August 2001

Mapping SF-36 to Utilities: How can it be done?

PAULA K. LORGELLY

School of Economics and Trent Institute for Health Services Research
University of Nottingham

Abstract: Recently there has been a growing need to map health status profiles, which are favoured by clinicians but difficult to use in economic evaluations, to preference-based utility measures, which allow for cost-utility analyses to be undertaken. A literature search revealed a vast amount of evidence suggesting there exists some relationship between profiles and preferences, however, few researchers have estimated empirical equations which allow for utility scores to be predicted. With regard to one of the more common generic health status measures, the SF-36, there are currently half a dozen articles which use various multiple regression techniques to derive utility scores. Each of these, in their own right, establishes a quantitative link between the SF-36 and a preference-based measure of their choice, including the Quality of Well-being index (QWB), the Health Utility Index (HUI), EQ-5D, and visual analogue scale and standard gamble measures of utility. This paper compares each of these approaches using SF-36 data from a wide-range of population groups (including data from knee pain, depression and MS sufferers) to derive health utility summary scores in accordance with the pre-determined coefficients and estimating equations as reported in this previous research. The results provide some insight regarding whether it is possible to use "off-the-shelf" predictive equations to generate utility values, or whether it is more appropriate to undertake individual mappings such that they are specific to the population cohort and health issue under consideration.

Paper prepared for the Health Economists' Study Group, Summer 2001, City University

I. INTRODUCTION

The continued favouritism by clinicians to use generic health measures rather than preference based economic measures has resulted in an upsurge of articles which attempt to map such generic health profiles to preference-based utility measures. A number of these carry out this mapping specifically to undertake cost-utility analyses using QALYs. However, other investigators have attempted to derive predictive equations (e.g. Fryback *et al.*, 1997; Brazier *et al.*, 1998; Nichol *et al.*, 2001), thereby providing an "off-the-shelf" tool for use by researchers who require utility values but were not elicited from patients within the original trial setting. If such a mapping function or predictive equation was robust then this has considerable implications for economic evaluations; consider how convenient it would be if just one instrument could give both health status and utility scores, or consider the potential increase in the number of formal cost-utility analyses which could be undertaken retrospectively.

With regard to the most common generic health status instrument, the Short-Form 36 (SF-36), there are currently half a dozen attempts which estimate utility scores and report predictive equations. Each of these vary in their approach. One maps the SF-36 to an additional health status classification and then estimates the relationship between this and elicited utility scores (Brazier *et al.*, 1998), while others simply estimate multiple regression equations using the SF-36 subscale scores as explanatory variables (e.g. Fryback *et al.*, 1997; Nichol *et al.*, 2001). Further, each researcher approaches the problem using different measures of utility; some elicit 'utility' using the standard gamble (Brazier *et al.*, 1998) and others elicit 'values' using the rating scale approach (Shmueli, 1998; Bartman *et al.*, 1998); while a number use 'preference-based' measures – pre-scored multi-attribute health status classification systems – like the Quality of Well-being index (QWB), the McMaster Health Utilities Index (HUI) or the EQ-5D (Fryback *et al.*, 1997; Nichol *et al.*, 2001; Bartman *et al.*, 1998; Longo *et al.*, 2000).

If a researcher was required to undertake an economic evaluation in a trial whose only outcome measure was the SF-36 then, at first glance, it would be difficult to decide which (if any) of these numerous predictive 'tools' to use. Meletiche *et al.* (1999) reviews three of the approaches, but as yet no one has attempted to compare the predictive equations empirically. This article, using a variety of different patient data, derives summary utility scores according to each model and compares and contrasts the results, in an attempt to provide some insight regarding how to map SF-36 profiles to utilities.

The next section provides a review of the literature with regard to the relationship between SF-36 and utility, including evidence on the simple pairwise correlation's that have been reported, as well as detailing each of the six attempts which estimate the complex relationship

and report predictive equations. Section Three, with the aid of available SF-36 data and each predictive regression equation, estimates utility scores, then analyses and compares the resulting values. The paper then concludes with Section Four, which details points for further discussion, and possible future work.

II. LITERATURE REVIEW

A number of researchers have examined the relationship between generic health measures and utility scores. Revicki and Kaplan (1993) review 15 studies comparing health status and utility-based measures and find that the correlation's are generally quite low. Similar poor-to-moderate correlation's have been reported with specific regard to the SF-36 and preference-based measures. Andresen *et al.* (1998), studying the health of older adults, finds a correlation of 0.47 between Quality of Well-being (QWB) scores and the SF-36 physical summary scale and an even lower correlation with the mental summary score (r=0.22).

One suggestion why the correlation's are so weak is that the relationship may be heterogeneous in nature, in that patients' risk attitudes and time preferences mean they assign varying weights to different health dimensions. Bult *et al.* (1998) invesigates this further, a discussion of which follows later.

Despite this low correlation several researchers have attempted to estimate utility scores from SF-36 data, thereby reporting predictive equations, which in theory can be applied to other SF-36 data to obtain utility values. (See Appendix One for a table detailing each article and Appendix Two for details of each predictive equation.) One of the first and most prominent attempts is that of Brazier *et al.* (1998) (which has recently been revised). They modify the SF-36, selecting a number of items and combining others, to obtain a six-dimension health classification, SF-6D. Each question has between two (now revised to four) and six ranking statements, for which profiles were elicited from a sample including health professionals and students (later work uses a more representative sample) who also expressed their preferences for the health states using the visual analogue scale (VAS) and standard gamble (SG) techniques (later work just uses the standard gamble). Health state values are then estimated using OLS with some fixed effects adjustment. The resulting regression coefficients can then be used, together with the SF-36 item responses mapped onto the SF-6D, to give predictive VAS or SG utilities.

Other researchers have taken a more simplified approach to the development of estimating equations, using the subscale scores from each dimension of the SF-36. Fryback *et al.* (1997) undertake analyses in order to predict QWB scores from SF-36 data. Using the eight SF-36 subscale scores as the explanatory variables they attempt to find the "most predictive and most parismonious mutliple linear regression equation" (Fryback *et al.*, 1997, p.3). This

entailed searching first- and second-degree polynomials and cross products for the best fit. The resulting regression coefficients together with summary subscale scores then allow QWB scores to be estimated.

