Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek
Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt

2011
2012

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Assessing the health related quality of life in the general population

Promotor :
Prof. dr. Niel HENS

Promotor :
Prof. PH. BEUTELS
Dr. JOKE BILCKE

Lucas Malla

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

2011
2012

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Assessing the health related quality of life in the general population

Promotor :
Prof. dr. Niel HENS

Promotor :
Prof. PH. BEUTELS
Dr. JOKE BILCKE

Lucas Malla

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Declaration

This is to certify that this report was written by Lucas Malla under our supervision.

Signature................................... Date.......................

Prof.dr. Niel Hens Internal Supervisor

Signature.................................... Date.......................

Prof. dr. Beutels Phillipe External Supervisor

Signature................................... Date.......................

dr. Joke Bilcke External Supervisor

Signature.................................. Date.......................

Lucas Malla Student

Thesis submitted in partial fulfilment of the requirements for the degree of Master of Statistics: Biostatistics

Acknowledgement

Successful completion of this thesis has been through the support of a number of individuals. First of all, I acknowledge my supervisors; dr. Joke Bilcke, Prof. dr. Beutels Philippe and Prof. dr. Niel Hens who continually guided me through this work, by supplying thoughtful suggestions for improving this report.

I am deeply grateful to the Flemish Interuniversity Council (VLIR) for the scholarship which has enabled me to pursue this valuable Master program.

I acknowledge all my lecturers as well for they have made a (bio)statistician in me.

Most of all, I thank God for giving me wisdom, strength and life to successfully complete this program.

Lucas Malla
University of Hasselt, Belgium, September, 2012

List of Abbreviations

Table 1: Abbreviations

AIC	-	Akaike Information Criteria
CDC	-	Center for Disease Control
EQ5D	-	Euroqol 5 Dimensions
HRQoL	-	Health Related Quality of Life
Lasso	-	Least Absolute Shrinkage and Selector Operator
VAS	-	Visual Analogue Scale
WHO	-	World Health Organisation

Abstract

Background

Health refers to an individual's mental, physical and social well-being. Every individual strives to improve his/her quality of health despite the odds like diseases that are quite often encountered in our daily life. The actual feelings such as stress, pain, anxiety experienced by an individual are used to quantify and rate quality of health. Therefore, special instruments widely used to measure quality of life include Visual Analogue Scale(VAS) and EQ5D.

Objectives

A survey was conducted in the general population residing in Flanders (Belgium) in order to determine and explain their HRQoL. In particular, it was of interest to identify factors that were significantly associated to their HRQoL, and also develop a statistical model to explain the relationship between these factors and HRQoL.

Methodology

Since there were many covariates from the study; regression trees, random forest and lasso regression were used as preliminary tools to reduce the number of variables. Thereafter, relationships between the responses and these factors were modelled using beta regression and one inflated beta regression (for VAS and EQ5D outcomes respectively). Linear predictors of these models were extended to polynomials and fractional polynomials.

Results and Conclusions

Age was significantly associated to the responses and was found to be a major factor in explaining Health Related Quality of Life for individuals in Flanders. Moreover, individuals who had suffered severe illnesses before had significantly lower HRQoL as compared to corresponding individuals of the same age who had not suffered any severe illnesses before.

Keywords: beta regression, EQ5D, fractional polynomials, HRQoL, one inflated beta regression, polynomials, VAS.

Contents

List of Tables

List of Figures

1 Introduction

The World Health Organisation's constitution (1948) defines health as : 'A state of complete physical, mental and social well-being and not merely the absence of disease or infirmity '. While Health Related Quality of Life (HRQoL) - refers to - wholesome quality of life with respect to physical and mental health, which can be influenced over time by diseases, ageing process among other risk factors. It is also possible to understand this concept at a population level, and can be viewed as a population's health and functional status influenced by policies, conditions and resources (WHO, 1948).

A number of factors such as education, technological and medical advancements in the modern society are perceived to have contributed both positively and negatively to quality of health in a population, but these merits and demerits are debatable. Considering merits of such advancements, it is true that research conducted in medical science have resulted to cures of chronic diseases which have given a boost to quality of health in the general population (Singh & Dixit, 2010).

Aspects of health such as pain, anxiety/depression among others are considered important. These health aspects are experienced by an individual and may be quite difficult to measure externally with an instrument. Therefore, it is increasingly becoming acceptable in clinical and health services research that information concerning health state be provided by an individual himself/herself , and not provided on his/her behalf so that the actual experience is captured (Brazier, Ratcliffe, Salomon & Tsuchiya, 2007).

