QORU quality and outcomes of person-centred care policy research unit

sepho South East Public Health Observatory

Report on using the GPPS to assess trends in EQ-5D scores for people with long-term conditions

July 2013

SEPHO
Andrew Hughes
Kevin Watson

QORU
Jose-Luis Fernandez
Julien Forder
Ray Fitzpatrick

Executive summary

Background

Estimating the extent to which NHS services are contributing to improving the health-related quality of life (HRQoL) of people with long-term conditions is an important (if challenging) objective. Its importance is reflected in domain 2 of the NHS Outcomes Framework. Understanding whether this goal is being achieved requires methods which help the interpretation of the role of services on observed trends in HRQoL. Controlling for the influence of external factors, such as the severity of the underlying condition – or 'need' – on quality of life, is particularly crucial because NHS and care activity levels increase with need-related factors (NRFs), but otherwise NRFs are strongly associated with worse HRQoL. Failing to control for NRFs makes it therefore very difficult to interpret observed changes in quality of life, and in particular to appraise the role that NHS and care services might play in improving the well-being of people with long-term conditions.

This report aims to develop a methodology which is easy to implement and which standardises for changes in NRFs when investigating changes through time in the HRQoL of people with long-term conditions.

Analysis data

The analysis used data about the 1,037,946 participants in the 2011/12 GP Patient Survey (GPPS) survey. We randomly split the survey dataset in two halves. Group 0 formed the basis for all modelling work, and Group 1 was reserved for testing and predictive work. There were 518,808 participants in Group 0, of whom 62.5% reported a long-term condition (after those whose did not answer were removed).

In line with the NHS Outcomes Framework, we used the EQ-5D™ as the overarching indicator of HRQoL. EQ-5D is a standardised instrument which provides an overall measure of a person's health status across five domains: mobility; self-care; usual activities; pain/discomfort; and anxiety/depression.

The following indicators available in GPPS were chosen as possible predictors of quality of life: age, gender, ethnic group, smoking, the geographical classification of deprivation status (IMD), self-reported working status, each of 15 named long-term conditions, and a count of the number of long-term conditions reported by the individual.

Relationship between need-related factors and health-related quality of life

Using regression analysis, we first examined the relationship between NRFs and HRQoL among people with long-term conditions to assess (1) the extent to which a significant proportion of the variability in EQ-5D can be controlled for with the indicators in GPPS; and (2) which NRF indicators are most important in doing so. Separate analyses were carried out for each of the five health dimensions within the EQ-5D.

Two main sets of regression models were developed:

- A set of 'full' models, using all relevant indicators of NRFs in GPPS.
- A set of 'simple' models including only those indicators with the strongest explanatory power, to test the impact of using a small number of indicators on the capacity of the model to predict EQ-5D levels. The predictors in the 'simple' models included the number of long-term conditions present (capped at five); age groups (seven categories); the presence of permanently sick or disabled from the working status variable; IMD, with two categories (the most affluent two-thirds of the population and the most deprived one-third).

As expected given the large size of the sample, the majority of the relationships tested in the 'full' and 'simple' models were found to be statistically significant. Also as expected, given the greater number of indicators they contained, the set of 'full' models predicted better EQ-5D than the set of 'simple' models.

A subset of models examining people with specific diseases and in particular age groups were not found to improve the results of the analysis.

Replicating and predicting EQ-5D scores

We tested three prediction methods for the 'full' and 'simple' models to determine how closely the predicted distributions of EQ-5D scores matched the actual scores.

The results of the 'simple' regression model from group 0 were used to predict EQ-5D scores in Group 0 and in Group 1 (test sample). The 'full' models were more accurate in predicting individual level scores, but the 'simple' models produced a closer mean value to the actual observed means. In both cases, the mean errors suggested that individual EQ-5D score predictions were limited in their accuracy.

Overall, the use of the dimension level regression models for predicting EQ-5D scores had limited success. Although the replication of the distribution of actual EQ-5D scores was reasonably good, the predicted individual and mean scores were not accurate. These findings are in line with previous work by Ara et al[1] whereby using regression models for case-mix adjustments can be problematic.

Derivation of a weighted health-related quality of life index

Given the limited success of the regression-based method, we sought an alternative approach to standardise the NRFs of people with long-term conditions in GPPS. We developed a needs-weighted health-related quality of life index using a set of weights which reflected the prevalence in GPPS of the key factors linked to variations in EQ-5D.

Based on the population weights thus derived, subsequent waves of GPPS could be reweighted to give people with different characteristics the same relative importance as in the base year when estimating (weighted) average EQ-5D scores.

By allowing average HRQoL levels to vary for the combination of characteristics represented in each cell, the weighted index approach is more flexible than the regression method in capturing the interaction effects between predictors of EQ-5D. However, the derivation of weights limited the number of factors that could be taken into account in the reweighting process because of the effect of including further indicators on the number of cases in each cell, and thus on the reliability of the weights.

We thus defined an index based on 172 cells grouping individuals with long-term conditions in terms of the number of LTCs, age bands, IMD binary variable, a 'permanently sick or disabled' working status, and presence of musculoskeletal conditions.

We tested the sensitivity of the index to changes in demographic patterns and disease prevalence by constructing artificial modelling scenarios. Overall, we found that the index successfully controlled for up to 20% increases in disease prevalence in most conditions and for increases in mean age of up to five years.

We developed additional scenarios to test the capacity of the index to detect changes in the underlying health status of the population. The index was found to be sensitive to very small changes in EQ-5D levels.

Estimates of the weighted index in Group 0 and Group 1 were found to be very similar. The analysis could not examine issues such as seasonality effects across GPPS waves.

Conclusion

The overall results of the study suggest that a weighted index methodology could be used to control for changes in NRFs while remaining sensitive to true changes in the HRQoL of people with long-term conditions.

However, the use of the index and in particular its interpretation can be challenging. Further work needs to be carried out to help with the interpretation of what can be very small differences in average HRQoL between GPPS waves, to examine what constitutes a policy-meaningful improvement or deterioration.

From a different perspective, the fact that even significant changes in the HRQoL of given population groups (for instance, people with a given condition) result in very small average changes in the index suggests the importance of carrying out sub-group analyses by condition. These sub-group analyses would be particularly useful from the point of view of understanding why changes in average HRQoL might occur and of identifying key areas for policy intervention.

Whereas most of the variations in NRFs through time in the GPPS might be due to a combination of sampling differences in the GPPS and changes not directly the result of NHS services (such as for instance changes in individuals' behaviour), controlling for variations in NRFs through time precludes the analysis from evaluating the impact of services on the actual prevalence of long-term conditions. As a result, this weighted index methodology is unable to capture the impact that NHS investments might have in terms of reducing the prevalence and/or severity of long-term conditions (and therefore to capture improvements in the quality of life of the people that otherwise would have developed such conditions). Assessing different strategies for doing so should be a priority objective for further research in this area. As it stands, this weighted index approach could be used as part of a portfolio of indicators to assess the performance of the NHS with regard to chronic illness.

Table of contents

Table of figures

Tables in report

Introduction

The NHS Outcomes Framework identifies as domain 2: 'enhancing the quality of life for people with long term conditions'. It will use EQ-5D as the overarching indicator of Health Related Quality of Life (HRQoL). It is important to anticipate methodological issues that may arise in the interpretation of observed trends over time in EQ-5D, and to agree as far as possible the principles and methods that should be used to interpret trends.

The analysis included in this report contributes to the development of methodologies for understanding the role of the NHS in improving through time in the health-related quality of life (HRQoL) of people with long-term conditions. The report develops a methodology which is easy to implement and which ***standardises for differences in need-related factors*** when investigating ***changes through time*** in the HRQoL of people with long-term conditions. By need-related factors (NRFs) we understand the range of characteristics (e.g. demographic, socio-economic and health-related which might be associated with the use of NHS services). In other words, the analysis develops methods which capture changes in HRQoL through time other than those brought about by changes in case-mix in the datasets used for the analysis (for example, because of sampling problems, or changes in the prevalence of long-term conditions in society). By controlling for case-mix through time, the proposed methods do not account for possible impact of the care system on HRQoL through prevention effects resulting for instance in changes in the prevalence of long-term conditions in the population. Instead the focus is on the impact of the care system on people who already have long-term conditions.

Due to limitations in the available evidence and the estimation challenges involved, the analysis does not attempt to estimate explicitly the relationship between particular NHS services and improvements in the HRQoL of people with long-term conditions, or to decompose levels of HRQoL between the relative contributions of services and of NRFs. A further report considers the methodological requirements and the range of improvements in the existing evidence necessary to link reliably actual NHS services with variations in HRQoL among people with long-term conditions.

Evaluating the contributions of the NHS to health-related quality of life for people with long-term conditions

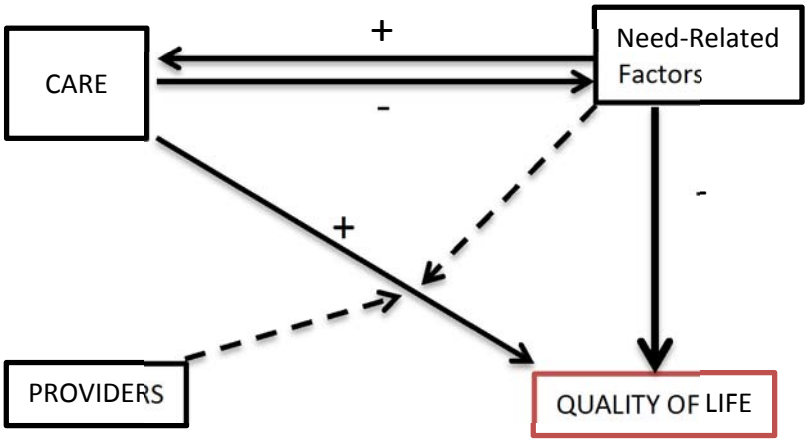
Figure 1 summarises the key relationships involved in the 'production' of care outcomes (e.g. improvements in HRQoL) for people with long-term conditions. In the figure, we hypothesise that the HRQoL of people with long-term conditions will vary with their NRFs and other external factors, and with the nature and intensity of support from NHS services. In addition to the direct effects illustrated by the continuous lines, Figure 1 represents with dotted lines the mediating effect that factors such as the quality of providers and patients' NRFs will also have on the relationship between care services and HRQoL. Importantly, Figure 1 reflects the endogenous nature of the relationship between care and NRFs. Hence, whereas we can expect a positive relationship between services and NRFs because more services tend to be provided to people with more intense NRFs, a negative relationship through time between care and NRFs could also arise if the provision of services reduces the intensity of NRFs by preventing further deterioration in particular conditions or the development of new conditions.

Ideally, estimating the relationship between NHS services and HRQoL requires information about each of the factors included in Figure 1. Controlling for the influence of NRFs and other external factors on HRQoL is particularly important because NHS activity levels are highly correlated with NRFs, which themselves are very significantly negatively correlated with HRQoL. Failing to control for NRFs and other external factors in the analysis would therefore make it very difficult to interpret observed changes in HRQoL, and in particular to appraise the role that care services might have played in improving them. However, as intimated in Figure 1, changes through time in NRFs are themselves likely to be affected by services. Whether and how NRFs are controlled for in the analysis affects the

interpretation of the changes in HRQoL observed. In particular, fully controlling for changes in NRFs when estimating changes through time in HRQoL for people with long-term conditions means that the analysis would not be able to account for improvements in HRQoL that would arise from possible preventative effects of services.

Disentangling the endogenous relationship between NRFs and services is particularly challenging because of the complex and longitudinal nature of the effects involved (for instance, preventative effects are likely to take a number of years to have full effect), and the significant data requirements, in particular about the levels and patterns of targeting of care resources over time. In the analysis below, we therefore concentrate on developing a method which controls for all changes in NRFs through time without attempting to disentangle those changes in NRFs which are and are not the product of NHS services. The methods developed do not attempt to account for the impact of preventative effects on the HRQoL of people with long-term conditions. Instead, they should be interpreted as providing a methodology for understanding changes in the HRQoL of patients controlling as far as possible for all the key NRFs driving variations in HRQoL for people with long-term conditions.

Figure 1: The 'production' of care outcomes



As is further discussed below, the analysis in this report is based on the GP Patient Survey (GPPS) because of its unique characteristics in terms of its size and the availability of indicators of HRQoL and NRFs for people with long-term conditions. Relative to Figure 1, however, the GPPS does not include direct indicators of service use. This means that analyses based on the GPPS cannot aim to disentangle the relative contributions of services and NRFs to variations in the HRQoL of people with long-term conditions.

Evidence from the GPPS can be used to control for changes in the case-mix of the population of people with long-term conditions to explore changes through time in HRQoL for such population, but NRFs to be taken into account in its interpretation. In particular, an important implication of the lack of service receipt indicators in the GPPS is that estimates of the correlation between NRFs (e.g. long-term conditions) and HRQoL will represent the combination of the negative effect of the NRFs and the (hopefully) beneficial effect of the treatment provided by the NHS. The coefficients of health conditions in hypothetical regression equations predicting HRQoL should not therefore be interpreted as illustrating purely the shift in HRQoL associated with the presence of the condition, but rather as the net effect of services and NRFs.

This is not to say that the GPPS does not constitute a useful source of evidence for assessing the contribution of the NHS to the HRQoL of people with long-term conditions. By examining successive waves of the GPPS whilst controlling for changes in the case-mix of the population with long-term conditions, it is theoretically possible to examine whether the HRQoL of people with long-term conditions improves ***relative to*** the HRQoL of people ***with the same characteristics in the past***. Assuming the analysis is able to control effectively for NRFs, and that the influence of non-policy factors such as changes through time in the behaviour of patients is relatively small, improvements or deteriorations year-on-year in the 'NRFs-standardised' HRQoL of people with long-term conditions could be interpreted as resulting from improvements or deteriorations in NHS services. In the analyses that follow, we explore the potential to develop such a standardised HRQoL indicator. Using regression analysis, we first examine the

relationship between NRFs and variations in EQ-5D to assess (1) the extent to which a significant proportion of the variability in EQ-5D can be controlled for with the indicators available in GPPS; and (2) which NRFs indicators are most important in doing so. The regression models were created for each of the five health dimensions within the EQ-5D separately using a generalised ordered logistic approach.

We then use the evidence produced to group people with long-term conditions into 'cells' based on the strongest predictors of EQ-5D, in order to develop a NRFs-weighted quality of life index to standardise the NRFs profile of successive GPPS waves and compare average quality of life for people with long-term conditions through time. Finally, we test the performance of the weighted index by checking its capacity to control for modelled changes in the case mix of people with long-term conditions in the GPPS, and discuss the limitations and caveats of the proposed methods.

Analysis data: the GP Patient Survey

The analyses in this report are based on the GP Patient Survey, which represents by far the largest source of evidence in England on HRQoL and NRFs among people with long-term conditions. The GPPS was designed to give patients the opportunity to comment on their experience of their GP practice. It includes questions about how easy or difficult it is for patients to make an appointment at their surgery, satisfaction with opening hours, and quality of care. The Department of Health has carried out the GPPS since 2007.

The 2011/12 survey uses a quantitative postal methodology with questionnaires sent to almost 2.8 million patients across two waves, from July – September 2011 and January – March 2012. This questionnaire was the first time that the EQ-5D measure was included as a question.

A total of 8,271 practices were sent questionnaires over the course of the year. At least one completed questionnaire was received from patients in 8,258 practices. Patients were eligible if they were aged 18 and over, had a valid NHS number, had been registered with the practice for at least six months and had not received a GPPS questionnaire in the previous 12 months. Sample selection was proportionately stratified by age and gender, and unclustered within each practice. Adjustments to the sample size were made depending upon eligibility to receive the questionnaire.

We were provided by IPSOS Mori (who carry out the survey on behalf of the Department of Health) with data about the total participants of the 2011/12 survey. We randomly split the survey dataset in two, called Group 0 and Group 1. Group 0 forms the basis for all modelling work, and Group 1 was reserved for later testing of any assumptions or predictive work. Sophisticated stratification techniques for splitting the data were not needed as the dataset was large enough to be similar in each group by chance after a random split (see Table 14 and Table 15). Total participants in year 6 of the survey numbered 1,037,946. After implementing the random split there were 518,808 participants in Group 0. All analyses were undertaken on Group 0. Of this group, 62.5% reported a long-term condition (after those whose did not answer were removed).

The GPPS has a number of variables that can be used as possible predictors of well-being scores. For the purposes of the current analysis, discrete dispositional and lifestyle variables were chosen: age, gender, ethnic group, smoking, as well as the geographical classification of deprivation status (IMD). In addition to these relatively stable variables, we included self-report of working status and each of the 15 named long-term conditions, due to the clear relationship with both of these factors and health status.[3] As well as the type of long-term condition reported, we included a count of the number of any long-term conditions reported by the individual. Initial descriptive analysis also looked at people without long-term conditions to observe the whole distribution of EQ-5D scores within the survey, but all further analysis only used individuals reporting a long-term condition.