A similar approach was taken by Nichol *et al.* (2001) who models the relationship between SF-36 and the Health Utility Index (HUI). In a cohort of HMO members they collected socio-demographic data, measured generic QoL using the SF-36 and determined preference values using the HUI Mark II. An ordinary least squares relationship, with the subscale scores and socio-demographic variables as independent variables, is then estimated. Using those variables that are found to be statistically significant (the eight subscale scores and age) utility scores can be obtained by combining SF-36 data with the estimating regression coefficients.

Bartman *et al.* (1998) also estimate the relationship between the HUI (and in addition a rating scale (RS) measure of utility) and the SF-36. For a small sample of elderly patients they collected data on health status and health state preferences and then use a stepwise linear regression model to find which of the eight SF-36 subscales and age, gender and race are the best predictors of utility scores.

Another study which, like those discussed previously, uses SF-36 subscales as predictors of utility is that of Longo *et al.* (2000). This research, however, differs in that the preference-based measure employed is the EQ-5D or EuroQol. Within a randomised controlled clinical trial examining the management of breast disorders in primary care, health status and preference values were collected. A multiple regression model is then estimated where the eight summary domains of the SF-36 were regressed against transformed EuroQol scores (transformed due to their known skewness). Insignificant explanatory variables are then omitted to produce what they regard as an optimal model.

Finally, Shmueli (1997) approaches the estimation problem from a different tract entirely. Rather than use the eight subscales as predictors, three summary measures, the overall scale, a physical health measure and a mental health measures are employed. Each is included separately in a bivariate estimating equation (such that three separate equations are estimated) which is based on the power curve relationship between von Neumann-Morgenstern utilities and measured values (in this instance the rating scale) as developed by Torrance *et al.* (1982), $\ln[1-u] = c\ln[1-v]$.

This short review has been ordered by simplicity. Brazier *et al.* undertakes complicated mapping onto an entirely new health state classification system and then estimates fixed effects regression equations, while Shmueli, at the other extreme, estimates simple bi-variate equations using summary SF-36 scores. Admittedly, it is only Brazier *et al.*, Fryback *et al.*, and Nichol *et al.* who explicitly set out to estimate the relationship between utility and the SF-36 with the aim of providing a tool for other researcher to use. Brazier *et al.* argue that "[t]he results of this study can be applied to any existing SF-36 data set" (Brazier *et al.*, 1998, p.1125); while Fryback *et al.* believe that "the equation we present here allows analysts who have SF-36 data about patient groups to translate those numeric profiles to predicted QWB scores …" (Fryback *et al.*, 1997, p.8). In a similar vain Nichol *et al.* state that "[t]he methodological framework presented in this study provides researchers with a tool to obtain an estimate of summary utility scores from secondary health status data using the SF-36." (Nichol *et al.*, 2001, p.110). In contrast, the research of Longo *et al.* and Bartman *et al.* was conducted in very condition specific environments. In addition, the work of Longo *et al.* is undergoing further development (Longo, 2001); while Shmueli admits that "[t]he relations stated in this paper are only preliminary" (Shmueli, 1998, p.195). Therefore, in the next section, while each attempt to derive utilities from the SF-36 (as described in Appendix Two) is undertaken and the results reported, in-depth discussion will focus more on the work of Brazier *et al.*, Fryback *et al.* and Nichol *et al.* (this also necessary given space limitations).

III. EMPIRICAL ANALYSIS

Aside from the work of Bartman *et al.* and Longo *et al.*, each predictive model appears to have been tested on a reasonably "representative" sample; however, in their early review of a number of these approaches, Meletiche *et al.* (1999) argued that "[t]o develop a reliable method to convert SF-36 scores into utilities, further studies in large, more diverse populations are needed" (Meletiche *et al.*, 1999, p.2023). The empirical analysis that follows is an attempt at this; diversity comes in the form of four different sources of SF-36 data which have been chosen, in part, to reflect various levels of debilitation, both physical and emotional. The data are the result of research into knee pain, depression and multiple sclerosis (MS). Knee pain is a fairly common complaint, the majority of which is due to trauma, and while it is generally acute, it can often be resolved/relieved non-invasively or with medication. However, if the pain is severe and prolonged it may affect other areas of sufferers lives, thereby impairing their emotional and mental state. The converse is the case with depression, which is by its nature an emotional and mental condition, however, if severe enough can have physical manifestations, like weight loss, sleep disturbances and fatigue. MS is different again, it is a neurological disease which has diverse and unpredictable symptoms (both physical and mental), and an uncertain rate of progression; there is also no known cure and limited evidence exists as to the success of newly developed treatments.

The knee pain patient data is from a study assessing the burden of knee pain and the success of a primary care-based exercise programme (Doherty *et al.*, 1999); while the data on depression is a by-product of the Counselling versus Antidepressants in Primary Care (CAPC) Study Group (see Chilvers *et al.*, 2001), where the efficacy of antidepressant drugs and generic counselling for treating depression in a general practice setting were compared. Two different data sets on multiple sclerosis were used, one from a study (Nicholl *et al.*, 2001) which assessed the quality of life in MS patients (SF-36 was one QoL instrument used) and the other which evaluated the benefits of providing psychology services, including cognitive assessment and invention to MS sufferers (Lincoln *et al.*, 2001).

Table 1 reports the mean values for the eight domains of the SF-36 for each data set. It is interesting to note, that as expected, the knee pain sufferers scored low on the bodily pain domain compared to the other illnesses, while the depression patients had comparatively lower scores on the emotional, mental and vitality subscales relative to the other domains. In contrast, the SF-36 scores for the MS sufferers are a mixed bag of high and low relative scores, though most are below those reported by people experiencing knee pain or depression, reflecting the chronic nature of the illness.

The estimation of utility scores was undertaken using SPSS. Each data set was subjected to estimation using the ten different predictive equations described in Appendix Two; Tables 2 to 5 report the average utility scores and other standard descriptives. From an initial visual appraisal of the averages it would appear that each predictive equation produces similar utility scores within each data set; although the Brazier *et al.* VAS values are considerably lower than the other predictions, however, this is similar to what Brazier *et al.* found when comparing their SG and VAS results. Aside from this, there are two predictive results that are quite striking; the very high values produced by the Nichol *et al.* HUI predictive equation and the very low values produced by the Bartman *et al.* HUI predictive equation.