Surveillance of a population's health have become a common and important practice among governments and health agencies since 1980s. It forms valid scientific evidence and basis under which health improvements and interventions are carried out. Particularly, captured HRQoL measures help in determining burden of diseases which are preventable, disabilities and might result to valuable information concerning association between risk factors and HRQoL(CDC). Moreover, the HRQoL data facilitates the identification of subgroups with poorer health status and this may point out the need of curbing serious subsequent eventualities in such particular subgroups.

HRQoL study results may help in enforcing health policy needs, allocation of resources, formulation of best health strategic plans and monitor intervention mechanisms and their effectiveness. In health economics and epidemiology, HRQoL is used to estimate Quality Adjusted Life Years achieved through medical interventions. These are used in turn in cost effectiveness analyses to inform policy makers on how to prioritize between different interventions in health care (CDC).

1.1 Measurement of HRQoL

Presented here are two generic instruments which were used to measure the HRQoL in Flanders. Studies concerning HRQoL have resulted to instruments that are able to detect minimal effects that are important in clinical trials and for investigating the quality of health in a population, the most important being EuroQoL 5 Dimensions (EQ5D) and the Visual Analogue Scale (VAS), (Guyatt, Feeny & Patrick, 1993).

1.1.1 EuroQol 5 Dimensions (EQ5D)

This instrument was developed by international collaboration of health researchers from Europe (EuroQoL). The group was established to design non-disease specific , simple , comprehensive and standardized quality of health measure to be used by the entire European community (EuroQol Group, 1990) . EQ5D captures health profiles and also generate health utilities. It has five questions referred to as dimensions, each expressing a given health dimension : mobility, capability to carry out daily activities, pain, anxiety/ depression and self -care. Each question has three possible responses which yield different possible states for health (see appendix table 8). The instrument has different modes of administration : telephone interviews, face to face, proxies, self-complete et cetera (Frayback, 2010).
Cleemput(2010) derived a preference valuation set for EQ5D Health states from the general population in Flanders. This was expressed as: $EQ5D = 1-(0.152+0.074MO+0.083SC+0.031UA+0.084PD+0.103AD+0.253N3)$,where : MO-Mobility, SC-Selfcare, UA-Daily

Activities, PD-Pain/Discomfort, AD- Anxiety/Depression , N3-penalty term if any of the dimensions has a score of 3. Each of the scores is down weighed by subtracting 1. For instance if an individual filled 22113 in the EQ5D questionnaire (shown in appendix table 8) , then that person had an EQ5D score equal to:$1 - (0.152 + (0.074 * 1) + (0.083 * 1) + (0.031 * 0) + (0.084 * 0) + (0.103 * 2) + 0.253) = 0.485$. The intercept in this algorithm was interpreted as a decrement due to any move from perfect health. Therefore, 0.152 was not subtracted for individuals who filled 11111 since they had perfect health. However, negative scores are also possible which would represent an unimaginable health condition worse than death.

1.1.2 Visual Analogue Scale (VAS)

VAS refers to a psychometric scale used to measure response , and this can be included in a questionnaire. It is also known as thermometer scale whose range lies between 0 ('worst imaginable health state ') and 100 ('best imaginable health state ') both inclusive. This instrument has been widely used to collect HRQoL data due to its simplicity and practical applicability (Shmueli, 2005). It is also noted that VAS results to high response rate and high levels of completion (Brazier, Ratcliffe, Salomon & Tsuchiya, 2007).

1.2 Description of Survey Data

This survey was conducted in a random sample of 2204 individuals of all ages (22.38% relates to children less than 12 years, 61.17% relates to adults and 16.45% to elderly individuals of 60 years and above). For a subsample of 430 individuals (19.5%), information was collected for all members of their household whereas the other 1777 respondents each belonged to a unique household. There were also specific questions which were asked to each of three age categories (Children, Adults and Elderly).

The mode of data collection was through self-administered questionnaires, where these questionnaires were sent by post to selected individuals to participate in the study and sampling was done through random digit dialing. See below and appendix table 10 for detailed description of the measured variables.

1.3 Response

Respondents were asked to fill in both the EQ5D and the VAS. Health utility scores were calculated based on Cleemput algorithm assuming these utilities were additive. VAS instrument gave a single overall score for the HRQoL. Hence in total two different response variables are considered.

