QALY algorithm alongside values for risks in each stage of the profile (see equations 7.5 to 7.10). This way of incorporating risk is *ex post* (see Chapter 4). In other words, the risk is entered into the equation after the valuation process. Risks are treated as objective probabilities, which will be the same for all respondents (all other things being equal). Subjective attitudes to the possible adverse risks associated with parts of the health profile are not dealt with.

In the case of the holistic method, respondents value the whole profile, including any risky aspects. This is dealing with risk in the *ex ante* perspective. In other words, respondents are introduced to the risks before the valuation process, and allowed to make their own subjective judgements.

The only significant difference between holistic valuations of profiles was for Moderate-Surgery-Mild without risk and this profile containing risk. The risks involved were the risk of dying under the general anaesthetic and the risk of the veins recurring (see above). The inclusion of risk into the profile resulted in a mean holistic valuation of 0.57 healthy years equivalent less than when risk was not included (17.37 versus

16.80, see Table 7.10). Alternatively, this can be translated to a mean drop of 0.029 on the 0 to 1 scale (see Table 7.11). It should be noted that this lower value due to the inclusion of *ex ante* risk is less than the MEID of 0.05, although it was a statistically significant difference.

When the risk of recurrence was added to the sclerotherapy profile, there was a drop in mean holistic valuations of 0.13 healthy years equivalent (Table 7.10) or 0.007 on the 0 to 1 scale (Table 7.11). The risk of recurrence was sited as 75% for sclerotherapy, and 20% for surgery. The greater decline in the mean holistic value for the surgery profile incorporating risk therefore suggests that it was the risk of death (1 in 10,000) that caused the drop in the mean value.

According to the holistic valuations, the non-risky surgery and sclerotherapy profiles leading to the mild state were valued equally, but when risk was added the sclerotherapy profile was preferred.

The incorporation of the *ex post* risks to the QALY algorithm has an interesting effect on QALY valuations of the profiles in question. Before the incorporation of risk, the sclerotherapy profile is preferred by 0.03 of a year (Table 7.10) or by 0.002 on the 0 to 1 scale (Table 7.11). However, upon incorporation of the risk factors, there is a reversal of preferences over treatment. The risky surgery profile is rated higher than the risky sclerotherapy profile by 0.18 out of 20 years (Table 7.10) or 0.009 on the 0 to 1 scale (Table 7.11). Although these differences are small, they are statistically significant (Table 7.12). The differences between the sclerotherapy profiles without and with risk are 0.3 (Table 7.10) or 0.015 on the 0 to 1 scale (Table 7.11). For the surgery profiles these differences are 0.09 (Table 7.10) or 0.004 on the 0 to 1 scale (Table 7.11). The drop in mean value is therefore greater for the sclerotherapy profile. This indicates that, when risks are incorporated *ex post*, the risk of recurrence is the more important factor. This is not surprising, because the risk of recurrence for sclerotherapy is so much higher than the risks of recurrence or death for surgery.

These results clearly show that the way in which risk is incorporated is important to the results of economic evaluations. If the QALY method is used to value these health profiles, the result of incorporating risks would suggest that surgery was the preferred treatment option. However, according to the holistic valuations of these risky health profiles, sclerotherapy is the preferred option.

These findings suggest that *ex ante* risks do play a part in decision-making. However, the differences between the surgery and sclerotherapy profiles in which risks were incorporated were greater for the holistic method, which may suggest that *ex ante* risks might be more heavily weighted than is allowed for in the QALY algorithm. The finding that QALY values for the surgery profile containing risk were valued higher than the sclerotherapy profile, whereas the holistic value for the surgery profile was valued lower than the sclerotherapy profile (Table 7.10) is also an important finding. This lends further credence to the evidence that *ex ante* risk is an important factor in patient decision-making. This factor is not commonly taken into account by the QALY algorithm.

7.11.7 Discount rates

QALY valuations were greatly affected by the discount rate used (Table 7.17). The higher the discount rate, the lower the values given to the profiles. The discount rate currently recommended by the UK government is 0.035 (HM Treasury, 2003). As shown in Table 7.17, applying this discount rate to the QALY values leads to preference reversals, such that the holistic valuations are significantly higher than the discounted QALY values. This demonstrates that choice of discount rate can have a significant impact on CEAs.

It is assumed that the holistic valuation takes into account discounting. However, the comment of one respondent suggests that there may be cognitive difficulties in trying to make judgements about preferences for futuristic profiles of health:

> "The other problem is one of relating your age now with your feelings in 20 years time when this might be just one problem that ????? you".

This is a concern that has been voiced in the literature (Buckingham, 1993; Kahneman and Snell, 1990; Boyd *et al*, 1990). These concerns will be addressed in greater detail in the Discussion in Chapter 9. However, the results of using the discount rate of 0.035 suggested by the UK government is to lower the QALY values of the health profiles compared to the holistic values. If discounted QALY values are used, the implications may be that the condition would receive higher priority than if holistic values were used. It is evident that, if respondents are discounting their values for the holistic valuations, they are not using the discount rates used in the QALY algorithm. However, it is possible that respondents are not tending to discount their values due to considerations of the future values of any benefits or dis-benefits. It may simply not have occurred to them.

7.11.8 Preferences over treatment process

One aim of this study was to explore preferences of varicose veins patients for the different methods of treatment: surgery and sclerotherapy. In order to do this, the QALY and holistic valuation methods were both used in an attempt to determine the effect on HRQoL of treatment for moderate varicose veins by either sclerotherapy or surgery. Obviously either treatment could result in failure (and therefore remaining with moderate varicose veins), or success (in which case the condition would improve to the mild state).

As discussed in Section 7.10.9, differences between failure and success for surgery and sclerotherapy for each of the two valuation methods are virtually the same when profiles are valued in a riskless context. For the QALY method, the differences (on a scale of 0 to 1) between failure and success for sclerotherapy and surgery are approximately 0.04 (see Table 7.11). For the holistic method the differences for sclerotherapy and surgery are equal at 0.012. This finding is no surprise for the QALY method, which applies a mathematical algorithm. However, it is slightly surprising to find that the respondents placed an equal value on success of treatment for the two treatment methods using the holistic method. One possible implication of this is that the respondents were not excessively concerned about which treatment they received, but more about whether their treatment succeeded or failed. The differences between the health profiles in terms of the treatment involved was in the region of 0.002 according to the QALY valuations (Table 7.11).

In a riskless context, sclerotherapy was given a higher mean value by the QALY method. When the risks of recurrence and mortality were factored in, however, surgery was given a higher mean value by the QALY method. According to the holistic method, there was indifference between the two treatments in a riskless context, but sclerotherapy was preferred to surgery when the risks were incorporated.

In summary, taking into account risks of recurrence and mortality, according to the QALY method of valuation surgery is the preferred treatment option according to the sample means, and according to the holistic method of valuation sclerotherapy is the preferred treatment option according to the sample means.

7.11.9 Effects of previous treatment on valuations

The 19 respondents who had already received treatment gave lower valuations for most profiles than the non-treated group of 40. It has been suggested previously that people who suffer from a condition adapt to it, and their valuations of the associated states improve over time even when the state does not (Tsevat et al, 1995). Perhaps the treated group no longer suffered the symptoms so badly and had forgotten how well they used to cope prior to treatment. However, it is possible that respondents who had experienced the treatment had been disappointed in their expectations of improvements in HRQoL after treatment, and were rating their experiences rather than the profile presented to them. It would have been useful to conduct this comparison on a larger sample, but the problems in recruitment and the limitations to resources prohibited this.

7.12 Conclusions

This study provides evidence that varicose veins can have more than a cosmetic effect on people, and in fact can cause a significant detriment to HRQoL even at relatively mild levels of symptoms.

Initial t-tests showed no significant differences between valuations by QALY and holistic methods, although there were significant differences for three of the profiles according to the Wilcoxon-sign test.

The QALY method appeared to be more sensitive to the severity of the health profile than the holistic method. The QALY performed better in tests of logical consistency, possibly reflecting the effect of unwillingness to trade on holistic valuations.

For the profiles containing risk, the QALY valuations were sensitive to time of recurrence. The incorporation of risk caused a reversal of preferences in the QALY valuations, such that when risk was incorporated surgery was the preferred treatment of choice. Holistic valuations appeared to be more sensitive to the small risk of mortality than time of recurrence. According to holistic valuations, sclerotherapy was the preferred treatment of choice when risk was taken into account. The evidence from the holistic valuations suggests that *ex ante* risks can have a significant effect upon medical decision-making, and this is not taken into account by the QALY model.

The choice of discount rate has a profound effect on QALY valuations of health profiles. If a discount rate of 0.035 is applied, as proposed by the UK government, QALY values for the health profiles are greatly reduced, leading to holistic values being significantly higher.

Many of the differences between profile values were below the MEID of 0.05 which this study was powered to detect. However, there were difficulties in recruitment. and the resources in terms of time and funding necessary to increase the sample of respondents were not available.

An important issue is whether the QALY model can be used to infer patient treatment choices. This study offers no evidence to suggest that the holistic method is any better than the QALY method at inferring treatment choices. The values from both methods were close and there was little disagreement. However, the results from the QALY method showed greater logical consistency.

Figure 7.1 The varicose veins health states.

Severe varicose veins

☹ You have **big** veins, which **stick up** and look **lumpy** and **unsightly**. Your veins are **noticeable.**

☹ Your legs or ankles **often become very swollen**, so it's difficult to put your shoes on. Elasticated socks and stockings are **uncomfortable**, because the leg swells up around the elastic.

☹ Your legs **often ache** and feel **painful.** You **often get cramp** in your legs. You have to keep **moving around** to avoid **cramps** and **aches.**

☹ You get a lot of **irritation** and **itching** on your legs.

☹ You have to keep your **weight** down to avoid problems with your legs.

☹ You may **worry** about the possibility of getting an **ulcer**.

☹ You may find that you are **organising your life around your symptoms.**

Mild varicose veins

☹ Your veins are **noticeable**.

Your legs or ankles do not become **swollen**, so it is not difficult to put your shoes on.

Your legs do not **ache** or feel **painful**. You do not get **cramp** in your legs. You do not have to keep **moving around** to avoid **cramps** and **aches**.

You do not get **irritation** and **itching** on your legs.

You do not have to keep your **weight** down to avoid problems with your legs.

☹ You may **worry** about the possibility of getting an **ulcer**.

You do not find that you are **organising your life around your symptoms**.

Moderate varicose veins

☹ Your veins are **noticeable**.

☹ Your legs or ankles **sometimes become swollen**, so it is difficult to put your shoes on.

☹ Your legs **sometimes ache** or feel **painful**. You **sometimes get cramp** in your legs. You have to keep **moving around** to avoid **cramps** and **aches**.

☹ You get some **irritation** and **itching** on your legs.

You do not have to keep your **weight** down to avoid problems with your legs.

☹ You may **worry** about the possibility of getting an **ulcer**.

You do not find that you are **organising your life around your symptoms**.

Sclerotherapy

You will go to clinic and receive injections into your varicose veins. Each injected area will be covered with a pad, and a tight stocking will be put on your leg and left on for **48 hours**. During these **48 hours**, you are advised:

- to walk around regularly
- to do most of your usual activities, including light sports
- to avoid standing still or sitting with your affected foot on the floor for more than half an hour
- to avoid particularly strenuous activities which may result in the stocking or pads moving out of place
- to avoid having a bath or a shower and other activities which would get the stocking wet
- to take leave of absence from your job

Your legs may feel inflamed, swollen and painful for a week after the injections.

Surgery

You will go into hospital for surgery. It will be done under general anaesthetic. You will go home the same day. You are advised over the next **10 to 14 days**:

- to avoid standing still or sitting with your affected foot on the floor for more than half an hour, to avoid having a bath or a shower, to avoid particularly strenuous activities, and to avoid driving
- to take short walks at frequent intervals, to participate in most of your usual activities, and to wear support stockings

You may feel tired for the next week, and you may need plenty of rest for the next few days.

Your legs may have large bruises initially, and the scars may remain noticeable for a few months. Your legs may feel painful for 2 weeks after surgery.

You are advised to take leave of absence from your job for **3 to 6 weeks**

Figure 7.3 Predicted values for TTO for each value of VAS from 0 to 1 using the MVH transformation method.

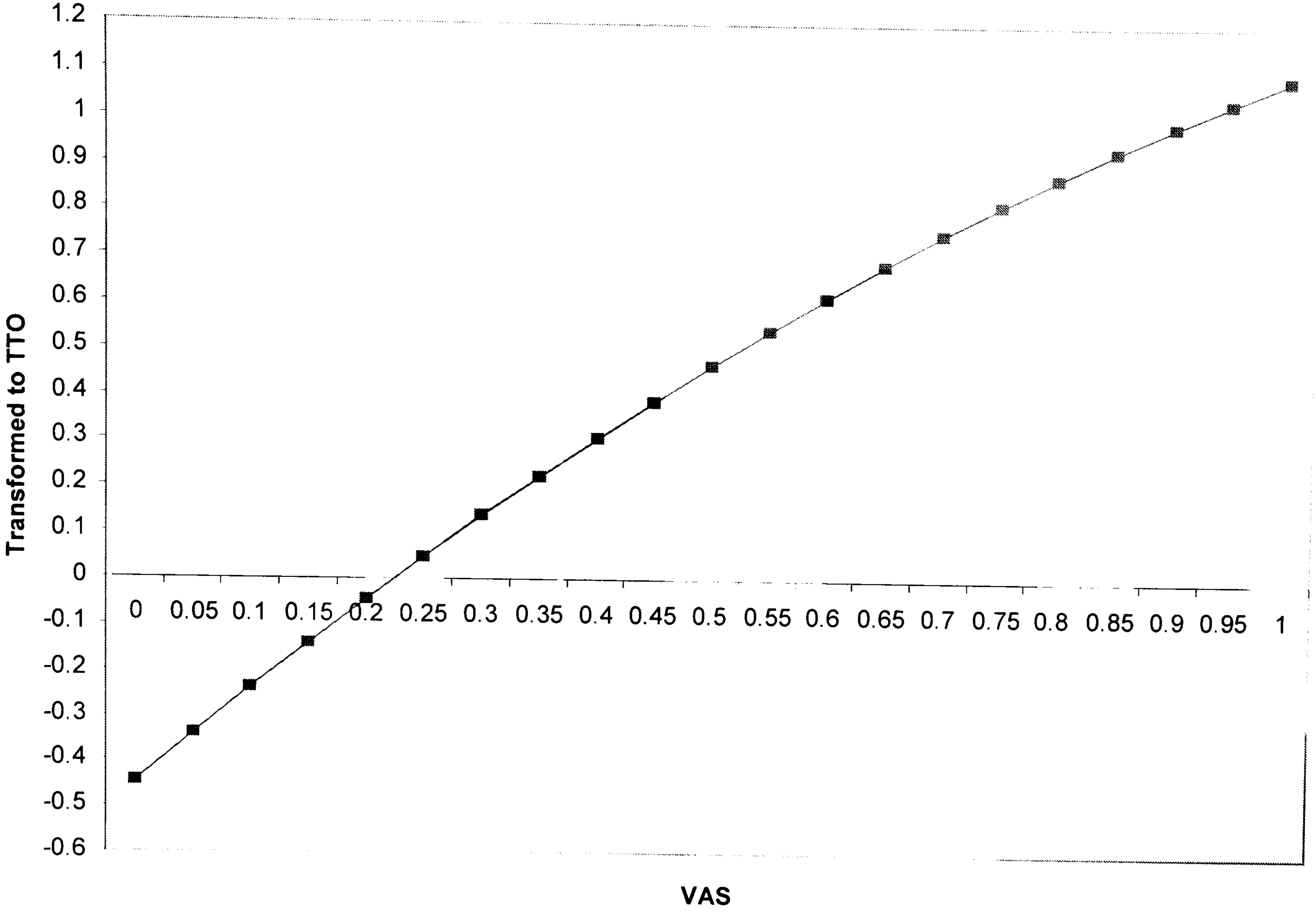


Figure 7.4 Holistic and QALY means for profiles.

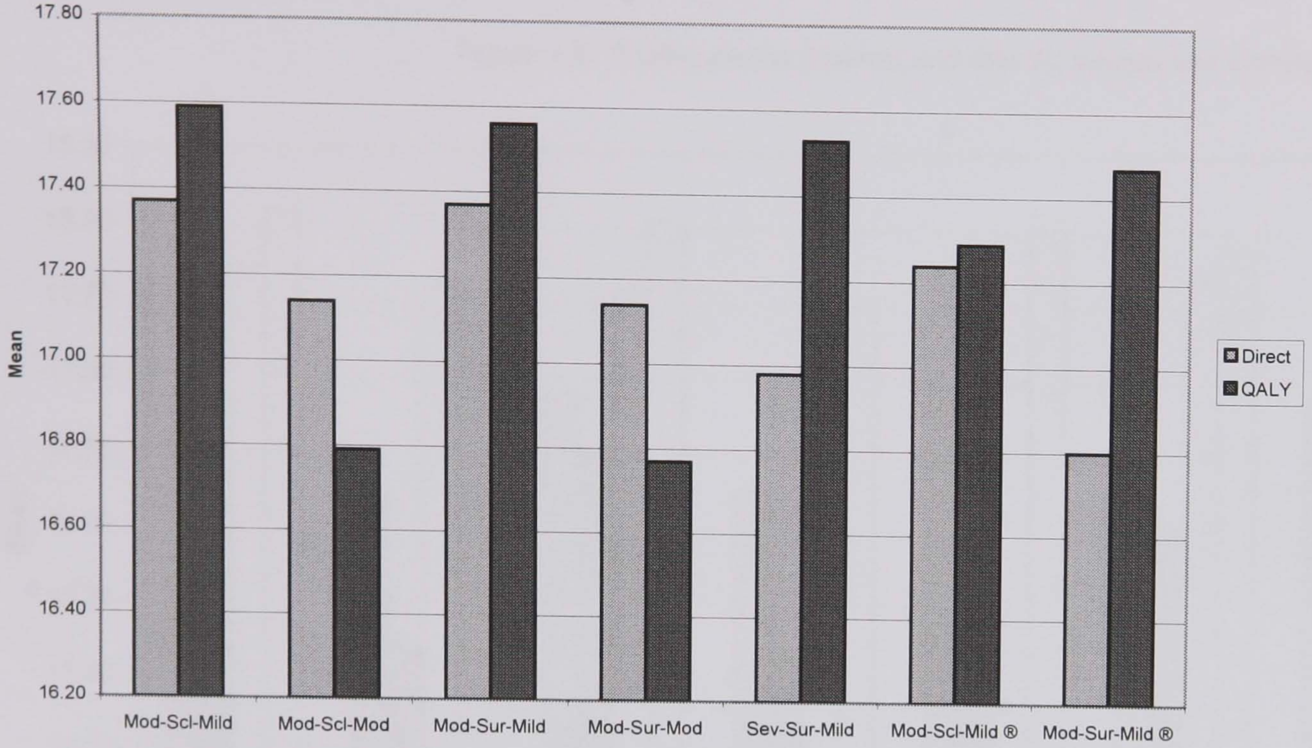


Figure 7.5 Profile means (holistic and QALY): treated and untreated.

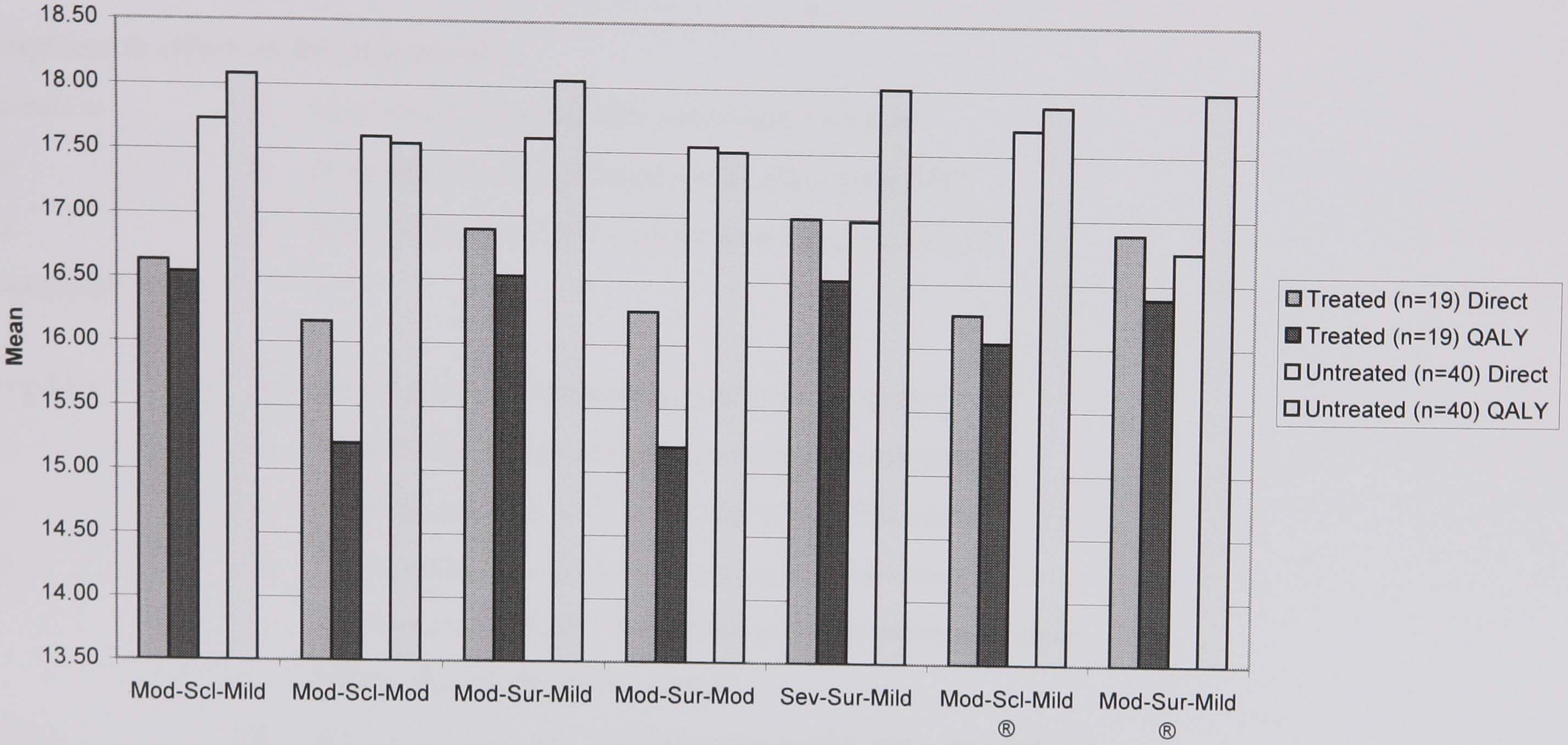


Table 7.1 Focus group interview plan for the pre-treatment group

Opening	1.	Tell us your first name, and one of your interests or hobbies outside work.
Introduction	2.	What does good health mean to you?
Symptoms & effect on life in general		
Transition	3.	How would you describe your varicose veins?
Key	4.	How does having varicose veins affect your life?
Key	5.	What is the worst thing about your varicose veins?
Treatment		
Transition	6.	Which type of treatment do you hope to receive?
Key	7.	What made you decide on that form of treatment?
Key	8.	How did medical staff assist you in your decision?
Key	9.	Do you think you have been provided with enough information about the different types of treatment, and the effects they might have on you?
Ending	10.	We are going to put together some health state descriptions for varicose veins. This discussion has been very helpful, but are there any other aspects of varicose veins which we should have talked about but didn't?

Table 7.2 Focus group interview plan for the post-treatment group

Opening	1.	Tell us your first name, and one of your interests or hobbies outside work.
Introduction	2.	What does good health mean to you?
Symptoms & effect on life in general		
Transition	3.	How would you describe your varicose veins?
Key	4.	How does having varicose veins affect your life?
Key	5.	What is the worst thing about your varicose veins?
Treatment		
Transition	6.	Which type of treatment did you receive?
Key	7.	What made you decide on that form of treatment?
Key	8.	How did medical staff assist you in your decision?
Key	9.	Thinking back, do you now believe that you were provided with enough information about the different types of treatment, and the effects they would have on you?
Ending	10.	We are going to put together some health state descriptions for varicose veins. This discussion has been very helpful, but are there any other aspects of varicose veins which we should have talked about but didn't?

Table 7.3 Matrix of possible health profiles. The shaded cells are those that were included in the study.

		Outcome			
		No symptoms	Mild	Moderate	Severe
Start state	Mild	Conservative Sclerotherapy	Conservative Sclerotherapy		
	Moderate	Sclerotherapy Surgery	Sclerotherapy Surgery	Sclerotherapy Surgery	
	Severe	Conservative Surgery	Conservative Surgery	Conservative Surgery	Conservative Surgery

Table 7.4 Opinions about treatment.

Treated	Non treated			
	Surgery	Sclerotherapy	Other	Indifferent
22	13	0	8	24

Table 7.5 Frequency of respondents' self-reported symptoms.

Respondents' self-reported current symptoms		**Yes**	**No**	**Missing**
Your veins are **noticeable**		66 (98.5)		1 (1.5)
Your veins **stick up** and look **lumpy** and **unsightly**		62 (92.5)	5 (7.5)	
Your legs or ankles become **swollen**, making it difficult to put your shoes on		17 (25.4)	49 (73.1)	1 (1.5)
Your veins **sometimes become swollen**		58 (86.6)	7 (10.4)	2 (3.0)
Your veins **often become very swollen**		38 (56.7)	26 (38.8)	3 (4.5)
Elasticated socks and stockings are **uncomfortable**, because the leg swells up around the elastic		36 (53.7)	27 (40.3)	4 (6.0)
Your legs **ache** or feel **painful**	Often 30 (44.8)	Sometimes 33 (49.3)	4 (6.0)	
You get **cramp** in your legs	Often 18 (26.9)	Sometimes 33 (49.3)	15 (22.4)	1 (1.5)
You have to keep **moving around** to avoid **cramps** and **aches**		34 (50.7)	32 (47.8)	1 (1.5)
You get **irritation** and **itching** on your legs	A lot 18 (26.9)	Some 37 (55.2)	11 (16.4)	1 (1.5)
You have to keep your **weight** down to avoid problems with your legs		27 (40.3)	37 (55.2)	2 (3.0) Not sure 1 (1.5)
You **worry** about the possibility of getting an **ulcer**		32 (47.8)	33 (49.3)	2 (3.0)
You find that you are **organising your life around your symptoms**		11 (16.4)	55 (82.1)	1 (1.5)

Table 7.6 Transformed and non-transformed VAS values for treatments.

	Non-transformed (VAS)		Transformed (TTO)	
	Sclerotherapy	Surgery	Sclerotherapy	Surgery
Mean	0.679	0.605	0.696	0.599
SD	0.212	0.192	0.290	0.283
Min	0.200	0.100	-0.046	-0.240
Max	1.000	1.000	1.087	1.087
25th %	0.520	0.500	0.496	0.466
50th %	0.700	0.600	0.749	0.613
75th %	0.840	0.730	0.920	0.788

Table 7.7 Health state valuations (n = 59 except for current health where n = 58 due to one unclear response).

	Current health	Mild	Moderate	Severe
Mean	0.86	0.88	0.84	0.79
SD	0.19	0.19	0.18	0.21
25th %ile	0.85	0.95	0.75	0.65
50th %ile	0.95	0.95	0.95	0.85
75th %ile	0.95	0.95	0.95	0.95
Minimum	0.15	0.10	0.20	0.15
Maximum	1.00	1.00	0.95	0.95

Table 7.8 Wilcoxon-sign and t-test results for comparisons between valuations of different health states.

	Difference between means	CI of mean difference	*p* (t-test)	*p* (Wilcoxon)
Mild - Current health	0.03	-0.01 to 0.07	0.140	0.213
Current health - Moderate	0.02	-0.02 to 0.05	0.365	0.073
Current health - Severe	0.07	0.03 to 0.12	0.003	0.002
Mild – Moderate	0.04	0.01 to -0.07	0.008	0.003
Mild - Severe	0.09	0.04 to -0.14	0.001	0.001
Moderate - Severe	0.05	0.02 to 0.09	0.003	0.003
Full health – Current health	0.14	0.08 to 0.19	0.000	0.000
Full health – Mild	0.12	0.07 to 0.17	0.000	0.000
Full health – Moderate	0.16	0.11 to 0.21	0.000	0.000
Full health - Severe	0.21	0.16 to 0.27	0.000	0.000

Table 7.9 Health state valuations for people classified as severe, moderate and mild (n = 59).

	States being valued											
	Own current health			Mild			Moderate			Severe		
Classified as:	Sev	Mod	Mild	Sev	Mod	Mild	Sev	Mod	Mild	Sev	Mod	Mild
Mean	0.842	0.871	0.975	0.882	0.875	0.950	0.834	0.844	0.850	0.804	0.775	0.800
SD	0.172	0.208	0.035	0.172	0.208	0.000	0.180	0.192	0.141	0.173	0.247	0.212
25th %ile	0.750	0.950	0.950	0.950	0.950	0.950	0.775	0.775	0.750	0.700	0.613	0.650
50th %ile	0.950	0.950	0.975	0.950	0.950	0.950	0.950	0.950	0.850	0.850	0.950	0.800
75th %ile	0.950	0.950	1.000	0.950	0.950	0.950	0.950	0.950	0.950	0.950	0.950	0.950
Min	0.450	0.150	0.950	0.250	0.100	0.950	0.350	0.200	0.750	0.350	0.150	0.650
Max	1.000	0.950	1.000	1.000	1.000	0.950	0.950	0.950	0.950	0.950	0.950	0.950

Table 7.10 Health profile valuations (n = 59). Paired t-test and Wilcoxon-sign test were used to test for differences between the holistic and QALY valuations of the profiles. The t-test provided 95% confidence intervals for the mean differences between the two methods.