Applying the estimating equation as proposed by Nichol *et al.* results in very large utility scores, each table reports maximum values greater than one, while the mean HUI utility score for depression sufferers is also larger than one (see Table 3). While some health states can be valued worst than death (<0) it is not possible or logical to have values greater than perfect health (1). An initial investigation into why this occurred found that the population cohort sampled by Nichol *et al.* had relatively low averages for the eight SF-36 subscales (in many instances lower than those reported in Table 1, suggesting their sample may not be as representative as first thought). Because, in this instance, the sample of knee pain, depression and MS sufferers have higher health profiles, it has lead to higher health utility values, some of which are greater than one. This obviously places considerable doubt on the reliability of this model to predict utility values should other researchers wish to use it.

In contrast, the predictive equation for HUI scores as described by Bartman *et al.* results in very low mean values, especially for knee pain and MS sufferers (see row 6, Tables 2 and 4). This is even more puzzling given that their rating scale approach produces values which are highly comparable to the rest of the reported predictions. One possible explanation may be that the transformation they applied to their HUI scores, in an attempt to rectify a non-normal distribution, is having some undue influence. Bartman *et al.* take the arcsin square root (see Appendix Two), which means that for replication it is necessary to impose the inverse (the sine squared), this transformation leads to small values since the numbers are (in theory) bounded between zero and one.

Shifting the focus towards the three explicitly predictive models, Brazier *et al.* (the revised version), Fryback *et al.* and Nichol *et al.* (despite the reservations discussed above) finds that, unsurprisingly, when comparing mean values (not reported) that they are all significantly different. Since concentrating solely on averages can disguise what is happening at the individual level, scatterplots comparing individual utility scores across predictive equations were produced, see Figures 1 to 12. These scatterplots show that the pairwise relationships are generally linear in nature, though most do not fit the 45° line of perfect correlation. Further graphical analysis of these three predictive approaches, by way of histograms, suggests there is no general consensus regarding normality (Figures 13 to 24). The estimated utility values for knee pain sufferers appear to be normal across the three predictive models, although this is probably helped by the large sample size. In general, the Brazier *et al.* predictions would appear to perform best in this regard.

It is important to admit that these analyses have been relatively simple and further detailed investigations are required to inform researchers which (if any) predictive equation should be employed. It would appear, however, from the results reported above that the mapping approach suggested by Brazier *et al.* would be sufficient if a researcher was required to undertake a CUA but only had available health profiles in the form of SF-36 data. Further support for choosing the Brazier *et al.* approach is given by the fact that it is being continually refined and is a major research project which includes collaboration with SF-36 development team, whereas the other attempts have been one-off pieces of research (aside from that of Longo *et al.*). However, one draw back of using their mappings is that it requires raw SF-36 profiles, which may not be available to secondary researchers, those attempts that make use of the SF-36 subscales would, therefore, be more practical. The model proposed by Fryback *et al.* would in such instances be appropriate (as would Nichol *et al.*, if it did not produce such illogical values) because, in addition to using the subscales, they also produce confidence intervals for researchers who only have average component-wise scores.

IV. DISCUSSION

This paper set out to investigate how health state preferences could be derived from SF-36 health profiles. However, as expected, while providing some insight on the current state of play, as this research progressed a number of other issues were highlighted, these are discussed below.

The first question that many may wish to ask is whether it is valid to undertake such a comparison in the first place. A point-of-fact often addressed to researchers who elicit utility from subjects using a number of approaches (standard gamble, time trade-off, and preference-based multi-attribute classification systems) and upon comparing the results find differences, is that they should not be surprised as each measurement involves a different valuation task and different attributes. Following this argument it is not surprising that the results reported in Tables 2 to 5 vary and that perfect correlation's are not shown in Figures 1 to 12. A corollary to this is provided by Drummond *et al.* (1997), "users of economic evaluation studies and preference-scored health status classification systems should be aware that all of these methods are in use. Users should check carefully to determine what method was used in studies or pre-scored instruments of interest to them, and to ensure that the method suits their purpose" (Drummond *et al.*, 1997, p.150). So just as a tradesman should choose the most appropriate tool, as should we. Therefore, the simple response to "how" perhaps should not be based on predictability and validity, but on an approach's resulting predicted preference score (standard gamble utilities, rating scale values or QWB/HUI/EQ-5D scores).

Further, it is interesting to note that in all of the approaches reviewed and replicated above that the focus was purely one of statistical correlation and maximum predictive power. There appears to be little regard for theory and no discussion of the resulting theoretical implications. This is supported by Meletiche *et al.* who argued that "[f]uture research should be focused on solving the theoretical puzzle governing this relationship" (Meletiche *et al.*, 1999, p.2024). Brazier *et al.* does make an attempt by mapping onto another classification system, but still "plays god" in order to get the best fit; bearing in mind, however, this "best fit" is for their data only, what about other data like those presented here? A similar argument can be made with respect to those researchers who include interaction terms and undertake stepwise regressions, methods whose sole aim is to find statistically significant variables, not necessarily theoretically appropriate relationships. Further to this argument, it does not appear that these researchers have considered whether the SF-36, its components and domains would be expected to be related to preference-based health states and, therefore, whether they should be included in an estimating equation, this point is highlighted by Lundberg *et al.* (1999). They estimate the relationship between health state utilities and the SF-12, but chose not to include the SF-12 question on overall health status as it "is in itself a single measure of overall health status, it is not appropriate to treat than term in the same way

as the other items" (Lundberg *et al.*, p.130). A similar argument could be made in relation to the appropriateness of including the SF-36 subscale summary score on general health perceptions, interestingly all the previous accepts, bar Brazier *et al.*, have included it in some form or other, see Appendix Two.

Also as eluded to earlier many of these relationships and estimations may be heterogeneous. Bult *et al.* (1998) argue that the reason reported correlation's between descriptive and valuation measures are low is because heterogeneity exits. They argue that mapping health states to preferences ignores important external influences which may not be reflected in generic health profiles. They test their hypothesis by comparing a simple (one class) regression to a latent variable analysis (four classes), the latter of which they find has greater explanatory power. Simply put, this supports the inclusion of socio-demographic variables (like age, gender and race) and disease specific variables in predictive equations. The former can easily be accommodated (see Nichol *et al.* and Bartman *et al.*), however, the latter implies that an off-the-shelf predictive equation would be inappropriate unless it had been validated within the specific health issue in question. What implications does this have for this field of research?