Mobility categories		Presence of diabetes categories	
-3	Not answered	-3	Not answered

-1	Multi-coded	-1	Multi-coded
1	I have no problems in walking about	0	No diabetes
2	I have some problems in walking about	1	Diabetes
3	I am confined to bed		

Table 1: Breakdown of how responses were coded

All missing data within the survey were recorded as such; 'not answered' and 'multicoded' were re-coded as 'missing'. Technically, the only conflicting variables were where the individual had responded that they had no long-term conditions under 'do you have a long-term condition?' but reported having a specific long-term condition. Another variable had already been produced in the dataset to describe all people who had either replied 'yes' to 'do you have a long-term condition?' or had reported having a specific long-term condition. In summary, the data appeared to have already been cleaned well.

Figure 2 to Figure 6 show the distribution, by age group, of the ratings of the population for the five dimensions of quality of life in EQ-5D.

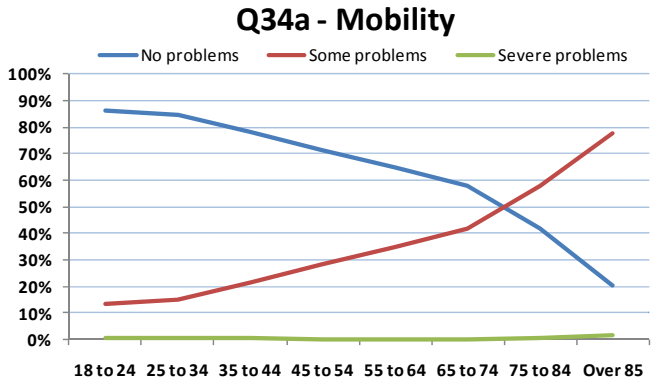


Figure 2: Percentage of age group giving each dimension response

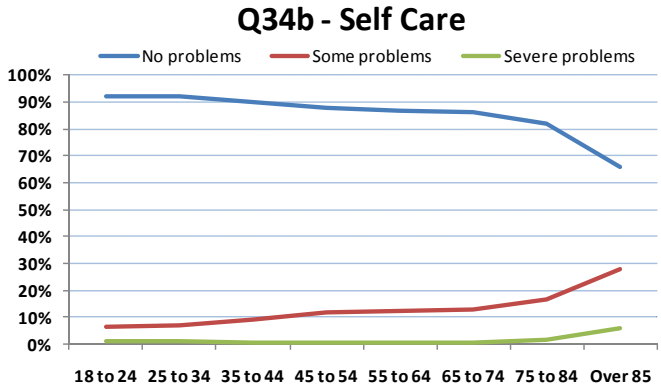


Figure 3: Percentage of age group giving each dimension response

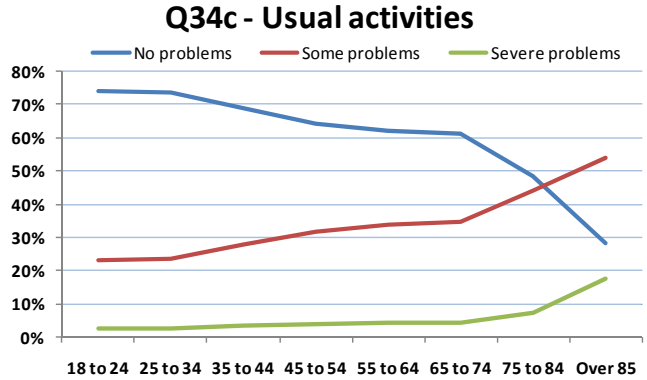


Figure 4: Percentage of age group giving each dimension response

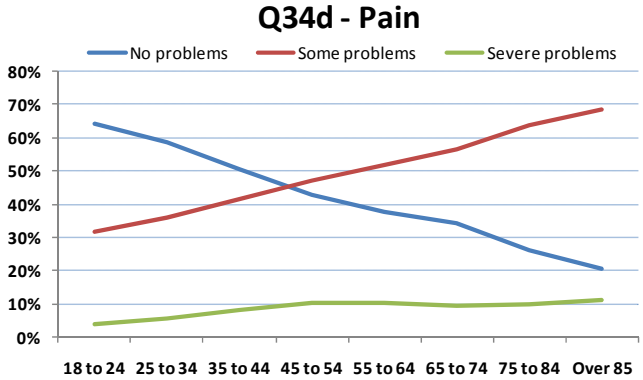


Figure 5: Percentage of age group giving each dimension response

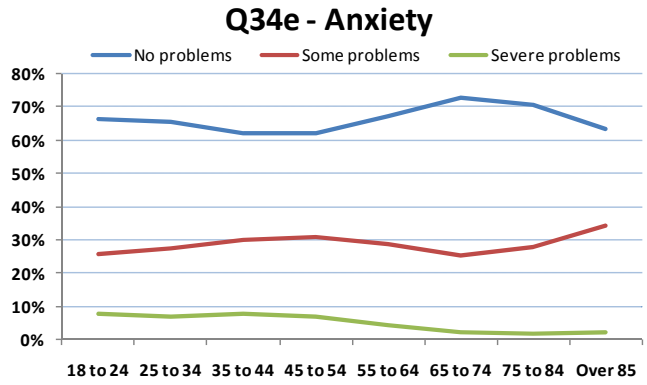


Figure 6: Percentage of age group giving each dimension response

Drivers of quality of life among people with long-term conditions

Appendix 1 provides a summary of key descriptive statistics for the indicators of NRFs used in the analysis, and a basic bivariate analysis of their relationship with EQ-5D.

As stated above, the first stage of analysis was to fit regression models to explore the correlation between NRFs and variations in EQ-5D among people with long-term conditions. Using Year 6 of the GP Patient Survey, we identified individuals reporting long-term conditions and used the EQ-5D as dependent variable with the core NRF indicators in GPPS: 15 long-term conditions, gender, age, ethnicity and deprivation as independent variables. In addition, working status and limiting long-standing illness were included in the model as proxies for health status. Different model specifications were explored to account for the distributional characteristics of EQ-5D (see Appendix 3). Given evidence of different relationships between long-term conditions and physical compared to psychosocial aspects of quality of life, separate analyses were performed of the five dimensions that make up the overall EQ-5D score.[2] A study undertaken by the Policy Unit in Economic Evaluation in Health and Care Interventions in 2011 on long-term conditions and the EQ-5D with the Health Survey for England[1] suggested that further work could be undertaken researching significant predictors to each of the five health dimensions of the EQ-5D. The work in this report includes the results of models exploring the impact of available predictors within the GPPS. Further work explored the impact of matched area-level socio-economic indicators, including indicators of deprivation, demographic patterns, mortality, and local authority typology (ONS classification) but could not identify a significant effect for other than the deprivation indicator.

As further described below, three sets of regressions models were developed: a full model, including all relevant NRFs-related predictors in GPPS; a simple set of regression models including those indicators with the strongest explanatory power; and a subset of models of specific diseases and age, to determine whether model fit could be improved from the full model within sub-categories.

The ‘full’ model

A first model specification was fitted for each of the five EQ-5D dimensions, including all the core NRF variables in GPPS, as noted above. This specification is referred to as the ‘full model’.

In line with our initial assumptions that different relationships exist between specific long-term conditions and physical compared to psychosocial aspects of quality of life, five models were specified for each of the five health dimensions within the EQ-5D. As the score for each dimension was recorded as 1, 2 or 3 (Level 1: indicating no problem, Level 2: indicating some problems, Level 3: indicating extreme problems), ordered logistic regression was adopted – see Box 1. Previous work by Ara et al[1] also undertook a two-part model, where the probability of scoring

full health (score of 1) was undertaken in conjunction with health dimension models. We discounted this method due to Ara et al reporting that the two-part models did not add to the predictive power.

Initial exploratory analysis of all the selected independent variables using logistic regression was carried for each of the five EQ-5D dimensions of health: mobility, self-care, usual activities, pain/discomfit and anxiety/depression. Logistic regression was performed between the first and second category in the EQ-5D dimensions and compared with the relationships between the first and third. For example, the difference in odds ratios of people reporting 'I have no problems in walking about' and 'I have some problems in walking about' against Alzheimer's were compared with odds ratios of people reporting 'I have no problems in walking about' and 'I am confined to bed' against Alzheimer's. It was found that the odds ratio did not always increase by the same proportion between each response category.

Box 1. Ordered logistic model

The *ologit* command in STATA applies an ordered logit model under the proportional odds assumption - that the relationship between any two pairs of outcome groups is statistically no different. Effectively, this means that the coefficients for 'no problems' versus the response 'some problems' on each dimension are the same as those that describe the relationship between 'some problems' the worst outcome. We used a user-written command *omodel*[4] in STATA (but not an official STATA command) that tests the proportional odds assumption after looking at the initial exploratory analysis. The test result was significant, indicating that the model may break the proportional odds assumption. This suggested that a generalised ordered logistic regression may be more appropriate, where the assumptions are less restrictive than the proportional odds model, but take into account the ordering of the categories, unlike a multinomial method.[5] There are no in-built STATA commands that use this method, so a user-created command was downloaded called 'gologit 2'[6] that met requirements. This model was adopted for all regression models. This can be written as:

$$P(Yi > j) = g(X\beta j) = \frac{\exp(\alpha j + Xi\beta j)}{1 + \{\exp(\alpha j + Xi\beta j)\}}, \quad j = 2, \ldots, M-1$$

where M is the number of categories of the ordinal dependent variable. From the above, it can be determined that the probabilities that Y will take on each of the values 1, . . . , M are equal to

$$P(Yi = 1) = 1 - g(Xi\beta 1)$$
$$P(Yi = j) = g(Xi\beta j{-}1) - g(Xi\beta j)\ j = 2, \ldots, M-1$$
$$P(Yi = M) = g(Xi\beta M{-}1)$$

Not surprisingly given the size of the sample, the majority of coefficients in the full model for each of the dimensions tested were statistically significant at $p<0.01$. Of the disease groups, diabetes and kidney/liver disease were not significant for mobility and cancer was not significant for the usual activities dimension. Unfortunately, the generalised model outputs two coefficients per variable (with three possible responses), so interpreting the table is difficult. The complete coefficients (expressed as log odds) of the full model can be seen in Table 16.

The goodness of fit of the regression models was assessed using the output from STATA generalised ordered logit model commands using McFaddens Pseudo R^2 (see Table 3 and Table 4). While pseudo R-squareds cannot be interpreted independently or compared across datasets, they are valid and useful in evaluating multiple models predicting the same outcome on the same dataset.

Subset disease-specific regression models

Specific diseases and age bands were analysed separately to test whether or not a better 'fit' could be obtained by using disease specific or age-specific models, using the remaining variables within the full model. The disease specific models used were:

- Diabetes

- Cancer
- Mental health
- Arthritis & back problems

An age-specific model was also constructed by using only those people aged 65 years or older and removing age band from the underlying model.

Disease
Alzheimer's disease or dementia
Angina or long-term heart problem
Arthritis or long-term joint problem
Asthma or long-term chest problem
Blindness or severe visual impairment
Cancer in the last five years
Deafness or severe hearing impairment
Diabetes
Epilepsy
High blood pressure
Kidney or liver disease
Learning difficulty
Long-term back problem
Long-term mental health problem
Long-term neurological problem
Another long-term condition

Table 2: Subset disease grouping by long-term condition

For the disease specific models, grouping of the diseases was undertaken to increase the power of the model. The justification for grouping was based on simple condition similarity and can be seen in Table 2. The fit of the sub-group analysis, including diabetes, cancer, mental health, arthritis & back problems and 'older people – over 65' can be seen in Table 3.

Dimension model	'Full' model	Diabetes	Cancer	Mental health	Arthritis & back problems	Over 65 only
Mobility	0.2947	0.3139	0.2818	0.2747	0.1942	0.2055
Self-care	0.2873	0.2725	0.2343	0.2210	0.2153	0.1983
Usual activities	0.2321	0.2597	0.1950	0.1873	0.1486	0.1806
Pain	0.2550	0.2563	0.2109	0.2295	0.1436	0.2282
Anxiety	0.1493	0.1432	0.0975	0.0957	0.1391	0.0874

Table 3: McFadden's Pseudo R^2 for subcategory regressions

Only the sub-group diabetes model appeared to have a slightly better fit when analysed separately, with other disease groups losing some fit. This is likely due to the greater power of the dataset with more people included. When regression analysis of the remaining people with long-term conditions without diabetes was carried out, the overall fit was slightly worse than the 'full model' – results are displayed in Table 4.

	No diabetes
Mobility	0.2885
Self-care	0.2852
Usual activities	0.2254
Pain	0.2531
Anxiety	0.1508

Table 4: McFadden's Pseudo R^2 for the group excluding diabetes

‘Simple’ model indicators

A ‘simple model’ was developed to test the impact on the explanatory power of the model of reducing the number of covariates to a level that would allow the specification of meaningfully large groups of cases (including 50 or more respondents) in the GGPS using key drivers of HRQoL. These groups formed the basis of the weighted HRQoL index introduced in later sections. The objective of the specification of the ‘simple’ model was therefore to achieve the best fit and acceptable predictions of EQ-5D while creating the smallest possible number of unique ‘cells’ or ‘categories’ for each of the five EQ-5D dimensions. The choice of variables in the simple version of the model was driven by the estimated strength of the effects identified, in terms of their statistical significance but more importantly given the very large size of the sample in terms of their contribution to the explanatory power of the model.

The selection of variables for the simple model was iterative, using the indicators with the largest effects from the full model. The final categorical variables included in the model specification were:

- The ***number of long-term conditions*** present (capped at five) – five categories: The number of long-term conditions reported by respondents was capped at five, as only 1% of people with long-term conditions had more than five conditions from the list.
- ***Age groups*** – seven categories: as shown in figures 2 through 6, ‘aged 18 – 24’ had a similar EQ-5D profiles to ‘aged 25 – 34’. Age groups were therefore reduced to seven categories, which reduced the number of cells defined on the basis of the indicators in the simple model by one eighth.
- The presence of ***permanently sick or disabled*** from the working status variable – two categories
- ***IMD*** – two categories: the main IMD indicator was found to be a strong predictor but, as an ordinal categorical broken down by decile rank, it would have added too many categories to the index. If split into a binary variable, it predicted well without having a negative impact on the cell count or cell populations. The decision to use a binary IMD variable was made shortly after the initial tests to determine the strongest contenders for the simple model. A decision was taken to use a split between the most affluent two thirds of the population in one half of the IMD binary variable and the most deprived third of the population in the other, based on the IMD rank included with survey dataset.

All indicators included in the simple model were found to be statistically significant at the 0.001 level in each EQ-5D dimension. A comparison of the McFaddens Pseudo R2 for the Full and Simple model can be seen in Table 5. In all cases the Full model outperforms the simple model, as expected. The increase in fit is a similar change in all dimensions, except pain.

	Mobility	Self-Care	Usual Activities	Pain and discomfort	Anxiety & depression
‘Full’ model	0.2947	0.2873	0.2321	0.2550	0.1493
‘Simple model’	0.2324	0.2350	0.1748	0.1347	0.0855

Table 5 McFaddens Pseudo R2 for the full model and simple models

Predicting EQ-5D scores

One of the goals of the regression analysis was to predict a single EQ-5D score to compare against actual scores in the survey dataset. The five-dimension specific models produced can be used to predict probabilities for each of the three possible outcomes using a simple predict command in STATA. Unfortunately, creating a single EQ-5D index score from these combined probabilities is not straightforward, and there are different ways to create an index. Although we are interested in comparing an average EQ-5D score for the dataset against the predicted average EQ-5D score, a key goal was to replicate the observed distribution of EQ-5D scores. We therefore tested three prediction methods for both the full model and the simple model to determine how closely the distribution of EQ-5D scores

matched the actual scores, whilst comparing the mean errors and mean absolute errors as an indication of the individual level predictions.

Method 1 applied five individual random numbers to each respondent, ranging between 0 and 1 – one random number to each of the EQ-5D dimensions. If, for example, the probabilities of answering 'no problems', 'some problems' and 'severe problems' were 0.7, 0.2 and 0.1 respectively and the random number was lower than 0.7, this record would be assigned a response of 'no problems'; if the random number was between 0.7 and 0.9, then a response of 'some problems' would be recorded; otherwise a response of 'severe problems' would be recorded. The dimension predicted probabilities were applied to each random number.