	Mod-Scl-Mild		Mod-Scl-Mod		Mod-Sur-Mild		Mod-Sur-Mod		Sev-Sur-Mild		Mod-Scl-Mild (risk)		Mod-Sur-Mild (risk)	
	Holistic	QALY	Holistic	QALY	Holistic	QALY	Holistic	QALY	Holistic	QALY	Holistic	QALY	Holistic	QALY
Mean	17.37	17.59	17.14	16.79	17.37	17.56	17.14	16.77	16.98	17.53	17.24	17.29	16.80	17.47
Mean holistic-QALY (95% CIs)	-0.21 +/- 0.70 (-0.92 to 0.49)		0.34 +/- 0.60 (-0.26 to 0.95)		-0.18 +/- 0.84 (-1.02 to 0.65)		0.37 +/- 0.62 (-0.25 to 0.99)		-0.55 +/- 1.04 (-1.59 to 0.50)		-0.05 +/- 0.75 (-0.80 to 0.70)		-0.68 +/- 1.02 (-1.70 to 0.34)	
SD	3.54	3.75	3.64	3.66	3.42	3.73	3.48	3.63	3.99	3.71	3.16	3.56	3.54	3.66
25th %ile	17.00	18.90	17.00	15.00	17.00	18.86	17.00	15.02	17.00	18.79	17.00	17.44	15.00	18.49
50th %ile	19.00	18.99	19.00	18.99	19.00	18.94	19.00	18.92	19.00	18.91	19.00	18.99	19.00	18.94
75th %ile	19.00	19.00	19.00	19.00	19.00	18.97	19.00	18.97	19.00	18.97	19.00	19.00	19.00	18.97
Min	6.00	2.21	5.00	4.01	5.00	2.29	3.00	4.09	1.00	2.34	5.00	3.40	3.00	3.07
Max	20.00	19.97	19.00	19.00	20.00	19.95	19.00	19.02	20.00	19.93	19.00	19.61	20.00	19.85
p (t-test)	0.547		0.262		0.663		0.238		0.298		0.891		0.188	
p (Wilcoxon)	0.021		0.004		0.113		0.006		0.319		0.587		0.768	

Table 7.11 Health profile valuations transformed to a scale of 0 to 1 (n = 59).

	Mod-Scl-Mild		Mod-Scl-Mod		Mod-Sur-Mild		Mod-Sur-Mod		Sev-Sur-Mild		Mod-Scl-Mild (risk)		Mod-Sur-Mild (risk)	
	Holistic	QALY	Holistic	QALY	Holistic	QALY	Holistic	QALY	Holistic	QALY	Holistic	QALY	Holistic	QALY
Mean	0.869	0.880	0.857	0.840	0.869	0.878	0.857	0.839	0.849	0.877	0.862	0.865	0.840	0.874
QALY-holistic	0.011		-0.017		0.009		-0.018		0.028		0.003		0.034	
SD	0.177	0.188	0.182	0.183	0.171	0.187	0.174	0.182	0.200	0.186	0.158	0.178	0.177	0.183
25th %ile	0.85	0.945	0.85	0.75	0.85	0.943	0.85	0.751	0.85	0.940	0.85	0.872	0.75	0.925
50th %ile	0.950	0.950	0.950	0.950	0.950	0.947	0.950	0.946	0.950	0.946	0.950	0.950	0.950	0.947
75th %ile	0.950	0.950	0.950	0.950	0.950	0.949	0.950	0.949	0.950	0.949	0.950	0.950	0.950	0.949
Min	0.300	0.111	0.250	0.201	0.250	0.115	0.150	0.205	0.050	0.117	0.250	0.170	0.150	0.154
Max	1.000	0.999	0.950	0.950	1.000	0.998	0.950	0.951	1.000	0.997	0.950	0.981	1.000	0.993

Table 7.12 Comparisons between QALY valuations of all the health profiles.

Profile comparisons			Mean Difference	95% CIs	*p* (t-test)	*p* (Wilcoxon-sign test)
Mod-Scl-Mild	>	Mod-Scl-Mod	0.792 +/- 0.577	0.216 to 1.369	0.008	0.002
Mod-Scl-Mild	>	Mod-Sur-Mild	0.030 +/- 0.011	0.019 to 0.041	0.000	0.000
Mod-Scl-Mild	>	Mod-Sur-Mod	0.818 +/- 0.576	0.243 to 1.394	0.006	0.000
Mod-Scl-Mild	>	Sev-Sur-Mild	0.056 +/- 0.021	0.035 to 0.077	0.000	0.000
Mod-Scl-Mild	>	Mod-Scl-Mild (risk)	0.297 +/- 0.216	0.081 to 0.513	0.008	0.003
Mod-Scl-Mild	>	Mod-Sur-Mild (risk)	0.111 +/- 0.061	0.051 to 0.172	0.000	0.000
Mod-Scl-Mod	<	Mod-Sur-Mild	0.762 +/- 0.574	0.188 to 1.337	0.010	0.780
Mod-Scl-Mod	>	Mod-Sur-Mod	0.026 +/- 0.011	0.015 to 0.037	0.000	0.000
Mod-Scl-Mod	<	Sev-Sur-Mild	0.736 +/- 0.569	0.168 to 1.305	0.012	0.821
Mod-Scl-Mod	<	Mod-Scl-Mild (risk)	0.495 +/- 0.360	0.135 to 0.855	0.008	0.002
Mod-Scl-Mod	<	Mod-Sur-Mild (risk)	0.681 +/- 0.517	0.164 to 1.198	0.011	0.833
Mod-Sur-Mild	>	Mod-Sur-Mod	0.788 +/- 0.574	0.215 to 1.362	0.008	0.002
Mod-Sur-Mild	>	Sev-Sur-Mild	0.026 +/- 0.017	0.009 to 0.043	0.003	0.003
Mod-Sur-Mild	>	Mod-Scl-Mild (risk)	0.267 +/- 0.214	0.053 to 0.481	0.016	0.780
Mod-Sur-Mild	>	Mod-Sur-Mild (risk)	0.081 +/- 0.057	0.024 to 0.138	0.006	0.000
Mod-Sur-Mod	<	Sev-Sur-Mild	0.762 +/- 0.568	0.195 to 1.330	0.009	0.004
Mod-Sur-Mod	<	Mod-Scl-Mild (risk)	0.521 +/- 0.359	0.162 to 0.881	0.005	0.000
Mod-Sur-Mod	<	Mod-Sur-Mild (risk)	0.707 +/- 0.516	0.191 to 1.223	0.008	0.970
Sev-Sur-Mild	>	Mod-Scl-Mild (risk)	0.241 +/- 0.209	0.032 to 0.450	0.024	0.821
Sev-Sur-Mild	>	Mod-Sur-Mild (risk)	0.055 +/- 0.053	0.002 to 0.109	0.044	0.001
Mod-Scl-Mild (risk)	<	Mod-Sur-Mild (risk)	0.186 +/- 0.157	0.029 to 0.343	0.021	0.833

Table 7.13 Logical consistency of health profiles for each valuation method.

	Strongly consistent		Including weakly consistent		Non-consistent	
	Holistic	QALY	Holistic	QALY	Holistic	QALY
Mod-Scl-Mild > Mod-Scl-Mod	11 (18.6%)	18 (30.5%)	52 (88.1%)	56 (94.9%)	7 (11.9%)	3 (5.1%)
Mod-Sur-Mild > Mod-Sur-Mod	7 (11.9%)	19 (32.2%)	55 (93.2%)	56 (94.9%)	4 (6.8%)	3 (5.1%)
Mod-Sur-Mild > Sev-Sur-Mild	10 (16.9%)	19 (32.2%)	52 (88.1%)	55 (93.2%)	7 (11.9%)	4 (6.8%)
Mod-Scl-Mild > Mod-Scl-Mild (risk)	15 (25.4%)	18 (30.5%)	53 (89.8%)	56 (94.9%)	6 (10.2%)	3 (5.1%)
Mod-Sur-Mild > Mod-Sur-Mild (risk)	15 (25.4%)	56 (94.9%)	55 (93.2%)	56 (94.9%)	4 (6.8%)	3 (5.1%)

Table 7.14 Health profile valuations for those willing to trade (n = 38). Paired t-test and Wilcoxon-sign test were used to test for differences between the holistic and QALY valuations of the profiles. The t-test provided 95% confidence intervals for the mean differences between the two methods.

	Mod-Scl-Mild		Mod-Scl-Mod		Mod-Sur-Mild		Mod-Sur-Mod		Sev-Sur-Mild		Mod-Scl-Mild (risk)		Mod-Sur-Mild (risk)	
	Holistic	QALY	Holistic	QALY	Holistic	QALY	Holistic	QALY	Holistic	QALY	Holistic	QALY	Holistic	QALY
Mean	16.47	16.81	16.11	15.58	16.47	16.78	16.11	15.56	15.87	16.74	16.26	16.35	15.58	16.66
Mean holistic-QALY (95% CIs)	-0.33 +/- 1.11 (-1.44 to 0.78)		0.53 +/- 0.95 (-0.42 to 1.48)		-0.31 +/- 1.32 (-1.63 to 1.01)		0.55 +/- 0.97 (-0.42 to 1.52)		-0.87 +/- 1.64 (-2.51 to 0.76)		-0.08 +/- 1.18 (-1.26 to 1.10)		-1.08 +/- 1.59 (-2.67 to 0.51)	
SD	4.16	4.51	4.21	4.09	4.00	4.48	3.99	4.06	4.62	4.44	3.60	4.16	3.92	4.36
25th %ile	15.00	16.46	15.00	15.00	15.00	16.44	14.75	14.98	15.00	16.29	15.00	15.66	13.00	16.29
50th %ile	19.00	18.94	18.50	17.00	19.00	18.89	17.00	16.96	17.00	18.83	17.00	18.21	17.00	18.70
75th %ile	19.00	19.00	19.00	18.99	19.00	18.96	19.00	18.94	19.00	18.94	19.00	18.99	19.00	18.96
Min	6.00	2.21	5.00	4.01	5.00	2.29	3.00	4.09	1.00	2.34	3.40	3.40	3.00	3.07
Max	20.00	19.97	19.00	19.00	20.00	19.95	19.00	19.00	20.00	19.93	19.61	19.61	20.00	19.85
p (t-test)	0.546		0.266		0.638		0.261		0.286		0.888		0.178	
p (Wilcoxon)	0.388		0.086		0.936		0.141		0.538		0.587		0.119	
Lower than whole sample mean by:	0.90	0.78	1.03	1.21	0.90	0.78	1.03	1.21	1.11	0.79	0.98	0.94	1.22	0.81

Table 7.15 Sensitivity analysis of time of recurrence.

	Mod-Scl-Mild (risk)				Mod-Sur-Mild (risk)			
	Holistic	QALY			Holistic		QALY	
		1 year	Middle	End of life		1 year	Middle	End of life
Mean	17.24	17.02	17.29	17.56	16.80	17.40	17.47	17.55
SD	3.16	3.55	3.56	3.72	3.54	3.61	3.66	3.72
25th %ile	17.00	16.12	17.44	18.75	15.00	18.14	18.49	18.82
50th %ile	19.00	18.99	18.99	18.99	19.00	18.94	18.94	18.94
75th %ile	19.00	19.00	19.00	19.00	19.00	18.97	18.97	18.97
Min	5.00	3.73	3.40	2.50	3.00	3.31	3.07	2.37
Max	19.00	19.28	19.61	19.94	20.00	19.77	19.85	19.94

Table 7.16 Sensitivity to treatment process. QALY results from setting values for sclerotherapy and surgery at 0 and 1.

	Mod-Scl-Mild		Mod-Scl-Mod		Mod-Sur-Mild		Mod-Sur-Mod		Sev-Sur-Mild		Mod-Scl-Mild (risk)		Mod-Sur-Mild (risk)	
	Sclerotherapy=0	Sclerotherapy=1	Sclerotherapy=0	Sclerotherapy=1	Surgery=1	Surgery=0	Surgery=1	Surgery=0	Surgery=1	Surgery=0	Sclerotherapy=0	Sclerotherapy=1	Surgery=1	Surgery=0
Mean	17.57	17.59	16.78	16.80	17.60	17.48	16.82	16.70	17.58	17.46	17.26	17.30	17.52	17.40
SD	3.75	3.75	3.66	3.66	3.73	3.73	3.64	3.64	3.71	3.71	3.56	3.56	3.66	3.66
25th %ile	18.88	18.90	14.99	15.01	18.91	18.79	15.03	14.91	18.81	18.69	17.42	17.44	18.52	18.40
50th %ile	18.98	19.00	18.98	19.00	19.01	18.89	19.01	18.89	19.01	18.89	18.98	19.00	19.00	18.88
75th %ile	18.98	19.00	18.98	19.00	19.01	18.89	19.01	18.89	19.01	18.89	18.98	19.00	19.00	18.88
Min	2.20	2.22	4.00	4.02	2.31	2.19	4.10	3.98	2.36	2.24	3.39	3.41	3.08	2.96
Max	19.96	19.98	18.98	19.00	19.98	19.86	19.01	18.89	19.98	19.86	19.59	19.61	19.88	19.76

Table 7.17 Discounted QALY values for the health profiles, using each of the four discount rates. Undiscounted QALY and holistic values are shown for comparison.

	Discount rate	Mean	SD	25th %ile	median	75th %ile	min	max
Mod-Scl-Mild	Holistic	17.37	3.54	17.00	19.00	19.00	6.00	20.00
	Undiscounted	17.59	3.75	18.90	18.99	19.00	2.21	19.97
	-0.029	23.60	5.041	25.389	25.482	25.491	2.888	26.807
	0.035	12.928	2.753	13.872	13.965	13.973	1.676	14.683
	0.073	9.756	2.074	10.449	10.542	10.550	1.316	11.080
	1.240	1.567	0.323	1.613	1.706	1.715	0.308	1.779
Mod-Scl-Mod	Holistic	17.14	3.64	17.00	19.00	19.00	5.00	19.00
	Undiscounted	16.79	3.66	15.00	18.99	19.00	4.01	19.00
	-0.029	22.533	4.910	20.130	25.480	25.491	5.378	25.495
	0.035	12.351	2.690	11.037	13.962	13.973	2.953	13.977
	0.073	9.325	2.030	8.335	10.540	10.550	2.232	10.554
	1.240	1.514	0.328	1.360	1.704	1.715	0.372	1.719
Mod-Sur-Mild	Holistic	17.37	3.42	17.00	19.00	19.00	5.00	20.00
	Undiscounted	17.56	3.73	18.86	18.94	18.97	2.29	19.95
	-0.029	23.573	5.015	25.349	25.432	25.466	2.978	26.786
	0.035	12.898	2.727	13.832	13.914	13.949	1.765	14.662
	0.073	9.726	2.048	10.409	10.492	10.526	1.405	11.060
	1.240	1.537	0.298	1.573	1.656	1.684	0.394	1.759
Mod-Sur-Mod	Holistic	17.14	3.48	17.00	19.00	19.00	3.00	19.00
	Undiscounted	16.77	3.63	15.02	18.92	18.97	4.09	19.02
	-0.029	22.507	4.885	20.147	25.415	25.458	5.458	25.509
	0.035	12.325	2.665	11.054	13.897	13.941	3.034	13.991
	0.073	9.299	2.006	8.352	10.475	10.518	2.313	10.568
	1.240	1.488	0.304	1.366	1.639	1.682	0.453	1.733
Sev-Sur-Mild	Holistic	16.98	3.99	17.00	19.00	19.00	1.00	20.00
	Undiscounted	17.53	3.71	18.79	18.91	18.97	2.34	19.93
	-0.029	23.547	4.996	25.280	25.400	25.458	3.028	26.762
	0.035	12.872	2.709	13.762	13.882	13.941	1.815	14.638
	0.073	9.700	2.030	10.340	10.460	10.518	1.455	11.036
	1.240	1.511	0.288	1.504	1.619	1.676	0.369	1.735
Mod-Scl-Mild (risk)	Holistic	17.24	3.16	17.00	19.00	19.00	5.00	19.00
	Undiscounted	17.29	3.56	17.44	18.99	19.00	3.40	19.61
	-0.029	23.056	4.745	22.699	25.480	25.491	4.735	26.134
	0.035	12.696	2.611	12.732	13.962	13.973	2.528	14.398
	0.073	9.610	1.979	9.733	10.540	10.550	1.882	10.901
	1.240	1.566	0.322	1.613	1.706	1.715	0.308	1.779
Mod-Sur-Mild (risk)	Holistic	16.80	3.54	15.00	19.00	19.00	3.00	20.00
	Undiscounted	17.47	3.66	18.49	18.94	18.97	3.07	19.85
	-0.029	23.427	4.891	24.650	25.432	25.466	4.327	26.607
	0.035	12.836	2.675	13.546	13.914	13.949	2.373	14.586
	0.073	9.687	2.015	10.236	10.492	10.526	1.787	11.012
	1.240	1.536	0.298	1.573	1.656	1.684	0.394	1.759

Table 7.18 Health profile valuations split by whether respondents have recently received surgery or sclerotherapy.

	Mod-Scl-Mild		Mod-Scl-Mod		Mod-Sur-Mild		Mod-Sur-Mod		Sev-Sur-Mild		Mod-Scl-Mild (risk)		Mod-Sur-Mild (risk)	
	Holistic	QALY	Holistic	QALY	Holistic	QALY	Holistic	QALY	Holistic	QALY	Holistic	QALY	Holistic	QALY
Treated (n=19)														
Mean (SD)	16.63 (4.76)	16.54 (4.87)	16.16 (5.30)	15.21 (5.14)	16.89 (4.89)	16.53 (4.83)	16.26 (5.00)	15.20 (5.11)	17.00 (4.23)	16.52 (4.81)	16.26 (4.33)	16.04 (4.82)	16.89 (4.48)	16.39 (4.81)
Median (IQR)	19.00 (15.00-19.00)	18.95 (14.84-19.00)	19.00 (17.00-19.00)	17.00 (13.00-19.00)	19.00 (19.00-19.00)	18.92 (14.88-18.96)	19.00 (15.00-19.00)	17.00 (13.01-18.96)	19.00 (17.00-19.00)	18.94 (14.88-18.95)	19.00 (13.00-19.00)	18.22 (13.00-19.00)	19.00 (16.00-19.00)	18.74 (14.29-18.96)
Mean holistic – QALY (95% CIs)	0.09 +/- 1.40 (-1.31 to 1.49)		0.95 +/- 0.92 (0.03 to 1.86)		0.37 +/- 1.23 (-0.86 to 1.60)		1.06 +/- 0.97 (0.10 to 2.03)		0.48 +/- 1.57 (-1.09 to 2.05)		0.22 +/- 1.21 (-0.99 to 1.43)		0.50 +/- 2.00 (-1.49 to 2.50)	
p (t-test)	0.894		0.043		0.536		0.032		0.530		0.704		0.603	
(Wilcoxon)	0.091		0.003		0.014		0.011		0.136		0.184		0.117	
Untreated (n=40)														
Mean (SD)	17.73 (2.79)	18.08 (3.04)	17.60 (2.46)	17.55 (2.43)	17.60 (2.50)	18.05 (3.01)	17.55 (2.43)	17.51 (2.41)	16.98 (3.93)	18.01 (3.01)	17.70 (2.36)	17.88 (2.64)	16.75 (3.06)	17.99 (2.90)
Median (IQR)	19.00 (17.50-19.00)	18.99 (18.94-19.00)	19.00 (17.00-19.00)	18.99 (17.00-19.00)	19.00 (17.00-19.00)	18.94 (18.86-18.98)	19.00 (7.00-19.00)	18.93 (16.95-18.97)	19.00 (17.00-19.00)	18.90 (18.80-18.98)	19.00 (17.00-19.00)	18.99 (18.21-19.00)	19.00 (14.25-19.00)	18.94 (18.69-18.97)
Mean holistic - QALY (95% CIs)	-0.36 +/- 0.84 (-1.20 to 0.48)		0.05 +/- 0.79 (-0.73 to 0.84)		-0.45 +/- 1.11 (-1.56 to 0.67)		0.04 +/- 0.79 (-0.76 to 0.83)		-1.03 +/- 1.37 (-2.40 to 0.33)		-0.18 +/- 0.98 (-1.16 to 0.79)		-1.24 +/- 1.19 (-2.43 to –0.05)	
p (t-test)	0.395		0.892		0.424		0.924		0.132		0.709		0.041	
p (Wilcoxon)	0.113		0.154		0.737		0.104		0.809		0.819		0.192	

Chapter 8

Study 4: A test of additivity over probability in treatment decisions for unfit patients with abdominal aortic aneurysms

8.1 Introduction

This study sought to value the health profiles related to different treatment pathways for abdominal aortic aneurysms (AAAs), using both the traditional quality-adjusted life year (QALY) algorithm and methods involving the holistic valuation of the profiles themselves. The treatment pathways for AAA were chosen as a case study since the choice involves a trade-off between short-term risk for longer-term survival. The results from these alternative methods were compared to determine whether there were systematic differences.

McNeil *et al* (1981) did a study of 37 volunteers to explore their preferences between receiving radiotherapy for laryngeal cancer and receiving surgery. Because surgery involved removal of the voice box, there would be a varying degree of speech impairment after the procedure. The risk attitude of respondents was first ascertained over an expected survival of either 10 or 25 years (depending on the age of the respondent, this was meant to represent a normal life expectancy). Respondents were then asked to state how many years they would be willing to forgo in order to have normal speech for the remainder of their life. This study involved a trade-off of quantity and quality of life. Surgery was the option preferred by doctors. The results suggested that a higher number of respondents than expected preferred radiotherapy, thus preferring not to take the risk of decreased quality of life. However, the McNeil *et al* study did not undertake a comparison of QALY versus holistic valuation approaches.

Whereas McNeil *et al* explored trade-offs between quality and quantity of life over a normal life expectancy, this present study looks at risk attitude over very short life expectancies such as might be predicted for the terminally ill. Within this short survival duration choices between short-term mortality and risks of longer-term severe morbidity are explored. Thus a significant dimension of risk is added to the trade-off between quality and quantity.

The concept of HYEs is about finding a way to obtain valuations for health profiles holistically, rather than relying on an algorithm dependent on valuations of discrete health states. Empirical work to date on holistic valuation of scenarios in health has found that in some cases holistic valuations differ from valuations obtained from the QALY algorithm (Richardson *et al*, 1996). The differing findings of the studies presented in Chapters 5, 6 and 7 of this thesis show that the relationship between holistic and QALY valuations is not straightforward, but is dependent upon the context in which they are employed. It is important therefore not only to discover the extent to which holistic valuations differ, but also to explore the reasons behind the differences in this study.

8.2 Abdominal aortic aneurysms

An aneurysm occurs when the wall of an artery becomes dilated (Loyola University. 1999). AAAs are dilations greater than 30 mm in diameter, and mostly occur within the infrarenal aorta (Calvert *et al*, 1999). If left untreated, they are likely to expand and eventually rupture. Rupture rates are related to AAA size. It is estimated that the rate of rupture of AAAs of 40 – 50 mm is 3 – 12% over five years, whereas the 5-year rupture rate for AAAs of over 50mm is estimated to be within the range of 25 – 41% (Calvert *et al*, 1999).

Once the AAA has ruptured, the individual has a very high probability of bleeding to death. Studies have reported hospital post-rupture mortality rates range from 40-80%, but since many cases of rupture occur outside hospital this figure may be as high as 95% (Thomas, 1999; Thomas and Stewart, 1988; Budd *et al*, 1989; Ingoldby *et al*, 1986; Johansson and Swedenborg, 1986; Campbell *et al*, 1986; Bengtsson and Bergqvist, 1993; Kantonene *et al*, 1999; Eskandari *et al*, 1998).

The condition is more common in men than women, and age is a principal risk factor. Ruptured AAAs are rare in people under 50 years of age, but the incidence of ruptures increases in frequency in men aged 55 and over (Collin *et al*, 1988). Wilmink and Quick (1998) estimate that 1.36% and 0.45% of deaths in men and women respectively over the age of 65 in England and Wales are due to ruptured AAAs. In men aged 70-74 this cause accounts for 1.8% of deaths, compared to a peak of 0.6% of deaths in women aged 80-84 (Collin *et* al, 1988). The incidence of ruptured AAAs was recently reported as between 1 and 21 per 100,000 per year (Calvert *et al*, 1999), although it is likely that many go unreported due to the asymptomatic characteristic of smaller AAAs.

There has been recent interest in comparisons between different treatments for AAAs. Two clinical trials began at the end of 1999, which are examining the cost-effectiveness of different treatment options (Calvert *et al*, 1999). The two trials are called EVAR (endovascular aneurysm repair) I and II, with EVAR I comparing conventional open surgical repair to endovascular stent grafting in patients fit for open surgery and EVAR II comparing endovascular repair (EVAR) with best medical treatment (BMT) in patients deemed unfit for surgery. The endovascular procedure is a minimally-invasive technique.

Thomas (1999) used a Markov design study to examine the cost effectiveness (in terms of life years gained) of placing stent-grafts in unfit patients as a compliment to the EVAR II trial. Thomas found BMT to be the less expensive treatment, with an average cost of £113 per life year gained (LYG), whereas the average cost of EVAR was £2154 per LYG (Thomas, 1999). However, his results also suggested that EVAR would generate more LYG than BMT. The results of the cost-effectiveness analysis indicated that EVAR was cost-effective, with a marginal cost of £6717 per LYG. This finding is interesting and possibly open to question, because the unfit patients would be taking an increased risk of death in the near future for a moderate gain in life expectancy. Although Thomas did not conduct a cost utility analysis, he looked at the possible effects of utility values on cost-effectiveness during his sensitivity analysis. He found that cost per QALY could be significantly altered depending on utility values following EVAR or best medical treatment (BMT). Thomas expressed the necessity to obtain utility values for unfit patients with AAA.

The present study focused on patients who form a similar group to those entering the EVAR II trial: patients with large AAAs (over 55 mm), who are unsuitable for open repair surgery. AAA is a particularly good condition upon which to test the QALY assumptions. It is an asymptomatic condition until rapid expansion or rupture occurs. Once aware of their AAAs, unfit patients are faced with different risks of mortality depending on the treatment. The choice faced by this group of patients is the increased risk of immediate death and morbidity due to a procedure to exclude the aortic aneurysm versus a lower life expectancy from doing nothing. If successful, however, the treatment is likely to result in a longer life expectancy. The question to be answered is how people balance their desire to maximise life expectancy against different levels of risk of more immediate mortality or morbidity. It is worth noting that the EVAR II trial deals with an unfit population, who are therefore likely to have a reduced life expectancy anyway due to other causes. The decision to concentrate on the unfit population was made so that the focus would be on the trade-off between longer life expectancy against short-term risks arising from the EVAR treatment.

8.3 Assumptions of the QALY model

If the axioms underlying the QALY algorithm do not accurately reflect the preferences of individuals, then QALYs may be an inaccurate measure of utility. The degree of inaccuracy attached to QALYs would depend upon the extent to which the underlying axioms are violated. The QALY axioms are discussed in depth in Chapter 2. However, the following are a basic list of the fundamental QALY assumptions which are tested in this study.

8.3.1 Attitude to risk

This study tests the assumption that attitude to risk is constant no matter what the level of risk or the context of the risk. It also attempts to estimate risk attitudes in order to be able to calculate risk-adjusted QALYs in addition to the more commonly used risk-neutral QALYs. Estimations of risk attitude make it possible to also test whether respondents demonstrated neutrality over risk, as is commonly assumed in many studies.

8.3.2 Linearity of preferences through time

In concentrating on choices between immediate increases in risk of morbidity and mortality for a longer life expectancy versus a shorter life expectancy, this study also attempts to explore the concept of time preferences over short life expectancies. The TTO-QALY requires the assumption of zero time preference (Dolan, 2000). However, there is plenty of evidence that the zero time preference axiom is violated (Loewenstein and Prelec, 1993; Chapman and Elstein, 1995; Dolan and Gudex, 1995; Chapman and Coups, 1999). This study aims to test the axiom of zero time preference by eliciting individual time preferences. QALY valuations will be adjusted for time preference rates to examine the effects of time preferences on differences between QALY valuations and holistic valuations of scenarios.

8.3.3 Summary of QALY assumptions dealt with in this chapter

This study examined the QALY assumptions of:

- Constant attitude to risk

- Zero time preference

8.4 Aims of this study

This study follows on from Chapters 5, 6, and 7 in testing the hypothesis that valuations of profiles by the QALY algorithm and a holistic method are equal. If they are not equal, a key test of validity for each method is the extent to which the results match the initial rankings of the profiles. However, as discussed in previous chapters, profiles are ranked before the valuation exercises, and the ranking procedure is used as a "warming up" exercise. Thus a flaw in using original rankings as the test of convergent validity is the possibility that the original rankings may not be the true preferences (see Chapter 4).

Holistic valuation incorporates individuals' risk attitudes, the quantity effect, and time preferences. However, the QALY method restrictive assumptions about each of these. If the two methods are not equal, the QALY values will be adjusted for risk attitude and time preferences. The extent to which this equalizes the valuations from the two methods will be explored.

8.5 Methods

8.5.1 Development of the scenarios

Respondents were asked to value two health-related scenarios related to different treatment pathways for AAAs. These scenarios were drafted based upon the results of existing reviews of the evidence on the outcomes of clinical management of AAAs (Thomas, 1999) and discussions with clinicians. The scenarios are shown diagrammatically in Figure 8.1.

Thomas (1999) used a Monte Carlo model to estimate risks of AAA ruptures. These risks were then entered into a Markov model, which modelled outcomes for BMT and EVAR for a hypothetical cohort of 70-year-old men with AAAs over 6 cm in diameter. The model estimated probabilities and costs over a 20-year period. Using this model, Thomas estimated that patients undergoing EVAR would have an average life expectancy of 3.09 years, and patients undergoing BMT would have an average life expectancy of 2.14 years. Thomas based the probabilities of mortality and morbidity upon literature sources.

The EVAR 30-day mortality figures for this study were found by a systematic review of the clinical outcomes literature conducted by Thomas (1999), and used in his Markov model. The probabilities of suffering post-EVAR chronic morbidity were estimated with reference to Thomas (2000), Cuypers *et al* (1999), and Walker *et al* (1998). The results of the Cuypers *et al* study of 64 AAA patients who underwent EVAR indicated that 3% suffered chronic renal failure, and a further 3% suffered temporary renal dysfunction following the procedure. These were a mixed group of patients in terms of post-EVAR fitness, with 42% falling into groups 3 and 4 of the American Society of Anesthesiology Physical Status Classification, where physical status is best in group 1 and worst in group 4. Walker *et al* studied a sample of 164 AAA patients who underwent EVAR, and found that approximately 3% developed renal failure post-operatively. A further 6% developed some renal dysfunction, almost half of whom died. Thomas (2000) estimated that around 10% of patients would develop chronic renal failure after EVAR, and 10% would suffer a major stroke. The probability of surviving 3 years after EVAR was derived from a review by Thomas (1999), and was similar to the probabilities entered into his Markov design.

The EVAR scenario in Figure 8.1 describes the probability tree for an unfit AAA patient facing treatment by EVAR. If the patient undergoes EVAR, the evidence suggests that there is an approximately 20% chance that the patient will be dead within 30 days; a 20% chance that the patient will suffer either chronic renal failure or stroke and will die after 3 years; and there is a 60% chance that the patient will survive in his/her current health for 3 years and then die. This figure is based on average life expectancies in those treated by EVAR (Thomas, 1999).