1.4 Covariates

The following general socio-demographic factors were considered in the exploratory analysis in order to examine their impact on an individual's HRQoL ; age, gender, household size, previous sickness status, nationality, province, number of animals kept, number of parents. Besides, the data presented additional sub group covariates as follows: (1) Children : mother's education (2) Adults : smoking status, profession, education level, whether the adults worked/had worked for a health care facility and (3) Elderly : alcohol consumption frequency, frequencies children and grandchildren visited them, work status, whether the elderly had worked for a health care facility, education level and smoke status.

1.5 Objectives

i. To determine which covariates are significantly associated to either or both HRQoL outcome measures.
ii. To develop a statistical model describing the relationship between characteristics of respondents and their HRQoL experience.

2 Methods and Results

2.1 Variable Selection Methods

A combination of techniques was used to explore the response and possible covariates from the data, including: graphical techniques (boxplots, scatterplots and histograms), regression trees, random forest and lasso regression. These methods were used in order to address primary objective 1. Regression trees, random forests and lasso regression were used for selection since there was need to obtain a subset of covariates from the many variables in the data; in which some variables had many levels. Below is a brief overview of the stated techniques.

2.1.1 Regression trees

These are models which predict numeric responses to the given data. Predictions are based on decisions from the root node and down to subsequent daughter nodes. Suppose data contains a response y_i and t inputs x_i for each observation; that is to say, (x_i, y_i) for i=1,. . . ,N ,with $x_i = (x_{i1}, \ldots, x_{it})$, then the algorithm decides automatically on the split points , splitting variables and shape of the tree, this may be based on variable that gives minimum impurity. The response is modeled as a constant c_m in each of the partitioned M regions $(R_1, \ldots, R_M) : f(x) = \sum_{m=1}^{M} c_m I\{x \epsilon R_m\}$. Best estimate $\widehat{c}_m$ for least square criterion is the average of y_i in region R_m: $c_m = (y_i|x_i \epsilon R_m)$, (Hastie, Tibshirani & Friedman, 2009).

Regression trees represent information in a way that is intuitive and easily visualized. Major advantages of using this technique include: insensitivity to outliers, automatic modeling of interactions as implied by the hierarchical structure of the trees, and the fact that model outcomes are not affected by different scales and transformations of predictors (Elith, Leathwick & Hastle, 2008).

2.1.2 Random forest

This algorithm uses an ensemble of regression trees (Breiman, 2001). It implements bootstrapping approach for data sampling. Based on the bootstrapped data samples, each of the regression trees is generated and at each split the candidate set of variables is a random subset of the variables; which as well can be specified (Diaz-Uriarte & Andres, 2005). The correlation between trees is reduced without too much increment in the variability, which is made possible by the process of growing trees through randomly selecting the input variables. These trees are allowed to grow fully in order to obtain trees with low bias (Hastie, Tibshirani & Friedman, 2009) . The algorithm results to an ensemble obtained through averaging over lowly correlated trees with low bias.

For this analysis, a list was generated from random forest showing importance of variables based on node impurity and mean square error. The importance measures show how much impurity or mean square error increase when that variable is randomly permuted. If a variable is permuted and predictions done, then some change in impurity or mean square error will be observed, otherwise no much change will be noticed.

2.1.3 Least Absolute Shrinkage and Selector Operator (Lasso) Regression

Lasso is considered as a shrinkage technique and a selection method in linear regression analysis. Some coefficients are shrunken and others set to zero under this regression method, thus interplaying between ridge regression and subset selection (Tibshirani et al., 1996). The lasso estimate expressed in Lagrangian form is defined by :
$\hat{\beta}^{lasso} = \overset{\text{argmin}}{\beta} \quad \{\frac{1}{2}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|\}$. It minimizes the usual error sum of squares, with a restricted sum of the absolute coefficients (Hastie, Tibshirani & Friedman, 2009). Tibshirani (1996) states that both ridge regression and subset selection methods for improving the ordinary least squares have limitations. Subset selection can be extremely variable since small changes in the data can result to different selected models thus lowering its prediction accuracy . While ridge regression does not set any coefficient to zero hence resulting to a model which is not easily interpretable.