One final point for discussion is whether any of these approaches have been used and validated by other researchers. As described in the introduction this is the first attempt which compares all the approaches, although there is one published study which compares the results of Fryback *et al.*'s predictions with actual utility values. (Note there are undoubtedly a number of studies which use the Brazier *et al.* SF-6D but none that are currently published.) Gabreil *et al.* (1999), in analysing the differences between patient and community elicited preferences, compared directly measured utility (time trade-off and the rating scale) to those estimated using preference-classification systems (HUI Mark 2 and the QWB using Fryback *et al.*'s estimating equation). Interestingly, they found the estimated QWB scores were consistently lower that all the other elicited utility values, however, they did not pursue the issue further. This provides a lead into future work; perhaps the way ahead is not to compare each approach against all others, given that they use a variety of different dependent variables (utilities, values, scores etc), but instead to compare real elicited values with those derived from SF-36 data, that is compare actual numbers with predictions.

To conclude, it is poignant to quote Revicki and Kaplan (1993), who argue that "utility measures and psychometric health status measures are constructed to address different purposes" (Revicki and Kaplan, 1993, p.??). Then perhaps a more appropriate title for this paper is "Mapping SF-36 to utilities: *should* it be done?"

REFERENCES

Andresen EM, Rothenberg BM and Kaplan RM (1998). "Performance of a self-administered mailed version of quality of well-being (QWB-SA) questionnaire among older adults." *Medical Care*, **36**, 1349-1360.

Bartman BA, Rosen MJ, Bradham DD *et al.* (1998). Relationship between health status and utility measures in older claudicants." *Quality of Life Research*, 7, 67-73.

Chilvers C, Dewey M, Fielding K, *et al.* (2001). "Antidepressant drugs and generic counselling for treatment of major depression in primary care: randomised trial with patient preference arms." *British Medical Journal*, **322**, 722-725.

Brazier J, Usherwood T, Harper R and Thomas K (1998). "Deriving a preference-based single index from the UK SF-36 health survey." *Journal of Clinical Epidemiology*, **51**, 1115-1128.

Bult JR, Hunink MGM, Tsevat J and Weinstein MC (1998). "Heterogeneity in the relationship between the time tradeoff and Short Form-36 for HIV-infected and primary care patients." *Medical Care*, **36**, 523-532.

Doherty M, Jones A, Muir K, *et al.* (1999). "A community based randomised intervention study examining the effects of exercise on knee pain and its associated disability." Final Report to the Department of Health.

Drummond MF, O'Brien B, Stoddart GL and Torrance GW (1997). *Methods for the Economic Evaluation of Health Care Programmes* (2nd ed), Oxford University Press, New York.

Fryback DG, Lawrence WF, Martin PA, *et al.* (1997). "Predicting Quality of Well-being Scores from the SF-36: results from the Beaver Dam Health Outcomes Study." *Medical Decision Making*, **17**, 1-9.

Gabriel SE, Kneeland TS, Melton LJ, *et al.* (1999). "Health-related quality of life in economic evaluations for osteoporosis: whose values should we use?" *Medical Decision Making*, **19**,141-148.

Lincoln NB, Dent A, Harding J, *et al.* (2001). "Evaluation of cognitive assessment and cognitive intervention for people with multiple sclerosis." Working Paper, School of Psychology, University of Nottingham.

Longo MF (2001). Personal communication.

Longo MF, Cohen D, Hood K and Robling M (2000). "Deriving an 'enhanced' EuroQol from SF-36." Paper presented at HESG, Nottingham, July 2000.

Lungberg L, Johannesson M, Isacson DGL and Borgquist L (1999). "The relationship between health-state utilities and the SF-12 in a general population." *Medical Decision Making*, **19**, 128-140.

Meletiche DM, Doshi D and Lofland JH (1999). "Medical outcomes study Short Form 36: a possible source of utilities." *Clinical Therapeutics*, **21**, 2016-2026.

Nichol MB, Sengupta N and Globe DR (2001). "Evaluating quality-adjusted life years: estimating of the Health Utility Index from the SF-36." *Medical Decision Making*, **21**, 105-112.

Nicholl CR, Lincoln NB, Francis VM and Stephan TF (2001). **"**Assessing quality of life in people with Multiple Sclerosis." Working Paper, School of Psychology, University of Nottingham.

Revicki DA and Kaplan RM (1993). "Relationship between psychometric and utility-based approaches to the measurement of health-related quality of life." *Quality of Life Research*, **2**, 477-487.

Shmueli A (1997). "The SF-36 profile and health-related quality of life: an interpretative analysis." *Quality of Life Research*, **7**, 187-195.

Torrance GW, Boyle MH and Horwood SP (1982). "Application of multi attribute utility theory to measure social preferences for health states." *Operations Research*, **30**, 1043-1069.

ACKNOWLEDGEMENTS

I would like to thank Prof Nadina Lincoln, Mr Paul Miller and Mr Ben Palmer for providing the data used in this analysis and Prof John Brazier for supplying the necessary SPSS syntax and answering numerous queries.

Table 1: Mean scores for SF-36 subscales

Subscale	Knee Pain	Depression	MS*	MS**
Physical function	57.85	81.26	10.51	34.95
Physical role limitation	52.21	63.93	19.60	32.58
Emotional role limitation	70.09	59.93	51.89	59.70
Social function	54.65	72.86	51.14	53.73
Mental health	64.01	58.95	60.14	63.12
Energy/Vitality	55.55	46.69	32.50	33.73
Bodily Pain	45.11	69.94	53.66	59.98
Health perceptions	58.59	60.22	40.22	38.63
Sample size	565	183	88	201

* From Nicholl *et al.* (2001)
** From Lincoln *et al.* (2001)

Table 2: Descriptive statistics for predicted utility scores for knee pain sufferers (N=565)

Author and utility measure	Mean	Median	Standard Deviation	Minimum	Maximum
Brazier *et al.* (1998) VAS	0.4221	0.4140	0.1023	0.14	0.78
Brazier *et al.* (1998) Standard Gamble	0.8150	0.8100	0.0050	0.58	0.94
Brazier *et al.* (revised) Standard Gamble	0.6971	0.6970	0.0074	0.40	0.87
Fryback *et al.* (1997) QWB	0.6278	0.6343	0.0049	0.48	0.73
Nichol *et al.* (2001) HUI	0.9388	0.9430	0.1043	0.49	1.26
Bartman *et al.* (1998) HUI	0.0055	0.0030	0.0066	0.00	0.45
Bartman *et al.* (1998) Rating Scale	0.8647	0.8620	0.0083	0.65	1.12
Longo *et al.* (2000) EuroQol	0.6900	0.6878	0.0053	0.51	0.91
Shumueli (1998) Rating Scale (v.1)	0.8203	0.8264	0.0033	0.73	0.88
Shumueli (1998) Rating Scale (v.2)	0.7471	0.7467	0.0052	0.62	0.87
Shumueli (1998) Rating Scale (v.3)	0.7805	0.8106	0.0064	0.62	0.90