Method 2 used a single random number and applied this to each of the dimension probabilities, essentially making the random number constant across dimensions. This method assumes that the choice of outcome in one dimension is dependent to scores in the other dimensions. This increased the chances of scoring 'full health' but also increased the number in very poor health, simply because if the random number happened to be particularly high (e.g. above 0.99 which it will be in 1 case in 100) then poor health is likely to be recorded across all five dimensions.

Method 3 simply applied a dimension score based on the highest predicted probability of each of the three possible answers. This is logically the best tool to predict an individual score. However, if everyone in the group had between 50% and 100% chance of an occurrence, with an average of 80%, a good model should predict that 80% would experience the occurrence whereas this method would predict 100%. Also, since the calculation of EQ-5D is set for the limited number of combinations (140 cells) the number of possible scores are limited to 140 (and likely much lower- in this example, there are just 18 different outputs) rather than the 243 possible scores further limiting the models ability to produce the desired distribution.

As well as the three methods described above, we produced an alternative weighted average method where the 15 probabilities (of giving each of the three responses for each of the five dimensions) were combined to produce 243 variables containing the probability of each individual giving each of the 243 possible responses. The method was produced by multiplying the probability of each of the 243 responses by the scores for the respective combination. All of the 243 products of these calculations were summed to create a score for each individual. It was expected that this method would not replicate the distribution of scores seen in the actual scores.

The weighted average method of predicting scores from the regression model produces a unimodal distribution and has no individuals in full health – the results can be seen in Figure 7. This is because there will always be at least a small probability that an individual will not respond with 'no problems' in at least one of the dimensions, no matter how healthy an individual. Figure 8 shows the distribution of predicted scores for full model using Method 1 against the actual EQ-5D scores for Group 0. Method 1 predicts across the whole range of scores, and replicates the observed distribution, but still has some error in the actual scores. Methods 2 and 3 also predict across the range of actual scores, but do not replicate the actual shape of the distribution. Table 6 shows how the weighted average method compares with the three methods where the distribution of individual predicted scores covers the full range of actual scores. Excluding the weighted average method, Method 1 appears to have the best distribution, closest mean and smallest error scores. The weighted average method does have a better mean and smaller individual errors, which suggests that this method has potential as a method of predicting an individual's score, but it does not produce a distribution which is as representative of the actual scores. For this reason, Method 1 was used and applied in all further predictions.

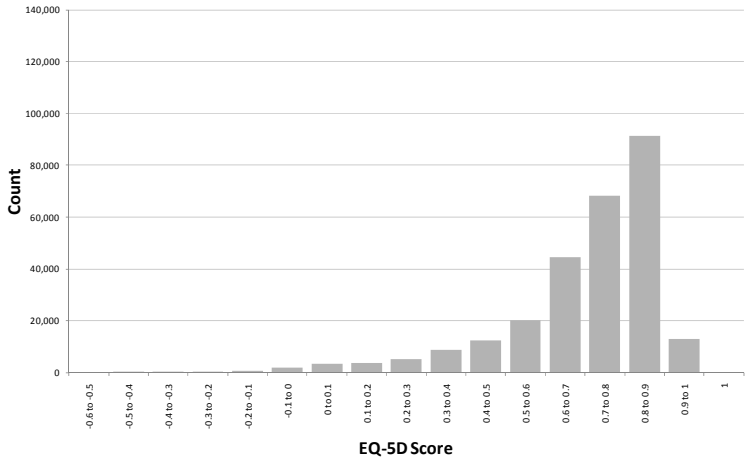


Figure 7: Weighted average prediction using the full model for Group 0

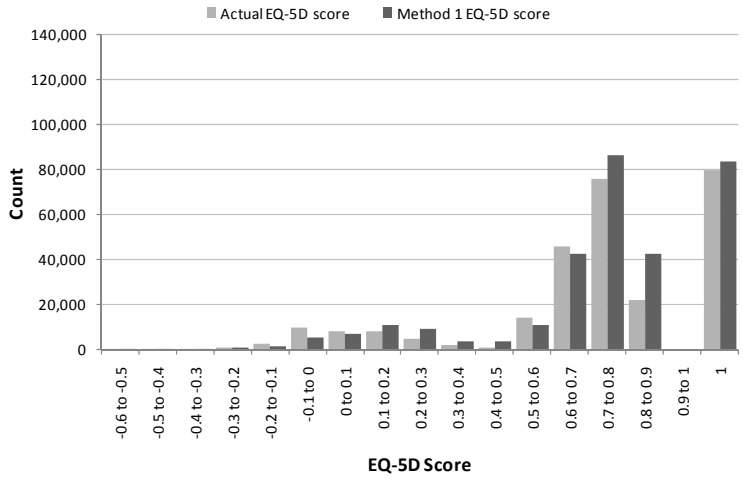


Figure 8: Actual EQ-5D scores and predicted scores using method 1 of the full model for Group 0

	Mean	Mean error	Mean absolute error	Root mean squared error
Observed	0.717			
Simple model, weighted average	0.698	0.022	0.172	0.234
Full model, weighted average	0.701	0.018	0.161	0.218
Simple model, method 1	0.724	-0.010	0.232	0.328
Simple model, method 2	0.758	-0.044	0.259	0.376
Simple model, method 3	0.678	0.012	0.262	0.448
Full model, method 1	0.735	0.019	0.214	0.312
Full model, method 2	0.761	0.045	0.235	0.354
Full model, method 3	0.652	0.035	0.260	0.465

Table 6: Mean error scores of four proposed prediction methods

Predictions with the simple model

The results of the simple regression model from group 0 were used to predict EQ-5D scores in Group 0 and in Group 1 (test sample). The observed mean EQ-5D score of people with long-term conditions, along with the mean error, mean absolute error and the root mean squared error is shown in Table 7. Predicted scores were calculated using Method 1. It is noticeable that the Full model is more accurate in predicting individual level scores, but interestingly the simple model produces a closer mean value to the actual observed means. In both cases, however, the mean errors suggested that individual EQ-5D score predictions were limited in their accuracy, especially compared to using the weighted average prediction methods in Table 6.

The results show almost the exact same means and error terms for predictions created for Group 0 and Group 1, based on the Group 0 model. This consistency suggests that the model applies very well to an independent cross-section of people answering the GPPS within the same time period. So although the individual-level predictions are limited, they are consistent in their output. This finding is important from the point of view of the final objective of the analysis, to find a method for comparing HRQoL indicators through time for people with long-term conditions.

	Mean	Mean error	Mean absolute error	Root mean squared error
Observed	0.717			
Simple model for Group 0	0.724	-0.010	0.231	0.328
Simple model for Group 1	0.724	-0.010	0.232	0.329
Full model for Group 0	0.735	0.019	0.214	0.312
Full model for Group 1	0.735	0.019	0.214	0.312

Table 7: Observed and predicted mean EQ-5D scores with error values

A comparison of the predicted scores between Group 0 and Group 1 are shown in Figure 9 for the simple model and in Figure 10 for the full model.

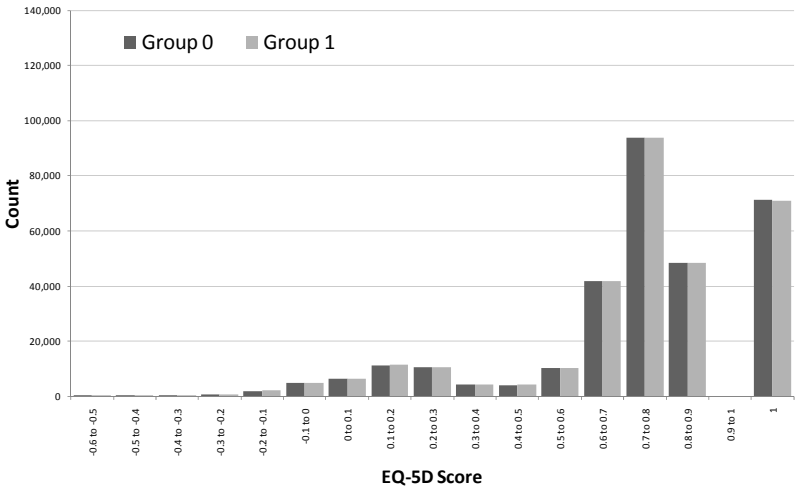


Figure 9 Distribution of the simple model predicted EQ-5D Scores for Group 0 and Group 1

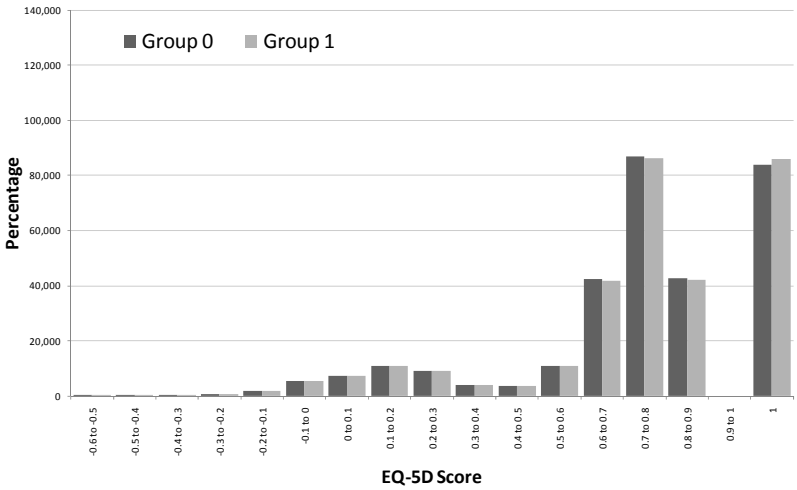


Figure 10: Distribution of the full model predicted EQ-5D Scores for Group 0 and Group 1

Figure 11 and Figure 12 show the correlations between the mean actual scores of the 140 unique cell groups of the simple model against the predicted scores using the simple model. Figure 11 shows the correlation with predicted scores in Group 0, whilst Figure 12 shows the predicted scores for the test sample in Group 1 (using the Group 0 model). The figures show very little difference in fit.

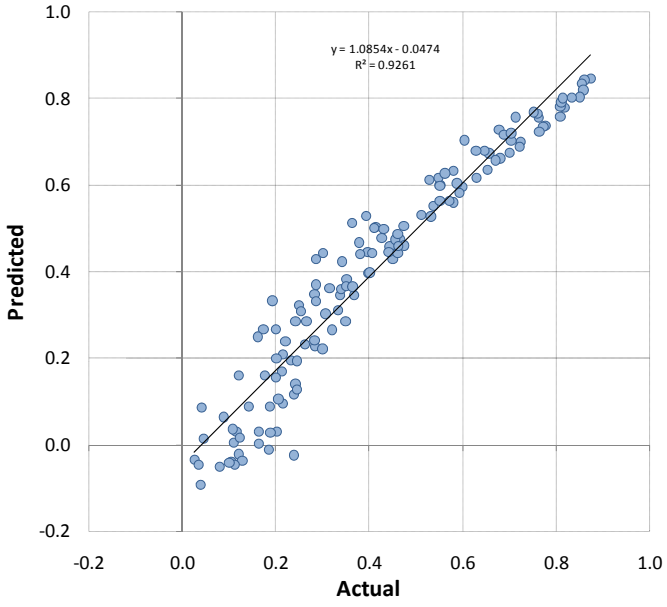


Figure 11: Mean scores of predicted and actual using Group 0

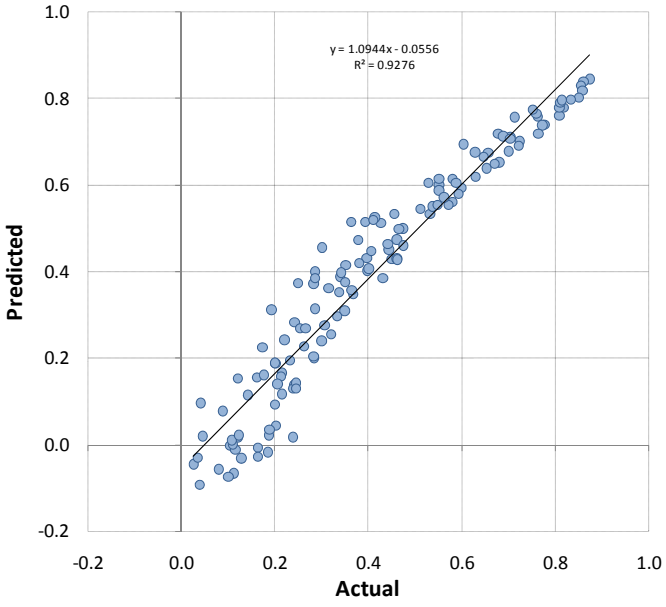


Figure 12: Mean scores of predicted and actual using Group 1

Controlling for case-mix in GPPS

The regression analyses described above identified the indicators available in GPPS most strongly associated with variability in EQ-5D among people with long-term conditions. Building on such results, we considered two methods for controlling for differences in case-mix between successive GPPS waves.

The first method would use the coefficients of the regression models to carry out adjustment analyses to estimate changes in EQ-5D in successive GPPS waves. Using these regression-based methods, changes in the contribution of the NHS to the HRQoL of people with long-term conditions could be tested by looking for differences in the parameters of identical models estimated in subsequent years, or by testing for significant differences in model residuals on the basis of coefficients estimated in precedent GPPS waves.

In general, regression-based methods are attractive because of the additional information conveyed by the parameters estimated, in this case in terms of the quantification of the relative effects of the predictors of EQ-5D. However, as noted in the introduction, the lack of information on service receipt in the GPPS meant that regression equations could not be used in order to decompose variations in EQ-5D between the contributions of services and those of NRFs. Furthermore, and in contrast with re-weighting methods, standardising for case-mix on the basis of regression models requires assumptions to be made about the distributional characteristics of the error term of the model, and about the functional form and nature of the interrelationship between predictors and dependent variable. Significant problems with such assumptions could lead to biases in the estimated coefficients and possibly in the estimates of changes in EQ-5D.

The second (and preferred) approach involved constructing a standardised EQ-5D index for the population of individuals with long-term conditions in GPPS. This index used a set of weights which reflected the composition of the GPPS on a given 'reference' year in terms of the key factors linked to variations in EQ-5D (the factors identified in the 'simple' model). Based on the population weights thus derived, subsequent waves of GPPS would be reweighted to give each cell (combination of NRFs) the same relative importance in determining the overall weighted-average EQ-5D score in subsequent years as in the base year. This method therefore removes the effect of any change in NRFs in the calculation of the overall (weighted) EQ-5D score. The idea is that any observed difference between overall weighted EQ-5D scores between years must come from changes in the mean EQ-5D score within the cell, and that this would reflect changes in other factors, including the contribution of NHS activity.

Arguably, the derivation of such weights is more transparent and easily implemented than the estimation and application of multivariate regression methods (particularly taking into account the non-normality of the EQ-5D distribution). Also, by allowing average score levels to vary for the combination of characteristics represented in each cell, the weighted index approach could be more flexible than regression methods in allowing for interaction effects between predictors of EQ-5D. On the negative side, the derivation of weights limits the number of factors that can be taken into account in the reweighting process because of the effect that including further indicators has on the number of cases in each cell, and thus on the reliability of the weights.

Derivation of the weighted HRQoL index

Given the relative benefits outlined above, we adopted and tested the reliability of the weighted index approach. The goal of the simple model was to predict the EQ-5D score as closely as possible, but using a small number of categorical variables so that the individuals could be categorised into cells using the unique combinations of each variable. Each cell represented a unique combination of the individual respondents and was created from a combination of independent variables in the model.

For example, if only gender and smoking status were used in the model, eight cells would be produced:

- males who had never smoked;
- males who are former smokers;
- males who are occasional smokers;
- males who are regular smokers;
- females who had never smoked;
- females who are former smokers;
- females who are occasional smokers;
- females who are regular smokers.

The mean actual EQ-5D score of the total number of unique cells would then be used to produce a weighted HRQoL index. This index is weighted by the population but, to ensure the robustness of the individual cell means, we set the minimum number of individuals in each cell to be 50. This becomes difficult as the number of cells exceeds 100. Therefore, a figure of 150 cells was used as a guide for the creation of the index using indicators in the simple model. Using all current categories of the variables present in the full model, excluding gender and grouping IMD into deciles, would define over 1.4 billion categories or cells. With a sample of fewer than 1 million, almost all of these would be empty, and even populated cells would contain a very small number of people. The aim of the 'simple' model was to show goodness of fit and prediction close to the full model in order to evidence the use of a smaller number of variables in the creation of a weighted HRQoL index. To achieve a reasonable minimum number of cases per cell, we grouped independent variables into fewer categories where possible, as the number of cells would increase dramatically after introducing more than a few variables. A model with three binary variables has eight cells, add one more binary variable and this becomes 16, add age (which has seven categories within it) and this becomes 112.