2.2 Variable Selection Results

2.2.1 Exploratory of the response (HRQoL)

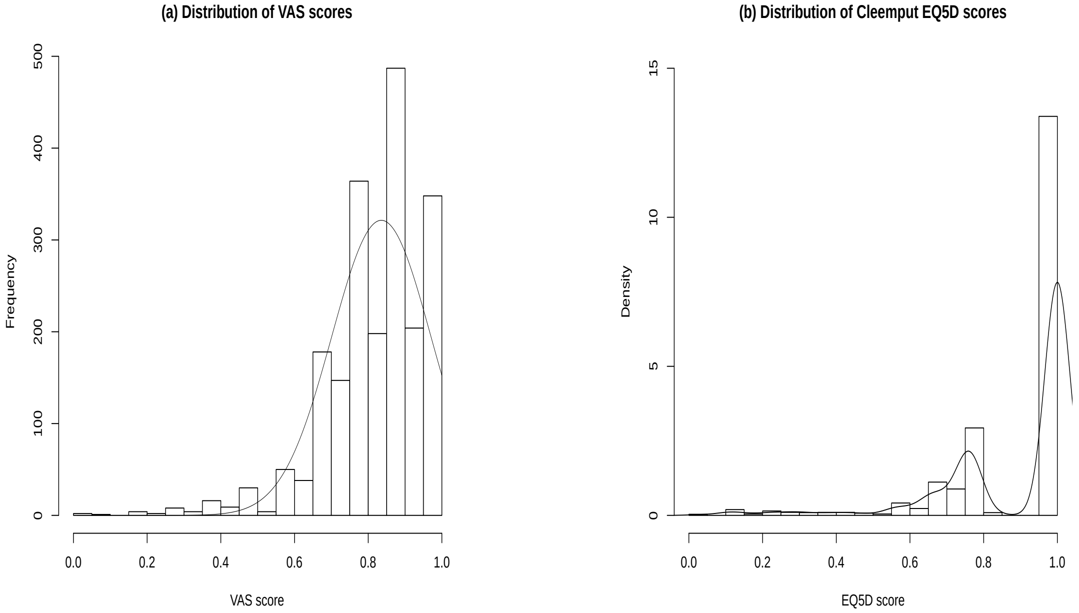

Figure 1: *Distribution of VAS and EQ5D outcomes respectively*

The distribution of HRQoL as captured by Visual Analogue Scale and Cleemput EQ5D showed high negative skewness (Figures 1 (a) and (b)). Highest peaks observed at 1 for the EQ5D criterion and at 0.8 to 1 for VAS. The distribution of VAS scores showed a unimodal density with one local maximum, whereas a gap was observed in the histogram of EQ5D outcomes with more than one local maxima. Thus implying existence of two sub populations, with one sub population dominated by HRQoL scores of 1. Similar distribution patterns of the response (HRQoL) exhibited in figures 1 (a) and (b) pointed out an underlying concordance between VAS and EQ5D (Cleemput) measurement instruments. Concordance correlation between these two techniques was estimated to be 51.17% with a 95% confidence interval of [0.4831, 0.5393], this indicated a moderate concordance as pertains to precision and accuracy of measurements.

2.2.2 Exploratory and Selection of Explanatory Variables

The regression trees considering VAS and EQ5D as response measures similarly selected two variables at the population level (Age and disease1), and some of the few splits shown were quite sensible since health related quality of life of an older respondent who had suffered from a severe illness before was lower than HRQoL of a respondent of the same age who had not suffered any severe illness (figure 2 (a) and (b)). Additionally, variable importance plots considered age and disease1 to be very important since poor predictions would be yielded without them in the model and both explained much of the reduction in node impurity (appendix figure 15).