The BMT option is best medical treatment for each individual who is unfit for open repair surgery or EVAR. Regular ultrasound scans indicate whether the AAA is expanding or remaining stable. Pharmacological means may be used to try to slow expansion and reduce risk of rupture, *e.g.* by lowering blood pressure. A corresponding 30-day death rate was needed for this group, but this information was not available from the medical literature. These patients may suffer mortality due to their AAA, their comorbid condition which is making them unfit, or any other cause. Death rates in Yorkshire and Humberside were used as a proxy for BMT 30-day mortality. These were 14/1000 (HMSO, 1992), and were rounded up to 20/1000.

As Figure 8.1 shows, in the BMT scenario used in this study the patient faces a 2% chance of death within 30 days; a 0% chance of chronic renal failure or stroke: and a 98% chance of current health for 2 years followed by death. Life expectancies of 3 years (EVAR) and 2 years (BMT) were derived from Thomas (1999), who obtained his data from the Registry of Endovascular Treatment of Aneurysms.

The above probabilities for EVAR and BMT are estimates rather than exact. It was necessary to keep figures as simple as possible in order to avoid confusing respondents. Also the estimates in the literature were very variable.

The scenarios describe hypothetical outcomes from the treatments EVAR and BMT. The main differences between the two scenarios are in terms of risks and life expectancy. EVAR has a longer life expectancy than BMT (3 years versus 2 years). However, the EVAR scenario carries a much higher level of risk than the BMT scenario. These risks are both in terms of serious morbidity and also mortality.

8.5.2 Developing the health states

The health states valued in the questionnaire were current health as described by EQ-5D, stroke, chronic renal failure, and immediate death.

Condition-specific descriptions of the health states of stroke and chronic renal failure were developed with reference to sources of information such as the available literature on these conditions (familydoctor.org, 2000; American Academy of Neurology, 2002; American Heart Association, 2002; National Institute of Diabetes & Digestive & Kidney Diseases, 2002; Kidney Patient Guide, 2002), and patient perspectives (Anonymous, 2000). The states of stroke and chronic renal failure were developed specifically for this questionnaire, and are shown in Figure 8.2.

All the health states were valued against full health as described by the EQ-5D state 11111.

8.5.3 The QALY method of valuing scenarios

QALY scores were derived for each scenario using the conventional assumption of linearity of preferences through time and risk. QALY scenario valuations were obtained using Equations (8.1) and (8.2) derived from Figure 8.1 (death is given a utility of 0):

$$\text{BMT}=(0.02 * 0.000) + \{0.98 * \text{U (current health)} * 24 \text{ months}\} \tag{8.1}$$

$$\text{EVAR}=(0.20 * 0.000) + \{0.10 * \text{U (chronic renal failure)} * 36 \text{ months}\} + \{0.10 * \text{U (stroke)} * 36 \text{ months}\} + \{0.60 * \text{U (current health)} * 36 \text{ months}\} \tag{8.2}$$

Respondents valued their current state of health and stroke and chronic renal failure states separately, and their utility values for each of these states were used in the above equations.

Adjusting for risk attitude

The QALY algorithm described above took an *ex post* perspective of probabilities. Individual risk attitudes of respondents were not incorporated. An attempt was made to incorporate individual risk attitudes into the QALY valuations as described below.

Risk attitudes were estimated for each individual by use of a set of standard gamble questions, modified from those used by McNeil *et al* (1978), Miyamoto and Eraker (1985), and Stiggelbout *et al* (1994). This method is a form of the standard gamble, but is synonymous with the TTO for valuing scenarios containing uncertainty. Whereas in the standard gamble the subject states the probability of x years in full health that is equivalent to the certainty of x years in a given health state, for the certainty equivalent the subject states the number of years y in full health for certain that is equivalent to the given probability of x years in full health. The basics of the two methods are demonstrated in Figure 8.5. This is a certainty equivalent (CE) method, in which respondents are presented with a gamble where there is probability p of surviving x months in a given health state, and a $1 - p$ probability of dying immediately. Each individual states the number of months lived in the given health state for certain which would be equivalent to the gamble (Figure 8.5). Certainty equivalents are obtained for $p = 0.25$, $p = 0.50$, and $p = 0.75$. These are referred to as CE25, CE50, and CE75 respectively. With this information it is possible to plot risk attitude curves.

Since what is being varied is time, the TTO is the appropriate valuation technique. Another reason to use TTO rather than standard gamble is that the scenarios contain fixed probabilities. It has been found that the cognitive burden for respondents to value two levels of risk (risk-risk questions) is too much for most respondents (Jones-Lee *et al*, 1995).

This method was adapted for the present study by letting x = 36 months, and getting respondents to state their CE25, CE50 and CE75 values. However, the method used in this study differed from that used by the authors cited above in one respect. McNeil *et al* (1978) suggested that gambles other than 50:50 gambles analogous to the toss of a coin were difficult for untrained individuals to value. They therefore made all their gambles equivalent to 50:50. McNeil *et al* (1978) first obtained the value of CE50 by eliciting the number of years "a" equivalent to a 50:50 gamble involving outcomes of 25 years or 0 years. To obtain CE25, the number of years "b" equivalent to a 50:50 gamble involving outcomes "a" years or 0 years were elicited. To obtain CE75, the number of years "c" equivalent to a 50:50 gamble involving outcomes 25 years or "a" years were elicited (see Figure 4.1). However, this present study asked respondents to value CE25 and CE75 directly in order to examine risk attitude over a range of risks in order to determine whether it would be constant.

Instead of choices over the state of full health, this study asked respondents to make decisions over choices involving their current state of health. This was intended to ease the cognitive process for AAA patients, who were likely to have comorbid conditions.

The risk attitude (*r*) for each individual was obtained using the method suggested by Miyamoto and Eraker (1985), where

n = the probabilities of 25, 50, 75 on a scale of 0 to 100

C_n = certainty equivalent (CE) over the probabilities of 25, 50, 75

Y_{max} = 36 months, the maximum life expectancy

$X_n = \ln(n / 100)$

$Z_n = \ln(C_n / Y_{max})$

The value of r is shown in Equation (8.3).

$$r = (\Sigma X_n^2) / (\Sigma X_n Z_n) \tag{8.3}$$

The proof behind this equation is presented in detail in Appendix 4.

Quantity effect

A question was drafted to measure quantity effect (Figure 8.3). However, it proved too cognitively demanding for respondents in the pilot and was excluded (see Section 8.5.7).

Time preferences

Time preferences were investigated using questions that were based upon the method introduced by Cairns (1991, 1992) and Cairns and van der Pol (2000). These authors devised an open-ended method of eliciting time preference for health, which involved asking respondents to imagine a scenario in which a given period *x* of ill health would occur *t* years in the future. Respondents were asked to state the maximum duration *y* in the ill health state that they would be willing to endure if this illness was in year *s* rather than year *l*. It is possible for *y* to be greater, lesser, or equal to *x*. Likewise, it is possible for *s* to be greater, lesser, or equal to *t*. The method was open-ended in that there were no limits to the values a respondent could place on *y*. The reference ill health was based upon the EQ-5D classification system, and described a state of severe depression.

Cairns (1991, 1992) used this method over a hypothetical life expectancy of 42 years. For example, respondents might have been told that they would experience 20 years of excellent health, spend *x* days in the health state, then experience another 22 years of excellent health followed by death. Cairns and van der Pol (2000) did not limit the life expectancy. They did not describe the duration of the period of full health following the state of ill health. In order to avoid respondents finding the scenarios of controlling the timing of ill health unrealistic, the authors suggested that the timing would be controlled by choices over accepting or rejecting a minor form of medical treatment.

The method described above was modified for the present study. The reference state of severe depression was replaced in this study by a moderately bad health state (22222), using the Euroqol descriptions. Depression was not of special relevance to this study, and it was felt that a more moderate all-round health state would be appropriate for use with this study.

This study looked at a far shorter time horizon than the studies cited above. The time preference questions were modified accordingly, so that the time preference question would be relevant to the short time horizons of the scenarios being valued. Because of the short time horizon involved, it was necessary to use a short period of ill health to avoid a large proportion of the life expectancy being taken up by it. Periods of ill health and excellent health were devised accordingly as shown in Figure 8.4. Following Cairns (1991), scenarios were chosen so as to allow choices to both pospone the ill health or expedite it.

The discounted utility (DU) model assumes that future benefits or losses are discounted at the same rate, no matter how distant into the future. It assumes stationarity (Cairns and van der Pol, 2000), which states that time preferences are based on absolute time intervals rather than relative intervals. Thus if a person prefers to receive £100 in 2 months than £90 in 1 month, he would also prefer £100 in 27 months to £90 in 26 months.

Individual discount rates were calculated, using the equation for the DU model described in equation (8.4)

$$r = (y/x)^{1/(s-t)} - 1 \qquad (8.4)$$

where y is the delayed duration of ill health, and s is the number of years in the future at which y would be equivalent to x days of ill health starting t years from the present.

Respondents in the present study were shown scenarios consisting of a time period of 36 months followed by death. The 36 months consisted of a period of ill health (the EQ-5D state 22222) sandwiched between periods of excellent health. The timing of the ill health differed between the scenarios. Respondents were asked to state the number of days of ill health at which they would feel indifferent between the scenarios. Figure 8.4 shows the two questions in diagrammatical form.

Two time preference questions were asked. In the first question, respondents were asked which they would prefer of Scenario A or Scenario B. If they preferred Scenario A, they were asked how many days of good health they would be prepared to give up (or "pay") in exchange for receiving their preference. Thus, from Equation (8.4), $y > 14$ days, $s = 3$ months, $x = 14$ days, and $t = 12$ months. If respondents preferred Scenario B, $y < 14$ days.

In the second time preference question, respondents were asked which they preferred of Scenarios B and C. If they preferred Scenario B, $y < 14$ days, $s = 30$ months, $x = 14$ days, and $t = 12$ months. If they preferred Scenario C, $y > 14$ days.

8.5.4 Holistic valuation of scenarios

The certainty equivalent (CE) method was used to obtain holistic values for the scenarios. Respondents were asked for their point of indifference between a life expectancy of y months in current health for certain and x months in each of the risky scenarios EVAR and BMT.

A CE method has been previously used by Sutherland *et al* (1982) to obtain values for CE25, CE50, and CE75 for three health states. However, the present study differs in two ways from that used by Sutherland *et al.* Firstly, this study does not seek to elicit values for CE25, CE50 and CE75 for the scenarios BMT and EVAR. Rather, the scenarios describe different levels of risk, and the CE method is used to value these scenarios with their variable levels of risk of mortality and morbidity. Secondly, the time horizon is shorter in the present study than that used by Sutherland *et al* (1982). In the study by Sutherland *et al*, x was a lifetime. In the present study, x is up to 36 months.

8.5.5 The TTO method of valuation

The Gudex (1994) self-completion titration version of TTO was used in this study (see Chapter 4). Since values were to be obtained for hypothetical health states and scenarios in which the maximum life expectancy would be 36 months, the present study used a time horizon of 36 months for all health state valuations. Durations of 36 months and 24 months were used for scenario valuations of EVAR and BMT respectively. The time horizons used for the scenario valuations were valid, because these were the estimated life expectancies for these scenarios. According to the QALY model, the use of different durations should not affect values. It is also valid for holistic valuations, because the nature of the holistic method is to value the profiles as they might occur in life. A scale interval of 2 months was used for the valuations throughout the questionnaire.

A modified version of that suggested by Gudex (1994) was used to value stroke and chronic renal failure if they were deemed worse than death (see Chapter 4). Whereas patients with renal failure may experience a change from the state described in Figure 8.2 to the EQ-5D state 11111 if treated with a kidney transplant, sufferers of stroke to the degree indicated in Figure 8.2 are unlikely to be transformed to full health again in their lifetime. It was therefore considered unfeasible to have the worse than death state followed by full health, and the order was reversed for both renal failure and stroke to be consistent. The values of *a* and *b* summed to 36 months.

The health states in the first part of the questionnaire and the worse than death ratings of scenarios were valued using full health as the comparator. However, for the better than death ratings of scenarios EVAR and BMT, the comparator was current health rather than full health. This was because the questionnaire was intended for patients who were unfit and may not be able to imagine full health. The scenarios were complex in themselves without the additional cognitive burden of attempting to imagine full health. It is possible to chain the valuations through current health so that the resulting valuations would be on a scale of death – full health. As it turned out, as will be explained later, the sample was relatively healthy with EQ-5D scores at or close to 1, so there would have been little difference between full health and current health. The reference states for each valuation exercise are summarized in Table 8.1.

There was an error in the design of the holistic valuation exercise for the EVAR scenario (see the questionnaire in Appendix 4). Although the EVAR scenario was expected to last 36 months, the reference state of current health was accidentally truncated to 24 months in the final production of the questionnaire. This resulted in respondents being unable to give values between 24 and 36 months for EVAR, though these were entirely valid values for the scenario. When the data collected from the survey was subsequently examined, it was discovered that the majority of the sample ranked BMT higher than EVAR, and mean values for BMT were higher than EVAR by the holistic method. The BMT scenario incorporates a 98% probability of living in current health for 24 months and a 2% probability of immediate death. The reference state used in the EVAR valuation was a 100% probability of current health for 24 months. As such, the reference state would be expected to be preferred to the BMT scenario. The fact that BMT was widely preferred to EVAR lends support to the probability that many respondents might have stated a value of 24 months or less in current health as equivalent to the EVAR scenario even had they been able to go above this value. The effect of this truncation was also subjected to a systematic sensitivity analysis. If respondents stated indifference between 24 months in current health and 36 months in the EVAR scenario it was assumed that they would have gone higher if they had been able. Values of 33, 34, 35 and 36 months were assigned to these respondents to see how the most extreme valuation they could have given if able would affect the average scores.

8.5.6 The pilot survey

It was intended that there should be three stages to piloting the questionnaire:

1)The draft questionnaire would be shown to medical and non-medical health professionals to gain input from their expertise

2)There would be an informal consultation exercise with a group of AAA patients who would discuss the presentation of the questionnaire and topics related to their experience of AAAs

3)There would be a formal piloting exercise, in which a group of AAA patients would complete the questionnaire

However, there were problems obtaining a patient sample, so it was not possible to conduct stages 2 and 3.

The pilot questionnaire contained the following sections:

- *Background information* - Age, sex, occupation, age at completion of full-time education

- *Current health* - EQ-5D questionnaire

- *Current health valuation* – TTO valuation of current health state, including questions on better than death and worse than death states

- *Ranking exercise* - The states of full health, current health, chronic renal failure, stroke, and immediate death were ranked in order of preference

- *TTO valuations of morbidity states* – Chronic renal failure and stroke were valued by TTO, with versions for rating them better or worse than death.

- *Ranking exercise* - The scenarios of full health, best medical treatment (BMT), EVAR, and immediate death were ranked in order of preference

- *TTO valuations of scenarios* – BMT and EVAR were valued. It was thought that no one would rate BMT as worse than immediate death because it was only a 2% chance of death, so a worse than death version was not included. However, respondents had the option of rating EVAR as worse than death

- *Valuation of EVAR in terms of BMT* – The scenarios EVAR and BMT were valued directly against each other using TTO, where respondents stated the number of months in BMT which were equivalent to 36 months in EVAR[7]

- *Risk attitude* – Risk attitudes were assessed using three certainty equivalent questions, beginning with a 50:50 gamble, followed by 25:75 ad 75:25 gambles.

- *Time preference* – The method introduced by Cairns and van der Pol (2000) was modified for this study and used to assess time preferences over a time horizon of 36 months

- *Quantity effect* – A question was included to assess the quantity effect, or strength of preference over a life expectancy of 36 months

- *Comments* – The final section gave respondents the opportunity to write their comments on the questionnaire, the way it was presented, or any other topic

The draft questionnaire and scenarios were shown to five key professionals at the Northern General Hospital to gain input from their expertise. These professionals were aged from 26 to 38. Two were nurses, one was a nurse specialist, and two were doctors. Three were female, and two were male. Three described themselves as in excellent health, one in good health, and one in fair health. They were asked for their comments and any suggestions for improvements on the drafts.

In addition to these health professionals, the drafts were shown to five people known to the author. Two of these were health economists employed at the School of Health and Related Research, the University of Sheffield (the supervisors of the author). The other three were siblings of the author.

After this stage the thermometer part of the EQ-5D questionnaire was removed from the questionnaire. It was felt that this was unnecessary to the analysis, and would only add to the cognitive load of the questionnaire.

The time preference question was altered according to suggestions made at this stage of the piloting process. Diagrams were designed to aid the cognitive process, and the wording was changed to add clarity (Figure 8.4).

[7] This valuation proved non-useful and was dropped from the analysis.

The medical professionals aired the view that the questionnaire did not specifically relate to AAA patients. In filling it in, they did not feel it was specific to any particular condition. They felt it could be generalized to another group of patients or the general population.

8.5.7 Main survey

Sample size

The estimate of an appropriate sample size was based upon variations in the valuations of health states and profiles related to IBS in the study reported in Chapter 5 (see also Chapter 4). Equation (8.5) was used to calculate power (Walters, 1999), where the mean difference is the minimum economically important difference (MEID). The standard deviation (SD) obtained in the earlier study was 0.13. A sample size of 56 would be sufficient to detect a difference of 0.05 in utility between the different scenarios with 80% power.

$$n = 2 + \{8 / (\text{mean difference} / \text{SD of difference})^2\} \qquad (8.5)$$

Recruitment: Design 1

It was intended that respondents with AAAs of less than 5.5 cm in diameter, who were not shortly to undergo surgery, would be identified from the appointment lists for ultrasound monitoring of their AAA in the department of Radiology at the Northern General Hospital, Sheffield. Ideally this study would have been conducted upon a sample of patients who were actually faced with the decisions set out in the questionnaire. However, due to ethical considerations patients would only have been asked to participate if they were not currently faced with these treatment decisions. During their attendance for the ultrasound scan, patients would have the opportunity to have any queries about the study answered. They would be asked if they wished to participate, and those who were willing would be interviewed at the hospital after their scans.

Patients invited in this way would benefit whether or not they chose to participate in the survey. The ultrasound scan appointments would be arranged for this study in addition to their regular scans. Whether patients participated in the study or not, it would allow an assessment of the size of their AAA. Any issues raised by the examination (*e.g.* if patients needed to be considered for surgery because their AAA has grown) would be dealt with then, thus saving patients from having to make an additional visit to the outpatient clinic.

A research proposal was presented to the North Sheffield Research Ethics Committee, and ethical approval was obtained.

There proved to be several problems with obtaining a sample of AAA patients. The consultant responsible for identifying AAA patients who would be suitable to participate in the study initially thought that 50 to 60 patients awaiting ultrasound scans had been identified. Of these, around 60% lived in Sheffield, and the remainder lived in surrounding areas. However, the identification of AAA patients was complicated by the fact that the hospital computer system was out of date, and there were complications with patient registration.

There would have been a large logistical workload accrued in recruiting AAA patients. The consultant did not believe patients would attend the interview unless they were attending for an ultrasound scan at the same time, because resources were not available to offer remuneration for their attendance. However, inviting them for a scan prior to the interview would mean that they were invited to a scan over and above the one they have routinely once a year. This would have involved a great deal of logistical work to arrange. There was no funding with which to pay the staff that would be needed to do this work.

There was also a disadvantage to using the small number of AAA patients available for this study in that they may not have been amenable to taking part in future studies planned by staff at the hospital.

A reassessment of the objectives of this study was performed at this stage. The purpose of this study was to determine whether holistic valuations of scenarios incorporating risks differ from valuations obtained by the traditional QALY algorithm, and to examine whether attitudes to risk and time preferences can explain any such differences. Ideally the questionnaire would have been completed by people who were facing the high levels of risk conjectured in the scenarios. However, due to ethical considerations, the aim was to only interview AAA patients who would not be facing those risks. Thus this sample of patients would no more be facing these risks than certain other patient groups. As pointed out by health professionals at the hospital (see Section 8.5.6), the scenarios described in the questionnaire were generic, and not specific to AAA patients.

Recruitment: Design 2

A suitable alternative to AAA patients could be patients with severe peripheral vascular disease, who face risks concerning their legs such as amputation. The questionnaire measured risk attitude, and these patients would be likely to have formed risk attitudes because of the risks of their condition. In-patients would be easy to access. They would have had a procedure such as an angiogram, and would be in the day ward recovering. They would not be allowed to sit up in bed for four to six hours, after which they would be allowed to sit up in bed for another couple of hours before being allowed to go home.

After obtaining ethical approval for this alteration in the study protocol, the author arranged with day ward staff to visit the hospital and ask patients with peripheral vascular disease who were recovering from a procedure to complete the questionnaire.

The questionnaire was piloted on three peripheral vascular disease patients initially. They were two men and one woman, and the ages were 40, 54, and 58 years. They suggested improvements to the questionnaire, including arranging the risk attitude questions in ascending order of risk rather than asking the 50:50 one first. They had great difficulty understanding the question on quantity effect (Figure 8.3). This was dropped with reluctance, on the basis that there was no point in asking it if it would not be answered properly. It evidently needed to have more attention drawn to it than one page of a long questionnaire which was mainly designed to ask for other information.

A further pilot was conducted on two patients. These were a man and a woman aged 45 and 49 years respectively. This final version was completed adequately, and was therefore accepted for use in the main exercise.

Although an attempt was made to use peripheral vascular disease patients in the main valuation exercise, difficulties were encountered in the process. This group of patients tended to be elderly, and they often had hearing problems and were partially sighted. Sight was needed for the questionnaire, and hearing was needed for the verbal explanations of the tasks. They had had a local anaesthetic and a procedure done earlier in the day, and were not feeling very well. The questionnaire is mentally taxing, and requires concentration. Because the interviews took place by the bedside of patients in the day ward, there were plenty of distractions: people passing by, talking, the television, *etc*. Many patients simply did not want to expend the effort needed to complete the questionnaire. Patients were being recruited at around the rate of one per week, so it would have taken around a year to collect the data using this method.

Bearing in mind the generic nature of the questionnaire, it was decided that a convenience sample of staff and students at the hospital and the School of Health and Related Research (ScHARR) was the only feasible option.

Recruitment: Design 3

The clinical consultant approached members of staff at the Sheffield Vascular Institute and asked them to participate in the survey. The author also sent an E-mail to members of staff at ScHARR, asking for volunteers to complete the questionnaire. She also approached a small number of personal acquaintances. There were no reward incentives to participate.

8.5.8 The interviews

The final version was very similar in format to the pilot version.

The ordering of the risk attitude questions was altered so that the 25:75 gamble of current health versus immediate death was assessed first, followed by a 50:50 gamble and then a 75:25 gamble. This was upon the suggestion of one of the respondents to the pilot, who thought this would make the assessments easier.

The time preference assessment method introduced by Cairns and van der Pol (2000) was modified for this study and used to assess time preferences over a time horizon of 36 months, as described in the previous section. However, it now incorporated diagrams of the time horizons to aid elicitation.

The quantity effect question was dropped from the final version of the questionnaire, because the pilot sample had too much difficulty with it.

The questionnaire was administered in groups of one to 10 respondents by a trained and experienced interviewer (the author).

8.5.9 Analysis

The data collected in this study was subjected to a rigorous analysis, as described in the following sub-sections.

Background characteristics

A descriptive analysis of respondents' background characteristics was carried out in terms of age, age at completion of education, gender, and occupation. EQ-5D data was also described across the sample.

Exclusion criteria

Exclusions were made on the basis of ambiguous responses to the valuation exercises, in which cases it was impossible to enter reliable valuation data.

Range of indifference values

The midpoint between the lowest value at which a respondent is willing to trade and the highest value at which she is unwilling to trade was taken as the indifference value in the analysis of valuation data. In some cases, respondents gave a range of indifference values of greater than two months, and the midpoint was still taken. In order to explore the effects of taking the lowest and highest values in the range as the indifference point, the average lowest willing to trade and highest non-willing to trade values were reported. The narrower the gap between these two averages, the less the data were affected by wide indifference ranges.

Ranking order of health states

Respondents were presented with an envelope containing the five hypothetical health states: full health, current health, chronic renal failure, stroke, and immediate death. Each state description was printed on a separate piece of paper. Respondents were asked to choose the ranking order of these health states, and write the order on the appropriate page of the questionnaire, such that the most preferred state was at the top, and the least preferred at the bottom.

Health state values

Respondents were asked to value the health states of current health, chronic renal failure, and stroke using the TTO. Values for each state were subsequently calculated using x t for states "better than death". The ($-x$ / t) formulation was used in the main analysis for states "worse than death". A secondary analysis was conducted using the formulation $-x$ / (t - x) (see Chapter 4). Statistics for the health state values were reported.

Duration t was 36 months for all health states, and this was followed by death. This duration was based upon unfit patients with large AAAs. The respondents actually recruited into the study would have on average a normal life expectancy, so a life expectancy of 36 months was not a realistic scenario for them. However, the same could be said for the health states being valued, which they might not have experienced. This "unreality" is common in such hypothetical scenario valuation exercises.

Ranking order of scenarios

Respondents were presented with the scenario descriptions in an envelope, as for the health states. There were four descriptions: full health, EVAR, BMT, and immediate death. Respondents ranked these four in their order of preference as for the health states.

Health scenario values

Respondents were asked to provide TTO values for the two scenarios of BMT and EVAR. It was not necessary to divide by t to obtain holistic scenario valuations, because they are a direct valuation of time and quality of life. In other words, if a respondent gave a value of 16 to BMT, this would mean that this person stated a preference for 16 months in current health to the probability of 0.98 of being in current health for 24 months. Thus holistic valuations were simply given by x. For BMT valuations $t = 24$, and for EVAR valuations $t = 36$ months (although, as already stated in Section 8.5.5, EVAR valuations were mistakenly truncated to 24 months). If respondents rated EVAR worse than death, the number of months in full health was presented in the negative.

QALY values for scenarios were obtained using the equations (8.1) and (8.2).

QALY and holistic valuations were compared. Valuation data from both methods were found to be non-normally distributed. It is advantageous to use the paired t-test to compare means between methods, because it provides 95% confidence intervals. However, because of the non-normality it was deemed necessary to use a suitable non-parametric test in addition. The Wilcoxon-Sign test was used. These tests were used to determine whether any significant differences between QALY and holistic values for BMT and EVAR existed. These tests were also used to look for differences between the scenario values within each valuation method.

Holistic values were on a scale of death to current health rather than full health. These values had to be chained through to full health before a true comparison could be made between holistic and QALY methods. The calculation for chaining to full health is shown in Equation (8.6), where U_1 is the value for current health, U_2 is the original value of the scenario, and U_3 is the chained value.

$$U_3 = U_1 * U_2 \tag{8.6}$$

The analysis was repeated, comparing chained holistic values with QALY values.

QALY and chained holistic values were entered into a scatter plot, and the level of correlation between the QALY and holistic values given by each respondent for each of the two scenarios was explored.

The statistical analysis was repeated using the worse than death equation that was not constrained to –1.

Analysis of risk attitudes

People should place lower values on scenarios that have greater risk of immediate death, all other things being equal. To answer these questions logically, respondents must have values of CE25 ≤ CE50 and CE50 ≤ CE75. All other responses would be illogical. The logical consistency of each respondent was explored in this respect.

Paired sample t-tests and Wilcoxon-sign tests were carried out to determine whether there were significant differences between the mean values for the three risk questions. Because there were three values, three pair differences were tested: CE25 v. CE50, CE25 v. CE75, and CE50 v. CE75. Mean values were plotted for CE25, CE50, and CE75.

Risk attitude (r) was calculated for each individual over expected survival of 36 months. The distribution of r was examined in terms of the numbers of the sample who had $r < 1$, $r = 1$, and $r > 1$. Respondents who were unwilling to trade for the risk questions were excluded from this part of the analysis at this stage. The geometric mean was used to produce an average value of r over the whole sample (see Appendix B, Miyamoto and

Eraker, 1985). QALY valuations for BMT and EVAR were recalculated with an adjustment for individual risk attitude according to the method suggested by Miyamoto and Eraker (1985). The equation $(x/t)^r$ was used, where x is the period of time in full health equivalent to the period in ill health (t). This method is not suitable for scenarios rated worse than death or negative, but only one response given by the 54 respondents was negative. This was for EVAR rated by the holistic method. A statistical comparison between risk-adjusted QALYs and holistic values was carried out using the paired t-test and Wilcoxon-sign test, and also between values of the two scenarios from the risk-adjusted QALYs.

Analysis of time preferences

If $y > 14$ for the first time preference question (see Figure 8.4), a negative time preference (*i.e.* preferring to have the bad earlier) is implied. If $y < 14$, a positive time preference is implied. Similarly, if $y > 14$ for the second time preference question, a preference for Scenario C is implied, thus suggesting a positive time preference. If $y < 14$, a negative time preference is suggested. Of course, a value of y that is equal to 14 implies a neutral time preference.

Two time preferences were calculated for each person, using the two time preference responses. Equations (8.7) was used to calculate time preference x from the first response, and (8.8) were used to calculate time preference y from the second response (Cairns and van der Pol, 2000):

$$x = (y / 14)^{(1/(3-12))} - 1 \qquad (8.7)$$

$$y = (\mathrm{y} / 14)^{(1/(30-12))} - 1 \qquad (8.8)$$

The correlation between time preferences x and y were examined.

QALY valuations of the BMT and EVAR scenarios were adjusted for each of the time preferences as shown for x in equations (8.9) and (8.10). These are equations (8.1) and (8.2) adjusted for time preference. In order to get values for the scenarios for each year, equations (8.1) and (8.2) were altered in that the time components were removed. Thus what remains is a QALY weight for one year.

$$\text{BMT} \qquad [\{1/(1+x)\}*\text{U(BMT) in 2}^{\text{nd}}\text{ year}]+[\text{U(BMT) in 1}^{\text{st}}\text{ year}] \qquad (8.9)$$

EVAR $[\{1/(1 + x)^2\}*U(EVAR)$ in 3rd year]+$[\{1/(1 + x)\}*U(EVAR)$ in 2nd year]+[U(EVAR) in 1st year] (8.10)

At this stage of the analysis outliers were excluded from the time preference part of the study.

The paired t-test and Wilcoxon-sign test were once again used to determine whether there were significant differences between valuation methods, this time between the holistic values and the QALY values adjusted for time preferences x and y. These tests were also used to explore the differences between QALY values for the two scenarios when adjusted for time preferences.