In the first instance, we defined an index based on 140 cells grouping individuals with long-term conditions in terms of the number of LTCs (capped at five), age bands (with the bottom two merged), IMD binary variable (most deprived third by rank /others) and a 'permanently sick or disabled' working status. An initial review of the performance of the index when introducing large changes in prevalence found that the index did not control sufficiently well for large changes in the prevalence of musculoskeletal conditions. The index was therefore modified to account for these effects by splitting cells by presence of musculoskeletal conditions where there was no recording of permanently sick or disabled, fewer than five diseases, and aged 35-84. The final index was thus built on the basis of 172 cells. The methods of increasing cell sizes to account for the musculoskeletal effects are provided in Appendix 2.

Table 8 shows the constructed index (EQ-5D mean score and Index value) for Group 0 baseline, as well as a baseline for the total sample (Group 0 and Group 1), using the 172-cell population split. When using Group 0 sample as a baseline index, the change was very small when comparing the index value in the Group 1 sample. There is in general agreement between the two groups when comparing the predictions of the regressions models. This does suggest that the index is reasonably stable when comparing across a similar cross-section of the population in a similar time period.

The weighted HRQoL index was built using Group 0 population as the baseline and measured against Group 1 to test the stability of the index. Further index sensitivity testing was carried out using the total survey sample of people with long-term conditions combined from Group 0 and Group 1.

The construction of bootstrapped confidence intervals revealed extremely narrow confidence limits. The mean EQ-5D score of people with long-term conditions was 0.7232; the upper and lower bootstrap confidence intervals were 0.7224 and 0.7241 respectively.

	Index (Mean score and Index value)	
Group 0 baseline	0.7237	100.00
Change to Group 1	0.7230	99.90
Baseline: All people with a recording of a long-term condition (Total sample of Group 0 and Group 1 combined)	0.7232 (0.7224, 0.7241)	100.00

Table 8: Change in the sample demographics and index scores

Index sensitivity testing

The aim of the weighted HRQoL index is to remain stable in the face of changes in demographics and in the prevalence of other predictive variables in the populations over time. The sensitivity of the index to changes in population of cells should show how well it will do this. If the mean score within a cell remains stable with fluctuations in disease prevalence and age, this means that the choice of cells works well. The index should also be sensitive to shifts in mean HRQoL scores with particular health groups, and therefore reflect changes in the EQ-5D scores of the respondents within the cells. This is likely to occur due to some form of health or other intervention. One of the ways of looking at any change in patient scores either cross-sectionally or longitudinally is to determine a threshold value in those scores that are meaningful for the respondent (which reflect some form of intervention). These changes are often referred to as minimally important differences (MID).

Two different scenario methods were applied to the underlying population in order to determine the sensitivity of the overall index to changes in the case-mix of the population. These were:

1. Demographic scenarios
2. Intervention scenarios

Confidence intervals using bootstrapping methods were put around the index mean to determine a statistical measurement with which the scenarios could be judged.

Demographic and co-morbidity scenarios

The prevalence of diabetes in England is predicted to have a relative increase of 11.2% by 2020[7] from the current year. The sample dataset was adjusted to reflect an increase of 5%, 10% and 20% for diabetes, but also cancer and the grouped diseases of musculoskeletal and mental health to observe how this affected the index. This was also done by age group, and by the number of diseases present. We only explored increases in prevalence as it is assumed that diseases will increase in an ageing population and, while it is possible that some diseases may plateau, they are unlikely to decrease dramatically.

This analysis was performed in two different ways. The first method was a sampling with replacement, drawing the new sample randomly from the current dataset that matched the expected increase stratum (e.g. people with diabetes) to replace a random sample outside of this stratum. All of the disease changes, age effects and increase in permanently sick and disabled group used this method. The only ‘scenario’ which was not produced this way was the ‘Decrease in permanently sick or disabled by 20%’. This was done by changing the classification of a proportion of people indicated as permanently sick or disabled to indicate that they were not permanently sick or disabled, without altering their well-being scores. It was done this way to replicate possible effects of a policy change rather than a physical or perceived physical change in individual respondents.

The number of diseases was increased by 5% by calculating 5% of the people with only one disease, and duplicating that number in the people who have two diseases. This process was repeated for the number of people with two, three and four diseases but could not be for the people with five or more diseases as they would still have five or more diseases. The sample size was then adjusted back down to the original size by dropping the appropriate number of records where the number of diseases was one.

The age bands were adjusted using an iterative process to reflect the possible age increase in the underlying population, constraining the total number of people within the dataset to the same number in the original sample. This effectively increased those people in the 75-84 and 85+ age categories and reduced the number in the younger age groups, as shown in Figure 13.

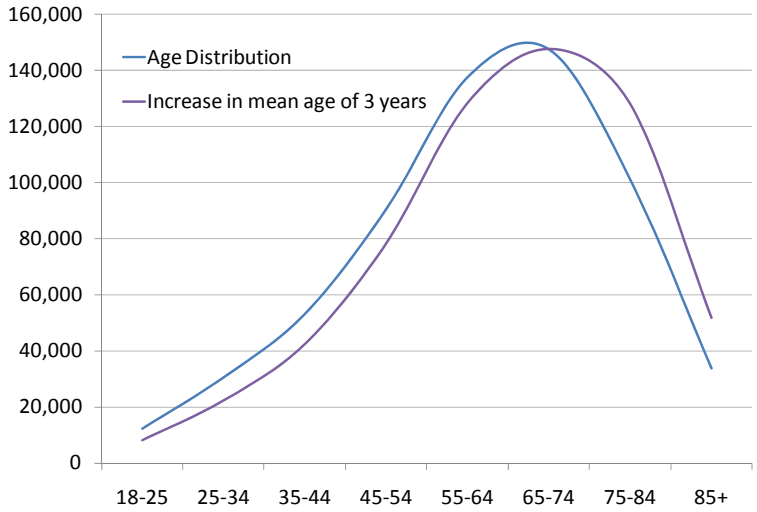


Figure 13: Age distribution before and after the sample was manually altered to reflect a sample with a mean age which is 3 years older

Care must be taken with these artificial increases because a group of people with a specific LTC in a sample of people who all have LTCs may have a better than average well-being score. As a result, increasing the prevalence of certain conditions could lead to an actual improvement in average HRQoL in the sample as a whole. The aim of this exercise was to ensure the index can handle demographic changes in the sample. For example, an increase in diabetics of 10% to show how the index will change by 2020 is done by adding duplicates randomly from the sample, and these people have a level of co-morbidities which is representative of now, not 2020.

The results of the demographic scenarios change can be seen in Table 9. Very large changes (20%) in the prevalence of musculoskeletal conditions and mental health conditions have a small but statistically significant effect on the index using the 95% confidence intervals, decreasing the index by more than 0.25 from the baseline 100. However, all other changes in the index from the baseline position were not significant. The decrease in permanently sick and disabled people by 20% led to a significant change and the largest decrease in the index from the baseline position. A 20% decrease in those who classify themselves as sick or disabled would only be caused by a dramatic change in policy, and a policy change would be likely to have an impact on the more healthy people within this group compared to this scenario where it was produced using a random sample.

Furthermore, it is important to note the very different nature of the scenario modelling decreases in the sick or disabled scenario, which rather than resampling cases changes the classification of existing cases from being rated as sick and/or disabled to not being so. As a result, the average HRQoL scores for people who are not sick or disabled are worsened, a fact that should be reflected in average index even after the reweighting is applied. So these results suggest that the index is on the one hand reasonably robust to changes in the composition of the sample, but sensitive to changes in the HRQoL within cells (i.e. controlling for NRFs).

	Index (Mean score and Index value)	
Baseline (total survey sample)	0.7232 (0.7224, 0.7241)	100.00
Increase in mean age of one year	0.7232	99.99
Increase in mean age of two years	0.7232	99.99
Increase in mean age of three years	0.7231	99.99
Increase in mean age of four years	0.7233	100.00
Increase in mean age of five years	0.7231	99.99
Increase in diabetes by 5%	0.7234	100.03
Increase in diabetes by 10%	0.7236	100.05
Increase in diabetes by 20%	0.7240	100.11
Increase in cancer by 5%	0.7233	100.01
Increase in cancer by 10%	0.7233	100.02
Increase in cancer by 20%	0.7235	100.04
Increase in musculoskeletal issues by 5%	0.7227	99.93
Increase in musculoskeletal issues by 10%	0.7220	99.84
Increase in musculoskeletal issues by 20%	0.7210*	*99.69*
Increase in mental health issues by 5%	0.7227	99.93
Increase in mental health issues by 10%	0.7222	99.85
Increase in mental health issues by 20%	0.7213*	*99.74*
Decrease in permanently sick or disabled by 20%	0.7174*	*99.19*
Increase in permanently sick or disabled by 5%	0.7232	100.00
Increase in permanently sick or disabled by 10%	0.7232	100.00
Increase in permanently sick or disabled by 20%	0.7231	99.98
Increase in the number of diseases by 5%	0.7232	100.00
Increase in the number of diseases by 10%	0.7233	100.01
Increase in the number of diseases by 20%	0.7231	99.99

Table 9: Change in the sample demographics and index scores

Intervention scenarios – using minimally important differences

Interpreting any changes in EQ-5D scores is challenging. Demonstrating responsiveness of a Patient Reported Outcome Measure (PROM) instrument to change is generally undertaken by longitudinal studies over time and can be calculated using statistical techniques. However, demonstrating the minimal important difference (MID) or clinically meaningful difference for any change in PROM outcome or score requires external data not inherent in the dataset itself. The MID is the smallest possible change in the PROM score that the patient/respondent would perceive as beneficial or that would result in a change of treatment.[8]

There are two broad ways in which to calculate a minimal important difference. The first is using anchor-based methods based on patient/user feedback and the second is using statistical distribution methods. In all cases, robust MID values are created from multiple studies and are specific to particular patient groups and disease severities. Therefore, a MID in one patient group characteristic cannot be generalised to different patient characteristics.[9] With this in mind, it is impossible to determine a single valid MID value for the GPPS with multiple disease groups and unknown respondent characteristics (such as disease stage). However, a 2005 a study by Walters and Brazier[10] reviewed eight longitudinal studies with 11 different patient groups looking at the range of MIDs. The results gave an EQ-5D mean MID of 0.074. The highest groups were those with AMI, osteoarthritis of the knee, early rheumatoid arthritis and leg ulcers, all higher than 0.10. The lowest MIDs were recorded in the patient groups with COPD, down to −0.011. In 2007, Pickard et al 2007 reported EQ-5D MIDs for cancer patients between 0.09 to 0.12 using different

anchor-based methods.[11] The study comprised advanced cancer patients, including bladder, brain, breast, colon/rectum, head/neck, liver/pancreas, kidney, lung, lymphoma, ovary and prostate.

Using a distributional approach, taking half the standard deviation of the baseline scores offers another method of defining a minimal important difference[12], but may over-estimate the smallest possible difference[9] in specific patient groups.

For the intervention scenarios, we artificially 'inflated' the individual scores of the specific disease groups in order to simulate a possible disease group intervention. In creating the inflated EQ-5D score scenarios for testing the index, we have used both the average MID reported by Walters and Brazier, Pickard et al, and the half standard deviation of Group 0 to indicate interpretations of the change against possible minimally important differences for each of the disease-specific changes. Scenarios were created by inflating the average score for those with cancer, diabetes, mental health and musculoskeletal conditions. Each scenario was tested separately on the assumption that disease interventions would not all occur simultaneously over a short time period.

For each of the disease specific groups, the half standard deviation thresholds were the following:

- Cancer **σ** = 0.3018451, **½σ** = 0.1509226 (mean = 0.6929575)
- Diabetes **σ** = 0.3260799, **½σ** = 0.1630400 (mean = 0.6725545)
- Mental health **σ** = 0.3710164, **½σ** = 0.1855082 (mean = 0.4706168)
- Musculoskeletal **σ** = 0.3171884, **½σ** = 0.1585942 (mean = 0.557089)

For the intervention scenarios, disease group EQ-5D scores were artificially inflated to represent a mean overall intervention change in that sub-group. The results of these inflations using the distributional thresholds (½σ) and the anchor thresholds can be seen in Table 10. The distributional method scenarios produce the largest increases in the index, with musculoskeletal having the biggest impact. The cancer scenarios have the smallest impact and this is likely to be due to the fact that the cancer group is the smallest proportion of respondents within the GP survey sample.

Although the change in the overall index mean score and the index value is extremely small, the effects are statistically significant against the 95% confidence intervals. The actual small change in value is due to the aggregate effect of the change against the total sample (approximately 600,000 people with long-term conditions). Minimally important shifts in EQ-5D scores in sub-groups, such as the diabetes group, only affect around 13% of the sample so, although there is a noticeable change in the scores of people with diabetes (0.6726 to 0.8356), the total sample mean score only shifts slightly. For this reason it is best to use the index value as an indication of change, rather than the index mean score.

	Baseline condition specific Mean	*'Inflated' condition specific Mean*	Index Mean score	Index value
Baseline			0.7232 (0.7224, 0.7241)	100.00
Change in the mean cancer score of ½ **σ**	*0.6930*	*0.8439*	0.7322	101.23
Change in the mean diabetes score of **½σ**	*0.6726*	*0.8356*	0.7449	102.99
Change in the mean mental health score of **½σ**	*0.4706*	*0.6561*	0.7399	102.30
Change in the mean musculoskeletal score of **½σ**	*0.5571*	*0.7157*	0.7795	107.78
Change in the mean cancer score of 0.11	*0.6930*	*0.8030*	0.7297	100.90
Change in the mean cancer score of 0.074	*0.6930*	*0.7670*	0.7277	100.61
Change in the mean diabetes score 0.074	*0.6726*	*0.7466*	0.7331	101.36
Change in the mean mental health score of 0.074	*0.4706*	*0.5446*	0.7299	100.92
Change in the mean musculoskeletal score of 0.074	*0.5571*	*0.6311*	0.7495	103.63

Table 10: Sensitivity of the index to changes in EQ-5D scores

Table 10 shows that for all the scenarios explored, modelled changes in the HRQoL of different patient groups are detected as statistically significant changes in the weighted index score, thus suggesting a reasonable sensitivity of the score to improvements (or in this case deteriorations) in HRQoL.

Discussion

Estimating the extent to which NHS services are contributing to improving the HRQoL of people with long-term conditions is an important (if not trivial) objective. A key challenge for attaining this goal is to develop methodologies which allow the analysis to control for variations in the NRFs of people with long-term conditions through time.

Previous work by the Department of Health [13] and Ara et al[1] has developed regression-based models that used survey respondent-level data to correct for case-mix issues. The latter study identified that errors in the predicted scores versus actual EQ-5D scores meant that using this method of case-mix approach could be inaccurate.

The analysis in this report has estimated regression models for each of the five EQ-5D dimensions in order to replicate the distribution of the observed EQ-5D scores of people with long-term conditions in the GPPS and to replicate their overall mean scores. However, the use of the dimension-level regression models in predicting EQ-5D scores had limited success. Although the replication of the distribution of actual EQ-5D scores was reasonably good, predicting the mean overall score was not close enough using a range of prediction methods. As well as a difference in mean scores, individual level predictions were limited in their ability to replicate actual individual scores – errors could be substantial at the individual level. These findings were in line with the conclusions by Ara et al, whereby using regression models with case-mix adjustments can be problematic.

We have provided an alternative approach to measuring potential change in the HRQoL of people with long-term conditions by creating a population-weighted HRQoL index based on the actual scores of individuals in the survey sample. The index methodology develops weights which allow the characteristics of GPPS samples to match the profile of a baseline wave in terms of key dimensions of NRFs included in the GPPS with the largest estimated impact on HRQoL. The index is designed to work well with a large sample such as the GPPS, which contains approximately 600,000 people with long-term conditions in a sample year that can contribute toward the index construction and potential change every year. The foundation for the index was a regression model predicting variations in HRQoL with a small number of variables in order to yield population cell sizes that were considered acceptable for a population weight. Comparing the predictive power of this simple model by itself and in comparison to a model using a larger number of variables, we concluded that the 'simple' model was a good base for the HRQoL index.

The key objective of the index was to identify change in HRQoL, whilst being insensitive to changes in demographic and disease prevalence issues - the case-mix - of the population. We tested this by constructing artificial scenarios of case-mix change and mean score changes in disease specific groups. Overall, the index did not change significantly for up to 20% increases in disease prevalence in most conditions and did not change significantly for increases in mean age of up to five years. However, changes in individual health score in specific disease groups to a minimally important degree for EQ-5D led to significant differences in the index for the total population.