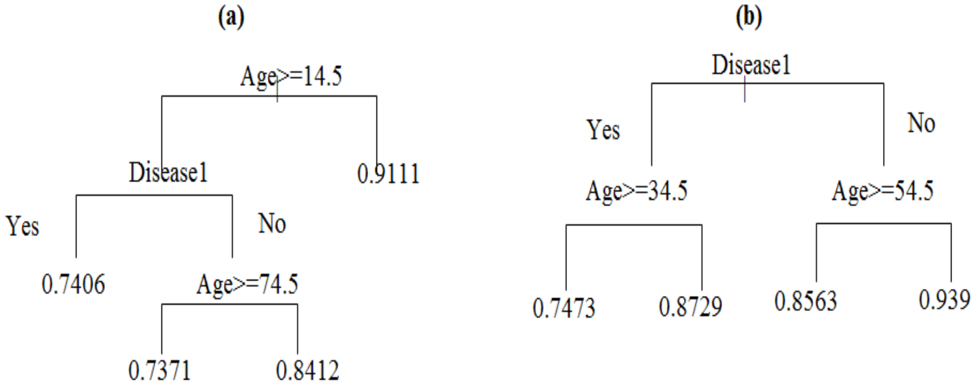

Figure 2: *Regression trees for VAS and EQ5D outcomes respectively*

Lasso regression assigned higher absolute coefficients to age and disease1, while other covariates had lower coefficients in each of the response cases (table 2). Generally, the three methods selected two mentioned covariates to be possible factors which would influence quality of life at the population level, and their association with the responses will further be investigated in the subsequent sections. Even though gender was accorded the least weight and thus dropped by these three techniques, a study by Kirchengast & Haslinger (2008) showed there could be possible gender differences in HRQoL among healthy aged and elderly in Austria, thus it was deemed appropriate to use it as a covariate in this analysis.

Table 2: Lasso regression coefficient estimates

Effect	Coefficient (VAS)	Coefficient (EQ5D)
Age	-0.0272	-0.0258
Disease1	-0.0761	-0.1181
Disease2	-0.0235	-0.0155
Gender	-0.00000	.
Householdsize	0.0027	0.0028
Parents	.	.
Animal	-0.0082	-0.0063
Normalday	-0.0148	-0.0006
Belgian	0.0232	.
Another EU country	.	.
Antwerp	.	.
Limburg	.	0.0092
Vlaams.Brabant	-0.00000	0.00004
West.Vlaandren	.	-0.005

Further trends were investigated for population level covariates (age and disease1) using scatterplots. Figures 3(a) and 4(a) are in agreement with results obtained using regression trees that individuals who had suffered severe illnesses before seemed to have lower health related quality of life as compared to corresponding individuals who had not suffered any severe illnesses. Gender on the other hand seem not to have any influence on VAS outcomes (figure 4(b)), while some pattern is observed for EQ5D outcomes as many males have HRQoL scores of 1(figure 4(b)).

Age is shown as a major predictor by all the techniques for variable selection under population covariates. The raw scatterplots in figures 3 and 4 do not show a clear trend of the health scores across the ages. Therefore, averages of VAS and EQ5D health scores were plotted by ages and loess curves fitted to the averages in order to clearly show trend and possible effect of age (figures 5 (a) and (b)). HRQoL decreases across the ages, younger respondents seem to have higher health scores compared to older respondents who have considerably lower health scores. Individuals above 60 years of age had a rapid decline in their HRQoL scores.

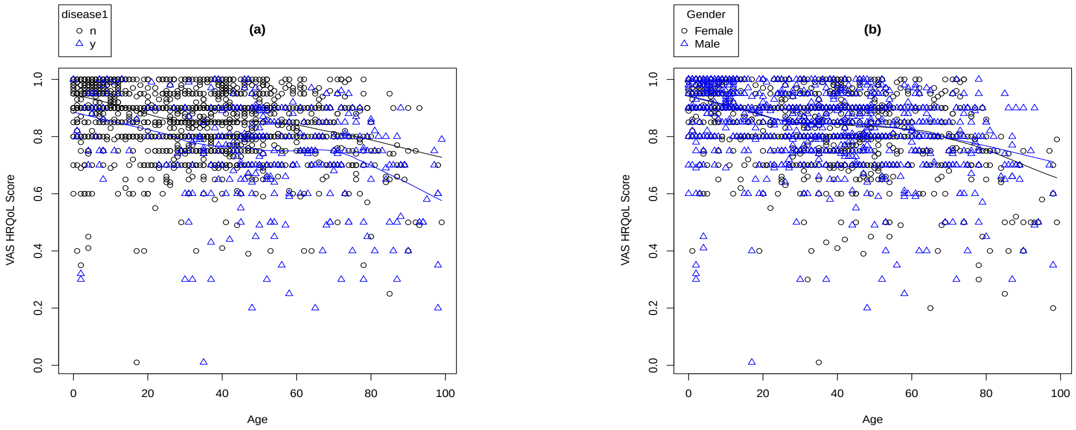

Figure 3: *Scatterplots for raw VAS HRQoL Scores vs. Age by Disease1 and Gender respectively*