Table 3: Descriptive statistics for predicted utility scores for depression sufferers (N=183)

Author and utility measure	Mean	Median	Standard Deviation	Minimum	Maximum
Brazier *et al.* (1998) VAS	0.5373	0.5430	0.1820	0.12	0.92
Brazier *et al.* (1998) Standard Gamble	0.8649	0.8840	0.0092	0.50	1.00
Brazier *et al.* (revised) Standard Gamble	0.7318	0.7470	0.1365	0.35	1.00
Fryback *et al.* (1997) QWB	0.7102	0.7125	0.0070	0.53	0.83
Nichol *et al.* (2001) HUI	1.0839	1.1267	0.3487	0.13	1.67
Bartman *et al.* (1998) HUI	0.4980	0.5355	0.2918	0.00	0.97
Bartman *et al.* (1998) Rating Scale	0.8401	0.8681	0.2297	0.22	1.23
Longo *et al.* (2000) EuroQol	0.7947	0.8344	0.1715	0.24	1.02
Shumueli (1998) Rating Scale (v.1)	0.8490	0.8582	0.0081	0.66	0.99
Shumueli (1998) Rating Scale (v.2)	0.8385	0.8432	0.1091	0.60	1.00
Shumueli (1998) Rating Scale (v.3)	0.7959	0.8059	0.1218	0.55	1.00

Table 4: Descriptive statistics for predicted utility scores for MS* sufferers (N=88)

Author and utility measure	Mean	Median	Standard Deviation	Minimum	Maximum
Brazier *et al.* (1998) VAS	0.2804	0.2600	0.1176	0.12	0.64
Brazier *et al.* (1998) Standard Gamble	0.7527	0.7740	0.0084	0.46	0.90
Brazier *et al.* (revised) Standard Gamble	0.5590	0.5605	0.1244	0.26	0.90
Fryback *et al.* (1997) QWB	0.5365	0.5278	0.0035	0.48	0.70
Nichol *et al.* (2001) HUI	0.7763	0.7588	0.2930	0.10	1.40
Bartman *et al.* (1998) HUI	0.1056	0.0068	0.1084	0.00	0.49
Bartman *et al.* (1998) Rating Scale	0.6106	0.5862	0.2065	0.28	1.06
Longo *et al.* (2000) EuroQol	0.5697	0.5663	0.1709	0.19	0.88
Shumueli (1998) Rating Scale (v.1)	0.7638	0.7493	0.0057	0.65	0.90
Shumueli (1998) Rating Scale (v.2)	0.6731	0.6651	0.0050	0.60	0.83
Shumueli (1998) Rating Scale (v.3)	0.7510	0.7431	0.1064	0.56	0.98

* From Nicholl *et al.* (2001)

Table 5: Descriptive statistics for predicted utility scores for MS** sufferers (N=201)

Author and utility measure	Mean	Median	Standard Deviation	Minimum	Maximum
Brazier *et al.* (1998) VAS	0.3676	0.3620	0.1250	0.12	0.73
Brazier *et al.* (1998) Standard Gamble	0.7927	0.8080	0.0079	0.50	0.95
Brazier *et al.* (revised) Standard Gamble	0.6362	0.639	0.1361	0.28	0.96
Fryback *et al.* (1997) QWB	0.5918	0.5732	0.0070	0.49	0.83
Nichol *et al.* (2001) HUI	0.8848	0.8802	0.3200	0.09	1.59
Bartman *et al.* (1998) HUI	0.1958	0.1101	0.2189	0.00	0.98
Bartman *et al.* (1998) Rating Scale	0.6645	0.6600	0.1995	0.27	1.18
Longo *et al.* (2000) EuroQol	0.6302	0.6168	0.1684	0.22	0.98
Shumueli (1998) Rating Scale (v.1)	0.7880	0.7795	0.0068	0.65	0.96
Shumueli (1998) Rating Scale (v.2)	0.7220	0.7003	0.0083	0.61	1.00
Shumueli (1998) Rating Scale (v.3)	0.7722	0.7689	0.1165	0.55	1.00

** From Lincoln *et al.* (2001)

Figure 1: Scatterplot of predicted QWB scores versus Standard Gamble utilities for knee pain sufferers

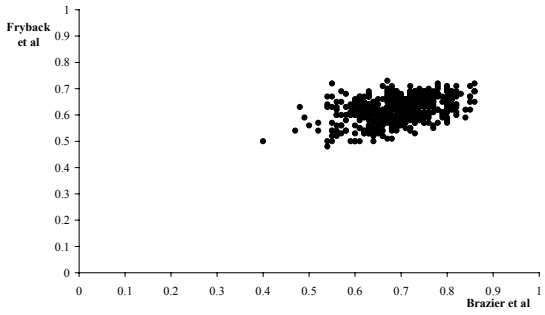


Figure 2: Scatterplot of predicted HUI scores versus QWB scores for knee pain sufferers

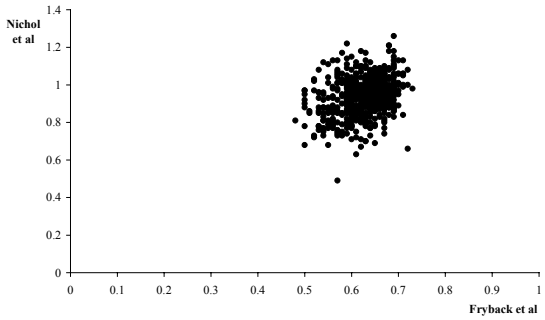


Figure 3: Scatterplot of predicted Standard Gamble utilities versus HUI scores for knee pain sufferers

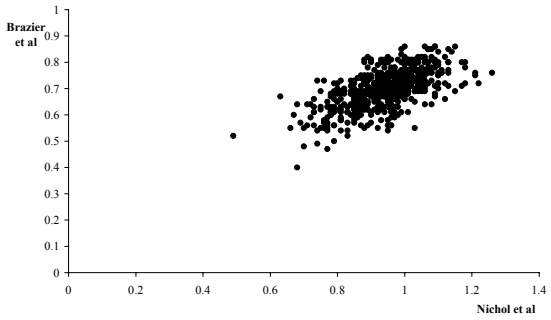


Figure 4: Scatterplot of predicted QWB scores versus Standard Gamble utilities for depression sufferers