Adjusting for time preferences and risk attitude

A simultaneous adjustment was made to QALY valuations taking into account individual time preference rates and risk attitude. QALY valuations for BMT and EVAR were first adjusted for time preference rates as in (8.9) and (8.10). This value was then raised to the power of the value of the individual's risk attitude. The paired t-test and Wilcoxon-sign test were used to explore differences between risk- and time preference-adjusted QALYs and holistic values, and also between values of EVAR and BMT for risk- and time-adjusted QALYs.

Convergent validity

Tests of convergent validity were carried out for the health states and scenario values. Rank orders for the five health states were compared to the rank orders implied by the TTO valuations for these states. Where a respondent ranked current health as preferred to full health, this was considered convergent, because it was assumed to the EQ-5D description of full health must have lacked completeness in the eyes of the respondent. Strong convergency (ranking order of health state values exactly the same as the original ranking order) was tested initially, and then followed by a test of weak convergency (an ordinal ranking might be replaced by equality of values).

Convergent validity was also checked for both QALY valuations and holistic valuations of the health scenarios. The ranking order was between full health, BMT, EVAR, and immediate death. There were therefore three possible pairwise comparisons, between the highest ranking and second-highest ranking, the second-highest ranking and the third-highest ranking, and the third-highest ranking and the lowest ranking. It was impossible to give any scenario a higher value than full health, but it was possible to value scenarios as worse than death. The results were examined to determine which valuation method achieved higher convergent validity.

A χ^2 test was carried out to determine whether there were any correlations between being non-convergent for health states and being non-convergent for scenarios. If there was a correlation between being non-convergent for states and scenarios such that the same people were being non-convergent all the time, it might imply that these people had difficulties with this type of valuation procedure. However, if no relationship was found, such that some people were non-convergent for states and others non-convergent for scenarios, it might imply either that states were more difficult to value than scenarios or *vice versa.*

Sensitivity analysis of wide ranges of indifference

A sensitivity analysis was included in order to examine the extent to which ambiguous responses to health state or scenario valuations and risk attitude questions might affect average values. Ambiguous responses were those which had gaps between where the respondent has stated a willingness to trade and an unwillingness to trade, or wide areas of indifference. The midpoint was taken in the above analyses. This sensitivity analysis addressed the concern that respondents' true values may not have been the midpoint in a wide range of indifference. It determined the extent to which results might vary if values were taken from the extremities of the range. A non-ambiguous response would either have a tick and a cross in adjacent boxes, one non-ticked box between ticks or crosses, or these with the "=" symbol instead of gaps. For the sake of this sensitivity analysis, responses were considered ambiguous if the indifference area was wider than this.

The comparisons of holistic and QALY methods and analysis of risk attitudes were repeated using the lowest and highest possible values for individuals who gave wide ranges of indifference. The purpose of this analysis was to determine the degree of possible error that might be made by assuming the midpoint for these somewhat ambiguous responses.

Sensitivity analysis in relation to truncation of the EVAR scale to 24 months

The above anlayses of holistic values was repeated with values of 24 for EVAR adjusted in turn to 33, 34, 35, and 36. The possible effects of truncation of the EVAR scale were examined in terms of comparisons between QALY and holistic values, and comparisons between holistic values for EVAR and BMT.

Unwillingness to trade

The data were examined by individual respondents in order to determine the extent to which members of the sample were unwilling to trade life months in exchange for an increased HRQoL.

Respondents' comments

The final page of the questionnaire gave respondents the opportunity to comment on aspects of the survey. Respondents' comments were examined and categorized according to content.

8.6 Results

There was a sample of 61 respondents, out of whom 40 were employees or students at ScHARR, 13 were angio day ward or secretarial staff from the Vascular Institute of the Northern General Hospital, 3 were patients with peripheral vascular disease who had just had a procedure on the angio day ward, and 5 were friends or relatives of the author.

8.6.1 Background characteristics

The mean age of the sample was 38.8 years (median 38.0), with a range of 19 – 70 years (Table 8.A.1). The mean age at completion of full-time education was 22.2 years (median 21.5), with a range of 15 – 39 years (Table 8.A.1). This question was intended as an indication of level of education. It is a very crude measure, because some respondents obtained education qualifications later in life. However, the results imply a relatively well-educated sample. A total of 42 (68.9%) of the sample were female, compared to 19 (31.1%) males.

Table 8.2 shows EQ-5D responses for the sample. The sample was relatively healthy overall. Over 90% reported no problems with mobility, self-care, and usual activities, and over 80% reported no problems with pain/discomfort and anxiety/depression. A total of 42 (68.9%) rated their current health as 11111 in the EQ-5D question. The mean EQ-5D score for current health was 0.949, median 1.000. This is higher than the general population score of 0.83 (Kind *et al*, 1998).

Occupation types have been grouped into four broad categories, as shown in Table 8.A.2. The majority of this sample were either semi-skilled or skilled. Since this sample was largely drawn from the university and the hospital, and was not particularly representative of the general population, this result is not surprising.

8.6.2 Exclusions

Out of the 61 respondents, seven people gave such unclear responses that it was impossible to ascertain their true meanings. Details of these respondents are given in Table 8.A.3. For three respondents the problem was that they answered the better than death question and also the worse than death question for the same state/scenario. For example, one respondent stated that she preferred 36 months with EVAR to 16 months in current health (EQ-5D state 11111), but in the worse-than-death option she stated that she preferred immediate death to 24 months in full health followed by 12 months with EVAR. This result is reminiscent of work by Dolan and Gudex (1995), who found that a small number of people state preferences for any state of health with a 10-year duration (including EQ-5D state 33333) to nine years in full health followed by a worse health state. Dolan and Gudex suggested that these respondents were not answering the questions being put to them, but rather incorporating assumptions into the equation that were known only to themselves. This may be the explanation of the preferences expressed by these three respondents. It may indicate a lack of understanding. The other four respondents in Table 8.A.3 completed one or more questions incorrectly. All seven of these respondents were excluded from further analysis.

8.6.3 Range of indifference values

Of the respondents remaining in the analysis, 18 (33.3%) gave a range of indifference (*i.e.* "=" in more than one box, or wide empty gaps between ticks and crosses). Table 8.3 describes the degree to which this occurred for each valuation exercise. The scale of the TTO exercise was from 0 to 36 months in 2-monthly intervals. Most indifference ranges covered either two or three intervals. However, indifference ranges extended to 19 intervals (*i.e.* stated indifference over the whole of the scale) for a response to the current health valuation. The midpoints were taken for responses with wide indifference ranges.

It is possible that the wide ranges of indifference may reflect difficulty with the valuation tasks. This difficulty could either be in understanding the task required, or in the process of making the decisions requested. In order to test the validity of these responses, the results were subjected to a sensitivity analysis to determine whether the results would have been significantly different over their range of answers. This is reported in Section 8.6.12.

In order to explore the extent to which this phenomenon might have affected average values across the sample, the mean and quartile values for the lowest value at which the respondents stated that they were willing to trade and the highest value at which they stated that they were not willing to trade were determined (Table 8.4). If the whole sample gave a clear point of indifference, the average values for lowest willing to trade values and highest not willing to trade would be close together. However, if there are wide indifferences ranges across the sample these values will be further apart. As can be seen from Table 8.4, the mean lowest willing to trade value and highest non-willing to trade values for current health are close (32.96 and 32.44 respectively), indicating that indifference ranges were narrow for valuations of current health. The state most affected by wide indifference ranges appears to be chronic renal failure, with mean values of 18.72 and 12.78 respectively.

8.6.4 Ranking of health states and scenarios

Respondents ranked five health states in order of preference. These were full health, current health, chronic renal failure, stroke, and immediate death. The results are presented in Table 8.5. The entire sample ranked either full health or current health top, with 29 (53.7%) rating full health as best, 7 (13.0%) stating a preference for their own current health state, and 18 (33.3%) rating full health and current health state equally. A total of 49 (90.7%) rated chronic renal failure as better than death. A total of 45 (83.3%) rated stroke better than death. Three (5.6%) respondents rated stroke equal to death. A total of 29 (53.7%) ranked full health and current health as the top two states, followed by chronic renal failure, then stroke, then death.

The preference for BMT over EVAR was overwhelming (Table 8.6), with 45 (83.3%) ranking full health first, then BMT followed by EVAR followed by death. A total of three (5.6%) thought EVAR was worse than death, and none thought BMT worse than death.

8.6.5 Health state values

Table 8.7 shows the statistics for the health states. Both the mean and median values indicate that the average order of preference is current health ≻ chronic renal failure ≻ stroke. Five respondents rated chronic renal failure worse than death, and ten rated stroke as worse than death.

Current health received a high average value, with a mean value of 0.936 (median 0.972). This is close to the mean EQ-5D score for the sample of 0.949 (median 1.000). The two morbidity states are valued markedly lower than current health. The mean for chronic renal failure is 0.459 (median 0.639), whereas the mean rating for stroke is 0.053 (median 0.194). Both these sets of statistics demonstrate negative skews, reflecting the negative values for these health states.

8.6.6 Health scenario values

Table 8.8 shows the statistics for the health scenarios BMT and EVAR calculated by holistic valuation and QALY valuation. Significant differences were found between holistic and QALY valuations of the BMT and EVAR scenarios by both the t-test and the Wilcoxon-sign test ($p < 0.01$). The QALY method consistently gave higher valuations for both scenarios. The mean values for EVAR were 22.07 (QALY) and 17.26 (holistic), and for BMT were 22.02 (QALY) and 19.94 (holistic).

The paired t-test and Wilcoxon-sign test were also used to determine whether there were significant differences between the valuations for the two scenarios. There were significant differences between the scenarios for the holistic method, with BMT scoring higher than EVAR ($p < 0.01$). However, for the QALY valuations the differences between scenarios were not significant ($p > 0.90$ for t-test, and $p > 0.45$ for Wilcoxon-sign test).

Table 8.8 shows holistic values on a scale of 0 to the individual's current health state. Table 8.9 shows the results of the scenario valuations when holistic values are chained through current health so that they are also on a death – full health scale. There is no change in the conclusion that QALY values are higher than holistic values, and BMT is valued significantly higher than EVAR by the holistic method. The holistic valuations are lower than non-chained valuations (EVAR mean 16.45; BMT mean 18.82), as would be expected. The QALY valuations were not affected.

Figure 8.6 is a scatter plot showing how QALY and chained holistic valuations for EVAR and BMT are related for individuals. The correlation between QALY and chained holistic valuations for BMT is 0.71. For EVAR the correlation between the two methods is 0.35. It is not surprising that there should be a greater correlation between QALY and holistic values for BMT. The BMT scenario is simpler, comprising a 98% chance of being in current health for two years. However, the EVAR scenario is more complex, and were the holistic and QALY methods to disagree, the differences would have been expected to show more for the EVAR scenario.

As discussed in Chapter 4, the primary analysis of scenarios used the equation $-x/t$ to calculate values worse than death. This formula constrains negative values to ≥ -1. However, BMT and EVAR scenario values were also calculated using the worse than death equation $-x / (t - x)$, which does not constrain values to a lower boundary of -1. One respondent rated EVAR worse than death by the holistic method. The results of this analysis are shown in Table 8.10. There are no changes in the BMT scores, because no one rated BMT worse than death. However, for EVAR the mean QALY score is much lower than the holistic score. In fact, according to the QALY method EVAR is worse than death. There is a very wide SD for the QALY valuation of EVAR, and this result is probably due to some outliers who had scores of –35 for the states of chronic renal failure and stroke. The median values for EVAR are unaffected by which formula is used.

8.6.7 Analysis of risk attitudes

Table 8.11 shows the internal consistency conditions for the risk attitude questions, and the level of agreement among the sample. The majority of the sample were logically consistent, such that their values for CE25 were less than or equal to CE50, which in turn were less than or equal to CE75 (81.1%). Of the 10 illogical respondents, 7 (13.2%) valued CE25 as greater than CE50, but were logical in their valuations of CE50 compared to CE75. This could be because they only began to grasp the nature of the task when they got to the second and third risk attitude questions.

The statistics for the risk attitude questions are shown in Table 8.12. One person out of the sample size of 54 who had been included in the entire analysis had missing values for one of the risk attitude questions. Expected survival for CE25 is nine months, for CE50 is 18 months, and for CE75 is 27 months. A risk neutral person would be expected to give these values. However, the mean value given for CE25 is 14.3 months, indicating that on average the sample was risk seeking for the earlier part of the 36 months. The mean CE50 value was 18.5 months, which is close to risk neutrality midway through the 36 months. The mean CE75 value is 24.2 months, indicating that the sample were on average risk averse for the latter half of the 36 months. These results are plotted in Figure 8.7. The straight line is how a risk neutral individual would be plotted. At U = 0.25 expected survival would be 9 months; at U = 0.50 it would be 18 months; and at U = 0.75 it would be 27 months (McNeil *et al*, 1981).

There were 16 respondents with r < 1 (0.422 – 0.981), 3 respondents with r = 1 (1.000 – 1.002), and 31 respondents with r > 1 (1.012 – 3.525). There were 3 respondents with r = 37.255, who gave values of 35 to their risk attitude questions. One possible explanation for this was that they were not willing to trade at all for the risk attitude questions. Perhaps they did not like this kind of gamble, and refused to participiate. Another possibility is that they may not have fully understood the task. The reasons for these values are unknown, but these responses were excluded from further analysis of risk attitude on the grounds that they were invalid.

The geometric mean of *r* was 1.21 (where r<1 indicates risk aversion, r=1 indicates risk neutrality, and r>1 indicates risk seeking), indicating that on average this sample had a tendency to be risk-seeking.

Table 8.13 compares risk-adjusted scenarios for the QALY valuations with the holistic valuations chained to full health. The QALY values were adjusted by the value of r for each individual (see equation 8.3). N = 50 for risk-adjusted QALY scenarios because one respondent did not provide risk attitude information, and three had to be excluded because they had r = 37.255 (see above). Mean and median values for the risk-adjusted scenarios are shown. The mean results of adjusting for risk attitude look unbelievable. The extremely high mean values appear to be due to the unexpectedly high values of r obtained from many of the sample. Thus the mean values seem to be exaggerated by the strength of risk-seeking behaviour. In this instance the median values obviously made more sense.

In Table 8.9 the mean QALY value for BMT is 22.020 (median 22.867), and the mean QALY value for EVAR is 22.069 (median 23.000). There are no significant differences between these mean values, implying that by the QALY method the two scenarios are valued virtually equally. As shown in Table 8.9, and again in Table 8.13, the holistic mean value for BMT is 18.816 (median 20.889), and the mean holistic value for EVAR is 16.342 (median 17.250). Thus according to the holistic method the BMT scenario is preferred to EVAR, and this difference is significant. However, as shown in Table 8.13. adjusting QALY values for risk attitudes has the effect of making EVAR appear preferable to BMT (medians 34.075 and 31.007 respectively). This is not significant according to the Wilcoxon-sign test (p=0.49). If, as indicated by the risk attitude results, the sample were on average risk seeking, it would not be surprising that they would prefer EVAR to BMT. They may have been willing to take the risks involved in the EVAR scenario in order to extend their life expectancy by a year.

8.6.8 Analysis of time preferences

Figure 8.4 shows the two questions in diagrammatical form. The responses to the time preference questions are shown in Table 8.14. There were 53 respondents who provided time preference estimates. Each given value of x is listed in the first column. For each value of x, a value of y was provided. These y values are listed in the second column. For example, where a value of x of 1 was given, a value of y of 14 was given. Respondents who gave x a value of 10 variably gave y values of 10, 14, 18, and 24. The columns entitled "Time preference" show whether x and y indicated positive or negative time preferences. It is clear than many respondents demonstrated different time preferences for x and y.

Out of the 53 responces to this question, 29 (54.7%) demonstrated neutrality of time preference for x, compared to 26 (49.1%) for y. As expected, the relationship between x and y does not appear to be linear (see Figure 8.8). Supposing this question measures pure time preference, it would be expected that where $x < 14$ then $y > 14$ and *vice versa*. Likewise, when $x = 14$ then $y = 14$ also. The correlation between x and y is only 0.07.

Table 8.15 shows the time preference statistics. As shown in Table 8.14, the range of values given for x and y lies between 1 and 30 for all bar one. There is one outlier value of 365 for x. If this person's value is included, there is an average strong negative time preference value for x (mean 21.7358). However, when this outlier is excluded the mean value for x is 15.1346. The 50th percentile value (median) is 14 whether or not the outlier is included. as expected from Table 8.14. in which the most frequent values for x and y are shown to be 14.

Applying equations 8.7 and 8.8 produced the time preference statistics presented in Table 8.16. Both x and y produce negative rates. thus indicating that respondents preferred to get the bad over with quickly. This was more marked for y (mean –0.026) than x (mean –0.002). However, the medians were zero for both measures.

Individual values for x and y were applied to QALY valuations of the scenarios, and the results are shown in Table 8.17. There is clearly little difference between the effects of the two discount rates when compared to each other. However, discounting results in higher QALY values. The overall results are still the same. The QALY and holistic scenario valuations are still significantly different from each other, and the QALY valuations of the scenarios do not differ significantly from each other.

8.6.9 Adjusting for both risk attitude and time preferences

Adjusting discounted QALY valuations for risk attitude gave the results shown in Table 8.18. Using time preference rate x, the median risk-adjusted QALY values are 30.7 for BMT and 35.3 for EVAR. The equivalent values for time preference rate y are 29.8 and 31.7 respectively. The differences between the median values for BMT and EVAR are wider when adjusted by time preference rate x. However, the differences between EVAR and BMT were not statistically significant according to the Wilcoxon-sign or the paired t-test.

8.6.10 Convergent validity for health states

Rank orders and TTO scores for the states of full health, current health, chronic renal failure, stroke and immediate death were compared within individuals. Some respondents ranked current health as better than full health in the ranking exercise. Although illogical on the face of it, it is possible that these respondents did not feel that the EQ-5D state 11111 adequately described their state of good health. However, they could not rate current health as greater than full health in the TTO exercise. In the analysis of convergent validity, therefore, these respondents were considered convergent. For the other health states, if the difference between ranking and TTO orderings was in terms of being equal to one where the other measure rated one greater than the other, this was described as weakly convergent. Thus strong convergency and weak convergency are reported in this section.

The results of the analysis of convergency for health states is reported in Table 8.19. There were four pairwise comparisons. A total of 23 (42.6%) were strongly convergent for 4/4 pairwise comparisons. A total of 24 (44.4%) were strongly convergent for 3/4 pairwise comparisons, and a total of seven (13.0%) were strongly convergent for 2/4 pairwise comparisons. If weak convergency is included, a total of 38 (70.4%) were weakly convergent for 4/4 pairwise comparisons. A total of 12 (22.2%) were weakly convergent for 3/4 pairwise comparisons, and a total of 4 (7.4%) were weakly convergent for 2/4 pairwise comparisons.

Details of people who gave non-convergent or weakly convergent responses for health states are given in Table 8.A.4. In their comments, some of these respondents indicated a lack of understanding of the TTO procedure, which may have led to their non-convergent responses. Some respondents said that they had difficulty imagining the hypothetical health states.

The level of strong convergency between implied ranking from TTO valuations and the original ranking order of the health states is relatively high, with only seven respondents demonstrating more than one non-convergency out of the four pairwise comparisons. If weak convergency is allowed, the results are even better, with only 16 (29.6%) respondents showing any non-convergency at all.

8.6.11 Convergent validity for health scenarios

The term strong convergency will be used to refer to those cases in which the same rank order for the scenarios is implied by the TTO valuations of those scenarios as is stated in the original ranking exercise. The term weak convergency will be used to refer to those cases in which the only difference between the rank by the two methods is that scenarios are equal by one method where one was preferred to the other by the other method. The term non-convergency will be used to describe those cases for which there are preference reversals implied by the original ranking method and the TTO valuations (*i.e.* one scenario is preferred according to the original state rank order, but the other scenario is preferred according to the TTO valuations).

The results of the convergent validity tests between QALY and holistic (chained to full health) valuations of the scenarios are presented in Table 8.20, and by individual in Table 8.A.5. A total of 27 (50%) respondents were strongly convergent for 3/3 pairwise comparisons by the holistic method, compared to 22 (40.7%) for QALY valuations. A total of 27 (50) were strongly convergent for 2/3 pairwise comparisons for the holistic method, compared to 32 (59.3%) for the QALY method. A further 14 (25.9%) of holistic valuations were weakly convergent in that these respondents rated BMT and EVAR equally by the holistic method while previously ranking BMT higher than

EVAR. If weak convergency is allowed, 41 (75.9%) respondents achieved weak convergency for 3/3 pairwise comparisons, and 13(24.1%) achieved weak convergency for 2/3 pairwise comparisons.

The convergent validity is good for both valuations methods in this study, with all respondents achieving a minimum convergence of 2/3 pairwise comparisons. Five more respondents achieved strong convergence for 3/3 pairwise comparisons for the holistic method than the QALY method, indicating slightly higher validity for this method. However, if weak convergency is allowed, this number increases to 19 more for the holistic method.

A χ^2 test was carried out to determine whether there were any correlations between being non-convergent for health states and being non-convergent for scenarios. The test was done for both QALY and holistic valuations (Table 8.21). The observed values are close to the expected values. The Yates corrected χ^2 value is 1.16 (p = 0.28) for the QALY method, and 2.08 (p = 0.15) for the holistic method. There is therefore no association between non-convergent responses for health states and scenarios for either method of valuing scenarios.

8.6.12 Analysis of sensitivity to wide ranges of indifference

Eighteen (33.3%) respondents provided wide ranges of indifference (see Table 8.3). The maximum possible range of indifference is if respondents placed the "=" symbol in all 19 boxes, as one respondent did for the state of current health. In the majority of cases in which there was a range of indifference, the range covered two or three boxes (four to six months). The above analyses of QALY and holistic valuations and attitude to risk were repeated with ambiguous responses set at maximum and minimum possible values.

Comparison of holistic and QALY methods

The results of the sensitivity analysis of QALY and holistic values are shown in Table 8.22. There is a greater difference between highest and lowest values for the holistic method, with mean differences of 0.90 months (BMT) and 0.78 months (EVAR). The equivalent mean differences for the QALY method are both 0.45 months. The differences between the holistic and QALY methods are still highly significant by both the paired t-test and the Wilcoxon-sign test (p < 0.001) and the ranking order of the scenarios was not affected by the lowest or highest possible indifference values being used.

Sensitivity analysis of risk attitude

Sensitivity analysis of risk attitude showed that r has a lowest possible geometric mean of 1.16, a midpoint geometric mean of 1.21, and a highest possible geometric mean of 1.28. Thus these results suggest that the sample are, on average, risk-seeking overall no matter whether we take the lowest, midpoint or highest possible value for r.

The results of these sensitivity analyses suggest that the findings of this study were not unduly affected by the 18 respondents who gave wide ranges of indifference.

8.6.13 Analysis of sensitivity to truncation of the scale for EVAR

The TTO question for valuing EVAR had an upper limit of 24 months. There was a concern that those respondents who gave the uppermost value of 24 months might have given a higher value if they had been able. This section reports a sensitivity analysis to determine the extent to which the results might have been different if respondents had given values of 33 to 36 for EVAR.

Six (11.11%) respondents stated an indifference value of 24 months for EVAR. This analysis allowed the possibility that these respondents who gave the maximum possible value for EVAR may have wished to give a higher value. The statistics for EVAR are reported for the sample when these six respondents are given values of 33 to 36 (Table 8.23). The results suggest that it makes no difference to the overall results in terms of significant differences between methods for EVAR if those with values of 24 months are adjusted up to values of 33 to 36 months. As expected, the mean values for EVAR are higher under these assumptions, ranging from 17.2 to 17.5 months as opposed to a mean value of 16.3 in the original analysis. However, these values remain lower than the QALY value for EVAR (which was not truncated) of 22.1 months, and the differences remain statistically significant ($p < 0.001$).

As far as comparisons between EVAR and BMT for the holistic method, there were no reversals of preference if values of 24 for EVAR were adjusted upwards to 33 to 36 months. Values for EVAR were consistently lower than BMT, and these differences were statistically significant by both the t-test and the Wilcoxon-sign test ($p < 0.05$). The results are reported in Table 8.24.

The fact that there were only six respondents who gave the maximum possible value for EVAR, and the results of this sensitivity analysis are reassuring as regards the main analysis. It would appear that the accidental truncation of the EVAR scale to 24 months has had little effect on the results.

8.6.14 Unwillingness to trade

Analysis of the data by individual respondents indicated that four (7.4%) respondents were unwilling to trade for both health states and scenarios. These respondents gave maximum values for all states and scenarios.

8.6.15 Respondents' comments

The last page of the questionnaire gave respondents the opportunity to make their own comments about the study, the health states and scenarios, or whatever they felt was relevant. The comments are described in detail in Table 8.A.6, and are categorised in Table 8.25.

A total of 19 (31%) out of 61 respondents provided written comments. The largest category of comments regarded difficulty with the questionnaire tasks, with 11 respondents saying they found the valuation tasks difficult and/or confusing. Four respondents gave suggestions for possible improvements to the questionnaire to aid better understanding for the future. Four respondents made comments to the effect that it was difficult to accurately value states and scenarios of which they had no experience. Three respondents made comments in praise of the questionnaire, and one respondent commented that his preferences might change with time.

Some respondents made verbal comments. Some of these were also regarding the difficulty in completing the questionnaire. One respondent said of the time preference questions that he preferred the ill health state to be later, but not right near the end of his life.

8.7 Discussion

8.7.1 Strengths and weaknesses of the study

One of the strengths of this study was that it explored areas which have previously received little attention in the literature. Attitudes to risk were explored over risks of morbidity and mortality over a short-term life expectancy. Attitudes to *ex ante* risks were examined, as were methods of adjusting individuals' QALY valuations for risk *ex post*. Time preferences were also measured, which has not previously been done over such a short life expectancy. This study revealed the complexities involved in assessing time preferences over such a short life expectancy. Difficulties in using the Miyamoto and Eraker (1985) method of adjusting QALY values for risk attitude were also brought to light, as the risk attitudes for members of this sample were frequently non-linear.

A weakness in the design of the study was the accidental truncation of EVAR values to 24 months. However, this problem was spotted and a sensitivity analysis was conducted to determine the possible effects of this error on the results of the study. The sensitivity analysis involved examining comparisons between QALY and holistic values of EVAR, and holistic values of EVAR and BMT, when the six respondents who gave values of 24 months to EVAR had their values changed in turn to 33, 34, 35, or 36 months. The results were encouraging in that mean holistic values for BMT remained significantly greater than mean values for EVAR, and holistic values for EVAR remained significantly lower than QALY values for EVAR.

Finally, current health was used as the reference state in the holistic TTO valuations of BMT and EVAR, whereas full health was used as the reference state in the valuations of the health states and the worse than death valuation of EVAR. However, in most members of the sample current health was given a value very close to full health, so the effects of chaining on the results were negligible. The holistic values were chained through individual values for current health to make them comparable to QALY values for the scenarios.

8.7.2 Implications for the QALY

This study explored the issues involved with adjusting QALY values for individual risk attitudes and time preferences. The results and implications in relation to the QALY are discussed below.

Effects of risk attitude

The QALY valuations for this study were initially constructed by multiplying the utility for each health state by the duration of that health state, and multiplying the product by the probability of entering the health state. Thus no account was taken, in this initial analysis, of the possible disutility associated with the risks of each procedure (*i.e.* the possibility of risk aversity).

Adjusting QALY valuations for risk attitude gave the results shown in Table 8.13. The median value for EVAR was 34.1 compared to a median value of 31.0 for BMT. This was a large increase from the non-risk adjusted medians of 23.0 and 22.9 respectively (Table 8.9), and resulted in a higher value for the EVAR scenario than the BMT scenario. However, this difference did not achieve statistical significance (p = 0.49). By the holistic method, BMT was found to be valued significantly higher than EVAR (Table 8.9). Although the difference between median risk-adjusted QALY values for BMT and EVAR were not found to be statistically significant, these results suggest a reversal of preferences depending upon which valuation method is used.

Although a high percentage (80.8%) of respondents demonstrated a logical consistency in their responses to the risk attitude assessment questions (Table 8.11), the mean risk attitude for the sample was not linear (Table 8.12 and Figure 8.7). Thus these respondents were not constant in their attitude to risk for different levels of risk. In fact, the sample tended to be risk seeking over the very short term, and risk averse over the longer term of expected survival for this relatively short time horizon of 36 months. McNeil *et al* (1978) also found that risk attitude varied over expected survival for some respondents such that they were risk seeking over the earlier expected survival and risk averse later. The results from Spencer (2000) tentatively suggest that people may show aversion to risk towards the end of life. This agrees with Figure 8.7, which suggests that respondents are risk averse over the later period of expected survival.

It was the original intent to separate quantity effect from risk attitude in order to obtain pure risk attitude. However, as explained earlier in this chapter, the quantity effect question was excluded from the final version of the questionnaire because it proved too difficult for respondents to complete. It is therefore difficult to be certain from these results that risk attitude is not constant. It may be the quantity effect factor that is variable. If this is the case, it would seem that the sample of respondents on average placed a higher value on the latter half of the 36 months life expectancy, thus leading to the appearance of risk averse behaviour. There is the question of whether or not it is important to be able to separate the two factors. After all, it is unclear whether it matters if it is risk attitude or quantity effect that is variable when respondents do not separate them in their valuations.

The extreme mean results from adjusting QALY values for risk attitude throw doubt on the validity of using this method for a sample showing such widely differing risk attitudes. It would have been useful to be able to separate those with r > 1, r = 1. and r < 1. The analysis could then have been performed for each of these three groups. However, the number of respondents in each category was too small to do so (see Section 8.6.7).

This study used the method suggested by Miyamoto and Eraker (1985) to adjust QALY values for risk attitude. There was wide variation in values of r across the sample. Large values of r, which suggested respondents had risk seeking attitudes, caused massive boosting of QALY scores beyond what one could expect to be realistic. This is the explanation for the large sample means (Table 8.13).

It is clear from these results that the holistic method provides different results to the risk-adjusted QALY. Namely, the mean value for BMT is greater than that for EVAR according to the holistic method, whereas the results are opposite (though non-significant) for the risk-adjusted QALY method. It may be the result of the differences between valuing risk *ex ante* and *ex post*.

Further research is required into the extent to which risk attitude, as measured using the methods suggested by McNeil *et al* (1978), can be applied meaningfully to real life decision-making. This study has added to the research indicating that risk attitude is non-linear over expected survival. Research is also required into the extent to which risk attitude varies with context. For example, it is possible that an individual might demonstrate risk seeking behaviour by choosing EVAR over BMT, because he might be willing to take the risks involved in order to extend his life expectancy. However, this same individual might demonstrate risk averse behaviour by choosing not to use the London transport system for fear of terrorist attack.