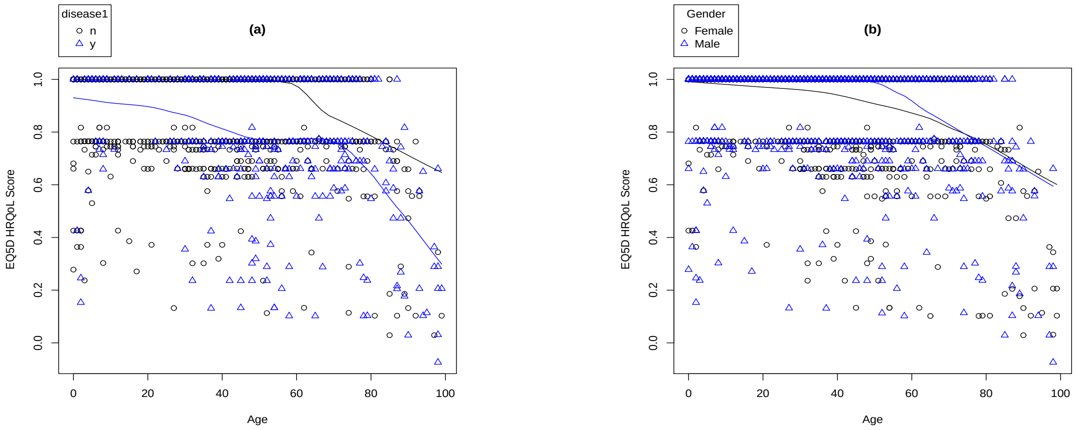

Figure 4: *Scatterplots for raw EQ5D HRQoL Scores vs. Age by Disease1 and Gender respectively*

Different plotted points in figures 5 (a) and (b) represented number of individuals replicating for a particular age, whose HRQoL scores were averaged. It is observed that for both VAS and EQ5D, most of the individuals considered for the analysis were between 0 and 60 years old.

Boxplots of EQ5D and VAS as a function of every covariate were made, and most of them did not seem to affect EQ5D and VAS (not included in the report). See appendix figures 11, 12 and 13 showing how the HRQoL scores varied by age categories, disease1 and gender.

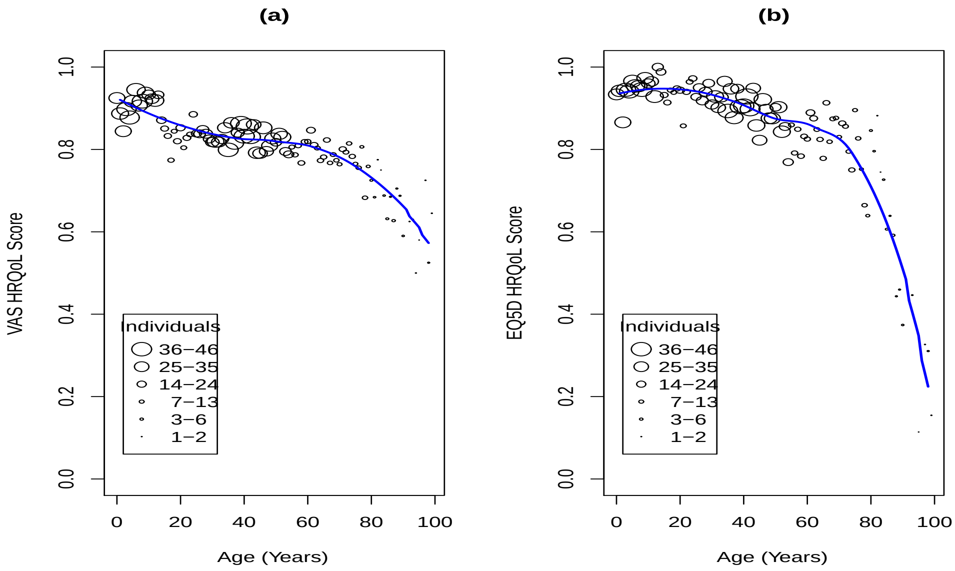

Figure 5: *Averages of VAS and EQ5D scores vs. age*

2.3 Statistical Modelling of VAS and EQ5D HRQoL outcomes : Methods

Modelling techniques discussed in subsections 2.3.1 and 2.3.2 below were deemed suitable to address primary objective 2.

2.3.1 Beta Regression

The Beta distribution is a continuous probability distribution that offers high flexibility to accommodate densities with varying skewness. This flexibility allows for the estimation of distributions with intractable skewness, making normalizing transformations impossible. Furthermore, the support for the Beta distribution lies within (0, 1) interval, making it suitable for modeling proportions, percentages and any form of continuous outcome that lies within the (0,1) interval.