Although statistical significance could occur through large changes in the prevalence of some diseases (for example, musculoskeletal disease), the index does remain quite stable and still quite sensitive to changes in EQ-5D scores resulting from factors outside changes in prevalence and demographics. It therefore seems possible to create an index which is sensitive to changes in EQ-5D while controlling for changes in prevalence and demographics.

A comparison of the index created on one random half of our sample showed extremely good agreement with the other random sample half, in line with the earlier regression models, suggesting the index works well in the same time period. However, it would be useful to carry out further analysis to check issues such as seasonality effects across GPPS waves.

Interpretation of the index

The results of the study suggest that the weighted index methodology is able to control for changes in case-mix while remaining sensitive to true changes, ceteris paribus, in the HRQoL of people with long-term conditions. However, the use of the index and in particular its interpretation can be challenging.

From a conceptual point of view, the fact that the index controls for changes in the case-mix indicators available in the GPPS (and in particular in health conditions) means that changes in the HRQoL identified using the proposed methodology do not reflect potential improvements (or deteriorations) in the HRQoL of people with long-term conditions associated with changes in NRFs brought about by NHS services. Improvements in HRQoL from prevention effects in particular cannot be accounted for with the proposed methodology, and the interpretation of the results obtained should reflect this. The proposed methodology contributes one element necessary for answering the question 'is the HRQoL of people with long-term conditions improving through time in England', and its results should therefore be interpreted in conjunction with other evidence about variations in the prevalence of conditions to provide a rounded picture of the performance of the NHS in improving HRQoL.

Given the very large size of GPPS, the analysis is powered to detect very small differences in average EQ-5D scores. In fact, with the data used in the analysis, any shifts in the mean index of more than 0.0008 would be identified in the analysis as statistically significant. Whereas the high statistical power exhibited by the approach using the GPPS is desirable, a key challenge for the application of the weighted index approach concerns the 'interpretability' from a policy point of view of observed changes in the weighted index. For an individual, an improvement of 0.0008 in the EQ-5D score is meaningless from a 'clinical' point of view, as it would not translate into a meaningful change in the HRQoL of the individual in question. At the population level, however, an average change of 0.0008 in EQ-5D scores could 'hide' a clinically significant change in the HRQoL of a particular group of patients (for instance, individuals with a particular long-term condition). Furthermore, a hypothetical improvement of 0.0008 in the average EQ-5D score for the 2.5 million individuals with long-term conditions could be aggregated up to represent a gain of approximately 2,000 QALYs. Whether such a gain should be considered as policy meaningful is difficult to say. Further work is required to help with the interpretation of differences in average HRQoL between GPPS waves, to examine what constitutes a policy-meaningful improvement or deterioration in average HRQoL. It is not clear, in particular, that this assessment should be carried out purely in terms of statistical significance.

A number of strategies could be used for conveying the results of the analysis of changes based on the weighted index. In addition to absolute changes in average index scores, it might be useful to present relative changes (proportional changes) in scores. Although intuitively appealing, measures of relative changes have, however, the limitation that the index itself is bound by 1 (representing perfect health) and therefore that equal proportional changes through time in HRQoL would be increasingly difficult to obtain and therefore should be valued differently. In other words, as the index mean is bound by 1, an increase from 0.72 to 0.75 is not the same as the increase from 0.75 to 0.78. It may be useful to set the base year score as 100 and plot change on an exponential scale to more appropriately reflect the extra effort involved in year-on-year change as the mean index scores increases.

As noted above, it might also be useful to translate average changes in the score into aggregate figures, in terms of the number of quality adjusted life years saved. Such a measure, however, might 'overemphasise' the size of the change observed in terms of how it might be perceived by the public.

From a different perspective, the fact that even significant changes in the HRQoL of given population groups (for instance, people with a given condition) result in very small average changes in the index suggests the importance of carrying out sub-group analyses by condition. These sub-group analyses would be particularly useful from the point of view of understanding why changes in average HRQoL might occur and of identifying key areas for policy intervention. A useful analytical approach would be therefore to interpret the index in conjunction with sub-group score changes. An increase or decrease in HRQoL scores may be statistically significant, but without referencing this to sub-group change (for example, people with diabetes), we would not know why or where there has been change in perceived HRQoL of the population of people with long-term conditions. Mean score changes may be noticeable within specific disease groups, or may be randomised across multiple groups. In fact, index change should be related to NHS or other health policy interventions where known. If any change in the index score appears to be independent of known policy interventions, then this should be questioned.

Caveats and limitations

The general limitations of the regression models in the current study suggests that there are other indicators, not included in the current analysis, that contribute to explaining self-reported health status and EQ-5D. Whether this unexplained proportion of the variation in EQ-5D stems mainly from random chance or other measureable factors, not included in the dataset, is not known from the evidence here. Adding additional linked variables from outside the dataset would appear to be worth pursuing in order to improve the fit of the regression model. Given the large proportion of variability in EQ-5D that remains unexplained by the factors included in the model, a given set of weights should only be used to investigate short-term changes in HRQoL, in the expectation that only small changes in uncontrolled factors would take place. Over time, the proposed methodology would require the periodic recalibration of the model, and the identification of any changes in the set of factors most strongly associated with variations in EQ-5D and thus to be included in the recalculation of the index.

Disease-specific regressions showed that only in the diabetes case would we see a potentially fit better than a multi-disease model. However, we did not take the disease specific models further, as our core goal was to find an approach to predicting HRQoL for the total sample of people with long-term conditions. Interpreting results of change in disease sub-categories using regression approaches may be useful in the overall interpretation of an aggregate HRQoL index over time.

The self-report measure of being sick and disabled was a very good predictor of HRQoL score in people with long-term conditions, but is a volatile variable. A change in policy with regard to sick and disabled people (i.e. benefits and classification issues) could affect the index in later years.

Regardless of the statistical or econometric technique used to control for case-mix, the success of the analysis will depend on the quality of the indicators available, and in particular on the extent to which they capture systematic differences in NRFs across waves. In this respect the GPPS may not be ideal: for instance, in the extent to which the indicators available within it are able to control for changes in the severity of the disease within each condition in the survey. Improvements in life expectancy of people with long-term conditions might lead, for instance, to increases in the intensity of NRFs within condition groups. Furthermore, changes in the public perception and awareness of conditions may change through time, which again might lead to differences in the nature of the NRFs of people declaring to have a certain condition. The lack of finely-graded indicators of condition and the self-reported nature of the data in the GPPS means that some differences in case-mix between waves could go undetected and result in significant differences in the weighted index. The significance of these problems is likely to increase if comparisons are made between GPPS waves several years apart.

The key objective of the analysis in this report was to develop a method for controlling for changes in the case-mix of people with long-term conditions. The assumption was made that changes through time in the standardised average HRQoL for people with long-term conditions would reflect the impact that improving (or deteriorating) services would have year on year. Whereas most of the variations in case-mix through time in the GPPS might be due to a combination of sampling differences and changes not directly the result of NHS services (such as, for instance, changes in individuals' behaviour), controlling for variations in case-mix through time precludes the analysis from evaluating the impact of services on the actual prevalence of long-term conditions. As a result, NHS investments leading to reductions in the prevalence of long-term conditions (and therefore to improvements in the HRQoL of the people that otherwise would have developed such conditions) are explicitly not taken into account in the analysis. Assessing different strategies for doing so should be a priority objective for further research in this area.

Overall, we believe that the proposed weighted HRQoL index makes a useful contribution to the assessment of changes in quality of life for people with long-term conditions. However, it should be used as part of a broader set of tools for policy assessment given that it is only able to capture a proportion of need-related variations in HRQoL, and because of its inability to account for preventative effects linked to NHS services. Indicators of perceived unmet need

for services such as the ones included in GPPS could be examined to assist the interpretation of changes through time in weighted HRQoL scores. Further work should be carried out to improve the available sources of evidence about NRFs (including, for instance, evidence about changes in the average severity of disease through time) and the methods used to identify the specific contribution of the NHS to improving the quality of life of people with long-term conditions.

References

1. Ara R, van Hout B, Kearns B, Brazier JE. *Case-Mix Methodology for the NHS Outcomes Framework GPPS Questionnaire Data*. Policy Research Unit in Economic Evaluation in Health and Care Interventions; 2011.

2. Fortin M, Lapointe L, Hudon C, et al. Multimorbidity and quality of life in primary care: a systematic review. *Health Qual Life Outcomes*. 2004;2:51.

3. Marmot M. *Fair Society, Healthy Lives: The Marmot Review ; Strategic Review of Health Inequalities in England Post-2010*. The Marmot Review; 2010.

4. Wolfe R. *STATA user command: omodel*. Victoria: Royal Children's Hospital, Parkville, Victoria; 1997. Available at: http://fmwww.bc.edu/RePEc/bocode/o.

5. Williams R. Generalized ordered logit/partial proportional-odds models for ordinal dependent variables. *STATA Journal*. 2006;6(1):58–82.

6. Williams R. *STATA user command: gologit2*. Notre Dame Department of Sociology; 2009. Available at: htthttp://www.stata-journal.com/software/sj7-2p://www.nd.edu/~rwilliam/gologit2/.

7. Anon. APHO Diabetes Prevalence Model. 2010. Available at: http://www.yhpho.org.uk/resource/view.aspx?RID=81123. Accessed July 6, 2012.

8. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis*. 1987;40(2):171–178.

9. Revicki DA, Cella D, Hays RD, et al. Responsiveness and minimal important differences for patient reported outcomes. *Health Qual Life Outcomes*. 2006;4:70.

10. Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. *Qual Life Res*. 2005;14(6):1523–1532.

11. Pickard AS, Neary MP, Cella D. Estimation of minimally important differences in EQ-5D utility and VAS scores in cancer. *Health Qual Life Outcomes*. 2007;5:70.

12. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care*. 2003;41(5):582–592.

13. Lees K. *Are people with long-term conditions reporting improving or declining health?* Department of Health; 2010.

Appendix 1: summary statistics and variables

The distributions of the EQ-5D scores were graphically displayed for both those who reported having a long-term condition and all respondents. The EQ-5D distribution is not normally distributed, as can be seen from the distribution charts (see Figures 21: EQ-5D score distributions). It is multimodal usually with two prominent peaks. For this reason, linear or other regression models based on the EQ-5D index score alone were discounted with the expectation that any predictions using these methods would not be able to replicate this distribution.

It was noted that the number of observations varied across the same disease group, meaning not everyone who filled an EQ-5D dimension filled out all dimensions. These people did not have an EQ-5D score, and therefore did not contribute to the distribution. Removing them for this reason was considered, but it was decided that they were still useful in predicting individual scores.

Over 50% of the sample had a recording of a long-term condition; in group 0 there were 539,464 long-term conditions recorded for 309,251 people. These were made up of a range of conditions including cancer, Alzheimer's, deafness, high blood pressure and learning difficulty. The most common conditions were 'high blood pressure' (116,626) and 'arthritis or long-term joint problem' (83,567), and the least common conditions were 'Alzheimer's disease or dementia' (3,439) and 'learning difficulty' (4,279).

Some of the predictive variables correlate with each other, although only deprivation (IMD) and ethnicity were higher than $R = 0.2$.

Age

From Figure 14, we can see that the relationship between age and EQ-5D is fairly progressive, so was treated as ordered categorical.

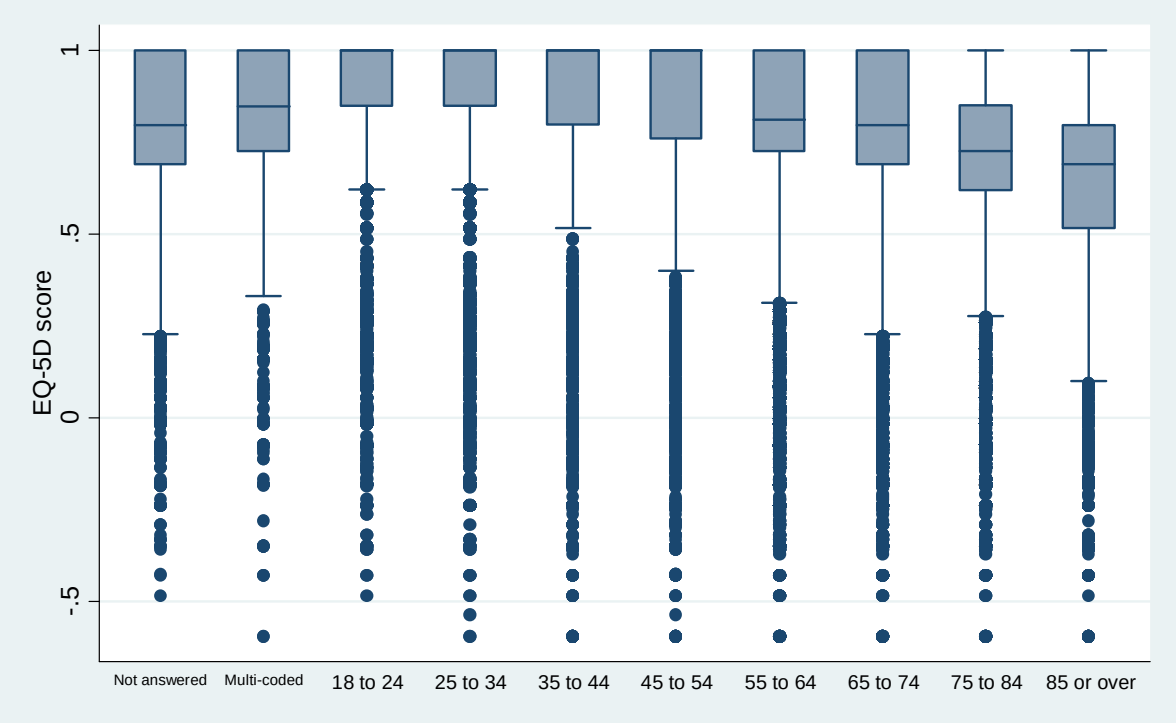


Figure 14: Age Box-Whisker chart

Smoking

From Figure 15 it appears this may be transformable into a binary variable showing 'never smoked' against 'ever smoked' (a grouped variable showing smokers, occasional smokers and ex-smokers), but on closer inspection (Figure 16) the categories within 'smoking status' appear to show different patterns so this variable was kept as categorical nominal.