Time preferences

Table 8.17 shows that this sample demonstrated a variable mean time preference. with a mean of -0.002 for x and –0.026 for y. Median values were zero for both x and y. Cairns and van der Pol (2000) found a mean time preference of 0.07 in their random sample of 897, and a median of 0.06. Dolan and Gudex (1995), however. found that median time preferences were close to zero for all the seven health states in their study. However. mean time preference values ranged from –0.029 to 0.014. Only one-third of their sample demonstrated a positive, negative or neutral time preferences over all the states. Two-thirds demonstrated different time preferences for different health states.

The results of the present study more closely resembled those of Dolan and Gudex (1995) than Cairns and van der Pol (2000).

Table 8.14 shows that many respondents demonstrate different time preferences for *x* and *y*. It is likely that these values do not reflect pure time preference. The time period involved is very short (36 months), and there is a clear "death point" in view. These questions may be confounding time preference with context (Dolan and Gudex. 1995; Tsuchiya, 2001). This hypothesis is supported by the verbal comments of one respondent, who explained that he would prefer the ill health state to occur later, but he would not like it to occur too near to his death. It is likely that the time preference questions were not actually measuring time preference, but that the preferences obtained were a function of how proximate death was. It would be difficult to avoid this problem while assessing time preferences over short life expectancies.

It is doubtful that they reflect pure time preference. It is even likely that time preference is only a small component of the findings. The differences in sign for *x* and *y* may spring from sequence effects, or quantity effects, or context (Gafni, 1995). Adjusting QALY scenario values for time preferences using either *x* or *y* results in raising QALY values slightly (Table 8.17). However, there is no change in the overall result in terms of comparison between QALY and holistic valuations.

Implications

The implications of this research are that attempts to adjust QALYs for risk attitude may be fraught with problems. Attempting to adjust for time preferences over a short life expectancy has the associated problem of the proximity of death and the effects this has on the time preference results. This study has important implications for the validity of such measures over short-term life expectancies. Further research is required into methods for adjusting QALYs for risk attitude and time preference. However, it is notable that these factors are implicit in holistic valuations.

8.7.3 Differences between the holistic and QALY methods

The holistic and QALY methods gave significantly different results, both in the initial analysis and after adjustment of QALYs for risk attitude and time preferences (Tables 8.9, 8.13, 8.17, 8.18). The QALY method gave consistently higher results than the holistic method. For the holistic method EVAR was rated significantly lower than BMT ($p < 0.01$). For the QALY method the mean value was slightly higher for EVAR than

BMT, though this difference was not statistically significant. The implications for CUA are that both methods would suggest the use of BMT. BMT was preferred according to the holistic method, and there were no significant differences according to the QALY method. Given that the BMT scenario is cheaper (Thomas, 1999), these results therefore suggest that it would be more cost effective to use BMT than EVAR.

The results of this study were different from the results of the studies reported in Chapters 5, 6 and 7. In Chapter 6 QALY scores were lower than holistic values for most IBS profiles. The results of the varicose veins study reported in Chapter 7 showed that QALY and holistic scores did not differ significantly. It is important to remember that each of these studies used different types of health profiles. The profiles used in the IBS studies were described in terms of proportion of time in each health state. Several possible explanations have been discussed for the finding that QALYs tended to be lower than holistic scores, including the possibility that respondents may have been showing insensitivity to scope in the holistic valuations (Healey and Chrisholm, 1999). The study described in Chapter 7 asked respondents to value states and profiles associated with varicose veins. Some respondents commented that they felt these states and profiles were relatively mild on the scale of possible ill health (Table 7.A.9). It is possible that the TTO scale used in the varicose veins study was insensitive to very small amounts that respondents may have been willing to trade for these states and profiles. This present study was very different again, focussing on short-term scenarios and studying the effects of *ex ante* risks on preferences.

The results of this study were robust to sensitivity analyses over ranges of indifference. They were also robust to the sensitivity analysis of the effects of truncation of the EVAR scale to 24 months on comparisons between holistic and QALY valuations (Table 8.23). Truncation of EVAR to 24 months had no affect on comparisons between holistic valuations for EVAR and BMT when EVAR values of 24 months were adjusted to 33 to 36 months (Table 8.24).

8.7.4 Comparison of health states and scenarios with rank ordering

There is a high level of convergent validity between health state values and original ranking order of health states (Table 8.19). A total of 47 (87%) of the sample achieved strong convergency for three or more of the four pairwise comparisons. If weak consistency is allowed this rises to 49 (92.6%).

All respondents achieved strong convergency for two or more of the three possible pairwise comparisons of health scenarios for both holistic and QALY valuation methods (Table 8.20). However, 27 (50%) of the sample achieved strong convergency in all pairwise comparisons for the holistic method, whereas the corresponding figure was 22 (40.7%) for QALY valuations. If weak convergency is allowed, the holistic method achieved strong convergency for all three pairwise comparisons in 41 (75.9%) respondents. Thus, convergent validity was slightly higher for the holistic method with only strong convergency, and notably higher with weak convergency included.

It was hoped that a comparison of convergent validity between the two methods of valuing scenarios might help in determining which method was more likely to reflect respondents' true preferences. The fact that holistic TTO rankings of EVAR and BMT were closer to the original ranking order of these scenarios than the QALY valuations could suggest that the holistic method is more valid in this instance. However. there is the caveat that the original ranking might not be a gold standard for determining ordinal preferences, but a learning process by which respondents formed their preferences. In support of the validity test, however, the profile ranking took place later in the questionnaire than that of the health states, when respondents were already used to the idea of valuing health.

The results suggest that the holistic method is slightly better at reflecting ordinal preferences than the QALY method. The risk attitude data suggests that risk attitude is not linear over short-term expected survival, as is assumed by the QALY algorithm. It appears that the *ex ante* risks involved were given a heavier weight that assumed by the QALY.

A glance at Table 8.A.6 will reveal that the majority of respondents' comments related to the difficulty in completing the questionnaire. However, the finding that BMT was on average rated significantly higher than EVAR for the holistic method was too systematic to be due to a lack of understanding of the tasks on the part of the respondents. There is no reason to suppose that the difficulty associated with the questionnaire was in any way biased towards one scenario or the other.

8.7.5 Implications

The holistic valuations of scenarios allowed *ex ante* risks to be taken into account by the respondents. These risks may be the causal factor in EVAR being valued lower than

BMT by the holistic method. The finding that holistic valuations were significantly lower than QALY valuations suggests that there may be greater disutility attached to these scenarios than would be picked up by the QALY algorithm. This could have important implications for health care resource allocation.

In this study BMT and EVAR were found to be similar by the QALY algorithm. However, adjusting QALY valuations for median risk attitude resulted in EVAR being valued higher than BMT (although the difference did not reach statistical significance). This could have implications for choice of treatment in terms of cost-effectiveness. BMT is cheaper than EVAR (Thomas, 1999). If the holistic method of valuation were to be used, the findings would be that BMT was altogether more cost-effective in terms of monetary cost and HRQoL. However, if the risk-adjusted QALY algorithm was used, it is possible that EVAR might be found to be more cost-effective. depending on cost per QALY values. There may be a difference in treatment choice depending on the valuation method used. This could be potentially misleading, because according to the ranking procedure, most people preferred BMT. This is not reflected in QALYs nor in risk-adjusted QALYs. It is better reflected by the holistic valuations.

This reversal of preferences according to choice of valuation method is a very important issue. The fact that the two methods give opposite results in terms of comparisons between the two scenarios, twinned with the finding of slightly higher convergent validity from the holistic method, demonstrates the importance of further research into holistic valuations. Specifically, the issue of valuing *ex ante* risks requires further exploration. It would appear that people value scenarios differently according to the level of *ex ante* risks attached. This needs to be explored systematically over differing levels of risk, in different contexts, over different durations. In addition to acquiring a greater understanding of individual risk attitudes, it may be that such research allows the formation of a model of predicting preferences over *ex ante* risks that can be used to adjust QALY values.

Another issue requiring further research is that of time preferences. These are implicit in the holistic method of valuation. However, if improvements are to be made in the ways in which QALY values are adjusted for time preferences, further research into time preferences over different time horizons is required. Further research is also required into the relationship between time preferences and attitude to risk.

The results of this study could have important implications for resources allocation with regard to treatment for large AAAs in unfit patients. However, these results were obtained using a healthy population. The target group of AAA patients would be likely to have a significantly lower value for current health than the mean value for the sample used in this study. It is possible that the different perspective of patients might lead them to make different choices over BMT and EVAR. It is therefore important to obtain valuations for AAA scenarios from AAA patients themselves.

8.8 Conclusions

This study showed that there were significant differences between valuations obtained by the traditional *ex post* representation of preferences provided by the QALY algorithm and holistic valuations of profiles containing significant *ex ante* risks. The holistic method of valuation provided results that better matched the original ranking orders of the profiles. Although the original ranking of the profiles is thought to be a process in which preferences over the profiles are formed, the exercise was placed halfway through the questionnaire, when respondents were already used to the concept of valuing health states. There were only two scenarios to rank, furthermore facilitating the ease of the ranking procedure. It would seem therefore that in this study the holistic method better reflected respondents' preferences.

This study demonstrated the difficulties involved in obtaining personal time preferences for health over time horizons as short as 36 months, in which death is imminent.

After adjusting for risk attitude there was a non-significant preference for EVAR over BMT according to the QALY method. This is a clear indication of how the QALY algorithm and holistic methods of valuation may result in different values for health profiles, which may have important implications for the distribution of health care resources. The risk attitudes of individual members of the sample were frequently non-linear, and the overall average values shown in Figure 8.7 are non-linear. Mean risk-adjusted QALYs were unclear (Table 8.13), and it was necessary to use the median values. This led to the premise that it may not be legitimate to use mean values for r when r varies so much over the sample. Bearing this in mind, it would seem that the holistic valuations may be a more reliable estimate of preferences than risk-adjusted QALYs in this study.

The results of this study suggest that BMT is the preferred method of treatment for most people. Being the cheaper treatment option, it would also be more cost-effective. However, it must be remembered that these results come from a non-patient population with a high mean value for current health. In a patient population with poor health, such as unfit patients with large AAAs, the results of the valuations may be quite different.

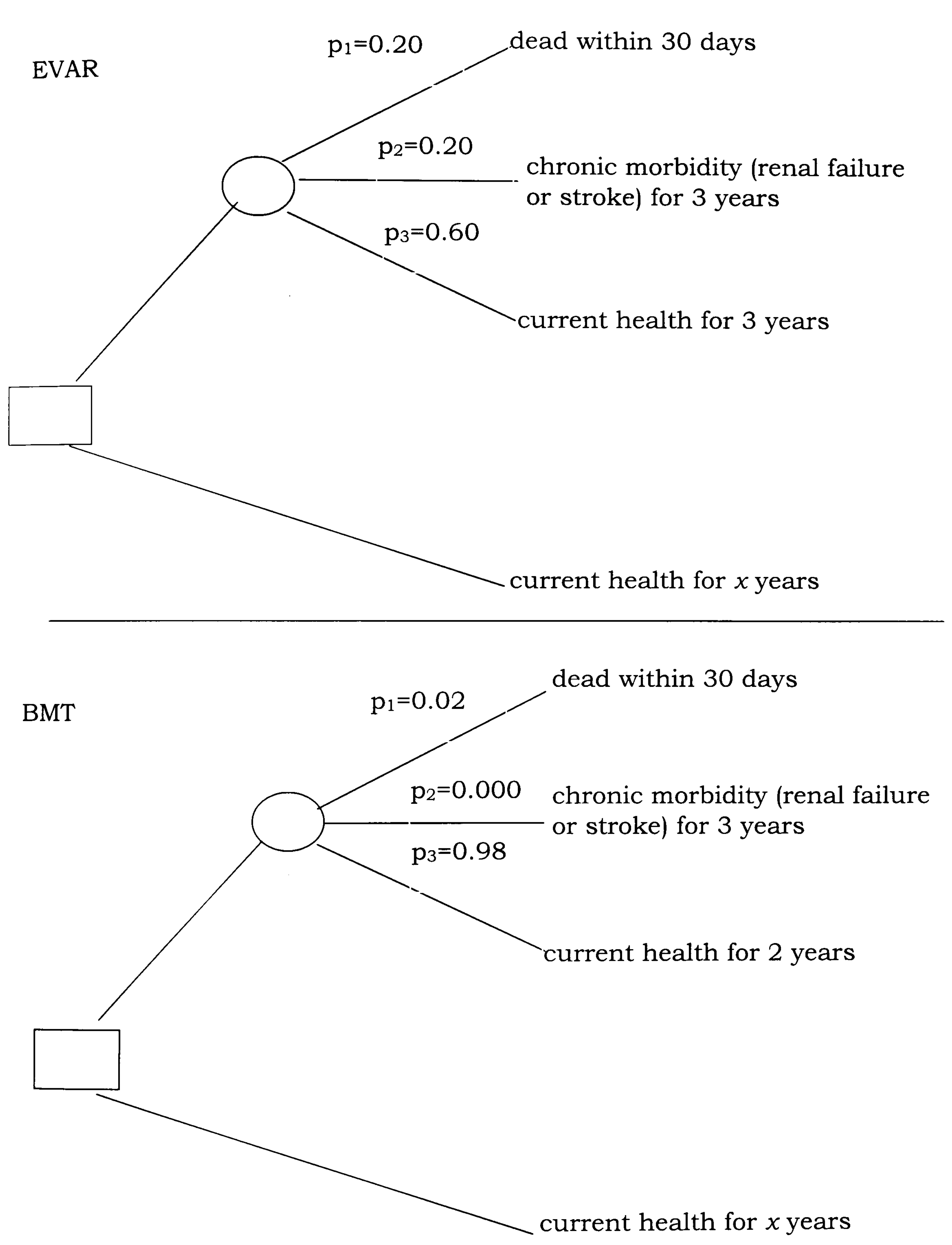

EVAR
p_1=0.20
dead within 30 days
p_2=0.20
chronic morbidity (renal failure or stroke) for 3 years
p_3=0.60
current health for 3 years
current health for *x* years
BMT
p_1=0.02
dead within 30 days
p_2=0.000
chronic morbidity (renal failure or stroke) for 3 years
p_3=0.98
current health for 2 years
current health for *x* years

Figure 8.1 EVAR and BMT.

Chronic renal failure

You have to undergo dialysis, which means that a machine takes over the role of the kidney. This involves spending 3 hours in hospital 3 times a week. Alternatively, you might do dialysis at home, in which case you need a large storage space in which to keep all the necessary materials. This method involves serious restrictions on your lifestyle. For example, you have to interrupt your normal daily activities to go on dialysis. You also face restrictions on taking holidays.

You have restrictions on what you can eat and drink. For example, you are able to drink only a very moderate amount of alcohol. You are instructed to moderate your intake of certain foods, such as bananas, cheese, milk, and meat.

You feel tired and depressed for much of the time.

Stroke

You have sensory loss, so that you no longer have a sense of touch. You are also unaware of the positions of your affected limbs when you are not looking at them.

You have significant loss of the ability to speak.

Your sight is affected, so that the affected eye is no longer able to recognise familiar objects.

You have lost some control over your movements. This means that the limbs on the affected side of the body seem clumsy, and no longer do exactly what you want them to do.

You are subject to mood changes.

You are considerably more dependent on the help of others than previously.

Figure 8.2 Health state descriptions used in the questionnaire for chronic renal failure and stroke.

Imagine that you will only live another 36 months (3 years). You are about to be asked to consider these 36 months.

Below is a scale of value rated from 0 to 100, representing the overall value of the remainder of your life. The first section of the scale considers a value ranging from 0 to 25. representing 25% of the value of your life. The next section considers the value ranging from 25 to 50, the third section considers the range of values from 50 to 75, and the last section considers the range of values from 75 to 100.

We all value time in different ways. We would like you to consider the next 36 months. Try to imagine how many months out of 36 would be worth 25% (one quarter) of the whole period. This might not be 25% of 36 (i.e. 9 months). For example, you might decide that the first 5 months hold 25% of the entire value of those 36 months. Please place against the first section (0 to 25) the number of months that you feel would have the first 25% of the value out of the 36 months.

Now please consider the next section of the scale (25 to 50). How many months would you feel have the next 25% of the value? Please place your answer beside this section of the scale.

Please do similarly for the remaining two sections of the value scale.

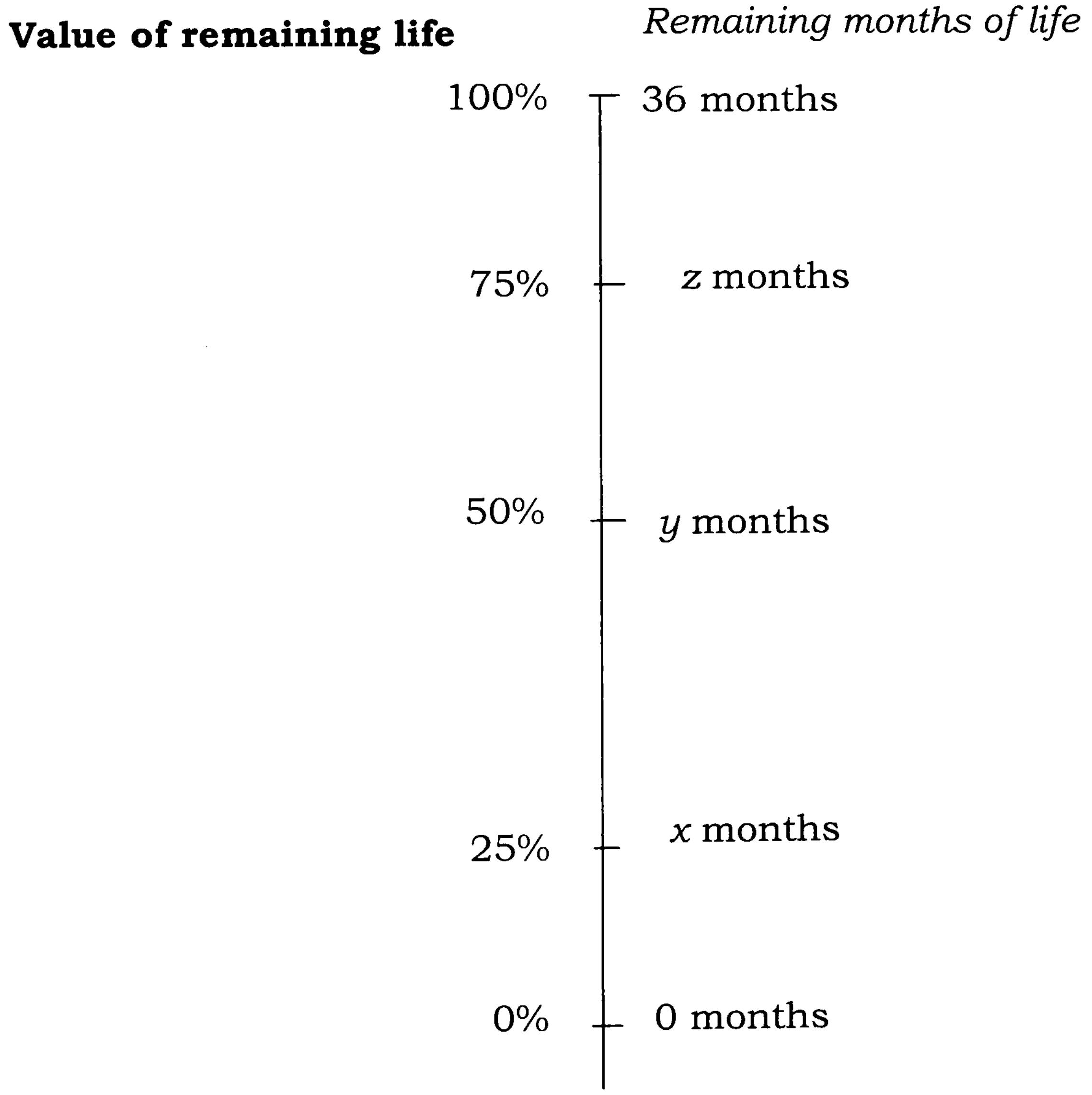
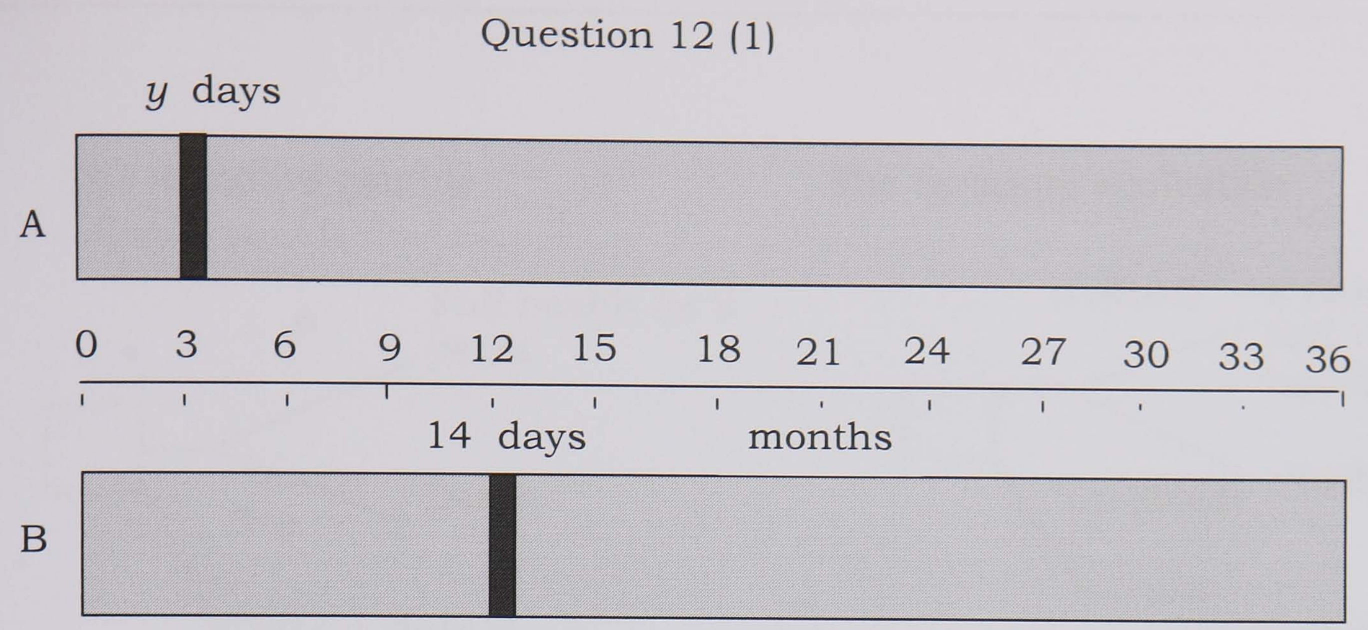


Figure 8.3 The quantity effect question.



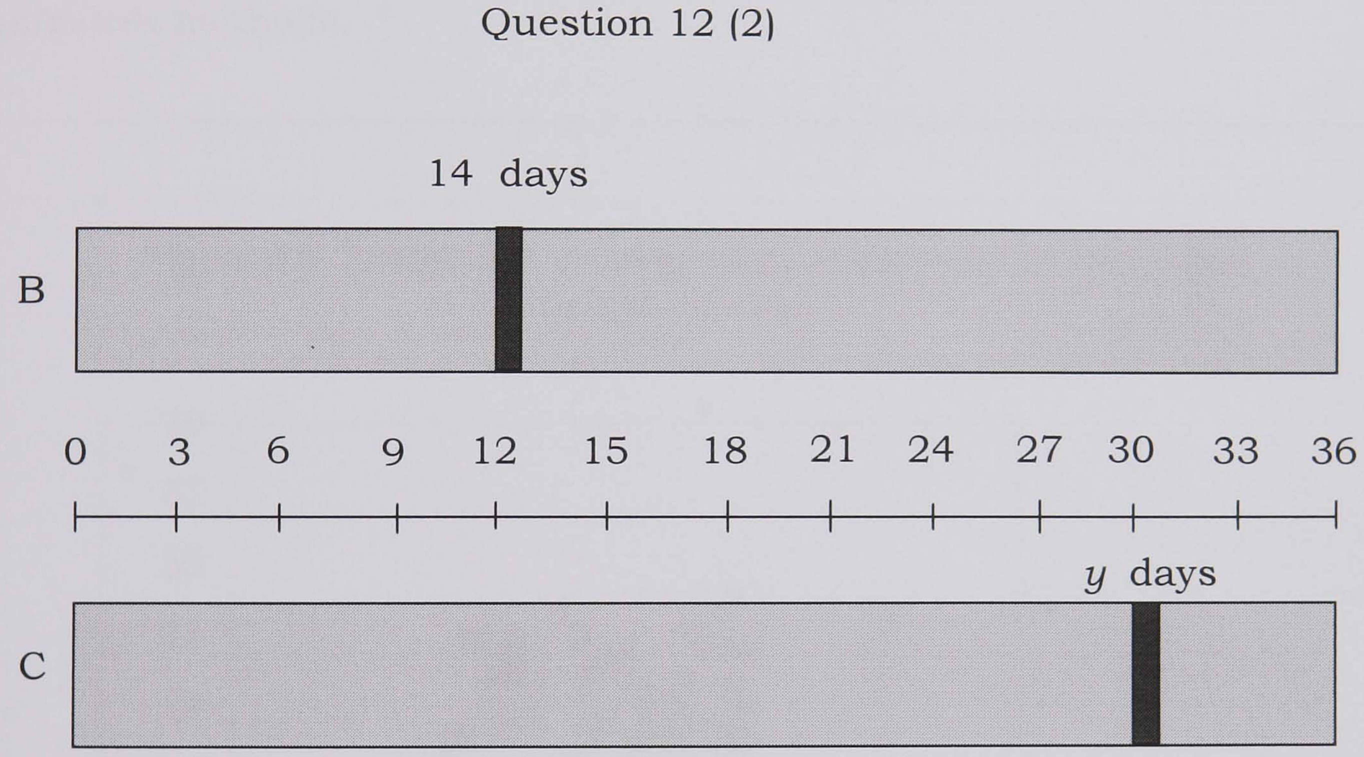


Figure 8.4 The two time preference diagrams used in the questionnaire. The red bar represents the EQ-5D ill health state 22222. The yellow bars represent excellent health.

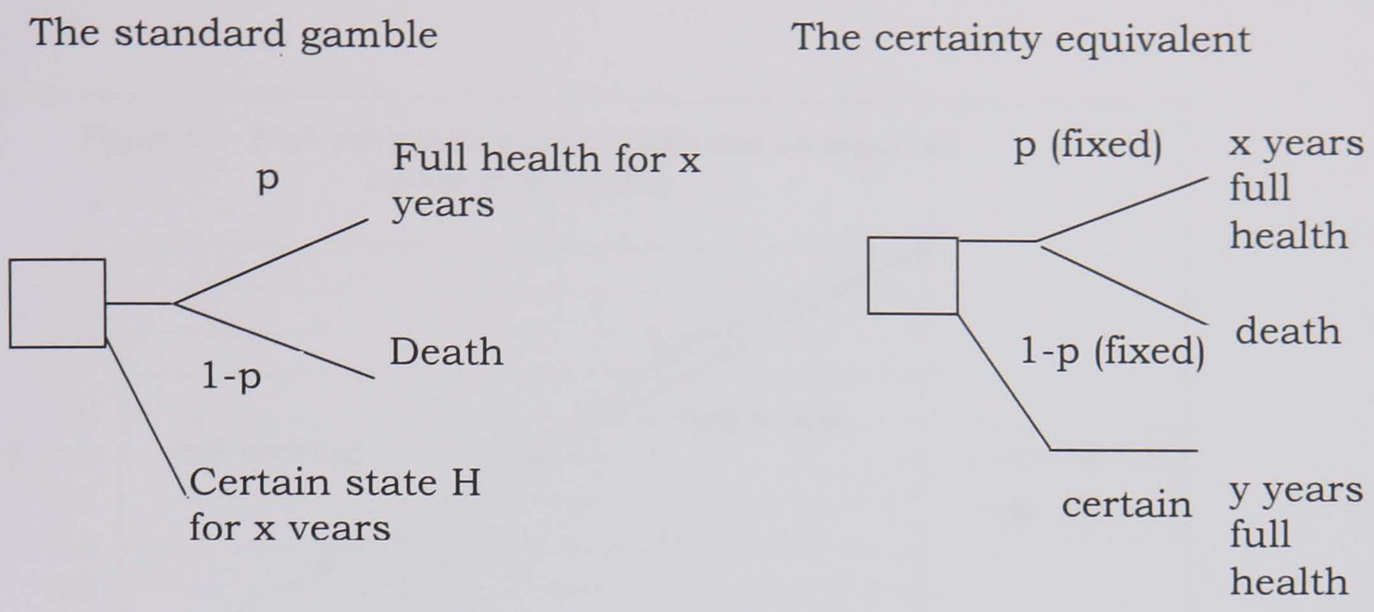


Figure 8.5 Comparisons between the standard gamble and certainty equivalent methods.

Figure 8.6 Scatter plot showing relationship between QALY and holistic valuations.