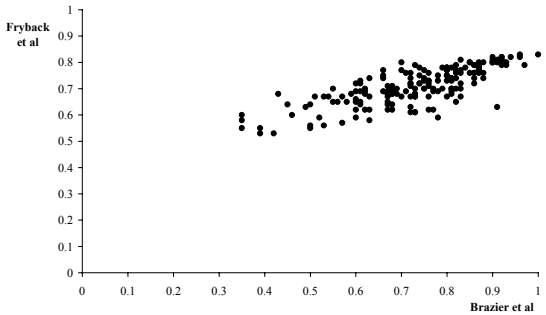


Figure 5: Scatterplot of predicted HUI scores versus QWB scores for depression sufferers

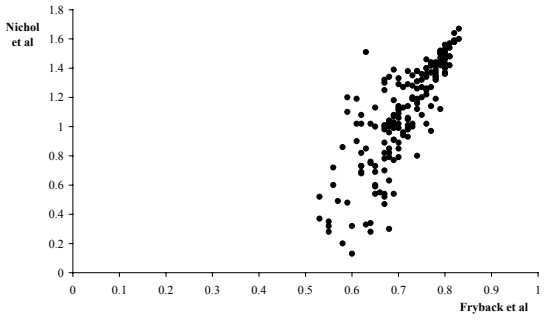


Figure 6: Scatterplot of predicted Standard Gamble utilities versus HUI scores for depression sufferers

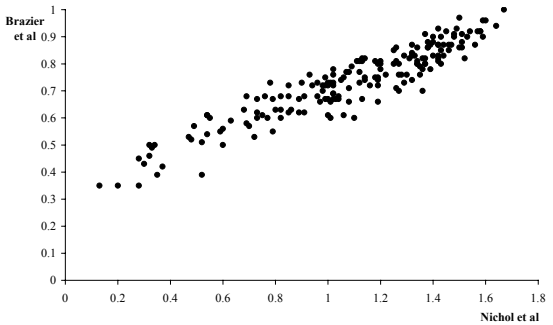


Figure 7: Scatterplot of predicted QWB scores versus Standard Gamble utilities for MS sufferers (Nicholl *et al.*, 2001)

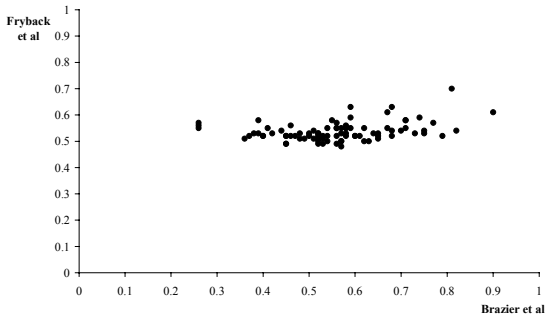


Figure 8: Scatterplot of predicted HUI scores versus QWB scores for MS sufferers (Nicholl *et al.*, 2001)

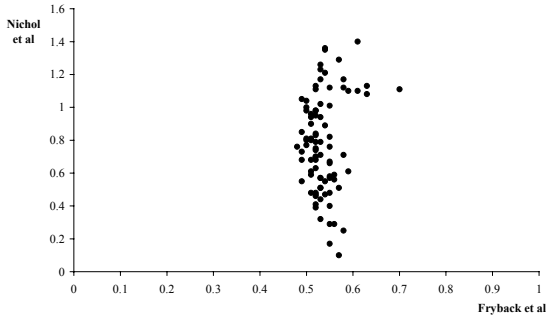


Figure 9: Scatterplot of predicted Standard Gamble utilities versus HUI scores for MS sufferers (Nicholl *et al.*, 2001)

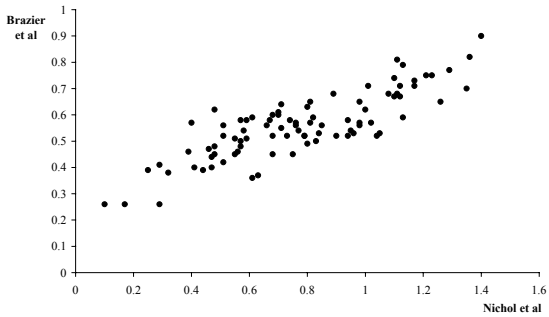


Figure 10: Scatterplot of predicted QWB scores versus Standard Gamble utilities for MS sufferers (Lincoln *et al.*, 2001)

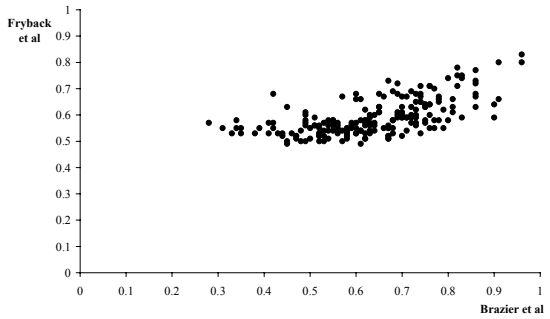


Figure 11: Scatterplot of predicted HUI scores versus QWB scores for MS sufferers (Lincoln *et al.*, 2001)

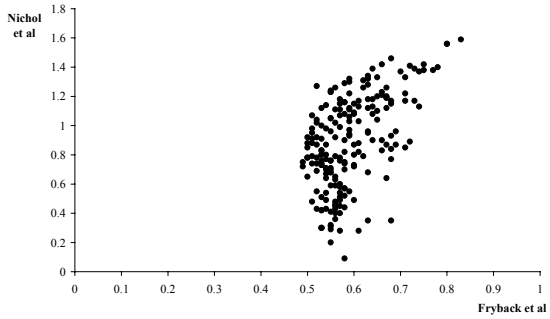


Figure 12: Scatterplot of predicted Standard Gamble utilities versus HUI scores for MS sufferers (Lincoln *et al.*, 2001)

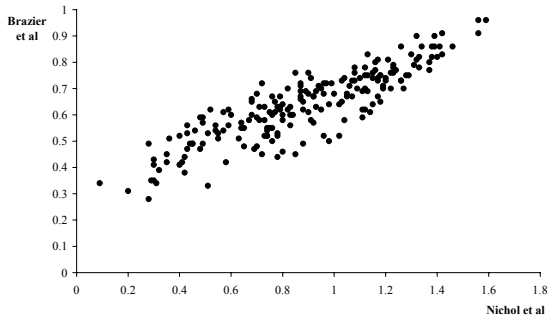


Figure 13: Histogram with normal curve for Brazier *et al.* standard gamble utility scores for knee pain sufferers