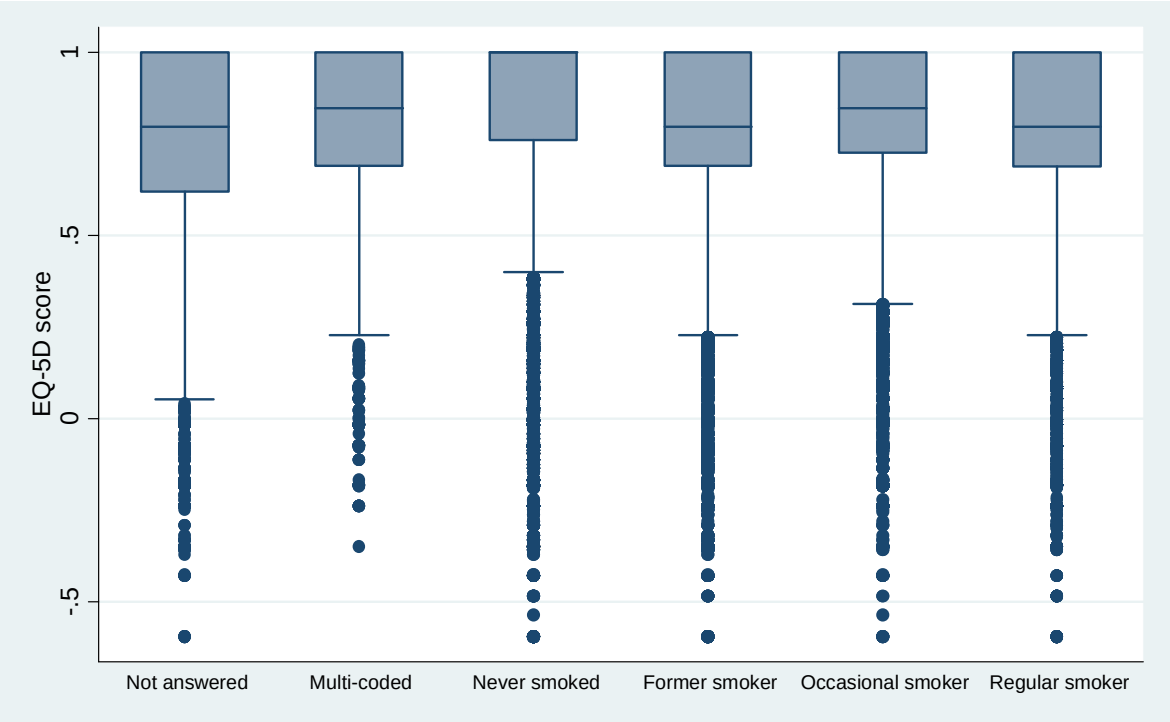


Figure 15: Smoking Box-Whisker chart

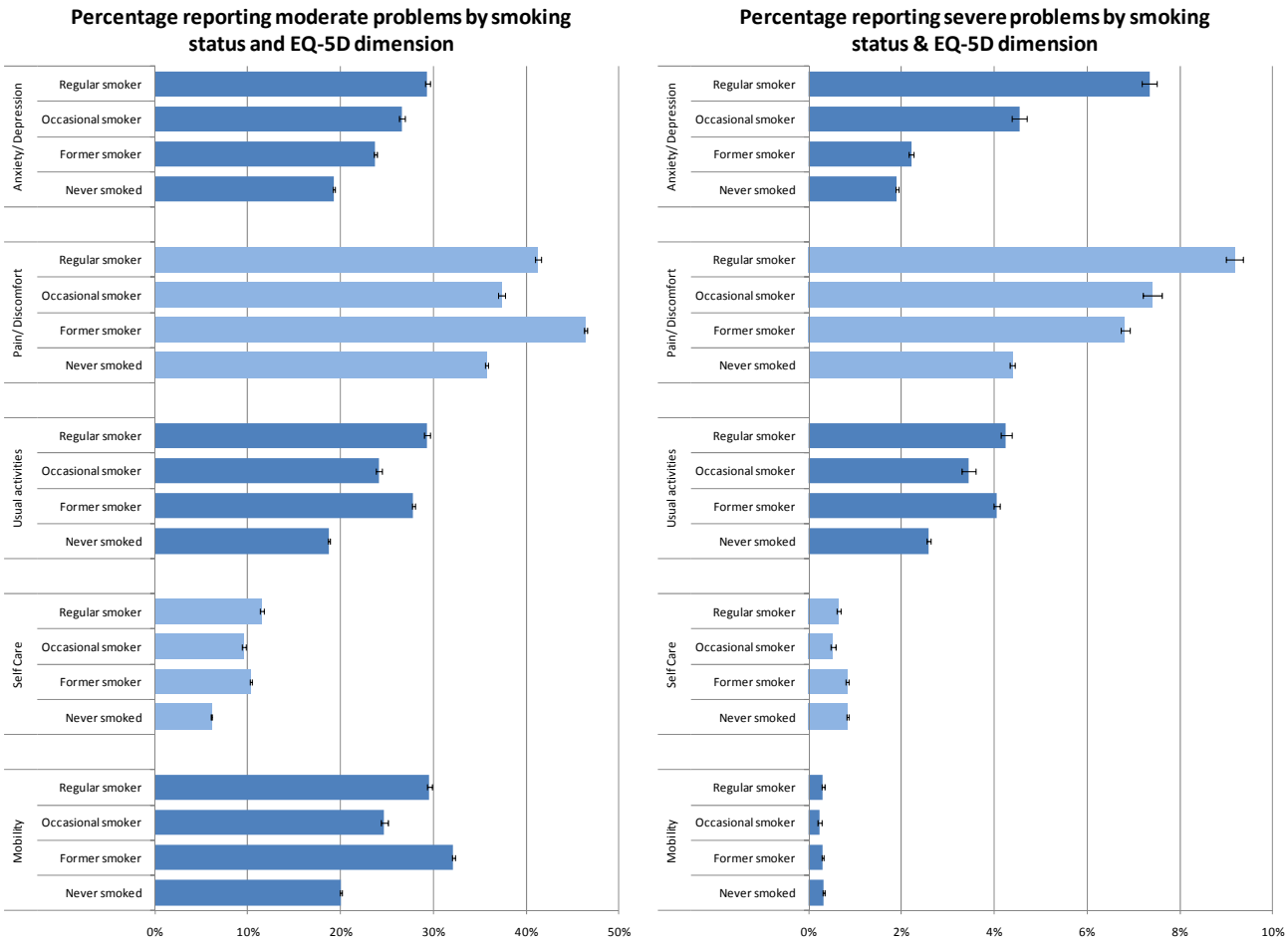


Figure 16: Distribution of where problems are reported by smoking status

Deprivation

It was decided to use IMD rank rather than IMD score for deprivation, since 'score' is not normally distributed or transformable but IMD rank is fairly evenly distributed, as shown in Figure 17 and Figure 18. Although highly significant, an increase of 1 in a scale which runs from 1 to over 32,000 is a very small number (large after multiplying) which is prone to rounding errors and slightly more difficult to interpret. Grouping of this variable was undertaken for the model.

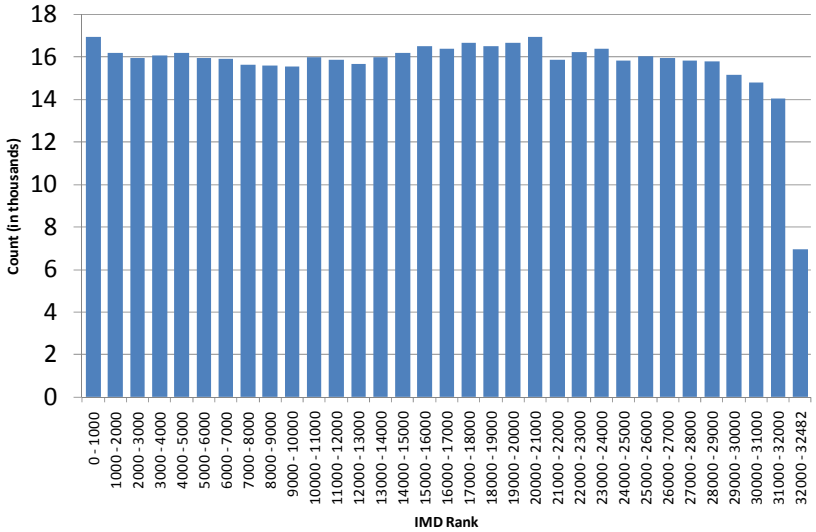


Figure 17: Deprivation rank distribution chart

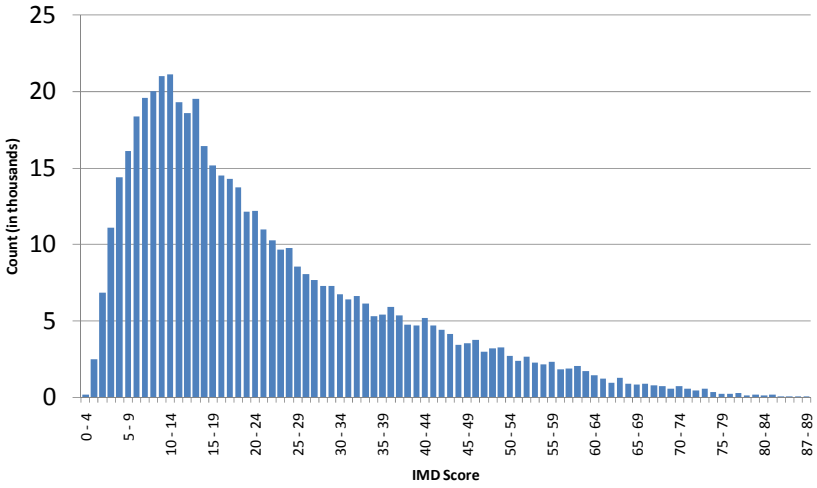


Figure 18: Deprivation score distribution chart

Working status

The working status variable is nominal in nature, with no clear ordinal element. There may be confounders between the presence of a long-term condition and having a job.

Ethnicity

The ethnicity variable is also nominal in nature and there is no clear ordinal element, although it may be possible to group some ethnicities if they have similar traits. Figure 19 shows that there are some categories that appear very similar for example:

- *Any other white background* with *white and black Caribbean,* and with *any other Asian background*
- *Pakistani* with *Bangladeshi*
- *Caribbean* with *other black/ African/ Caribbean*

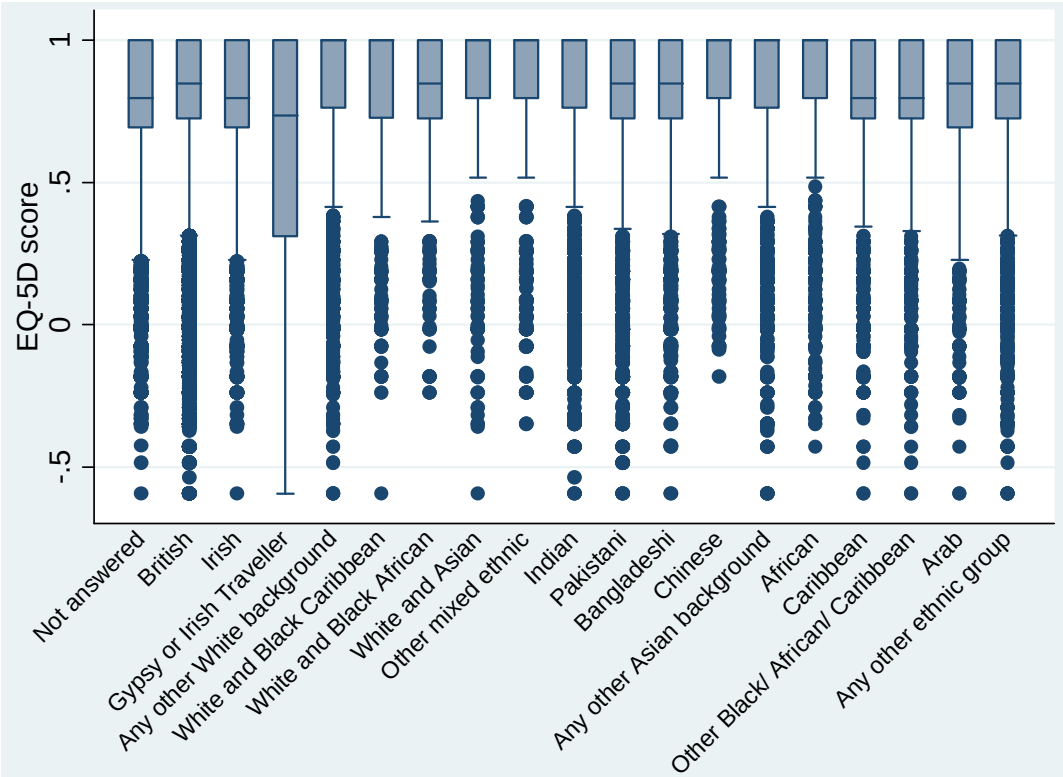


Figure 19: Ethnicity Box-Whisker chart

Table 11 shows the distribution of respondents stating which long-term condition they had. Note that individuals can appear in multiple conditions. People with high blood pressure made up the largest single condition, followed by arthritis or joint problems.

Disease	Count	Percent
Alzheimer's disease or dementia	3,439	1.1
Angina or long-term heart problem	33,556	10.3
Arthritis or long-term joint problem	83,567	25.8
Asthma or long-term chest problem	50,703	15.6
Blindness or severe visual impairment	6,542	2.0
Cancer in the last five years	18,822	5.8
Deafness or severe hearing impairment	24,539	7.6
Diabetes	42,788	13.2
Epilepsy	5,398	1.7
High blood pressure	116,626	36.0
Kidney or liver disease	9,029	2.8
Learning difficulty	4,279	1.3
Long-term back problem	54,337	16.8
Long-term mental health problem	17,571	5.4
Long-term neurological problem	9,158	2.8
Another long-term condition	59,110	18.2

Table 11: Sample size in Group 0 by long-term condition reported

Appendix 2: adding additional explanatory variable to the Index

Initial testing of the weighted HRQoL index showed good stability (within the 95% confidence limits) of introducing scenarios of change in mean age and disease prevalence of up to 10%. However, increasing musculoskeletal change prevalence by 20% showed a significant change in the index score. The poor control of large changes in musculoskeletal was demonstrated by the mean score dropping from 0.7233 to 0.7145 and the index dropping from 100 to 98.78.

Musculoskeletal is the largest group, and roughly half of all people with LTCs are in this group, so a 20% increase is a very large absolute increase and, to keep the sample size the same, a large number of healthier people must be removed from the sample. Nevertheless, because of the ability of musculoskeletal diseases to impact the index so much, modifications to the model were considered. First, the presence of this disease was added to the simple model to check the relative 'fit'. This required an increase in the number of cells to 280. Unfortunately 33 cells had a population below 50 (and four had fewer than 10).

Adding the binary variable musculoskeletal in to the model obviously increases the performance of the index with respect to controlling for changes in prevalence of musculoskeletal conditions; the score in this scenario drops from 0.7233 to 0.7232 and the index only drops from 100 to 99.98 with a simulated relative increase in prevalence of 20%. The 'fit' of the model compared to the simple model without musculoskeletal also improves, especially in the mobility and pain dimensions as can be seen in Table 12.

	Mobility	Self-care	Usual activities	Pain and discomfort	Anxiety & depression
Simple	0.2324	0.2350	0.1748	0.1347	0.0855
Simple including musculoskeletal	0.2735	0.2441	0.2013	0.2328	0.0860

Table 12: McFaddens Pseudo R^2 of the full model before and after the inclusion of musculoskeletal

As a solution to this issue, a 'hybrid' method was proposed. A subset of the original 140 cells would be split according to the presence of a musculoskeletal condition. Three different ways of identifying which cells should be split were investigated.

Split methods:

1. Take most populated 60 cells – end up with 200 cells
2. Split all where there is no recording of permanently sick or disabled – end up with 210 cells
3. Split all where there is no recording of permanently sick or disabled, fewer than five diseases, and aged 35-84 – end up with 172 cells

These different methods of constructing the index produced different levels of control for increases in musculoskeletal prevalence, the results can be seen in Table 13.

	Index	
140 cells	0.7145	98.78
280 cells	0.7232	99.98
Split method 1	0.7232	99.92
Split method 2	0.7223	99.88
Split method 3	0.7210	99.68

Table 13: Effect of a 20% increase in musculoskeletal conditions on a mean EQ-5D score of 0.7233

The reason for the decision to split those 32 cells in method 3 and not the others is because:

- The population in the cells with a recording of permanently sick or disabled is already quite small
- The population in the cells with five or more diseases is already quite small
- There are very few people under the age of 35 with musculoskeletal problems

- The population in the cells for the 85 and over is already quite small

Whilst the first two methods do give good results, the first method may be difficult to justify in the long term if the only justification is the selection based on cell size using an arbitrary rank cut off (60) as cell sizes will fluctuate and this means that a cell which is not split may become bigger than one that is. The second method is a simple method which is better than splitting all as there are far fewer small cells; however, there are still 14 cells with 50 or below and the lowest cell count is 6.

Using method 3, the control for such a large increase in musculoskeletal conditions is still very good: only 32 extra cells are produced for the index, of which 30 are in the most populated 60 cells and the lower end of the cell counts are no different from that of the 140 cell method. Figure 20 shows the relationship between 172 cells of the predicted and actual EQ-5D scores. All analysis of the weighted HRQoL index was finalised using the 172 cell population split.

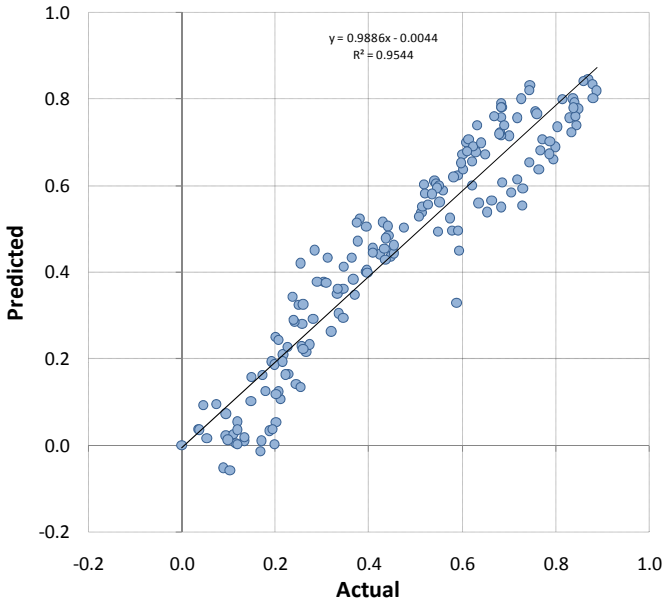


Figure 20: Mean scores actual and predicted EQ-5D scores using split method 3
note - the outlier (actual = 0.59, predicted = 0.33) has a sample size of 50.