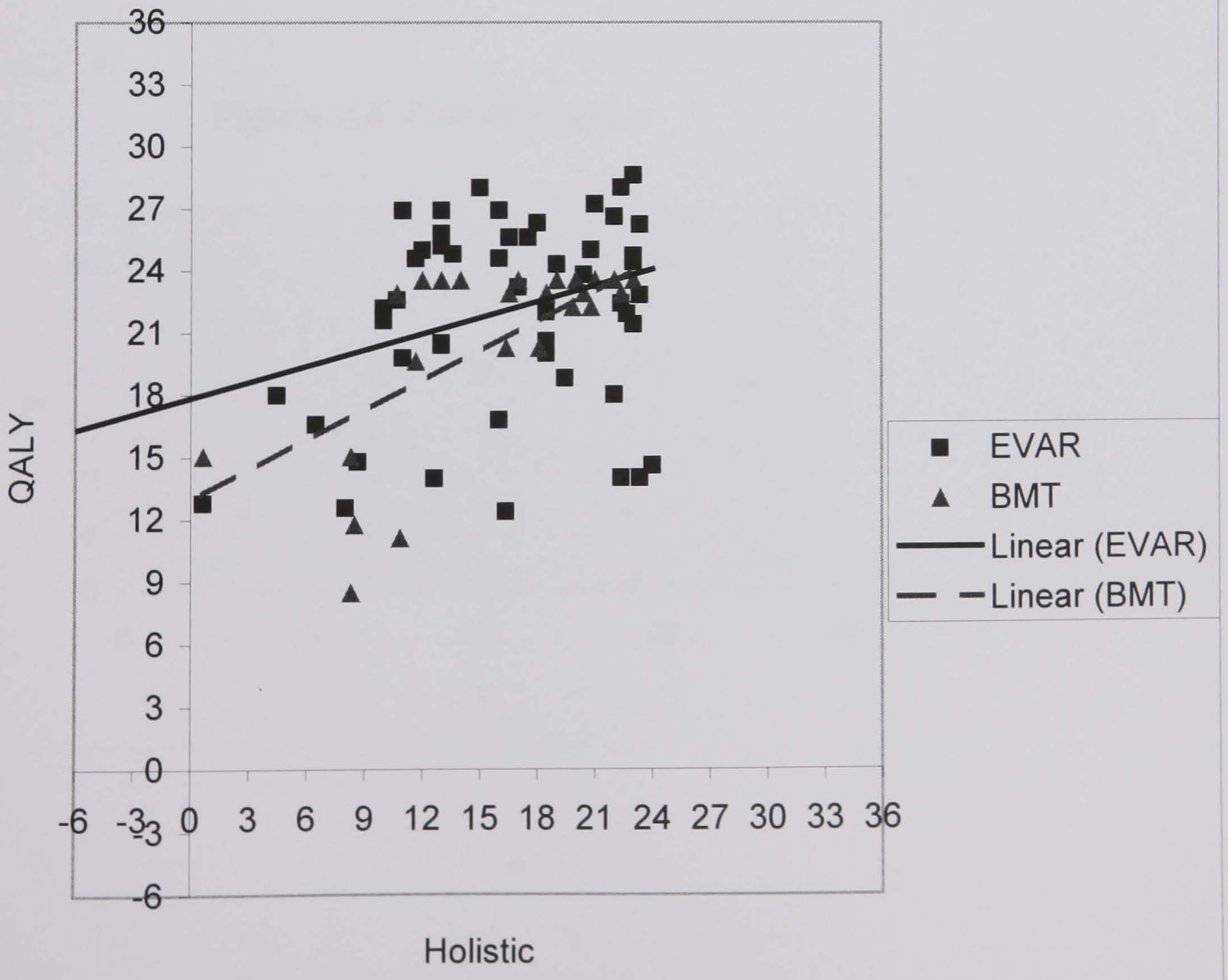


Figure 8.7 Risk attitude for current health over an expected survival of 36 months.

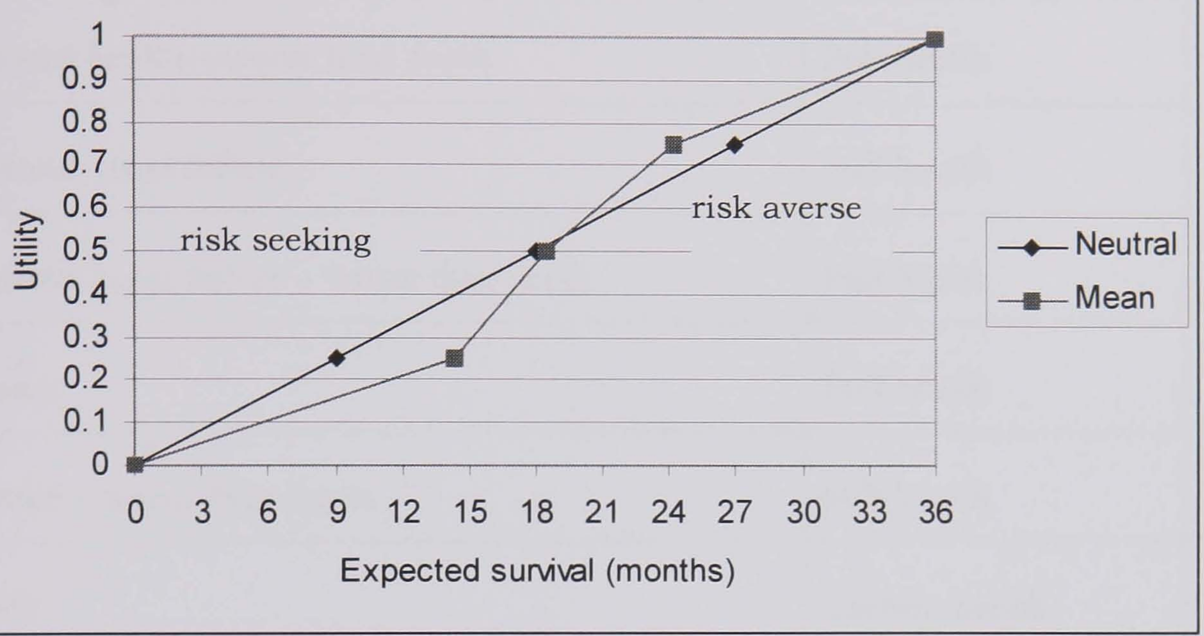


Figure 8.8 Plot of x and y

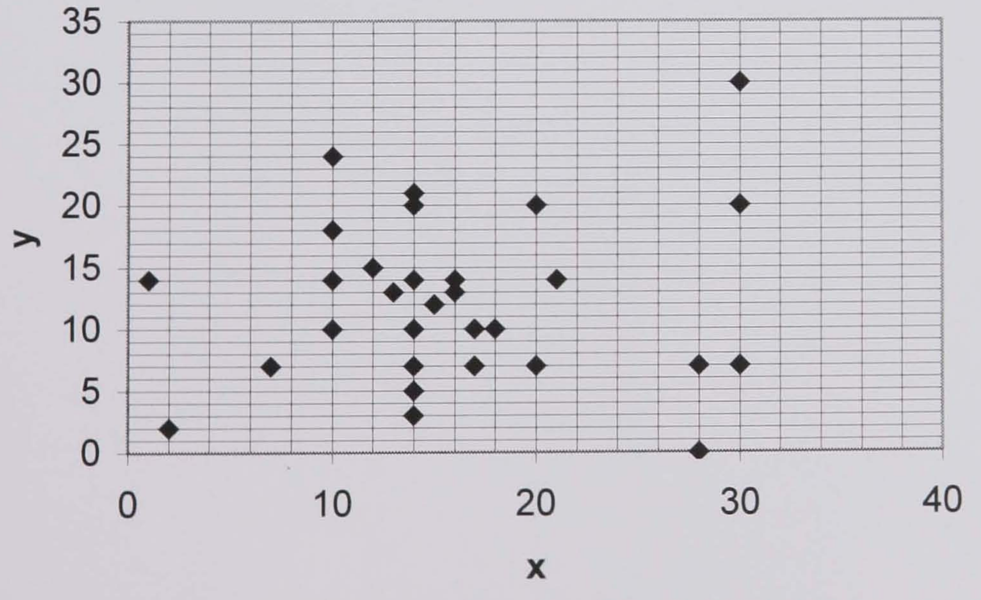


Table 8.1 Each valuation with its reference state.

Valuation	**Reference state**
Current health	Full health
Current health – worse than death	Full health
Chronic renal failure	Full health
Chronic renal failure – worse than death	Full health
Stroke	Full health
Stroke – worse than death	Full health
BMT	Current health
EVAR	Current health
EVAR – worse than death	Full health
EVAR	BMT
Three risk questions	Current health

Table 8.2 EQ-5D responses for the sample of 61.

	EQ-5D responses – n (%)		
	1	*2*	*3*
Mobility	56 (91.8)	5 (8.2)	0 (0.0)
Self-care	59 (96.7)	2 (3.3)	0 (0.0)
Usual activities	55 (90.2)	4 (6.6)	2 (3.3)
Pain/discomfort	52 (85.2)	8 (13.1)	1 (1.6)
Anxiety/depression	49 (80.3)	12 (19.7)	0 (0.0)

Table 8.3 Number of respondents showing wide indifference ranges for each question.

	Number of intervals covered by indifference range (each interval is two months)													
	2	3	4	5	6	7	8	9	11	12	17	18	19	Total for each question
Current health		1										1	1	3
Chronic renal failure		3					1	1	1		2			8
Stroke	1		1	1				1			1			5
BMT	1		1	1						1				4
EVAR		3			1	1								5
EVAR against BMT	2	1	2	1										6
Risk attitude 25:75 survival:death	2	1		1	1	1			1					7
Risk attitude 50:50 survival:death	2	1		1					2					6
Risk attitude 75:25 survival:death	2	1				1			1					5
Total with each indifference range	10	11	4	5	2	3	1	2	5	1	3	1	1	49

Table 8.4 Mean and quartile months for highest stated willingness to trade and lowest stated non-willingness to trade values.

		Mean	25th percentile	50th percentile	75th percentile
CH	Lowest willing to trade value	32.96	32	36	36
	Highest non-willing to trade value	32.44	34	34	34
CR	Lowest willing to trade value	18.72	12	24	35
	Highest non-willing to trade value	12.78	10	20	28
ST	Lowest willing to trade value	3.81	-26	8	24
	Highest non-willing to trade value	0.04	-30	6	18
BMT	Lowest willing to trade value	21.44	20	24	24
	Highest non-willing to trade value	18.59	18	20	22
EVAR	Lowest willing to trade value	17.91	14	20	24
	Highest non-willing to trade value	15.57	12	16	22

Table 8.5 Rank order of health states (FH = full health, CH = current health, CR = chronic renal failure, ST = stroke, and DD = death).

Ranking	n (%)
FH = CH > CR > ST > DD	11 (20.4)
FH > CH > CR > ST > DD	14 (25.9)
CH > FH > ST > CR > DD	1 (1.9)
FH > CH > DD > ST > CR	3 (5.6)
FH > CH > CR > DD > ST	2 (3.7)
FH = CH > CR > DD > ST	3 (5.6)
CH > FH > DD > CR > ST	1 (1.9)
FH = CH > ST > CR > DD	3 (5.6)
CH > FH > CR > ST > DD	4 (7.4)
FH > CH > CR = ST > DD	2 (3.7)
FH > CH > ST > CR > DD	2 (3.7)
FH > CH > DD > CR > ST	1 (1.9)
FH > CH > CR > ST = DD	3 (5.6)
CH > FH > CR > DD > ST	1 (1.9)
FH > CR > ST > CH > DD	1 (1.9)
FH = CH > CR > ST = DD	1 (1.9)
FH > CR = CH > DD > ST	1 (1.9)
Total	54 (100)

Table 8.6 Rank order of health scenarios (FH = full health, BMT = best medical treatment, EVAR = endovascular repair, and DD = death).

Ranking	*n (%)*
FH > BMT > EVAR > DD	45 (83.3)
FH > EVAR > BMT > DD	5 (9.3)
FH > BMT > DD > EVAR	3 (5.6)
FH > BMT = EVAR > DD	1 (1.9)
Total	54 (100)

Table 8.7 Health state valuations.

	Current health	Chronic renal failure	Stroke
Mean	0.93621	0.45936	0.053498
SD	0.14170	0.57411	0.68229
25th percentile	0.97222	0.30556	-0.76389
Median	0.97222	0.63889	0.19444
75th percentile	1.00000	0.84722	0.59722
Minimum	0.361	-0.972	-0.972
Maximum	1.000	0.972	0.972

Table 8.8 Scenario valuations (BMT 0 to 24 months, EVAR 0 to 36 months). Statistical tests compare holistic and QALY values.

	Scenarios			
	BMT		EVAR	
	Holistic	QALY	Holistic	QALY
Mean	19.944	22.020	17.385	22.069
SD	4.236	3.333	5.821	4.716
25th percentile	18.500	22.867	13.000	18.600
Median	21.500	22.867	18.500	23.000
75th percentile	23.000	23.520	23.000	25.650
Minimum	1.000	8.490	-0.194	12.400
Maximum	23.000	23.520	24.000	28.600
Mean difference (95% CIs)	2.075 +/- 1.285 (0.790 to 3.360)		4.683 +/- 1.848 (2.835 to 6.531)	
t-test *p*	0.002		0.000	
Wilcoxon-sign *p*	0.000		0.000	
N	54	54	54	54

Table 8.9 Scenario values with holistic values chained to full health. Statistical tests compare holistic and QALY values.

	Scenarios			
	BMT		EVAR	
	Holistic	QALY	Holistic	QALY
Mean	18.816	22.020	16.450	22.069
SD	5.010	3.333	6.125	4.716
25th percentile	16.528	22.867	12.479	18.600
Median	20.889	22.867	17.250	23.000
75th percentile	22.361	23.520	22.361	25.650
Minimum	0.640	8.490	-0.21	12.400
Maximum	23.000	23.520	24.000	28.600
Mean difference (95% CIs)	3.204 +/- 0.944 (2.260 to 4.148)		5.618 +/- 1.712 (3.907 to 7.330)	
t-test *p*	0.000		0.000	
Wilcoxon-sign *p*	0.000		0.000	
N	54	54	54	54

Table 8.10 Scenario values with holistic values chained to full health using the non-constrained worse than death equation.

	Scenarios			
	BMT		EVAR	
	Holistic	QALY	Holistic	QALY
Mean	18.816	22.020	16.449	-6.478
SD	5.010	3.333	6.125	71.095
25th percentile	16.528	22.867	12.479	10.124
Median	20.889	22.867	17.250	23.000
75th percentile	22.361	23.520	22.361	25.650
Minimum	0.640	8.490	-0.21	-231.00
Maximum	23.000	23.520	24.00	28.60
N	54	54	54	54

Table 8.11 Internal consistencies.

Condition	N (%)
Logical	
CE25 ≤ CE50 & CE50 ≤ CE75	43 (81.1)
Illogical	
CE25 > CE50 & CE50 ≤ CE75	7 (13.2)
CE25 ≤ CE50 & CE50 > CE75	2 (3.8)
CE25 > CE50 & CE50 > CE75	1 (1.9)
Total	53 (100)

Table 8.12 Statistics for risk attitude questions (number of months in current health for certain equivalent to the gamble).

Question	N	Mean (SD)	Median (IQR)	Comparison	Mean difference (95% CIs)	t-test	Wilcoxon- sign
						p	*p*
25% survival (CE25)	54	14.33 (8.15)	12.50 (7.75-19.00)	CE25 v. CE50	4.15 +/- 1.78 (2.37 to 5.93)	0.000	0.000
50% survival (CE50)	54	18.'48 (7.95)	17.00 (13.00-23.25)	CE25 v. CE75	9.79 +/- 2.3 (7.49 to 12.10)	0.000	0.000
75% survival (CE75)	53	24.23 (7.78)	25.00 (20.00-30.50)	CE50 v. CE75	5.72 +/- 1.42 (4.30 to 7.14)	0.000	0.000

Table 8.13 Health scenarios valuations with their risk-adjusted values.

		BMT	EVAR
	$(x/t)^r$		
Mean (SD)	Holistic (chained to full health)	18.816 (5.010)	16.342 (6.465)
Median (IQR)		20.889 (16.528-22.361)	17.250 (12.479-22.361)
Mean (SD)	QALY (risk-adjusted)	2576.659 (10442.230)	3206.510 (17864.896)
Median (IQR)		31.007 (12.534-108.880)	34.075 (15.239-132.286)
Wilcoxon-sign test		0.000	0.000

Table 8.14 Responses to time preference questions.

N = 53		Time preference	
x	*y*	*x*	*y*
1	14	Positive	Neutral
2	2	Positive	Negative
7	7	Positive	Negative
10	10	Positive	Negative
10	14	Positive	Neutral
10	18, 24	Positive	Positive
12	15	Positive	Positive
13	13	Positive	Negative
14	3, 5, 7 (N = 2), 10	Neutral	Negative
14	14 (N = 22)	Neutral	Neutral
14	20, 21	Neutral	Positive
15	12	Negative	Negative
16	13	Negative	Negative
16	14	Negative	Neutral
17	7,10	Negative	Negative
18	10	Negative	Negative
20	7	Negative	Negative
20	20	Negative	Positive
21	14	Negative	Neutral
28	0, 7	Negative	Negative
30	7	Negative	Negative
30	20, 30	Negative	Positive
365	28	Negative	Positive

Table 8.15 Statistics for time preferences.

	Including outlier ($x = 365$)					Excluding outlier ($x = 365$)				
			Percentiles					Percentiles		
	N	Mean (SD)	25	50	75	N	Mean (SD)	25	50	75
x	53	21.7358 (48.3943)	14	14	16	52	15.1346 (5.7532)	14	14	15.75
y	53	13.1509 (5.6207)	10	14	14	52	12.8654 (5.2730)	10	14	14

Table 8.16 Time preferences (N = 53).

	x	y
Mean	-0.0016	-0.0261
SD	0.0781	0.1388
25th percentile	-0.0147	-0.0185
50th percentile	0.0000	0.0000
75th percentile	0.0000	0.0000

Table 8.17 Scenario values with holistic values chained to full health with QALY values adjusted for time preferences.

Scenarios	BMT			EVAR		
	Holistic	QALY		Holistic	QALY	
		x	y		x	y
Mean	18.816	22.330	22.299	16.342	22.409	22.276
SD	5.010	3.261	3.142	6.465	5.074	4.721
25th percentile	16.528	22.867	22.867	12.479	19.219	19.010
Median	20.889	23.329	23.316	17.250	24.001	23.674
75th percentile	22.361	23.520	23.520	22.361	26.300	25.937
Minimum	0.640	7.414	8.493	-6.030	11.357	12.400
Maximum	23.000	28.655	24.863	24.000	30.604	28.718
t-test *p*	0.001			0.000		
Wilcoxon-sign *p*	0.000			0.000		
N	54	53	52	54	53	52

Table 8.18 Holistic values for BMT and EVAR compared with QALY values adjusted for time preference rates and risk attitude.

		BMT	EVAR
Mean (SD)	Holistic (chained to full health)	18.816 (5.010)	16.342 (6.465)
Median (IQR)		20.889 (16.528-22.361)	17.250 (12.479-22.361)
Mean (SD)	QALY (risk-adjusted with time x)	2609.496 (10106.661)	3044.491 (15915.934)
Median (IQR)		30.700 (12.265-113.717)	35.256 (14.030-137.811)
Mean (SD)	QALY (risk-adjusted with time y)	2746.929 (10987.919)	3543.749 (19488.047)
Median (IQR)		29.791 (12.088-106.307)	31.732 (14.724-126.488)

Table 8.19 Degree of non-convergency for health states.

	Strong	Weak
0/4	0	0
1/4	0	0
2/4	7 (13.0%)	4 (7.4%)
3/4	24 (44.4%)	12 (22.2%)
4/4	23 (42.6%)	38 (70.4%)

Table 8.20 Comparison of convergence between each valuation method and the original ranking of scenarios.

	Strong		Weak	
	Holistic	QALY	Holistic	QALY
0/3	0	0	0	0
1/3	0	0	0	0
2/3	27 (50.0%)	32 (59.3%)	13 (24.1%)	32 (59.3%)
3/3	27 (50.0%)	22 (40.7%)	41 (75.9%)	22 (40.7%)
Total	54	54	54	54

Table 8.21 The consistent and inconsistent results of state and scenario rankings compared to TTO valuations. E is the expected number if the effect is due to chance.

		States		
		Inconsistent	Consistent	Total
Scenarios QALYs	Inconsistent	19 (35.19%) E = 21.33	13 (24.07%) E = 10.67	32 (59.26%)
	Consistent	17 (31.48%) E = 14.67	5 (9.26%) E = 7.33	22 (40.74%)
	Total	36 (66.67%)	18 (33.33%)	54 (100%)
Scenarios Holistic	Inconsistent	15 (27.78%) E = 18.00	12 (22.22%) E = 9.00	27 (50%)
	Consistent	21 (38.89%) E = 18.00	6 (11.11%) E = 9.00	27 (50%)
	Total	36 (66.67%)	18 (33.33%)	54 (100%)

Table 8.22 The results of the sensitivity analysis, in which the lowest, midpoint, and highest possible values for ambiguous responses are compared.

		N	Lowest	Midpoint	Highest
Mean (SD)	Holistic BMT (chained to full health)	54	18.6 (5.8)	18.8 (5.0)	19.5 (4.6)
	QALY BMT	54	21.9 (4.1)	22.0 (3.3)	22.4 (2.9)
t-test	*p*		0.00	0.00	0.00
Wilcoxon	*p*		0.00	0.00	0.00
Mean (SD)	Holistic EVAR	54	15.9 (6.9)	16.3 (6.5)	16.7 (6.7)
	QALY EVAR	54	21.9 (5.1)	22.1 (4.7)	22.3 (4.8)
t-test	*p*		0.00	0.00	0.00
Wilcoxon	*p*		0.00	0.00	0.00

Table 8.23 Scenario values with holistic values for EVAR adjusted for truncation at 24 months, assuming values of EVAR = 33 to 36. Statistical tests compared holistic values with the equivalent QALY value.

	Scenarios				
	EVAR				
	Holistic				QALY
	33	34	35	36	
	EVAR=24	EVAR=24	EVAR=24	EVAR=24	
Mean	17.212	17.309	17.405	17.502	22.069
SD	7.685	7.861	8.044	8.233	4.716
25th percentile	12.479	12.549	12.639	12.910	18.600
Median	17.250	17.250	17.250	17.250	23.000
75th percentile	22.361	22.361	22.361	22.361	25.650
Minimum	-6.030	-6.030	-6.030	-6.030	12.400
Maximum	33.000	34.000	35.000	36.00	28.600
t-test *p*	0.000	0.000	0.000	0.000	
Wilcoxon-sign *p*	0.000	0.000	0.000	0.000	
N	54	54	54	54	54

Table 8.24 Holistic scenario values with EVAR adjusted for truncation at 24 months, assuming values of EVAR = 33 to 36. Statistical tests compared EVAR and BMT.

	Scenarios				
	BMT	EVAR			
		33	34	35	36
		EVAR=24	EVAR=24	EVAR=24	EVAR=24
Mean	18.816	17.212	17.309	17.405	17.502
SD	5.010	7.685	7.861	8.044	8.233
25th percentile	16.528	12.479	12.549	12.639	12.910
Median	20.889	17.250	17.250	17.250	17.250
75th percentile	22.361	22.361	22.361	22.361	22.361
Minimum	0.640	-6.030	-6.030	-6.030	-6.030
Maximum	23.000	33.000	34.000	35.000	36.00
t-test *p*		0.009	0.013	0.019	0.026
Wilcoxon-sign *p*		0.007	0.009	0.011	0.013
N	54	54	54	54	54

Table 8.25 Categories of respondents' comments.

	Number
Cognitive difficulties	11
General praise of the questionnaire	3
Suggestions of improvements	4
Difficulty with hypothetical states/scenarios	4
Personal preferences might change	1

Chapter 9

Discussion

One of the aims of this thesis was to explore the extent of violations of the QALY axioms. Specifically, the axioms of zero time preference, neutral risk attitude and constancy of risk attitude to survival duration, and aspects of the additive utility axiom were examined. The issue of time preferences was explored in Chapter 8, in which time preferences were measured over a short life expectancy. The study described in Chapter 8 also measured individual risk attitudes, and tested the axioms of risk neutrality and constancy of risk attitude to survival duration. Two aspects of the additive utility function were explored. Firstly, Chapters 5 and 6 described two studies looking at whether health profiles were valued proportionately according to the proportion of time in each health state. Chapter 7 explored the issue of whether small duration effects due to treatment are valued simply by multiplying duration by quality weight.

Another aim of this thesis was to explore issues relating to the construction of health profiles that could be valued by a holistic valuation approach. Three different types of health profile or scenario were constructed in the four empirical studies described in Chapters 5 to 8. As mentioned above, Chapters 5 and 6 explored the construction of health profiles which differed in the proportion of time in each health state, which occurred at random over the profile. In Chapter 7 health profiles consisted of a pre-treatment state, a treatment state, and a post-treatment rest-of-life period. Two of these profiles contained *ex ante* risks of recurrence and mortality. In Chapter 8 scenarios were constructed to describe different treatment outcomes over a very short life expectancy, and significant levels of *ex ante* risks were incorporated into these scenarios.

A third aim of this thesis was to explore different approaches to using the QALY. Methods of valuing short-term temporary treatment states were explored in Chapter 7. Chapter 8 explored ways of adjusting QALY values for individual risk attitude and time preferences over a short life expectancy.

Finally, this thesis aimed to compare QALY and holistic values for the same profiles against a set of pre-defined criteria. The results from the two methods were also compared more generally in terms of differences in results.

This chapter discusses how each of these aims were met. The chapter begins with a description of the limitations of the research. This is followed by the key findings of the research. The contribution to knowledge is then discussed, followed by a summary of the further research required in this area. The final section of this chapter outlines the conclusions of the discussion.

9.1 Limitations of the research

This section discusses various practical issues and limitations of the research contained in Chapters 5 to 8.

9.1.1 Difficulties with recruiting to the studies and small samples

For Study 1 (Chapter 5) 49 women who had taken part in the clinical trial of a GW drug were recruited from GW clinical trial centres. The completion rate was very high, with just one missing value datum. For Study 2 the sample consisted of male and female sufferers recruited from general practices. A total of 183 patients were invited to enter the study, and 56 (30.6%) chose to participate.

A total of 200 varicose veins patients were invited to take part in this study (Chapter 7). There was a response rate of 41 (20.5%). A further 26 patients were recruited from a weekend clinic, giving a final total of 67 participants.

The problems encountered in attempting to recruit a sample of AAA patients for the study in Chapter 8 proved insurmountable, to the extent that it was necessary to abandon the attempt. Instead, a convenience sample of 61 respondents were recruited, consisting of staff from the Vascular Institute of the Northern General Hospital, staff from the School of Health and Related Research, and a small number of people with peripheral vascular disease and friends/relatives of the author.

In the case of both the IBS study and the varicose veins study, it was suspected that some of the non-responses might be due to potential participants being too busy (as many people with these conditions work) and the appointments and venues not being convenient. If the funding and amount of time available for these studies had been greater, the author would have offered to interview people in their own homes or some other venue of convenience to them as standard. The times could then have been fitted more easily around the convenience of the potential participants.

The study that suffered most severely from recruitment difficulties was the AAA study described in Chapter 8. In order to obtain a sample of AAA patients the study may have benefited from a higher level of funding in order to offer some form of compensation for attending the interview, such as an interim ultrasound scan, which may have encouraged attendance. On the other hand, again, the author could have offered interviews in respondents' homes even though this would have involved travelling in a wider area surrounding Sheffield in order to visit participants.

The difficulties in recruiting patients to the AAA study were due to factors such as logistics at the hospital (AAA patients), and inability to complete the questionnaire by patients due to illness or impaired sight or hearing (this applied to the second sample of choice, which were peripheral vascular disease patients). This experience served to highlight the possible problems that might arise in attempting to elicit patient preferences. These problems may not apply to such an extent if values are sought from the general public, because there is a larger population from which to choose. As already explained, it was necessary to use a convenience sample for the AAA study. It was acceptable to use a potentially non-representative sample, because this was a methodological study. However, if a different sample had been selected, it is possible that the results of the research would have been different.

Due to these recruitment difficulties, the sample sizes were relatively small. By way of comparison, other studies of holistic methods had sample sizes ranging from 60 (Sculpher, 1996) to 194 (Llewellyn-Thomas *et al*, 2002). However, much of the previous research investigating the QALY axioms involved sample sizes smaller than those in the studies reported in this thesis, with Pliskin *et al* (1980) reporting results from a sample of just 10 (see Chapter 3).

A strength of the research reported in this thesis is the use of a sample size calculation (see equation 4.3, Chapter 4). This was based on an MEID of 0.05 and an SD of the difference of 0.13, which was obtained in the first IBS study. If it had been possible it would have been preferable to use an SD of the difference obtained from a TTO study to apply to sample size calculations for TTO studies, rather than using an SD of the difference from an SG study. However, the use of a formal sample size calculation appears to be very rare in this field.

A possible limitation of the research may have been that the studies were underpowered to detect the sizes of differences found between QALY and holistic methods in the case of the varicose veins study, in which the differences between the two methods of valuation were smaller than the MEID of 0.05.

9.1.2 Lack of generalisability of the samples

There was no random selection for the samples in any of the four studies reported in the thesis. The work was methodological in nature, and so a non-random sample was sufficient to test different methods and approaches to valuing health profiles holistically. However, it is clear that there is no gurarantee on the degree of representativeness of the samples in each study.

In the case of the second IBS study described in Chapter 6, one woman contacted the author to say that she suffered from IBS too badly to be able to attend the interview appointments set up at the suggested location. However, she was willing to be interviewed on a one-to-one basis in her home. The author did so. This responder explained that she was so afraid of suffering from urgency when outside the home that she found it difficult to go out. Due to limitations on the funding and time for this study, the author attempted to recruit IBS patients to attend group sessions either in their neighbourhood (*e.g.* at their GP surgery or a local church hall) or in a room at the hospital. However, after talking to this respondent, there arose a clear possibility that there may have been more potential responders who were in a similar position but would have been willing to be interviewed in their own home. If this was the case, this study may have suffered from wellness bias. In other words, patients may have only responded if they were well enough to attend the group sessions outside their home, and the study may have lost the valuations of those patients who suffered more severely from the condition.

The question arises as to how this "wellness" bias might have affected the results of the valuation study. All respondents valued the same health states and profiles, and each respondent valued hypothetical IBS states and profiles. Thus, in theory, it might be suggested that it should not make a difference to the valuations if they came mainly from a less ill sample of IBS sufferers. However, it is well-established that people who actually suffer from a condition are liable to give different values than non-sufferers (Salomon and Murray, 2002). Indeed this was one of the reasons for choosing to use patient samples where possible. It was thought that patients may have a deeper understanding of the impact of the condition upon quality of life. It may therefore also be the case that IBS patients with a greater degree of illness may have given different values to the hypothetical health states and profiles than the less ill IBS sufferers. For example, a more severely ill person may give lower values to the profiles containing greater frequencies of symptoms because they are more aware of the significance of this. However, one can only speculate on what differences might have existed between the less ill and more severely ill IBS sufferers. Suffice it to acknowledge there could have been significant differences in valuations between these two groups.

One thing that should be made clear is that the studies reported in this thesis are methodological in nature. They aimed to explore the methodology behind the construction of holistic health profiles, methods in which these profiles may be valued holistically, and compare holistic valuations with QALY valuations. The sample sizes were relatively small, and where patient groups were used these patients were not randomly selected. It is not assumed that members of the samples were representative of the larger population the sample comes from. For example, there is no evidence that the IBS sample in Chapter 5 is representative of IBS patients in general. This is not a problem as such for the purposes of this thesis. However, it should be recognised that a representative sample may have given different responses, and the overall results may have differed. Because of this limitation to the studies in this thesis, they cannot be taken and used more generally in the field of health care research to describe findings relating to each patient group. For example, the results reported in Chapters 5 and 6 cannot be taken to say that IBS patients behave in this manner generally, because this was not necessarily a representative sample of IBS patients. This is an important point, and the implication of this is that the results obtained in these four studies cannot be applied in economic evaluations.