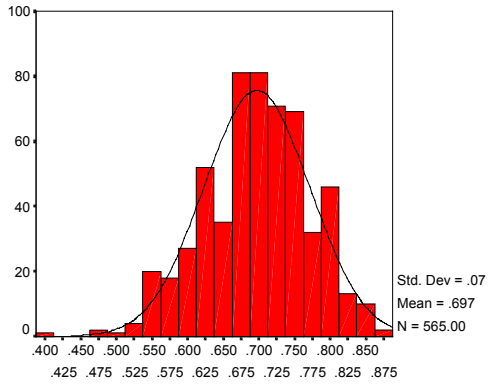


Figure 14: Histogram with normal curve for Fryback *et al.* QWB scores for knee pain sufferers

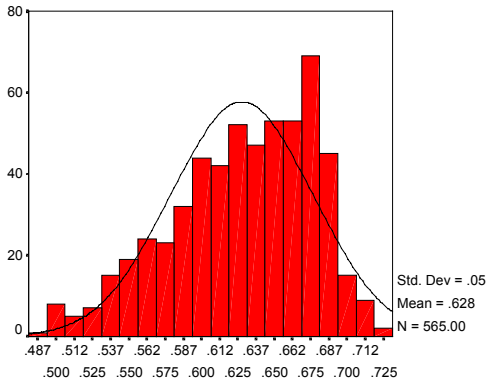


Figure 15: Histogram with normal curve for Nichol *et al.* HUI scores for knee pain sufferers

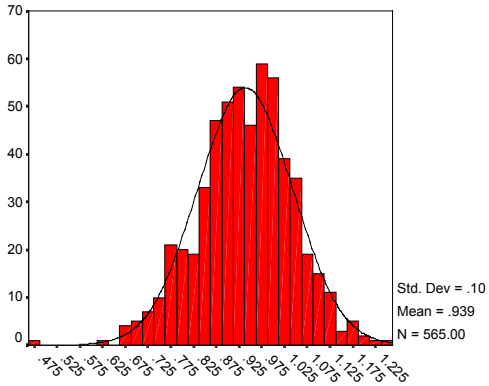


Figure 16: Histogram with normal curve for Brazier *et al.* standard gamble utility scores for depression sufferers

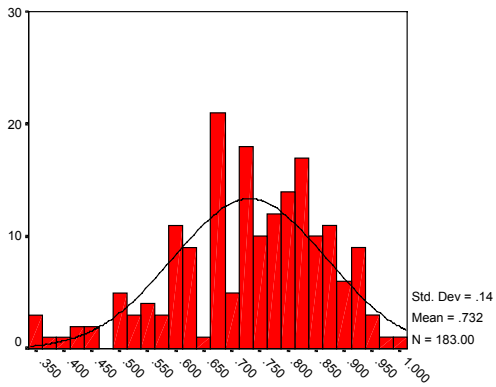


Figure 17: Histogram with normal curve for Fryback *et al.* QWB scores for depression sufferers

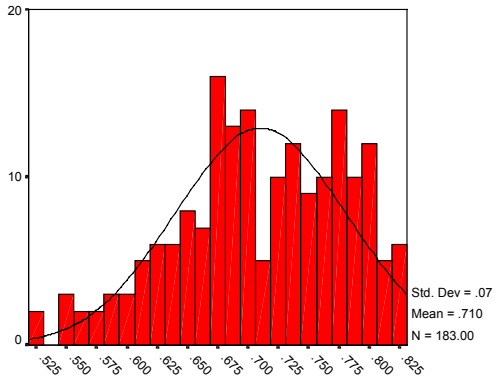


Figure 18: Histogram with normal curve for Nichol *et al.* HUI scores for depression sufferers

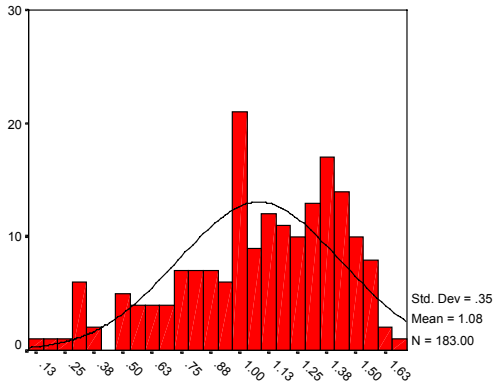


Figure 19: Histogram with normal curve for Brazier *et al.* standard gamble utility scores for MS sufferers (Nicholl *et al.*, 2001)

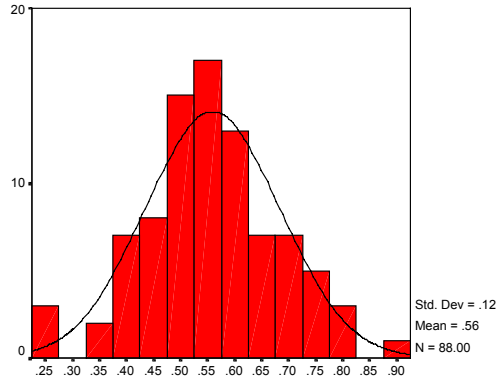


Figure 20: Histogram with normal curve for Fryback *et al.* QWB scores for MS sufferers (Nicholl *et al.*, 2001)

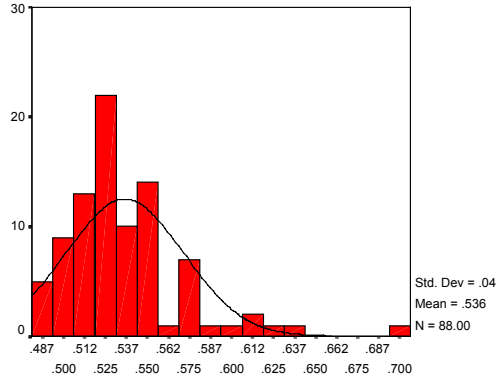


Figure 21: Histogram with normal curve for Nichol *et al.* HUI scores for MS sufferers (Nicholl *et al.*, 2001)