Appendix 3: Split of Survey sample into Group 0 and Group 1

		Group 0	Group 1
Participants	n (%)	518,808 (50.0)	519,138 (50.0)
Male	n (%)	219,308 (42.3)	219,414 (42.3)
Age band			
18 to 24	n (%)	22,974 (4.43)	22,872 (4.41)
25 to 34	n (%)	51398 (9.91)	51554 (9.93)
35 to 44	n (%)	69916 (13.48)	70276 (13.54)
45 to 54	n (%)	88795 (17.12)	89009 (17.15)
55 to 64	n (%)	102827 (19.82)	103054 (19.85)
65 to 74	n (%)	93788 (18.08)	93720 (18.05)
75 to 84	n (%)	59493 (11.47)	59243 (11.41)
85 or over	n (%)	19622 (3.78)	19413 (3.74)
Ethnicity			
Not answered	n (%)	10749 (2.07)	10812 (2.08)
UK	n (%)	419446 (80.85)	420155 (80.93)
Irish	n (%)	5797 (1.12)	5687 (1.1)
Gypsy or Irish Traveller	n (%)	148 (0.03)	141 (0.03)
Any other White background	n (%)	21754 (4.19)	21638 (4.17)
White and Black Caribbean	n (%)	1105 (0.21)	1111 (0.21)
White and Black African	n (%)	605 (0.12)	597 (0.11)
White and Asian	n (%)	1037 (0.2)	976 (0.19)
Any other Mixed / multiple ethnic	n (%)	1129 (0.22)	1139 (0.22)
Indian	n (%)	12459 (2.4)	12546 (2.42)
Pakistani	n (%)	6648 (1.28)	6608 (1.27)
Bangladeshi	n (%)	2219 (0.43)	2209 (0.43)
Chinese	n (%)	2576 (0.5)	2638 (0.51)
Any other Asian background	n (%)	6697 (1.29)	6577 (1.27)
African	n (%)	6733 (1.3)	6804 (1.31)
Caribbean	n (%)	4691 (0.9)	4708 (0.91)
Any other Black/African/Caribbean	n (%)	2174 (0.42)	2078 (0.4)
Arab	n (%)	994 (0.19)	976 (0.19)
Any other ethnic group	n (%)	11847 (2.28)	11738 (2.26)
Deprivation Score	mean (sd)	21.9 (15.6)	21.9 (15.6)
Smoking Status			
Never smoked	n (%)	260834 (50.28)	261548 (50.38)
Former smoker	n (%)	155713 (30.01)	155014 (29.86)
Occasional smoker	n (%)	33282 (6.42)	33440 (6.44)
Regular smoker	n (%)	51293 (9.89)	51515 (9.92)
Working Status			
Full-time paid work (30+ hours)	n (%)	167434 (32.27)	167770 (32.32)
Part-time paid work (<30 hours)	n (%)	64755 (12.48)	65083 (12.54)
Full-time education	n (%)	8568 (1.65)	8269 (1.59)
Unemployed	n (%)	21862 (4.21)	21777 (4.19)
Permanently sick or disabled	n (%)	24296 (4.68)	24564 (4.73)
Fully retired from work	n (%)	160746 (30.98)	160427 (30.9)
Looking after the home	n (%)	29601 (5.71)	29784 (5.74)
Doing something else	n (%)	10559 (2.04)	10716 (2.06)

Deaf	n (%)	2031 (0.39)	2009 (0.39)

Table 14: Confirmation of the equality of the split (1)

Q34a. State of health today...Mobility		Group 0	Group 1
Not answered	n (%)	20419 (3.94)	20220 (3.89)
Multi-coded	n (%)	1309 (0.25)	1285 (0.25)
I have no problems in walking about	n (%)	368439 (71.02)	368988 (71.08)
I have some problems in walking about	n (%)	127054 (24.49)	127006 (24.46)
I am confined to bed	n (%)	1587 (0.31)	1639 (0.32)

Q34b. State of health today...Self-Care		Group 0	Group 1
Not answered	n (%)	22350 (4.31)	22031 (4.24)
Multi-coded	n (%)	1209 (0.23)	1227 (0.24)
I have no problems with self-care	n (%)	449802 (86.7)	450067 (86.7)
I have some problems washing or dressing	n (%)	41311 (7.96)	41664 (8.03)
I am unable to wash or dress myself	n (%)	4136 (0.8)	4149 (0.8)

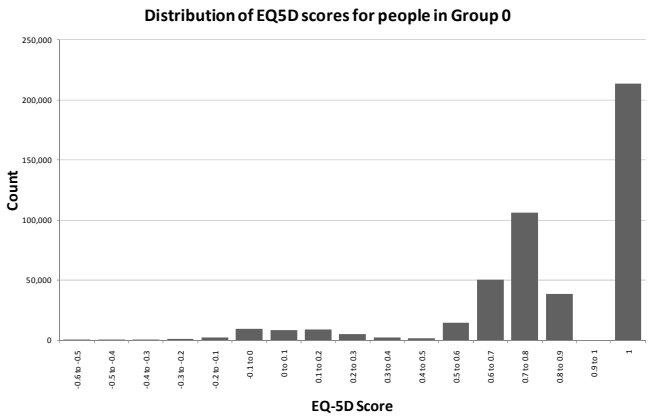
Q34c. State of health today...Usual Activities		Group 0	Group 1
Not answered	n (%)	20612 (3.97)	20604 (3.97)
Multi-coded	n (%)	1415 (0.27)	1429 (0.28)
I have no problems with performing my usual activities	n (%)	364578 (70.27)	364588 (70.23)
I have some problems with performing my usual activities	n (%)	115417 (22.25)	115685 (22.28)
I am unable to perform my usual activities	n (%)	16786 (3.24)	16832 (3.24)

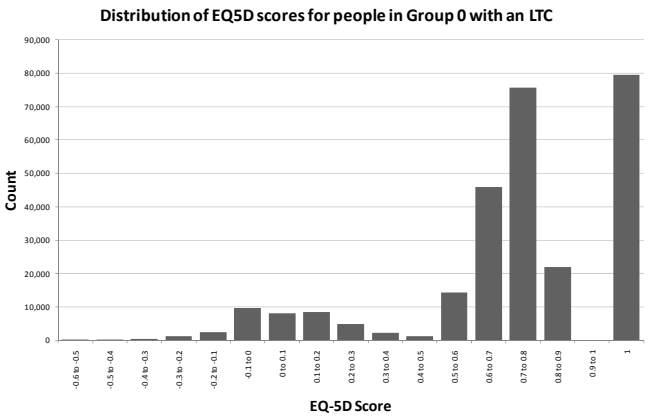
Q34d. State of health today...Pain/Discomfort		Group 0	Group 1
Not answered	n (%)	21610 (4.17)	21406 (4.12)
Multi-coded	n (%)	2455 (0.47)	2372 (0.46)
I have no pain or discomfort	n (%)	267753 (51.61)	267708 (51.57)
I have moderate pain or discomfort	n (%)	197556 (38.08)	197953 (38.13)
I have extreme pain or discomfort	n (%)	29434 (5.67)	29699 (5.72)

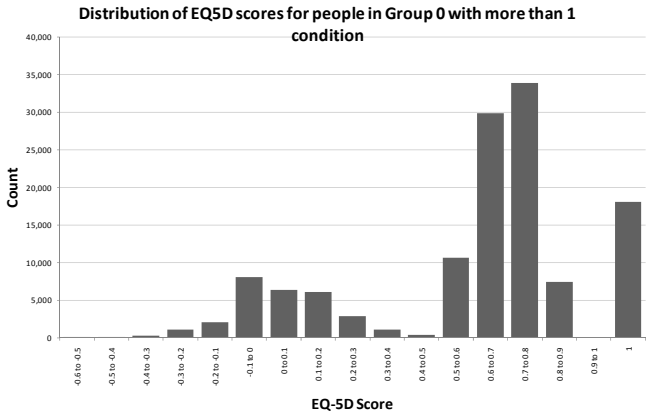
Q34e. State of health today...Anxiety/Depression		Group 0	Group 1
Not answered	n (%)	32090 (6.19)	31646 (6.1)
Multi-coded	n (%)	1123 (0.22)	1142 (0.22)
I am not anxious or depressed	n (%)	363607 (70.09)	364039 (70.12)
I am moderately anxious or depressed	n (%)	108566 (20.93)	108686 (20.94)
I am extremely anxious or depressed	n (%)	13422 (2.59)	13625 (2.62)

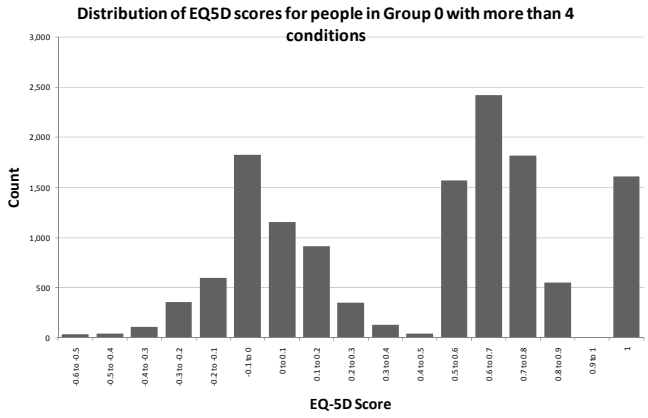
Table 15: Confirmation of the equality of the split (2)

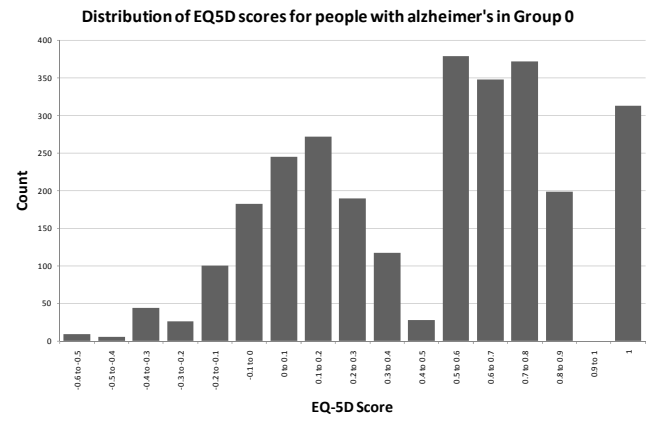
Appendix 4: EQ-5D distributions

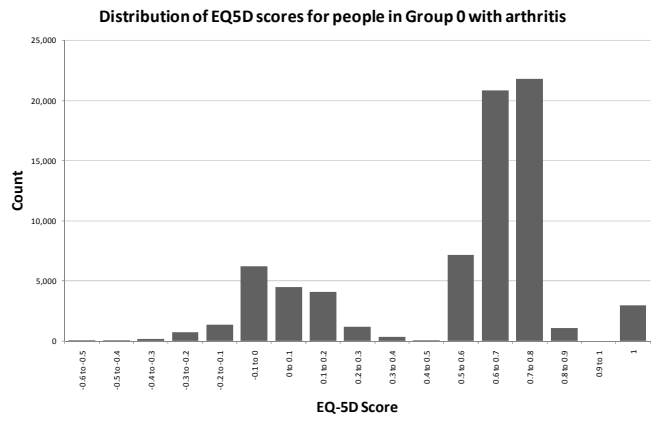


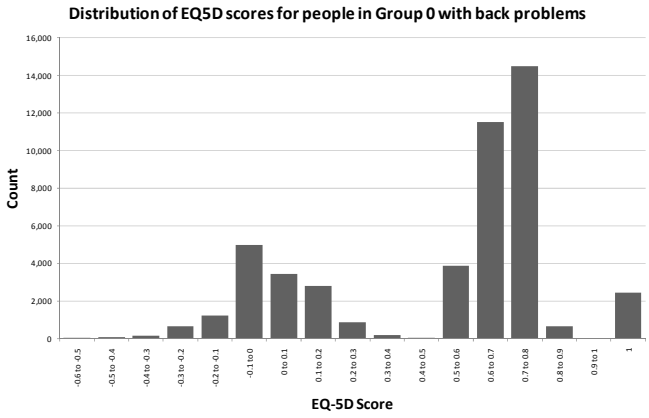


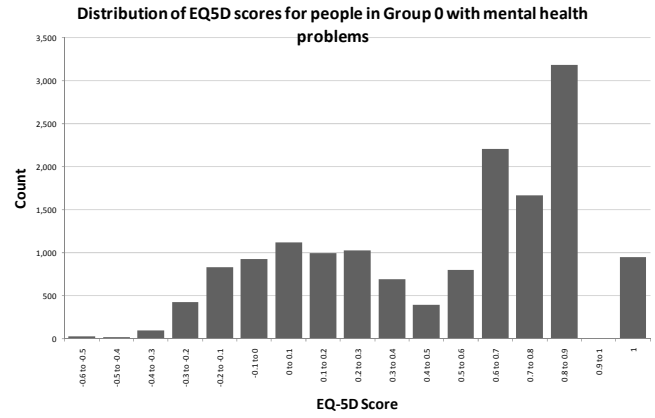












Figures 21: EQ-5D score distributions

Appendix 5: Full Model Coefficients (expressed as Log Odds)