If a dependent variable is presumed to follow beta distribution, then beta regression can be derived by expanding the generalized linear model (GLM) to regress predictor variables on that dependent variable (Swearing, Castro & Bursac, 2011). Paolin (2001) explains that beta regression provides efficient and more accurate parameter estimates when compared to ordinary least squares regression; when the underlying distribution for the dependent variable is skewed or when there is existence of heteroskedasticity. A random variable Y follows a beta distribution with parameters p,q >0, with the density denoted as : $f(y;p,q) = \frac{\Gamma(p+q)}{\Gamma p \Gamma q} y^{(p-1)}(1-y)^{q-1}$, $y \epsilon (0,1)$,where$\Gamma(.)$ is a gamma function and $E(Y) = \frac{p}{p+q}$ and $Var(Y) = \frac{pq}{(p+q)^2(p+q+1)}$. Since this distribution is characterized by two parameters p and q, it was necessary to re-parameterize it to allow the expression of this distribution in terms of its mean and scaling(precision) parameter. Therefore, Ferrari and Cribari-Neto (2004) defined the following re-parameterization; $\mu = \frac{p}{p+q}$ and $\phi = p + q$. Thus if $Y \sim B(\mu, \phi)$ then $E(Y) = \mu$ and $Var(Y) = \frac{\mu(1-\mu)}{(1+\phi)}$, where ϕ is the precision parameter. This allows for inferences to be drawn with respect to changes in the dependent variable's mean and precision (Simas, Souza & Rocha, 2008).

Although beta regression naturally models dispersion, it is necessary to consider further regressors in the dispersion sub model to explain heteroskedasticity (Simas et al., 2010). Thus implying the extension of beta regression to variable dispersion model which jointly models location and dispersion parameters. The VAS HRQoL scores ranged from 0 to 100 both inclusive. These were rescaled by dividing each of the scores by 100 in order to fit in the support of beta distribution. Beta regressions with and without dispersion covariates (as formulated in appendix table 9) were considered for analysis of VAS quality of health scores.

Maximum likelihood estimates from skewed distributions are not necessarily unbiased (Smithson & Verkuilen, 2005). Likelihood estimation degenerates at the boundaries for beta distribution and some corrections should be done for observations lying at the boundaries. Smithson & Verkuilen (2005) suggested proportional shrinkage of the outcome range to a sub-range nearly covering the unit interval or addition of a small value to 0 and subtracting the same from 1. Both methods are likely to bias the estimates towards no effect. The latter technique was used in this analysis and as such, there was need to validate the obtained estimates. Simas et al. (2010) derived general formulae for second-order biases of the maximum likelihood estimators and used them to define bias-corrected and reduced estimators for beta regression variable dispersion model. These were implemented together with non-parametric bootstrapping technique to investigate possible biases in the VAS outcome model.

2.3.2 One inflated Beta Regression

Data measured on a continuous scale between 0 and 1 (both inclusive) quite often contain non negligible number of observations at either 0 or 1 or at both points. Health related quality of life data almost always result to left skewed distribution , with a possibility of inflated observations at 1. This is indeed the case for the EQ5D survey data. As such it is worth accounting for this scenario in analysis. However, the beta distribution discussed in subsection 2.3.1 does not allow positive probability for observations lying at the boundaries. This might not be appropriate for modeling one inflated data since it does not satisfactorily model the entire outcome space and does not take into account the influence of the concentrated probability mass at 1.

Ospina & Ferrari (2007) considered a modeling approach using a mixture of continuous and discrete distributions. These two distributions result to a one inflated beta distribution which sufficiently models the entire outcome space without any adjustments to the original response. Therefore, mixture density is defined by:

$$f(y;\alpha,\mu,\phi) = \begin{cases} (1-\alpha)f(y;\mu,\phi), \text{if } 0<y<1 \\ \alpha, \qquad\qquad\qquad \text{if } y = c \end{cases}$$

where α is the mixing parameter which accounts for probability mass at 1 , $f(y;\mu,\phi)$ is the beta density and c=1. The mean and variance of the response in this case is given by : $E(y) = \alpha c + (1-\alpha)\mu$, which is a weighted average of the mean of a Bernoulli distribution at c=1 and the corresponding mean given by beta distribution with weights α and $(1-\alpha)$ respectively;- while : $Var(y) = (1-\alpha)V(\mu)/(\phi+1) + \alpha(1-\alpha)(c-\mu)^2$.