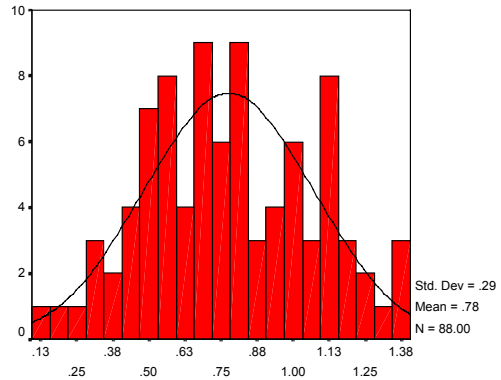


Figure 22: Histogram with normal curve for Brazier *et al.* standard gamble utility scores for MS sufferers (Lincoln *et al.*, 2001)

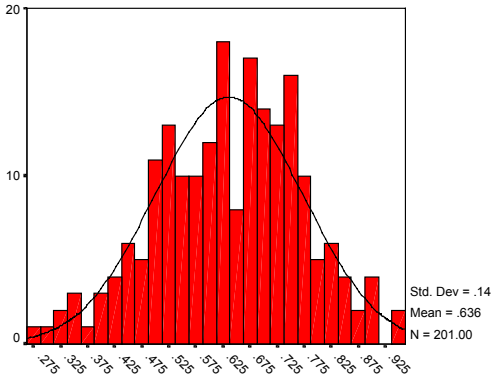


Figure 23: Histogram with normal curve for Fryback *et al.* QWB scores for MS sufferers (Lincoln *et al.*, 2001)

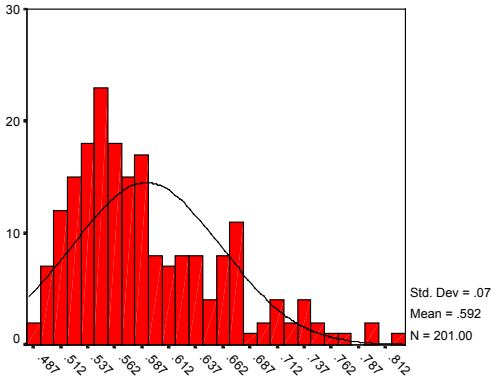
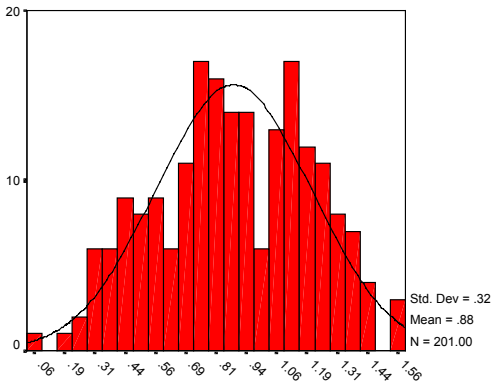


Figure 24: Histogram with normal curve for Nichol *et al.* HUI scores for MS sufferers (Lincoln *et al.*, 2001)



APPENDIX ONE

The following table summarises all known attempts to derive utilities from SF-36 data. It details the utility and SF-36 measures employed, the population cohort from which the data was gathered, as well as the predictive power of the estimating equation.

Author	Utility measure	SF-36 measure	R^2	Sample
Brazier *et al.* (1998)	Visual analogue scale (VAS) Standard gamble (SG)	SF-6D (revised SF-36)	0.680 0.495	Health professionals, health managers, staff and students of a medical school
Brazier *et al.* (revised)	Standard gamble (SG)	SF-6D (revised SF-36)	0.508	Representative sample of the UK general population
Fryback *et al.* (1997)	Quality of Wellbeing (QWB)	Individual summary scores	0.572	Random sample of adults older than 45 in Beaver Dam, Wisconsin
Nichol *et al.* (2001)	Health Utilities Index (HUI2)	Individual summary scores	0.500	Cohort of Southern California Kaiser Permanente members (HMO)
Bartman *et al.* (1998)	Health Utilties Index (HUI) Rating scale (RS)	Individual summary scores	0.530 0.590	Older patients with intermittent claudication
Longo *et al.* (2000)	EQ-5D (EuroQol)	Individual summary scores	0.677	Randomised control trial of women with breast problems
Shmueli (1998)	Rating scale (RS)	Overall summary score Physical health summary score Mental health summary score	0.367 0.317 0.258	Representative sample of urban Jewish Israelis

APPENDIX TWO

Below details each predictive equation employed to estimate utilities from SF-36 data. Note due the complexity of the SF-6D model, readers are asked to refer to Brazier *et al.* (1998).

Fryback *et al.* (1997):

$$QWB = 0.59196 + 0.0012588PhysFn - 0.0011709MentHlth - 0.0014261Pain + 0.00000705(GenHlth \times RolePhys) + 0.00001140(PhysFn \times Pain) + 0.00001931(MentHlth \times Pain)$$

Nichol *et al.* (2001):

$$HUI2 = 0.045 + 0.0009GenHlth + 0.00046RolePhys + 0.0043Pain + 0.0042MentHlth + 0.0018PhysFn + 0.0018Vital + 0.0015RoleEmot + 0.0015SocFn - 0.006Age$$

Bartman *et al.* (1998):

$$\arcsin(\sqrt{HUI}) = 0.006PhysFn + 0.004Pain + 0.005SocFn - 0.002RoleEmot + 0.005MentHlth - 0.014Age$$

$$RS/100 = 0.202PhysFn + 0.316Pain + 0.115GenHlth + 0.449Vital + 0.468Age$$

Longo *et al.* (2000):

$$\ln(2 - EuroQol) = 0.607 - 0.00204Pain - 0.0087GenHlth - 0.00113PhysFn - 0.00126SocFn - 0.00109Vital$$

Shmueli (1998):

$$\ln(1 - RS) = -1.021 + 0.814\ln(1 - TotalSF/100) \quad \text{(v.1)}$$

$$\ln(1 - RS) = -0.906 + 0.626\ln(1 - TotalPhys/100) \quad \text{(v.2)}$$

$$\ln(1 - RS) = -0.792 + 0.714\ln(1 - TotalMent/100) \quad \text{(v.3)}$$

Where
PhysFn = summary score of the physical function domain
RolePhys = summary score of the physical role limitation domain
RoleEmot = summary score of the mental/emotional role limitation domain
SocFn = summary score of the social function domain
MentHlth = summary score of the mental health domain
Vital = summary score of the vitality/energy domain
Pain = summary score of the bodily pain domain
GenHlth = summary score of the general health perceptions domain
TotalSF = overall summary score for all domains of the SF-36
TotalPhys = physical health summary score (from *PhysFn*, *RolePhys* and *Pain*)
TotalMent = mental health summary score (from *MentHlth*, *RoleEmot* and *SocFn*)