Independent variable	Log odds for mobility - Base: I have no problems in walking about		Log odds for Self-Care - Base: I have no problems with self-care		Log odds for usual activity-Base: I have no problems with performing my usual activities		Log odds for pain - Base: I have no pain or discomfort		Log odds for anxiety - Base: I am not anxious or depressed	
	I have some problems in walking about	I am confined to bed	I have some problems washing or dressing	I am unable to wash or dress myself	I have some problems with performing my usual activities	I am unable to perform my usual activities	I have moderate pain or discomfort	I have extreme pain or discomfort	I am moderately anxious or depressed	I am extremely anxious or depressed
Age band	0.40 (0.39, 0.41)	0.41 (0.36, 0.46)	0.23 (0.21, 0.24)	0.42 (0.39, 0.46)	0.19 (0.18, 0.19)	0.31 (0.29, 0.32)	0.20 (0.19, 0.21)	-0.06 (-0.08, -0.05)	-0.03 (-0.04, -0.02)	-0.18 (-0.20, -0.16)
IMD_Decile (by rank)	0.08 (0.08, 0.08)	0.01 (-0.01, 0.03)	0.08 (0.07, 0.08)	0.03 (0.01, 0.04)	0.06 (0.05, 0.06)	0.03 (0.03, 0.04)	0.06 (0.06, 0.07)	0.08 (0.07, 0.08)	0.04 (0.04, 0.04)	0.07 (0.06, 0.07)
Number of LTCs	0.55 (0.50, 0.61)	0.49 (0.21, 0.77)	0.67 (0.60, 0.74)	0.71 (0.52, 0.89)	0.56 (0.50, 0.61)	0.46 (0.36, 0.57)	0.73 (0.68, 0.78)	0.39 (0.29, 0.48)	0.57 (0.53, 0.62)	0.71 (0.60, 0.82)
Alzheimer's or dementia	0.27 (0.15, 0.38)	1.37 (1.04, 1.70)	1.40 (1.29, 1.51)	1.66 (1.45, 1.87)	0.97 (0.86, 1.09)	1.48 (1.34, 1.62)	-0.79 (-0.90, -0.69)	-0.41 (-0.57, -0.24)	0.46 (0.36, 0.55)	0.56 (0.37, 0.75)
Angina or heart problems	0.09 (0.03, 0.15)	-0.63 (-0.95, -0.31)	-0.37 (-0.44, -0.29)	-0.67 (-0.87, -0.47)	0.09 (0.03, 0.15)	-0.16 (-0.28, -0.04)	-0.21 (-0.26, -0.16)	-0.15 (-0.25, -0.04)	-0.28 (-0.33, -0.22)	-0.51 (-0.64, -0.38)
Arthritis or joint problems	1.16 (1.10, 1.22)	-0.76 (-1.06, -0.46)	0.32 (0.25, 0.39)	-0.61 (-0.80, -0.41)	0.87 (0.82, 0.93)	-0.02 (-0.13, 0.09)	2.10 (2.04, 2.15)	1.07 (0.98, 1.17)	-0.15 (-0.19, -0.10)	-0.50 (-0.62, -0.38)
Asthma or chest problems	-0.11 (-0.17, -0.05)	-0.47 (-0.78, -0.15)	-0.34 (-0.42, -0.27)	-0.61 (-0.81, -0.41)	-0.17 (-0.22, -0.11)	-0.20 (-0.31, -0.08)	-0.44 (-0.49, -0.40)	-0.19 (-0.29, -0.09)	-0.34 (-0.38, -0.29)	-0.54 (-0.66, -0.42)
Blindness or visual impairment	0.35 (0.26, 0.45)	0.22 (-0.13, 0.56)	0.11 (0.01, 0.20)	0.11 (-0.11, 0.33)	0.63 (0.54, 0.72)	0.33 (0.20, 0.46)	-0.59 (-0.68, -0.50)	-0.23 (-0.36, -0.11)	-0.12 (-0.20, -0.05)	-0.31 (-0.48, -0.14)
Cancer	-0.20 (-0.27, -0.13)	-0.29 (-0.63, 0.05)	-0.27 (-0.35, -0.19)	-0.66 (-0.88, -0.43)	-0.03 (-0.09, 0.04)	-0.05 (-0.18, 0.07)	-0.23 (-0.29, -0.18)	-0.18 (-0.29, -0.07)	-0.22 (-0.28, -0.17)	-0.56 (-0.71, -0.41)
Deafness or severe hearing impairment	-0.29 (-0.36, -0.22)	-0.67 (-1.01, -0.33)	-0.45 (-0.53, -0.37)	-0.62 (-0.83, -0.42)	-0.30 (-0.36, -0.23)	-0.26 (-0.38, -0.14)	-0.49 (-0.55, -0.44)	-0.26 (-0.37, -0.16)	-0.34 (-0.40, -0.29)	-0.48 (-0.62, -0.34)
Diabetes	0.00 (-0.06, 0.06)	-0.23 (-0.53, 0.08)	-0.31 (-0.38, -0.23)	-0.49 (-0.69, -0.29)	-0.18 (-0.24, -0.13)	-0.23 (-0.34, -0.11)	-0.40 (-0.45, -0.35)	-0.09 (-0.19, 0.01)	-0.44 (-0.49, -0.39)	-0.62 (-0.74, -0.49)
Epilepsy	-0.16 (-0.26, -0.06)	0.03 (-0.34, 0.40)	-0.29 (-0.40, -0.18)	0.03 (-0.21, 0.26)	-0.22 (-0.32, -0.13)	-0.20 (-0.35, -0.05)	-0.63 (-0.71, -0.54)	-0.45 (-0.59, -0.31)	-0.40 (-0.48, -0.32)	-0.39 (-0.56, -0.23)
High blood pressure	-0.47 (-0.53, -0.41)	-1.00 (-1.30, -0.70)	-0.75 (-0.82, -0.68)	-1.08 (-1.27, -0.88)	-0.58 (-0.64, -0.53)	-0.67 (-0.79, -0.56)	-0.59 (-0.64, -0.55)	-0.35 (-0.44, -0.25)	-0.46 (-0.50, -0.41)	-0.69 (-0.80, -0.57)
Kidney or liver disease	0.08 (-0.01, 0.16)	-0.19 (-0.56, 0.17)	-0.17 (-0.26, -0.08)	-0.30 (-0.53, -0.06)	0.02 (-0.06, 0.10)	-0.03 (-0.16, 0.10)	-0.09 (-0.17, -0.02)	-0.03 (-0.14, 0.09)	-0.25 (-0.32, -0.19)	-0.47 (-0.63, -0.32)
Learning difficulty	-0.30 (-0.41, -0.19)	0.09 (-0.29, 0.46)	0.46 (0.36, 0.57)	0.85 (0.63, 1.07)	-0.17 (-0.28, -0.07)	-0.11 (-0.27, 0.04)	-1.10 (-1.20, -1.00)	-0.80 (-0.95, -0.65)	-0.40 (-0.49, -0.31)	-0.81 (-0.97, -0.65)
Long-term back problem	0.71 (0.64, 0.77)	-0.68 (-0.99, -0.37)	0.13 (0.05, 0.20)	-0.70 (-0.90, -0.50)	0.77 (0.72, 0.83)	-0.02 (-0.13, 0.09)	1.94 (1.89, 2.00)	1.11 (1.01, 1.20)	-0.07 (-0.12, -0.02)	-0.31 (-0.42, -0.19)
Long-term mental health problem	-0.47 (-0.54, -0.39)	-0.68 (-1.02, -0.35)	-0.26 (-0.34, -0.18)	-0.62 (-0.83, -0.40)	0.35 (0.28, 0.41)	-0.12 (-0.24, 0.01)	-0.58 (-0.64, -0.52)	-0.28 (-0.39, -0.18)	2.24 (2.17, 2.31)	1.90 (1.79, 2.01)

Independent variable	Log odds for mobility - Base: I have no problems in walking about		Log odds for Self-Care - Base: I have no problems with self-care		Log odds for usual activity-Base: I have no problems with performing my usual activities		Log odds for pain - Base: I have no pain or discomfort		Log odds for anxiety - Base: I am not anxious or depressed	
Long-term neurological problem	1.15 (1.07, 1.23)	1.03 (0.73, 1.33)	0.78 (0.69, 0.87)	0.48 (0.28, 0.69)	1.11 (1.03, 1.19)	0.53 (0.40, 0.65)	0.83 (0.75, 0.91)	0.65 (0.54, 0.76)	0.21 (0.14, 0.28)	-0.23 (-0.37, -0.09)
Another long-term condition	0.06 (0.00, 0.12)	0.32 (0.03, 0.61)	-0.04 (-0.11, 0.03)	-0.06 (-0.26, 0.13)	0.21 (0.15, 0.26)	0.09 (-0.02, 0.20)	0.23 (0.18, 0.27)	0.31 (0.22, 0.41)	-0.05 (-0.10, -0.01)	-0.21 (-0.32, -0.09)
Part-time paid work	0.06 (0.02, 0.10)	-0.13 (-0.62, 0.37)	0.34 (0.26, 0.43)	-0.18 (-0.63, 0.27)	0.09 (0.05, 0.12)	-0.19 (-0.37, -0.02)	0.07 (0.03, 0.10)	0.16 (0.08, 0.24)	0.15 (0.11, 0.18)	0.17 (0.06, 0.27)
Full-time education at school, college	0.72 (0.57, 0.86)	2.06 (1.39, 2.72)	1.36 (1.17, 1.56)	1.96 (1.41, 2.50)	0.71 (0.60, 0.82)	0.75 (0.28, 1.23)	0.08 (-0.02, 0.18)	-0.18 (-0.46, 0.10)	0.16 (0.06, 0.26)	0.29 (0.05, 0.52)
Unemployed	1.07 (1.02, 1.12)	1.28 (0.84, 1.72)	1.66 (1.58, 1.74)	1.47 (1.12, 1.82)	1.03 (0.99, 1.08)	1.53 (1.39, 1.67)	0.47 (0.42, 0.51)	0.89 (0.81, 0.98)	0.96 (0.92, 1.01)	1.50 (1.41, 1.59)
Permanently sick or disabled	2.91 (2.86, 2.96)	3.01 (2.70, 3.32)	3.58 (3.51, 3.64)	3.49 (3.23, 3.76)	3.13 (3.07, 3.19)	3.19 (3.08, 3.30)	1.55 (1.49, 1.61)	2.40 (2.34, 2.46)	1.44 (1.40, 1.48)	2.08 (2.00, 2.16)
Fully retired from work	0.53 (0.49, 0.56)	0.37 (0.04, 0.70)	1.29 (1.22, 1.35)	1.14 (0.87, 1.41)	0.47 (0.43, 0.50)	1.12 (1.01, 1.23)	0.17 (0.14, 0.20)	0.76 (0.70, 0.82)	0.18 (0.15, 0.21)	0.56 (0.47, 0.65)
Looking after the home	0.46 (0.41, 0.51)	0.15 (-0.38, 0.67)	0.68 (0.59, 0.77)	0.18 (-0.25, 0.60)	0.46 (0.42, 0.51)	0.13 (-0.05, 0.32)	0.25 (0.21, 0.30)	0.43 (0.34, 0.52)	0.42 (0.38, 0.46)	0.66 (0.54, 0.78)
Doing something else	0.37 (0.30, 0.45)	0.51 (-0.15, 1.18)	1.09 (0.97, 1.21)	1.10 (0.68, 1.52)	0.45 (0.38, 0.52)	0.97 (0.77, 1.18)	0.19 (0.12, 0.26)	0.42 (0.29, 0.56)	0.39 (0.32, 0.46)	0.96 (0.81, 1.12)
Former smoker	0.19 (0.17, 0.21)	-0.17 (-0.31, -0.03)	0.10 (0.07, 0.13)	-0.31 (-0.39, -0.22)	0.13 (0.11, 0.16)	-0.02 (-0.06, 0.03)	0.17 (0.15, 0.19)	0.06 (0.02, 0.09)	0.06 (0.04, 0.08)	-0.03 (-0.09, 0.02)
Occasional smoker	0.39 (0.35, 0.44)	-0.51 (-0.81, -0.21)	0.24 (0.18, 0.29)	-0.51 (-0.69, -0.33)	0.36 (0.32, 0.40)	0.04 (-0.04, 0.12)	0.28 (0.24, 0.32)	0.26 (0.20, 0.32)	0.34 (0.30, 0.38)	0.29 (0.20, 0.37)
Regular smoker	0.43 (0.39, 0.46)	-0.23 (-0.44, -0.01)	0.16 (0.11, 0.20)	-0.53 (-0.67, -0.39)	0.41 (0.38, 0.44)	0.04 (-0.03, 0.10)	0.32 (0.29, 0.35)	0.26 (0.21, 0.31)	0.37 (0.34, 0.40)	0.44 (0.38, 0.50)
Irish	-0.14 (-0.23, -0.05)	-0.52 (-1.24, 0.20)	0.04 (-0.07, 0.15)	0.26 (-0.04, 0.56)	-0.20 (-0.29, -0.12)	0.32 (0.18, 0.47)	0.03 (-0.06, 0.12)	0.02 (-0.11, 0.16)	0.09 (0.01, 0.18)	0.02 (-0.18, 0.23)
Gypsy or Irish Traveller	0.41 (-0.19, 1.02)	2.01 (0.54, 3.47)	0.48 (-0.16, 1.13)	0.90 (-0.64, 2.44)	-0.18 (-0.76, 0.39)	0.56 (-0.22, 1.35)	0.03 (-0.54, 0.61)	1.10 (0.50, 1.70)	0.67 (0.14, 1.20)	0.50 (-0.25, 1.24)
Any other White background	-0.08 (-0.14, -0.02)	0.33 (-0.01, 0.66)	0.07 (-0.01, 0.15)	0.47 (0.27, 0.67)	-0.14 (-0.19, -0.08)	0.10 (-0.01, 0.22)	0.12 (0.06, 0.17)	0.04 (-0.05, 0.13)	0.05 (0.01, 0.11)	0.05 (-0.07, 0.17)
White and Black Caribbean	-0.17 (-0.43, 0.10)	0.82 (-0.34, 1.97)	-0.27 (-0.62, 0.09)	0.17 (-0.83, 1.17)	-0.03 (-0.26, 0.21)	0.36 (-0.10, 0.81)	0.10 (-0.12, 0.32)	0.36 (0.02, 0.69)	0.07 (-0.15, 0.29)	0.11 (-0.27, 0.49)
White and Black African	0.20 (-0.13, 0.53)	-11.94 (-874, 851)	0.04 (-0.44, 0.51)	-0.79 (-2.79, 1.21)	0.19 (-0.11, 0.50)	0.39 (-0.24, 1.03)	0.46 (0.14, 0.77)	0.10 (-0.39, 0.59)	0.00 (-0.29, 0.30)	0.34 (-0.18, 0.87)
White and Asian	0.07 (-0.19, 0.33)	-0.06 (-2.04, 1.90)	0.50 (0.16, 0.83)	1.15 (0.41, 1.90)	0.00 (-0.23, 0.24)	0.04 (-0.54, 0.62)	0.12 (-0.10, 0.34)	0.45 (0.08, 0.82)	0.06 (-0.17, 0.28)	0.47 (0.06, 0.89)
Any other Mixed / multiple ethnic	-0.19 (-0.46, 0.07)	0.22 (-1.22, 1.67)	-0.03 (-0.39, 0.33)	0.31 (-0.61, 1.23)	-0.14 (-0.38, 0.10)	0.24 (-0.25, 0.73)	0.11 (-0.11, 0.33)	0.34 (-0.01, 0.68)	0.16 (-0.06, 0.37)	0.41 (0.01, 0.80)
Indian	0.22 (0.15, 0.29)	0.75 (0.42, 1.09)	0.49 (0.40, 0.58)	0.56 (0.33, 0.78)	0.10 (0.03, 0.17)	0.22 (0.08, 0.36)	0.32 (0.25, 0.38)	0.23 (0.12, 0.34)	0.01 (-0.06, 0.07)	0.08 (-0.09, 0.25)
Pakistani	0.34 (0.23, 0.44)	1.02 (0.63, 1.42)	0.78 (0.65, 0.90)	1.01 (0.76, 1.27)	0.40 (0.30, 0.49)	0.57 (0.40, 0.74)	0.57 (0.47, 0.67)	0.44 (0.30, 0.57)	0.14 (0.05, 0.22)	0.32 (0.14, 0.50)
Bangladeshi	0.48 (0.31, 0.65)	1.12 (0.42, 1.81)	0.80 (0.59, 1.01)	1.48 (1.08, 1.88)	0.38 (0.23, 0.54)	0.58 (0.28, 0.89)	0.74 (0.57, 0.90)	0.40 (0.16, 0.64)	0.23 (0.08, 0.38)	0.17 (-0.15, 0.49)
Chinese	-0.51 (-0.72, -0.30)	0.16 (-1.04, 1.37)	-0.38 (-0.73, -0.04)	0.33 (-0.43, 1.09)	-0.56 (-0.76, -0.37)	-0.29 (-0.83, 0.24)	0.00 (-0.15, 0.16)	-0.91 (-1.41, -0.41)	-0.16 (-0.32, 0.01)	-0.17 (-0.65, 0.30)

Independent variable	Log odds for mobility - Base: I have no problems in walking about		Log odds for Self-Care - Base: I have no problems with self-care		Log odds for usual activity-Base: I have no problems with performing my usual activities		Log odds for pain - Base: I have no pain or discomfort		Log odds for anxiety - Base: I am not anxious or depressed	
Any other Asian background	0.08 (-0.03, 0.18)	0.72 (0.20, 1.24)	0.36 (0.22, 0.50)	0.93 (0.62, 1.23)	-0.07 (-0.17, 0.02)	0.20 (-0.01, 0.42)	0.25 (0.16, 0.34)	0.18 (0.02, 0.34)	0.04 (-0.05, 0.13)	0.09 (-0.12, 0.31)
African	-0.19 (-0.30, -0.07)	0.72 (0.10, 1.34)	-0.20 (-0.39, -0.02)	0.26 (-0.24, 0.76)	-0.40 (-0.52, -0.29)	-0.14 (-0.43, 0.15)	0.25 (0.15, 0.34)	0.31 (0.14, 0.48)	-0.25 (-0.35, -0.15)	-0.05 (-0.29, 0.18)
Caribbean	-0.10 (-0.20, 0.01)	0.41 (-0.20, 1.03)	0.03 (-0.11, 0.17)	0.25 (-0.13, 0.64)	-0.13 (-0.23, -0.03)	-0.05 (-0.27, 0.17)	0.48 (0.39, 0.58)	0.41 (0.26, 0.56)	-0.16 (-0.26, -0.06)	-0.22 (-0.47, 0.03)
Any other Black/African/Caribbean	0.15 (-0.01, 0.31)	1.27 (0.62, 1.92)	-0.23 (-0.47, 0.00)	0.37 (-0.21, 0.96)	0.03 (-0.12, 0.19)	0.08 (-0.25, 0.41)	0.57 (0.41, 0.73)	0.48 (0.26, 0.70)	0.00 (-0.15, 0.15)	0.01 (-0.32, 0.33)
Arab	0.18 (-0.07, 0.44)	0.86 (-0.05, 1.78)	0.58 (0.29, 0.87)	1.07 (0.52, 1.62)	0.14 (-0.10, 0.38)	0.46 (0.08, 0.85)	0.48 (0.23, 0.72)	0.47 (0.16, 0.77)	0.25 (0.03, 0.46)	0.60 (0.25, 0.96)
Any other ethnic group	0.11 (0.03, 0.18)	1.22 (0.93, 1.50)	0.33 (0.23, 0.42)	0.95 (0.75, 1.15)	-0.10 (-0.17, -0.03)	0.33 (0.20, 0.46)	0.21 (0.15, 0.28)	0.42 (0.32, 0.52)	0.12 (0.05, 0.18)	0.27 (0.13, 0.40)
_cons	-4.06 (-4.13, -3.99)	-8.24 (-8.67, -7.80)	-5.19 (-5.29, -5.09)	-7.83 (-8.16, -7.49)	-2.78 (-2.84, -2.72)	-6.18 (-6.33, -6.02)	-1.60 (-1.65, -1.55)	-4.52 (-4.64, -4.41)	-1.41 (-1.47, -1.36)	-4.15 (-4.29, -4.01)

Table 16: Log odds of full model