9.1.3 Information about responders and non-responders

It would have been very useful to have more information than was available about non-responders. As it stands there is no explanation for why non-responders chose not to respond. It would have been helpful to have access to such information as general demographic data in order to determine whether there were any significant differences between responders and non-responders. Due to ethical concerns regarding confidentiality, the information about patients being invited to take part in each study was limited. The first point of contact with the author was the return of the consent forms, basically showing an interest in participating in the study and agreeing to be contacted by the author. Information on age, sex, occupation, level of educational attainment from non-respondents would have been useful, and these data could have been compared with that of responders. However, none of this information about non-respondents is available.

It would be very useful to know whether certain sections of the population were not taking part in health care decision-making. Since health economics issues such as how quality of life is measured and included in resource allocation decisions affect the entire population, it is important that all sections of society have the opportunity to take part in consultations and be aware of how such decisions are made.

Since this research was methodological in nature, it was no less valid for the lack of information about non-responders. However, it is worth noting that the results may have been different if different people had taken part.

9.1.4 Unwillingness to trade or gamble

There is a problem with valuation studies using the SG and TTO if a significant proportion of respondents are unwilling to participate in the gambling or trading process. This problem has been found in previous research. For example, Llewellyn-Thomas *et al* (2002) found that 57% of their sample were unwilling to gamble in the first stage of the two-stage HYE procedure. The authors were unable to state a reason for this unwillingness to gamble. In Chapter 7 of this thesis, a large proportion (35.6%) of varicose veins respondents were unwilling to trade. Their reasons are unknown, and it would have been useful to know their reasons. It is possible that they believed the health states and profiles were too mild to warrant trading off years of their lives. It is unknown whether they would have been willing to trade smaller amounts of their lives, and this could be the subject of further research. Another possibility is that respondents who are unwilling to trade or gamble may have an aversion to making choices in the context of this type of choice involving their life expectancy and/or health.

The differences between whole sample means and the mean values of the seven health profiles for the sub-sample who were willing to trade in the varicose veins study (Chapter 7) ranged from 0.78 to 1.22. Compared to a MEID of 0.05, these differences are huge, and demonstrate the potential problems in using valuation methods that are not "friendly" to a significant proportion of the population.

The problem of unwillingness to give values below the maximum was most notable in Chapter 7. The AAA study demonstrated an unwillingness to trade on the part of just four (7.4%) respondents, and the second IBS study (Chapter 6) demonstrated an unwillingness to gamble on the part of just two (4.1%) respondents. None of the respondents in the first IBS study (Chapter 5) demonstrated unwillingness to gamble. Thus for three out of the four studies in this thesis, the extent to which respondents were unwilling to trade or gamble was negligible.

9.1.5 Abbreviation of descriptions within the health profile

One of the facts made clear from the research in this thesis is that there are many problems to be encountered in the construction of holistic health profiles. Some of the problems are quite basic. For example, when the profile consists of several health states, it may be necessary to alter the wording of profiles slightly from that of their constituent states in order to fit the profile description into a more compact space for ease of reading and understanding. The descriptions of varicose veins health profiles were abbreviated from their composite parts. The health states were shortened, such that for example the severe state was reduced to the words "severe varicose veins". This was because the pre-treatment and post-treatment state descriptions and the description of the treatment process could not be easily fitted onto one sheet of A4 within the questionnaire booklet. In order to get around any affect this may have had on valuations, respondents were reminded that this was the same state they had valued earlier on its own, and offered the opportunity to remind themselves of this state description. This was done for all the states within the profiles.

The health states were relatively straightforward to value compared to the profiles. For example, whereas the IBS health states on their own were quite straight forward, the respondents had to understand the proportionate nature of the health states within the profiles. Likewise, for the AAA profiles respondents had to understand the cumulative probabilistic nature.

It would be a very lengthy and complex profile description that accurately described all outcomes over time. Outcomes often have to be abbreviated or excluded from the description simply to make the profiles user-friendly and manageable. The risk estimates within the AAA profiles were rounded up so as to decrease the bulk of information being provided and to simplify the task for respondents. As described in Section 9.2.2.1 and Table 9.2, the IBS and varicose veins studies indicated that the holistic methods used were either unreliable or demonstrated high levels of inconsistency. This may have been due to them being overly demanding cognitively. It would therefore seem that keeping health profiles simple is of the essence.

It is difficult to distinguish the limitations of the research conducted in this thesis from the limitations of the holistic approach per se. This research aimed to develop the holistic approach, and investigate the methodological issues relating to the construction of health profiles which can be valued using the holistic approach. It may be that further research could overcome some of the limitations highlighted in this section. For example, perhaps descriptive methods could be developed to overcome the problem of cognitive overload without losing the important aspects that drive the valuation of the health profile.

9.1.6 Choice of health profiles

The process of constructing health profiles for valuation in this thesis is described in Chapters 5 to 8, and varied between studies. In summary, the profiles chosen were selected on the basis of covering the likely events for each condition studied. However, in some cases the choice of profiles to include was to some extent arbitrary. Since this research was methodological in nature, and one of the purposes was to examine issues relating to the construction of holistic health profiles, the choice of health profile descriptions does not detract from the research.

There are practical issues in constructing health profiles and obtaining valuations. One of the advantages of the QALY is that, once a number of states have been valued, it is possible to use these state values to calculate values for a large number of profiles involving these states. However, with holistic valuations it is only feasible to value a small proportion of the possible profiles. A choice therefore has to be made over which of the potentially possible health profiles relating to a particular condition should be included in the valuation process. It is a clear limitation of the holistic valuation approach that the number of health profiles valued must be limited. This must result in the elicitation of values for each different health profile that emerges as important.

9.1.7 Cognitive issues

One of the concerns over the holistic approach to the valuation of health profiles is the possibility that respondents may find aspects of the exercise too demanding cognitively (Buckingham, 1993), and this cognitive overload could lead to responses that do not truly reflect the individual's preferences.

At the end of each questionnaire was a section for respondents to write any comments they wished. The comments from the four studies in Chapters 5 to 8 have been categorized and tabulated in Table 9.1. The largest category was comments expressing difficulty with, or criticism of, the methods used to elicit data. There were 43 comments in this category out of a total of 233 respondents across the four studies (18%). A total of seven comments were made by the IBS patients in Study 1 (Chapter 5), and eight IBS patients in Study 2 (Chapter 6) also made comments along this line. A total of 12 varicose veins patients made similar comments (two of which were verbal). The largest group of people to comment on difficulties with the tasks were from the AAA study (a total of 16). Many of the comments just expressed how difficult respondents found the tasks. However, some comments were more specific. For example, some IBS patients had difficulty in dealing with death as a failure state with the SG procedure. Some respondents found the valuation methods of SG or TTO particularly difficult. Other respondents had difficulty with imagining being in the hypothetical health states and profiles they were asked to value.

The comments made by respondents in the four studies are tabulated verbatim in Appendices 1 to 4. When it comes to comments about the difficulty with the questionnaire, the strength of feeling in these comments seems greatest with the convenience sample used for the AAA study in Chapter 8. Some respondents expressed the belief that their responses to the exercises were meaningless, because they felt that the task was just so difficult that they were unable to give a meaningful answer (Table 8.A.6).

It is obviously a concern that, for policy decisions within the health care sector to be based on the preferences of either patients or the general public, these preferences must be elicited in an accurate and meaningful fashion. If the comments sited above are reflective of the whole sample, this is clearly a concern.

Another question that must be answered is whether the cognitive difficulties reflect a general concern with either the QALY or the holistic method of valuation, or whether it was due to the layout of the questionnaire, or the descriptive system used in the health states or profiles. If difficulties were with the valuation methods, research could be conducted into the use of different valuation methods. If the difficulties were with the descriptions of the health profiles, methods of making descriptions more accessible can be researched (see Section 9.4).

9.1.8 Implications of wide ranges of indifference

There were particularly wide indifference ranges in the AAA study in Chapter 8 for 18 (33.3%) respondents. Wide ranges of indifference for TTO valuations might suggest that the task was too difficult. This finding concurs with the above-reported comments about the difficulties respondents encountered in completing the AAA questionnaire.

9.1.9 The reading age of the questionnaires

One of the characteristics defining text readability is the "reading age" of the text. This is defined by the sentence structure of the text, and can be defined as chronological age of a reader who could just understand the text (Johnson, 2005). There are several methods for calculating the reading age of a piece of text. Some methods are used to assess the readability of material specifically for children, whereas other methods are designed to measure the readability of adult materials. Carr (2002) suggests that material written for the general public should have a maximum reading age of 15 years. This section examines the readability of the surveys used in this thesis in order to determine whether the level at which they were set may have influenced the level of demand upon cognition.

There appear to have been very few studies examining the reading age of literature used in health care. Bradley *et al* (1994) examined the reading age of patient information leaflets for over-the-counter medicines, and discovered that the reading ages for their sample of leaflets ranged from 10 to 20 years with a mean reading age of approximately 15 years. Conroy and Mulcahy (1985) examined the readability of 28 books and leaflets written for cardiac patients in Dublin, and found a mean reading age of 14 years with 21% of the literature demonstrating a reading age of 12 or less. However, a literature search did not reveal any studies relating to readability of valuation equipment used in health economics.

Subsequent to the research described in Chapters 5 to 8, a partial analysis of reading ages was carried out. The Flesch-Kincaid Formula, the standard test used by the United States Government Department of Defence, was used for this analysis (Ressler, 1993). This section describes the results of this partial analysis to test the reading ages of the questionnaires used in Chapters 6 to 8 (the two IBS questionnaires were very similar, and so only the second one was tested). The analysis was partial in that exclusive tests were not carried out on every bit of text, but were carried out on a representative sample of the text, including instructions for valuation exercises and descriptions of health states and profiles.

The reading ages for the three questionnaires ranged from 10.8-13.6 years (the second IBS study), 10.6-12.7 years (varicose veins), and 12.5-13.6 years (AAA). The highest reading age of 13.6 for the AAA study was obtained for the EVAR scenario. The health state descriptions for stroke and chronic renal failure had reading ages of 13.1 and 12.9 years respectively. The health profiles in the second IBS study had reading ages of 13.6 years.

In hindsight, a reading age analysis should have been conducted during the design of the questionnaires. However, the results are encouraging, because they all fall below the maximum reading age of 15 years suggested by Carr (2002).

9.1.10 A summary of the limitations of the research

Although these studies were among the few that use a sample size calculation to determine the appropriate sample size, there were difficulties in recruiting for three out of the four studies. In the AAA study, these problems were so severe that the attempt to recruit a patient sample was abandoned in favour of a convenience sample of colleagues.

The patient samples recruited to participate in the two IBS studies and the varicose veins study may not have been representative of these patient groups. Patients were not selected on a random basis. However, since these studies were methodological, the lack of representativeness was not a problem. Nonetheless, it means that the health state and health profile values cannot be generalised and used in economic evaluations.

There was no way of comparing respondents with non-responders, because of patient confidentiality. The researcher had no access to information from non-responders. Again, because of the methodological nature of this research, this does not invalidate the results. However, it is worth noting that the results may have been different if different people had participated.

In the varicose veins study, a large proportion of respondents (35.6%) showed unwillingness to trade life years. The reasons for this unwillingness to trade are unknown, although it is hypothesised that it may be due in part to an insensitivity of the trading scale, and that these respondents may have been willing to trade very small amounts of life. Encouragingly, most respondents in the other three studies showed willingess to trade or gamble, with the proportion who were unwilling being very low.

It was necessary to abbreviate health profile descriptions to some extent. For example, health state descriptions were shortened and numerical values were rounded up for simplicity. This might be a limitation to the holistic approach, but on the other hand further research may develop ways to overcome this limitation.

The choice of health profiles for these studies was, to some extent, arbitrary. Although this was not a problem as such for the methodological purposes of the research, it is a limitation of the holistic approach that only a selected number of profiles can be valued. This is in contrast to the QALY approach, for which any number of profiles can be valued once the constituent health states have been valued.

Respondents' comments demonstrated cognitive difficulties with the valuation tasks. This was particularly marked in the case of the AAA study. However, it is unclear whether either the holistic or the QALY method was more difficult congnitively than the other method.

The AAA study also demonstrated wide ranges of indifference in 33.3% of respondents. This may have reflected the cognitive difficulties sited by several respondents in their comments.

It would have been useful to test the questionnaires for reading age prior to each study. However, a partial analysis of reading age carried our after the research was completed showed that the reading age of all the questionnaires was below the maximum of 15 years advised by Carr (2002).

9.2 Key findings

This section discusses the key findings of the studies described in Chapters 5 to 8. First the findings are discussed with regard to tests of the QALY axioms. Secondly, comparisons between QALYs and holistic values are discussed with regard to pre-set criteria, and overall comparisons between results from the two health profiles valuation methods.

9.2.1 Tests of the QALY axioms

This thesis set out to test the axioms underlying the QALY algorithm. Specifically, the research in this thesis tested the axioms of:

- zero time preference
- linearity of risk attitude to survival duration
- risk neutrality
- additive separability over time

9.2.1.1 Research into individual time preferences

As the literature review in Chapter 3 has shown, it has already been established that people often demonstrate a non-zero time preference rate (*e.g.* Cairns and van der Pol, 2000). Although it is now readily accepted that people may have a positive time preference (HM Treasury, 2003), a large body of research has demonstrated that time preferences of individuals vary according to several factors. These factors include the context over which time preferences are elicited. For example, the work of Chapman and colleagues has demonstrated that time preferences may differ across the different domains of health and money (Chapman and Elstein, 1995; Chapman, 1996; Chapman *et al*, 1999). The evidence also suggests that individual time preferences are not necessarily stable within individuals, but may be a function of the magnitude of the outcome and the magnitude of the delay (Chapman and Elstein, 1995).

The only one of the four studies in this thesis to attempt to directly measure time preference rates was the AAA study in Chapter 8, using a modified version of the open-ended format introduced by Cairns (1991, 1992) and Cairns and van der Pol (2000). Cairns (1991, 1992) measured time preferences over a time horizon of 42 years. However, since the scenarios described and valued in the AAA study were terminal, and life expectancy was only two to three years, it was appropriate to attempt to adapt the Cairns method for this shorter time horizon. The modified method is shown in Appendix 4 and diagrammatically in Figure 8.4. Two questions were asked, obtaining time preference values x and y.

A varying time preference rate was indicated for many individuals within the sample, as suggested by Chapman and Coups (1999). This was to such an extent that many individuals, for example, indicated a positive time preference for x, but a negative time preference value for *y* or *vice versa*. Thus their time preference often varied according to whether the state of ill health they were asked to imagine occurred towards the beginning of the 36 months or towards the end. Overall sample means for exponential time preferences of zero and –0.03 were found for the two questions (Table 9.2). Thus a mean time preference rate of approximately zero was found when the ill health state occurred three months into the 36-month time horizon, and a slightly negative mean time preference rate was found when the state of ill health occurred 30 months into the 36-month time horizon. This sample may have been indicating that they had zero time preferences if a high proportion of the life expectancy would follow the state of ill health, but that they would prefer the state of ill health to occur earlier than close to the end of the life expectancy.

If time preferences are constant when isolated from other factors, it should be legitimate to measure them over a long period as was done by Cairns (1991, 1992). These time preference rates could then be applied to any valuations made by these individuals. However, since time preference has been shown in several cases to depend on duration (*e.g.* Sackett and Torrance, 1978; Dolan, 1996), time preference rates obtained using the lifetime duration of survival suggested by Cairns (1991, 1992) and Cairns and van der Pol (2000) may have been irrelevant in the AAA study.

The Cairns (1991, 1992) study used a life expectancy of 42 years. This was long enough for the respondent to answer a time preference question without giving undue attention to the point of death at the end of the 42-year period. The life expectancy for the AAA study was only 36 months. It is probable that respondents found it difficult to isolate time preferences from the context of proximity of death.

The findings of this study highlight the difficulties of measuring time preferences over a short life expectancy such as might be associated with terminal illness. The results may have reflected the fact that the point of death was so clearly in view with the time horizon being so short. The problem with trying to measure time preferences over such a short time horizon followed by death was that there was a perceived imminence of death. This definitely affected the answers of one man in the sample, who commented that he would rather have the ill health state later than stated, but he did not wish it to be too close to the point of his death.

It is possible that the imminence of death had such an effect on the results that the figures obtained were not really a measurement of pure time preference at all. Rather than merely considering the timing of the ill health event. respondents may have been placing it in context with the occurrences around it such as the proximity of death. Of course, it is possible that they would have thought about not wanting it to be close to their death even if the time horizon under consideration had been 42 years. However, if considering a time horizon in which death is 42 years from the present, the impact of death may not seem so immediate as it might if it is three years from the present.

As far as this author is aware, the AAA study described in Chapter 8 has incorporated the first attempt to measure individual time preferences over such a short life expectancy. The finding that it is difficult to differentiate preferences for timing from context (*e.g* proximity of death) is an important result. There are many instances of terminal conditions in which it may be desirable to take time preferences into account. It would appear that the method used by Cairns (1991, 1992) and Cairns and van der Pol (2000) is not adaptable to very short life expectancies of three years. Measuring time preferences over a longer time horizon such as was done by these authors may well produce time preference rates, but these may not be relevant to a shorter time horizon.

A subject of further research should be to attempt to elicit time preferences over short, intermediate, and long time horizons and determine the effect of the length of time considered. An answer is needed to the question of whether it is appropriate to apply time preference values obtained using a long time horizon to valuations of short-term terminal profiles. It may be that time preferences of individuals are fluid, changing over different periods of life or life stages as suggested by Pliskin *et al* (1980). If terminally ill patients are providing QALY values to be applied to terminal profiles, it may not be appropriate to ask them to provide time preferences over a longer time horizon than they can expect to live. Apart from for ethical considerations, there may be a lack of relevance.

To put the findings of the AAA study into context of the work of other researchers, the value of y (the state of ill health occurred at 30 months) was –0.026 (Table 8.15). The lowest mean discount rate found previously was –0.029 (Dolan and Gudex, 1995).

In summary, the AAA study attempted to measure individual time preferences directly. and found that time preferences varied within individuals according to the timing of the state of ill health. However. some doubt is cast over the validity of the results. because the proximity to death may have affected the values given by respondents. If this was the case, the results may not have reflected pure time preferences. This has implications for adjusting QALY values for time preferences when applying the QALY approach to short-term terminal health profiles. Since duration has already been shown to affect valuations using the TTO (Kirsch and McGuire, 2000), it may not be appropriate to apply time preference obtained over longer time horizons to short life expectancies.

9.2.1.2 Adjusting QALY values for individual attitudes to risk

The form of the QALY model most widely used in economic evaluation assumes risk neutrality. However, a risk-adjusted version of the QALY model has been developed (Miyamoto and Eraker, 1985) and assumes that risk attitude is constant with respect to survival duration. The AAA study described in Chapter 8 sought to test both these assumptions. A certainty equivalent method was used to assess individual risk attitudes, as described in Chapter 8. Risk attitude was measured using a method based on that used by McNeil et al (1978). This involved three TTO-based certainty equivalent questions, which asked respondents to state the number of certain years that would be equivalent to utilities of 0.25, 0.5, and 0.75 (on a scale of 0 to 1). In this study, the QALY scenario values were adjusted for individual risk attitudes in the *ex post*, and the holistic values of the scenarios were in the *ex ante* perspective.

The AAA study showed evidence that respondents were inconsistent in their risk attitude. When risk attitude was measured the sample were found to be, on average, risk seeking. However, risk attitude was non-constant over expected survival. They tended to be risk seeking over the short-term expected survival, and risk averse over the longer-term. This is a clear violation of the axiom of constancy of risk attitude to survival duration.

The primary analysis indicated that there were small, non-significant differences between QALY valuations of the two AAA scenarios of BMT and EVAR. However, upon adjustment of QALY values for risk attitude, the median risk-adjusted values showed a preference for the more risky EVAR profile (although it did not reach statistical significance). However, there were extensive violations of the assumption of non-constant attitude to risk, and many individuals demonstrated varying risk attitudes over expected survival. Because many individuals demonstrated large values for the risk attitude parameter (r), the mean values for risk-adjusted QALYs were driven up to extremely unrealistic values (Table 8.13).

As already explained in Chapter 8, it was originally the intention to also measure quantity effect in relation to risk attitude. Unfortunately the quantity effect question proved too cognitively demanding and had to be left out of the questionnaire. In theory, it should be possible to separate the effects of quantity and risk attitude. It is also likely that when risk attitude measurements are TTO-based they would incorporate time preferences. Thus it is probable that the risk attitude measurements obtained in this study were not of pure risk attitude. The QALY algorithm assumes zero time preference and zero quantity effect, but it is possible that the risk attitude measurements compounded these three effects. These factors cast doubt on the legitimacy of the risk-adjusted QALYs.

The findings of non-risk neutrality and non-constancy of risk attitude in the AAA study are not in themselves original as previous literature has suggested these findings also (Gaskin *et al*, 1998; Brealey and Myers, 1988; Boyd *et al*, 1982; O'Connor, 1989; Sackett and Torrance, 1978; Verhoef *et al*, 1994; Mehrez and Gafni, 1987). Verhoef *et al* (1994) found similarly to the AAA study in this thesis, that their sample (although risk averse overall) showed risk seeking behaviour over the short-term and risk averse behaviour over the longer term. Despite these previous findings, the findings of the AAA study have some important implications. Although there have been previous studies attempting to measure risk attitude (McNeil *et al*, 1978, 1981) and suggesting how QALY values may be adjusted for individuals' risk attitudes (Miyamoto and Eraker, 1985), this previous research has seen relatively little application in the economic evaluation literature. The AAA study reported in Chapter 8 of this thesis explored risk attitudes over a short-term life expectancy of two to three years, and therefore offers a new context for this area of research. Other studies have looked at longer life expectancies (McNeil *et al*, 1978). The AAA study demonstrated the practical difficulties to be encountered in attempting to measure individuals' risk attitudes over short life expectancies when individuals have non-linear risk attitude. This is a very important finding, as it is not made allowance for in the literature surrounding the issue of adjusting preferences for risk attitude.

In summary, this was a novel piece of research exploring the use of the Miyamoto and Eraker (1985) method for measuring and adjusting QALY values for individual risk attitudes over a short life-expectancy. The axioms of risk neutrality and constancy of risk attitude over expected survival were shown to be widely violated. It is likely that the methods used did not measure pure risk attitude, but that the results were influenced by such factors as time preference, quantity effect, and proximity of death. These results highlight the potential problems that may be encountered in attempting to adjust QALY values for risk attitude over short life expectancies.

9.2.1.3 Additive separability

This thesis explored two aspects relating to the additive utility function. The varicose veins study described in Chapter 7 explored issues relating to process utility and the assumption that small duration effects due to temporary treatment states are negligible to the patient. The two IBS studies described in Chapters 5 and 6 looked at the assumption that valuations of a profile should reflect the proportion of time spent in each health state within the profile.

The impact of short-term treatment states on valuations

The varicose veins study in Chapter 7 aimed to determine whether patients may place greater weight upon mode of treatment (*i.e.* short-term transient states) than would be predicted by additive separability. There has been debate in the literature about the differences between outcomes and process. Donaldson and Shackley (1997) refer to the "narrow consequentialist" view of health to which non-welfarists generally adhere. Rather than considering the entire utility function, as a welfarist may wish to do, the non-welfarist may wish to consider only health outcomes. Although it is possible to view aspects of treatment processes as health outcomes in themselves, in that they can have beneficial effects on mental well being in terms of factors such as increased happiness or reduced anxiety, non-welfarists frequently define health according to the attributes of the disease under study. Other potential outcomes, such as those resulting from treatment process, tend to be ignored.

In Chapter 7, the processes of treatment were treated as short-term temporary health states, and were valued as such. The descriptions of surgery and sclerotherapy were in terms of HRQoL rather than more general QoL. Whereas the constituent health states of the profiles were valued by TTO, the processes of surgery and sclerotherapy were valued on a VAS. The VAS values were transformed to TTO values using the MVH algorithm (MVH Group, 1995). As explained in more detail in Chapter 7, the reason for the use of the VAS was to avoid the potential confusion that might have arisen among respondents from the use of the two-stage TTO method suggested for valuing short-term health states. The values for the treatment processes were slotted into the

QALY algorithm to calculate values for profiles consisting of pre-treatment state. treatment, and a post-treatment "rest of life" state. These profiles were also valued holistically using a single-stage generalised TTO.

An objective of this study was to determine the role of treatment process in decision-making. The QALY algorithm assumes that experiences lasting a very short duration have little effect on the valuation of the profile. The varicose veins treatment processes described in the profiles each involved a matter of weeks: for surgery, respondents were told that they should expect to take leave of absence from their job for three to six weeks; for sclerotherapy, leave of absence was expected to last 48 hours. It was expected that the shorter duration of the effects of treatment might cause respondents to react favourably towards sclerotherapy. For the QALY method the differences between valuations of profiles in which the only difference was the treatment process were significant ($p < 0.05$). These differences were logical. For example, profiles ending in mild were preferred to the equivalent profile ending in moderate. However. the difference was only small (0.02 to 0.03 in favour of sclerotherapy).

There were no significant differences between holistic valuations for any of these profiles, regardless of the sequence of states and treatments being described. However, varicose veins is a relatively mild condition, and people were not willing to trade many years of their lives. A substantial portion of the sample (35.6%) was unwilling to trade for any of the profiles or states. This caused the mean valuations to be closer to the top of the utility scale. It is possible that any differences between profiles as measured by the holistic method therefore were undetectable. The high percentage of respondents who were unwilling to trade may have accounted for the lack of significant differences found between the profile valuations by the holistic method. For such a mild condition, the 0 (death) to 1 (full health) scale may not be sensitive enough to reveal differences between preferences for the profiles.

The results of the varicose veins study indicated that the QALY method of valuation was more sensitive to the different treatment processes than the holistic method. This may have been because the varicose veins profiles were too mild to warrant trading life years in the views of over a third of respondents. However, it is also possible that the apparent greater sensitivity of the QALY method is an artefact of the methods used. A review of methods used to value health states has found that values obtained from the VAS method show a greater correlation with changes in health status than values from either the SG or the TTO methods, which appear to be more highly correlated with changes in level of preferences (Brazier *et al*. 1999). The QALY values for varicose veins profiles were obtained using the TTO for health state valuations, and the VAS for valuing processes of treatment. It is possible that the small differences demonstrated by the QALY method between health profiles containing different treatment processes reflected a greater sensitivity to the differences in health status relating to surgery and sclerotherapy. Nonetheless, although there is evidence from the focus groups (Chapter 7) and the literature (Bradbury *et al* 1999, Tibbs 1992, Weiss 1999, Wyatt 1999) that varicose veins can have a significant impact on HRQoL, this was not reflected by the holistic valuations.

The results from this study do not refute the additive utility axiom, because the results from the holistic method did not suggest that more weight is placed on treatment states than would be suggested by the QALY algorithm. However, there are some doubts over the sensitivity of the holistic method.

Although there has been previous research using WTP and discrete choice experiments to explore values of treatment process (Shackley and Cairns, 1996; Sculpher *et al*. 2004), it is unusual to include a description of treatment process in the description of health profiles. This study was the first, to the knowledge of this author, to explore the issue of additive separability by looking specifically at differences between QALYs and holistic values for the same health profiles when the only difference was the short-term treatment state.

Do profile valuations reflect proportion of time in each health state?

The two studies reported in Chapters 5 and 6 of this thesis added to this research by taking a different approach to looking at the additive utility axiom. Previous work has largely been around sequences over a profile. These two studies examined the issue in terms of proportion of time in each health state over a health profile, such that the sequence over the profile was unknown and unpredictable. The IBS profiles of Chapters 5 and 6 were described in terms of the proportion of time in each IBS-related health state over a 12-week period that repeated over again for the rest of the respondent's life, the sequence and episode durations of which would be unpredictable. The single-stage SG method was used to elicit valuations.

The mean holistic valuations were found to differ significantly from the mean QALY valuations for most of the health profiles in Chapter 6, with holistic values being higher than QALY values in most cases (Table 6.9). However, the split-test of reliability used in Chapter 6 indicated a failure on the part of respondents to realise that the profile describing the state P+U+ should have the same value as P+U+ when described as a health state. This suggests that respondents did not fully understand the valuation task, and that this holistic method may not be a reliable way of eliciting preferences for IBS profiles.

One possible explanation for the differences between the QALY and holistic scores for the profiles is that the profiles may have been too difficult for the respondents to visualize. Thus respondents may have used heuristics to judge them, such as "I will feel bad most of the time" or "I will feel good most of the time". Gigerenzer *et al* (1999) suggest that people use "fast and frugal heuristics" to guide them in their choices over decisions. These may or may not be based on experience, and aid the decision-maker in the decision process by considering the minimum amount of information possible in order to make a choice. According to Gigerenzer *et al*, such heuristics do not rely on probabilities and utility calculations (Lloyd and Hutton, 2002). However, Cairns *et al* (2002) express concern that if decisions are based simply on heuristics, they may not reflect actual preferences.

Another possibility is that respondents were displaying insensitivity to scope. This has been frequently reported in the contingent valuation literature (Kahneman and Knetch (1992); Diamond *et al*, 1993; Kartman *et al*, 1996; Healey and Chrisholm, 1999; Carlsson and Johansson-Stenman, 2000), and refers to the situation when the value for the whole project is lower than the summed value of the individual parts of the project. This is also known as the embedding effect (see Chapter 2). Healey and Chrishom (1999) suggest that people might fall back on heuristics such as anchor points to assist in the decision-making process. In this case the IBS patients may have been insensitive to the precise proportions of time spent in each health state.

A third possibility is that the random nature of the occurrence of each health state meliorated the proportion of time in that health state. For example, suppose 50% of the time would be spent in the worst health state (P+U+), but the respondent was told that this might not occur in one block of time but rather occur randomly. The respondent might feel that the whole profile would not be so bad if P+U+ occurred now and then throughout, as if it occurred in one block. This possibility may be related to the concept of maximal endurable time (Sutherland *et al*, 1982) mentioned in Chapter 3.

These are the first studies to explore this issue in terms of proportion of time in each health state rather than the usual sequence approach. Although the QALY and holistic results were different in Chapter 6, it cannot be stated for sure at this point which is the more valid reflection of preferences. However, an effort was made to determine the degrees of validity for both methods, and the results from these tests are reported in Section 9.2.2.1.

9.2.1.4 Summary of testing the QALY axioms

The results from the four studies are summarised in Table 9.2. In terms of the QALY axioms tested the results are mixed. Time preferences have been shown to be non-zero when tested over an expected life expectancy of 36 months. Respondents in the AAA study gave a wide range of time preference values, and in many cases different time preferences were given by each respondent depending on whether the ill health state would occur closer to the beginning or the end of the time horizon. It is possible that the results did not reflect pure time preference, as suggested by the comment of one respondent who said he would rather have the ill health later, but not for it to be near the time of his death. This study brings to light difficulties with using this method to assess individual time preferences for short-term terminal scenarios.

Attitude to risk was measured using a certainty equivalent method, and it was shown to be non-linear and non-neutral over a hypothetical life expectancy of 36 months. This has implications for cost effectiveness analyses in which risk neutrality is commonly assumed, or at the very least a constant attitude to risk.

The QALY model was found to be more sensitive than the holistic method for detecting differences in values between profiles consisting of different treatment processes for varicose veins. The holistic results from the IBS studies suggest that utility may not be proportional to the time spent in each health state, but it is possible that respondents were showing insensitivity to scope of the type previously found in contingent valuation studies.

9.2.2 Comparisons between the QALY and holistic approaches

The results of comparisons between QALY and holistic valuations differed greatly between studies, reflecting the mixed picture in the previous empirical literature (see Chapter 3). Each study used a different method of valuation, and the health profiles were of a different nature in each study (except for the two IBS studies). Because of the different methods used in each study, there was no basis for external comparisons between studies.

9.2.2.1 Comparisons against pre-determined criteria

This section systematically compares the results from each of the four studies in terms of differences between QALY and holistic valuation methods using the criteria set out in Chapter 4 and Table 9.2. These were in terms of comparisons between QALYs and holistic valuations for feasibility in terms of completion rates, logical consistency, and convergent validity (Dolan, 2000).

Feasibility of methods of valuing health profiles

One way to assess feasibility is to determine the percentage of completed questionnaires and the number that had to be excluded from analysis due to unclear responses, which may have been symptomatic of a lack of understanding on behalf of the respondent.

The rate of completion of valuation exercises was compared between valuation methods. The data collected for the first IBS study was highly complete, with only one (2%) missing SG health state valuation out of 49 respondents. This translates to a completion rate of 98% for the QALY method and 100% for the holistic method.

For the second IBS study, seven (12.5%) were excluded due to missing valuation data. Four (7.1%) of these exclusions were due to missing health state valuation data, and four (7.1%) were due to missing health profile valuation data (one respondent had missing data for both). Thus there was a 93% completion rate for both the holistic and the QALY methods.

A total of eight (11.8%) varicose veins respondents were excluded due to incomplete valuation data. Of these eight, five were excluded due to missing data for both health states and profiles. A total of two were excluded due to missing health state valuation data only, and one respondent was excluded due to missing data for health profiles only. Thus there was a completion rate of 90% for the QALY method and 91% for the holistic method.

Out of a sample of 61 respondents to the AAA study, it was necessary to exclude eight (13.1%) due to unclear data (Table 8.A.3). Three respondents were unclear in their responses to health state valuations, and three were unclear with respect to scenario valuations. One respondent was unclear for health states and health scenarios. Thus there was a completion rate of 93% for both the QALY and holistic methods.

As shown in Table 9.2, completion rates were similar between studies and between the QALY and holistic methods of valuation. From these data it can be concluded that the three methods of valuing holistic profiles are all feasible, and all have a rate of completion of at least 90%.

Logical consistency

One way of looking at the meaningfulness of responses is to determine if they make logical sense. One way of determining this is to examine the logical consistency of respondents. Thus, where there is a logical ordering of health profiles, the valuations for these profiles should follow this logical order. For example, if profile A is logically preferable to profile B because it contains less time in an ill health state, this should be reflected in the valuations of the profiles, such that the value given to profile A should be higher than that given to profile B. Logical consistency was examined for the two IBS studies in Chapters 5 and 6, and for the varicose veins study in Chapter 7. The rates of logical consistency were compared between the QALY and holistic methods of valuation. The results are shown in Tables 5.14, 6.13, 7.13, and summarized in Table 9.2.

There are three levels of logical consistency. If all logically superior health profiles are ranked higher than logically inferior profiles, they are strongly consistent. If some profiles that are logically superior are ranked equally to logically inferior profiles, they are weakly consistent. Finally, if logically superior profiles are ranked below logically inferior profiles, they are non-consistent.

The results are clearcut for the first IBS study (Chapter 5). The QALY achieves a 100% rate of strong logical consistency for all pairwise comparisons between health profiles C, D and E (Table 5.14). The holistic method was less successful, with 43.8% responses showing strong consistency for comparisons between C and D, and D and E. Profile C was rated higher than E in 56.3% of responses. However, if weak consistency (where the logically preferred profile is rated greater than or equal to the logically less preferred profile) is allowed, the holistic method fairs better, though still less well than the QALY method. For comparisons between profiles C and D 81.3% of holistic valuations were weakly consistent, for comparisons between profiles D and E 75.1% were weakly consistent, and for comparisons between profiles C and E 77.1% were weakly consistent, compared to 100% in all cases for the QALY method.

The QALY did not fair so well in terms of logical consistency in the second IBS study (Chapter 6), in which the QALY never achieves 100% strong consistency (Table 6.13). However, if weak consistency is taken into account the QALY scores 100%. The QALY still performs better than the holistic method, which has rates of inconsistency (a logically preferable profile rated lower than one to which it should be preferred) ranging from 8.2 - 22.4% for pairwise comparisons between health profiles that had a clear logical order (*e.g.* A and B, B and C, *etc.*). Put conversely, the holistic method achieves weak consistency for 77.6 - 91.9% of responses across the pairwise comparisons between health profiles. In the first IBS study the range of weak consistency was 75.1 – 81.3% (see above). Thus the holistic method performed better in the second IBS study, whereas the QALY method performed less well in the second study.

The results from the varicose veins study in Chapter 7 indicate that, although the level of logical consistency between the two methods was similar at the level of weak consistency, the QALY method demonstrated a higher level of strong consistency (Table 7.13). However, unlike for the IBS studies, in the varicose veins study the QALY method demonstrated a small degree of non-consistency, ranging from 5.1-6.8% compared to a range of 6.8-11.9% for the holistic method.

On the face of it, these results could lend support to the QALY method. It is clear that the QALY valuation method produced greater logical consistency than the holistic method for all three of the above studies.

It is certainly worth noting that, even when the holistic method breached logical consistency, the breaches were often only by a small amount. For example, in the first IBS study most inconsistencies were in the order of magnitude of less than 0.05. In the second IBS studies these inconsistencies were all of less than 0.05 except for two responses which were approximately 0.07. In the varicose veins study, the QALY and holistic methods both produce some inconsistent responses. These inconsistencies were of greater magnitude than encountered in the IBS studies. For the QALY method they ranged from 0.003 to 0.4, and for the holistic method they ranged from 0.1 to 0.45. The inconsistencies were larger for the holistic method.

There is the question of how weakly consistent responses should be dealt with. These are the responses for which the logically less preferable outcome is valued equally to one that is logically preferred. Some may argue that weakly consistent responses should be grouped with the inconsistent responses, because they do not show strong consistency and are therefore not, strictly speaking, following the logical ranking order. However, they may be an indication of the occurrence of heuristics in decision-making. The respondent may not be making his judgements based on exact proportions of time in each health state, or exact TTO values, but more fuzzy estimates. The question is, should it be assumed that only strongly consistent results show that people understand the task, and should strong consistency therefore be used as a measure of how good a valuation method is? Or should allowance be made for the possibility that people may not make preference judgements precisely enough to follow logical ranking exactly? If the latter is allowed, the weakly consistent responses may be valid.

Richardson *et al* (1996) also examined responses for logical consistency. These authors used health states that had a non-ambiguous order of logical ranking, and because their health profiles consisted of a sequence of these health states they expected the holistic value of the profile to fall within the range of values for the discrete health states. Approximately 73% of their sample achieved logical consistency according to these criteria. The weak consistency rates are higher for the IBS and varicose veins studies in this thesis.

The two IBS studies and the varicose veins study demonstrated high levels of logical consistency at the level of weak consistency. The QALY method consistently out-performed the holistic method. This was most markedly so in the first IBS study, in which the QALY method achieved 100% strong consistency.

Convergent validity

Another proxy for validity is a measure of convergence between the ranking order of profile valuations and the direct ranking of the profiles (Streiner and Norman, 1989; Dolan, 2000; Bleichrodt and Johannesson, 1997). Thus the order of preference between profiles as given by the valuation method is compared with the order suggested by the straight forward ranking of those profiles. Convergent validity between the QALY and holistic profile valuations and the original ranking order was tested for the results from the two IBS studies and also the AAA study in Chapter 8.

For both the IBS studies, the QALY method performs better than the holistic method in terms of convergent validity. There were three pairwise comparisons in the first IBS study compared to seven in the second. In the first study, 60.4% of respondents achieved strong convergency with ranking by the QALY algorithm in two out of the three comparisons, and 37.5% of respondents achieved convergency with all three pairwise comparisons (Table 5.13). In the second IBS study 56.3% of respondents achieved strong convergency for all seven out of the pairwise comparisons for the QALY method (Table 6.12). This is over half the sample, and therefore seems to imply that the QALY performs better in terms of convergent validity in the second IBS study than the first IBS study. A further 22.9% achieved convergence for six out of seven comparisons, and 10.4% achieved convergency for five out of seven pairs.

In the first IBS study, for the holistic method 52.1% of respondents achieve strong convergent validity for one out of the three pairwise comparisons, and a further 45.8% achieve convergent validity for two out of the three pairwise comparisons (Table 5.13). For the second IBS study, 25% of respondents achieve no convergent validity at all for the holistic method of valuation (Table 6.12). Convergency was achieved for all seven comparisons by one respondent. Convergency was achieved for six out of the seven pairs by 6.3% of respondents, with 18.8% achieving convervency for five pairs, 12.5 achieving convergency for four pairs, 16.7% achieving convergency for three pairs, 10.4% achieving convergency for two pairs, and 8.3% achieving convergency for one pair. It is therefore a mixed picture in terms of convergent validity for the holistic method in the second IBS study. If the weak level of convergency was allowed, the level of convergency demonstrated by the holistic method rose considerably, with 33.3% of respondents achieving convergency in all seven pairwise comparisons in Chapter 6.

For the AAA study reported in Chapter 8, the holistic method performed better in terms of convergent validity (Table 8.20), with 50% of respondents showing strong convergence for all three pairwise comparisons in the case of the holistic method, and 50% showing strong convergency for two out of the three pairs. This compares to 40.7% respondents demonstrating strong convergency for all three pairs for the QALY method, and 59.9% showing strong convergency for two pairs. When weak convergency is included, the results are the same for the QALY method, but 75.9% respondents showed weak convergency for the holistic method.

These results suggest that there might be differences in validity measures between different types of health profiles or in different contexts. It may be due to the different nature of the AAA scenarios that the holistic method achieved greater convergency than the QALY, whereas the QALY achieved greater convergency in the IBS studies. It is possible that the holistic method better reflects preferences over *ex ante* risks.

Reliability

In the second IBS study the final question comprised a split-test reliability test (Dolan, 2000). The profile being valued and the worse state used as the failure branch of the gamble were essentially the same (although framed slightly differently), and it was therefore expected that a value of zero would be given to the profile. However, the majority of respondents gave it a value significantly greater than zero. It appears that this test of split-test reliability was failed by most respondents. This failure of the split-test of reliability suggests a lack of reliability in the holistic valuations.

It would have been useful to have information on respondents' reasons for giving this profile a value higher than zero. However, these were not ascertained due to the constraints of group interviews and time constraints of the interviews themselves. Various possible reasons could be speculated. Among the possible reasons for this failure of the reliability test is a lack of understanding of the task, implying cognitive difficulties. If this was found to be the reason the implications for holistic valuations would be discouraging.

Another possible explanation could be that respondents were showing an aversion to gambling with their life. They may have preferred the profile to the failure state because the latter implies a failure. However, this would still imply a failure to understand the question.

Another possibility is that respondents did not wish to take the hypothetical treatment mentioned in the SG instructions if the result would be a certainty of remaining in the same state or profile. To some people the process of receiving treatment in itself may involve some disutility, but this possibility was not accounted for in this study.

Of course, it is possible that the reason for the failure of the split-test of reliability was something completely different. The only way to find out would be to conduct a further study that included discussion with each individual of their responses and their reasons for their responses.

Summary

All four studies in this thesis demonstrated a high rate of completion, with both the QALY and holistic methods of valuation achieving greater than 90% completion rates. These results indicate that the valuation methods were highly feasible.

Two measures of validity were used: logical consistency of value rankings, and convergent validity of value ranking with the original ranking order of the profiles. The QALY performed better on both tests for the two IBS studies, and on the test of logical consistency for the varicose veins study. There was no logical ordering for the scenarios used in the AAA study, so convergent validity was used on its own. In the AAA study, the holistic method performed better than the QALY method, suggesting that the cognitive difficulties commented on by several respondents may have related more to the QALY method than the holistic approach. It may be that participants in the AAA study found the QALY method more demanding because they had to provide three risk attitude assessments and time preference values to be entered into the QALY algorithm.

The second IBS study contained a question which was used as a split-test of reliability. The majority of respondents failed this test, casting doubt over the reliability of the holistic approach as used in this study.

It is curious as to why the validity tests give different answers for different studies. Each study was very different in nature, with different valuation methods and different types of health state and profile. There were several comments from respondents, expressing difficulties with the surveys. However, in many cases there was no specific indication as to exactly what they found difficult. The results of the tests of validity suggest that, in the cases of the two IBS studies and the varicose veins study, the QALY method may have been less demanding cognitively. However, for the AAA study the results suggest that the holistic method may have been cognitively less demanding than the QALY method. However, these findings are not intuitively obvious. One might have supposed that the IBS health profiles may have been easier to rank than the health states. In the first IBS study there were only three profiles, whereas there were six states. In addition, for the health states respondents had to choose which attributes were more important to them. For example, in comparing P-U+ and U+P-, respondents had to decide whether they would prefer to suffer from urgency or abdominal pain. This kind of decision was not employed in ranking or valuing the health profiles. The health profiles contained all the health states, and differed only in the proportion of time in each health state. One might have expected the ranking order to be obvious for the health profiles.

In conclusion, whilst the tests of validity described above may give indications as to the validity of the two valuation methods, they are proxies in the absence of a "gold standard" for comparison (Dolan, 2000).

9.2.2.2 Overall comparisons between QALY and holistic values

One of the most important findings of the research in this thesis is that the results of comparing QALYs and holistic values vary considerably across the different studies. In trying to directly compare the results across studies, the author would not be comparing like with like. The four studies obtained valuations for different illnesses (IBS, varicose veins, and AAAs), and used different methods of valuation (the IBS studies used a single-stage SG, the varicose veins study used a single-stage TTO, and the AAA study used a single-stage TTO with *ex ante* scenarios).

In the case of the IBS studies, there were significant differences between QALY and holistic valuations (most notably in the second IBS study in which more measurements were taken), and the QALY values more closely reflected the increase in frequency of the worst health state within the profile. Within the varicose veins study there were few differences between QALY and holistic valuations of the health profiles. For the AAA study there were significant differences between values for scenarios from the two methods, and the holistic valuations were more convergent with the original rankings of the scenarios.

These findings would suggest that the differences between QALY and holistic values may depend in some way upon the context in which they are obtained. For example, they may be dependent upon the illness under study, or the method of obtaining valuations. As already cited in Chapter 3, the results of previous research into comparisons between QALYs and holistic values for health profiles has also been variable. It has been suggested that, when health profiles deteriorated markedly over time, the QALY values will be higher than the holistic values. For example, Richardson *et al* (1996) presented respondents with hypothetical breast cancer health profiles that deteriorated in a noticeable manner, and found that mean QALY values were higher than holistic values obtained by the single-stage TTO. However, MacKeigan *et al* (1999) investigated valuations of health profiles involving a gradual deterioriation of health for type II diabetes using the single-stage TTO, and found that the QALY and holistic methods gave similar results.

Like the study by MacKeigan *et al* (1999), the varicose veins study described in Chapter 7 found few differences between the values obtained by QALY and holistic methods. It may have been the case that the TTO scale chosen was not sensitive enough to detect differences in values which may have existed between health profiles. The condition of varicose veins is generally accepted to be comparatively mild, and as such respondents may not have wished to trade much of their life expectancy for improvements in health. This was suggested by the high proportion of respondents (35.6%) who were unwilling to trade life years for improvements to their varicose veins. Some of the comments (see Table 7.A.9) also support this hypothesis. However, if they had been offered the opportunity explicitly, they may have been willing to trade in very small units such as days or hours. As it was they were asked to imagine that they had a 20-year life expectancy, and the TTO scale offered trade-offs in two-year steps (see Appendix 3).

Risky profiles

The importance of how risks are incorporated into valuations is indicated by the results of the varicose veins study (Chapter 7) and the AAA study (Chapter 8). The final two questions in the varicose veins study examined the effects of stating the risks involved in the evaluations of the profiles. Surgery and sclerotherapy involve different degrees of risk. Sclerotherapy has a high recurrence rate compared to surgery (these were estimated at 75% and 20% respectively). However, surgery has a 1/10000 mortality rate. Thus there were two risk domains to be considered: the risk of recurrence and the risk of mortality.

The holistic valuation process appeared to lack sensitivity, as already discussed. The results are therefore somewhat unclear. However, it is noticeable that there is a greater effect upon holistic valuations of the Moderate → Surgery → Mild profile when the risks of mortality and recurrence are incorporated than for the equivalent sclerotherapy profile (a drop of 0.03 on the 0 to 1 scale). However, this difference is not significant according to the t-test. This meant that there was a non-significant mean preference for the sclerotherapy profile when risks were incorporated. However, according to the QALY results, there is a small but significant preference for the surgery profile when risks are included. These results suggest that, for *ex ante risks*, respondents may have been more concerned with the small risk of death than recurrence, but when the risks were considered *ex post* greater weight was given to the risks of recurrence. These findings support those of Cook *et al* (1994), who found that, when they incorporated an *ex ante* risk of death into their health states, the values were lowered significantly. The effect was enough to reverse health state rankings.

The only previous study into the use of holistic valuations to fully incorporate *ex ante* risks into health profiles was a study by Sculpher (1996) into different treatment pathways for menorrhagia, looking at risky scenarios relating to chances of death from surgery and recurrence of menorrhagia. Sculpher conducted a CUA and compared the results of QALY and holistic values for this analysis, and found that his holistic method gave a lower cost per incremental benefit for abdominal hysterectomy than the QALY algorithm. The less invasive form of treatment (transcervical endometrial resection) involved increased risks of recurrence, but no risk of mortality as was incorporated into the abdominal hysterectomy form of treatment. Sculpher's study indicated that his sample of respondents placed a higher importance on the risks of recurrence than on the small risk of mortality associated with abdominal hysterectomy. Risks of recurrence for the less invasive treatment were two-fold: a 12% chance was sited that she would have to undergo the same treatment again, and a 16% chance was sited that she would have to undergo a full abdominal hysterectomy as a result of recurrence. A risk of death of 1 in 1000 was sited for abdominal surgery.

In the AAA study, attitudes to risk were measured using a variant of the method suggested by McNeil *et al* (1978, 1981). One aim was to determine whether people would opt to maximize life expectancy while increasing short-term risks, or to choose a shorter life expectancy with fewer risks attached. Again, the results differed according to the way in which risks were taken into account. The QALY profile values were adjusted for individual risk attitudes using the method suggested by Miyamoto and Eraker (1988). Prior to adjustment of QALY values for risk attitude, the less risky but shorter-lived BMT was rated significantly higher for the holistic method, whereas there were no significant differences according to the QALY method. However, once QALY values were adjusted for risk attitude, the EVAR scenario was rated higher than the BMT scenario by the QALY (Table 8.13), although this difference did not achieve statistical significance. This is an important finding, because it indicates a preference reversal, dependent upon whether QALY values are adjusted for individual risk attitude. The way in which risks and attitudes to risk are taken into account can therefore have important implications to preferences and therefore could have implications in the long run for distribution of resources. In the case of AAAs, it could have implications for the type of treatments recommended.

The AAA study was very different to the study by Sculpher (1996), in that it explored choices over risky short-term terminal scenarios. This has never been previously done in a holistic valuation setting.

9.2.2.3 Implications from comparisons between QALYs and holistic valuations

It is clear from the work of Richardson *et al* (1996) and MacKeigan *et al* (1999) that the overall trend of health states within the sequential profile is important in determining differences between QALYs and holistic values. Similar findings have already been established in the psychological literature over non-health domains (*e.g.* Ariely and Zauberman, 2000).

Chapters 7 and 8 of this thesis have established that there are potentially very important differences between the two methods dependent upon the way in which risk is incorporated into the profiles. These studies suggest that *ex ante* holistic methods give greater weight to risks than *ex post* QALYs.

Chapters 5 and 6 are the only studies to compare holistic and QALY values for health profiles that differ in terms of proportion of time in the constituent health states. The more extensive study in Chapter 6 suggests that there are some forms of heuristic at work in respondent values by the holistic method. The implications of this are not completely clear, and further research into valuing this kind of profile is required.

This research puts holistic approaches forward in various different forms as a method of potential value in CUAs. There are questions about the validity of the holistic health profile measures, as demonstrated by the lower levels of convergent validity and logical consistency for the holistic methods when compared to QALYs in three out of four of the studies in this thesis. However, with further research (see Section 9.4), it may be possible to improve validity and produce a useful and viable measurement method.

Resource allocation issues (distributive issues) are only one application of preference measures. Another application is their use in clinical decision making to aid patients in making choices about their treatment. For this kind of use, it would be beneficial to utilise condition-specific measures such as were used in these studies. The studies described in this thesis explored the issues underlying the creation of condition-specific health state and health profile descriptions for IBS, varicose veins, and AAAs.

It can take a several years to complete research into a new method of valuing health profiles, such that a recommendation can be made regarding that method. If, after further research into holistic valuation of health profiles, the empirical evidence supports the use of this method either as a replacement for, or in conjunction with, the QALY, this could have significant impacts on health policy decision-making. In some cases the results from the two methods are different, and in these cases this could sometimes lead to different policy decisions being made. For example, there was a marked difference in the results of the profile valuation in Chapter 8. Using the QALY, the risk-adjusted EVAR scenario would have had a higher mean value, whereas under the holistic method the less risky BMT had a higher mean value. The results of a CUA using these data would depend upon the precise costs of EVAR and BMT. However, there could potentially be very different policy decisions made regarding treatment of AAAs, depending on the valuation method used. In the case of the IBS study in Chapter 6, both methods led to decreasing values as the proportion of time in the worse symptom increased. However, the decrease in value was more marked with the QALY method. Thus the QALY method gave greater weight to the increases of proportion of time in ill health than did the holistic method. This could potentially lead to more resources being diverted from other things to be invested in IBS, whereas less money might be spent on IBS if the holistic values were used.

Summary of comparisons between QALY and holistic values

There is evidence from Chapter 6 that respondents may not have fully understood the holistic exercises, because there was widespread failure of the split-test of reliability. In three out of the four studies in this thesis, the QALY performed better in terms of both logical consistency and convergent validity. In the AAA study, the holistic approach performed better in terms of convergent validity. The weight of evidence from the research reported here is that the QALY often performs better against criteria tests. However, this may depend on context, and results can be quite variable.

The overall differences between the values of health profiles as valued by QALYs or the holistic approach are also variable, and evidently depend upon the factors of the study. In the IBS studies the QALY values tended to be lower than holistic values, and although mean holistic values were lower when proportion of time in the worst health state increased, they showed insensitivity to the proportion of time in each health state in that they did not decline to the extent that would be expected if respondents had been employing additive separability of utility in their responses. The results for the varicose veins profiles were similar for both valuation methods. According to the holistic results, the BMT scenario was preferred in the AAA study, whereas the risk-adjusted EVAR scenario was rated higher (though not significantly) by the QALY method.

There are a number of questions raised by the research described in this thesis. For example, even if holistic valuations are theoretically better than QALYs, should this fact override empirical problems with deriving holistic values? Is it better to measure the right thing badly, or the wrong thing well? The proponents of the holistic methods of valuation might ask whether respondent understanding matters more than measuring the right thing. After all, they might argue, surely it is pointless to obtain accurate measures of the wrong thing! However, proponents of the QALY might reply with the argument that the QALY is a model. As with all models of reality, we do not expect it to get everything right every time, but to approximate the correct answer. These issues are discussed further in Section 9.2.3.

One of the possible limitations of holistic valuation methods is the possibility of cognitive overload of respondents. One way around this would be to devise ways of constructing profiles that are easier to process mentally.

9.2.3 Choosing whether to use QALY or holistic valuations of profiles

Whereas the results of the studies reported in this thesis demonstrated that the QALY and holistic methods may give rise to different results, and perform differently in terms to pre-defined criteria, these findings do not lead to a definitive conclusion as to which method is a more accurate reflection of preferences. The studies used a "positive economics" approach, *i.e.* an observational essay on how people value profiles of health (Friedman, 1953). However, the matter of which method is intrinsically "better" than the other is as much a normative or theoretical one (Keynes, 1917).

The decision as to which method to use may depend on context. For example, if, as in the varicose veins study, respondents would show little differences between the two methods, it may be better to use the QALY on the grounds that it is easier for both the researchers and the respondents (as suggested by the criteria tests). MacKeigan *et al* (1999) found that holistic and QALY methods gave the same results for graduall*y* deteriorating type 2 diabetes profiles. They concluded that, since either method could be used, it would essentially be better to use QALYs because of ease. However, this may not apply in other contexts, such as when the holistic method demonstrates greater convergent validity such as in the AAA study of Chapter 8.

One might ask if it is right to base the distribution of limited health care resources on the inability of people to judge between different choices. For example, the holistic values from the IBS study in Chapter 6 suggested that people may use heuristics to simplify complex cognitive tasks. However, it could be argued that the EUT assumptions and those underlying the QALY model should be used even if people do not obey them in real life, because these models are normative and state how we *should* make decisions. It should be realised that no descriptive model of decision-making will be entirely realistic (Friedman, 1953). What is important is the degree to which each model differs from realism. It is perhaps the case that most people find the cognitive tasks involved in rationally weighting the choices too complex. One possibility is that heuristics do reflect individual preferences to some extent.

Another question for consideration is that, if the logical consistency is less for the holistic method yet still reasonably high, could there still be an argument for it being more justifiable to use holistic valuations on the basis of their theoretical superiority? In support of this argument is the finding that responses were inconsistent by such tiny amounts. Also, there is a large body of evidence that the axioms of the QALY algorithm are violated, to which these studies add.

In three out of the four studies, the holistic approach faired less well than the QALY in terms of both convergent validity and logical consistency. The test of logical consistency is the stronger test of validity, because valuation rankings can be compared to a "gold standard" ranking that would be attained if respondents gave values to each health profile based on the contents of that profile. Thus a profile containing more illness would, logically, be given a lower value than one containing less illness. The extent to which responses differed from this logical ranking order might be seen as a reflection of how well respondents understood the exercise, the health profiles, or how accurate their individual internal value system was.

The weakly consistent responses are not necessarily invalid, because it is possible that individual internal value systems are not so precise as would be predicted by the QALY algorithm. There is no "gold standard" of permissible levels of inconsistency. Although levels of 25% (as were seen in the holistic results of Chapter 6) seem relatively high, there is no basis of comparison with other studies using holistic valuation methods.

The results of the attempt to measure individual time preferences over a three-year period in the AAA study (Chapter 8) were dubious, and it is likely that the proximity of death may have affected values. There was also doubt over the risk attitude assessments over such a short time period. Individual risk attitudes were very variable, to the extent that it was impossible to use mean risk-adjusted QALYs. The AAA study highlighted the practical difficulties encountered in the assessment of time preferences and risk attitudes for adjustment of QALYs, whereas time preferences and risk attitudes are directly incorporated into holistic valuations.

One aspect of the fact that there may be differences between QALY and holistic valuations is that it may lead to biases in results if some researchers opt for the method that provides the more desirable results. However, according to the reference case for the conduct of economic evaluations in the UK, QALYs should be used unless it is possible to justify an alternative approach (National Institute for Clinical Excellence, 2004). In the future, it would be possible to alter the reference case should an alternative approach be shown to be viable.

9.2.3.1 The contribution of this research to the welfarism versus non-welfarism debate

This thesis was empirical in nature, and it was not one of the main aims to provide input into the debate between the two viewpoints of welfarism versus non-welfarism. As discussed in Chapter 2, both the QALY and the holistic approach are based in non-welfarism. They are both concerned with valuing health outcomes. In the case of the QALY, these are in the form of health states whose values may be entered into an algorithm to obtain values for health profiles. In the case of the holistic valuation method, the health profiles are valued directly in their entirety. The empirical research presented in Chapters 5 to 8 of this thesis examine these two approaches to obtaining values for health profiles.

Neither of the two valuation approaches reported in this thesis were used to measure non-health aspects of the utility function. As such, neither method was measuring utility, but merely valuing health in different ways.

The non-welfarist is not constrained to the rules of welfarism, and as such may incorporate whatever non-health aspects of utility that are called for by decision-makers or society (Brouwer and Koopmanschap, 2000). In Chapter 7, treatment processes for varicose veins were valued. It is possible that respondents may have expressed preferences for treatments based on the procedural aspects of health care rather than health itself *per se*. However, a non-welfarist policy maker may choose to incorporate such preferences into the decision-making process. A welfarist would attempt to obtain utilities regardless of the source of utility, and the whole utility function would be incorporated.

If one wished to follow the welfarist rather than the non-welfarist school of thought, profiles and states would not relate merely to health, but would incorporate the whole of the utility function. If this route was the way forward, it may be preferable to devote resources to researching ways to improve CBA techniques rather than the holistic approach to valuing profiles that has been developed in this thesis.

9.3 Contribution to knowledge

This is one of the largest programmes of empirical research to explore the issues around holistic valuations of health profiles. This research had a broad scope, examining violations of the QALY axioms, construction of condition-specific health profiles, holistic valuations of health profiles, and comparisons between QALY and holistic methods.

The research carried out in this thesis has contributed to knowledge in terms of health state and profile valuations and also in terms of implications for mainstream economics as outlined below.

9.3.1 Valuations of profiles relating to specific conditions

<u>Valuation of IBS health profiles</u>