Inflated number of observations at 1 (shown in figure 1 (b)) motivated modelling of the EQ5D outcomes using a mixture of Bernoulli and beta distributions. First part modelled the ones using a logistic regression, while the second part modelled the outcomes less than one using beta regression. These two sub models were jointly modelled together with the dispersion sub model. Several models as outlined in appendix table 9 were formulated and fitted, with logistic and beta regressions having similar location covariates.

2.3.3 Link Functions

The VAS and EQ5D HRQoL responses lie in the 0 and 1 interval, as such indicating the necessity for link function(s). Therefore, various possible link functions were used; logit, probit and cloglog for location sub models, while log and identity link functions were used for dispersion sub models with and without covariates respectively. Log link was used principally in the dispersion sub models to avoid negative variances.

2.3.4 Polynomials and fractional polynomials

Linear predictors for both VAS and EQ5D outcomes were extended to polynomials and fractional polynomials in order to allow for more functional forms of the responses. Polynomials contain squared and higher order terms of the continuous predictor variable(s), making the response function curvilinear. But this does not interfere with the linearity of the parameters. Royston & Sauerbrei (2008) asserts that parsimonious higher order polynomials (models without over-fitting) result into better fits, but they will frequently deteriorate rapidly outside the range of the data (KUTNER,2005). Instead of using conventional polynomials (only integer powers), one can also use fractional polynomials. Although many different combinations of powers are possible in fractional polynomial models, it has been suggested that it is often adequate to consider only a subset of the powers, S = {-2, -1, -0.5, 0, 0.5, 1, 2, 3 }. This subset provides flexible shapes of curves for most practical model fitting purposes. Curves in this subset include linear, reciprocal, square root, square and logarithmic transformations. If the values of the powers are known, then fitting a fractional polynomial model is similar to fitting a conventional linear regression model. However, these powers are usually unknown and should be estimated from the data. The fractional polynomials differ from the conventional polynomials in that the power(s) can be a non integer number (Cui, et al., 2008).

2.3.5 Model Selection

Models under each response were compared based on Akaike Information Criteria (Akaike,1974) : $AIC = -2L(\widehat{\beta}(R)) + 2K$.

Because of the minus sign in the formula above, smaller AIC values imply better models. The last term on the right hand side is a penalization term for the number of parameters in the model. Likelihood ratio tests were performed to test the need for interactions and as well as inclusion of covariates in the dispersion sub models.

2.3.6 Software

SAS was used for data management and R (betareg and gamlss packages) used for the analyses.

2.4 Statistical Modeling Results

2.4.1 Modelling of VAS HRQoL outcomes:Results

Several models were fitted (as formulated in appendix table 9) and results for their comparisons in terms of AIC and likelihood ratio tests presented in table 3 below.

Table 3: Model comparisons based on AIC and Likelihood Ratio Tests

				polynomials				
Polynomial Order	Model	Model with interactions	Model with Dispersion covariates	AIC			Models compared Based on logit link function	Likelihood Ratio Test (p value)
				Logit	Probit	Cloglog		
1	1	Yes	Yes	-6773.087	-6774.32	-6775.35	(1) vs.(2)	<0.0001
	2	Yes	No	-6691.223	-6697.152	-6702.921	(1) vs.(3)	<0.0001
	3	No	Yes	-6741.802	-6742.034	-6741.366	(3) vs (4)	<0.0001
	4	No	No	-6679.052	-6685.900	-6691.465	(2) vs. (4)	0.0004
2	5	Yes	Yes	-6773.432	-6773.312	-6772.902	(5) vs. (6)	<0.0001
	6	Yes	No	-6708.495	-6709.152	-6709.459	(5) vs. (7)	<0.0001
	7	No	Yes	-6748.689	-6747.674	-6745.710	(7) vs (8)	<0.0001
	8	No	No	-6700.606	-6701.036	-6699.966	(6) vs. (8)	0.0031
3	9	Yes	Yes	-6787.665	-6785.900	-6783.952	(9) vs.(10)	<0.0001
	10	Yes	No	-6711.297	-6708.714	-6706.331	(9) vs.(11)	<0.0001
	11	No	Yes	-6758.620	-6756.233	-6752.760	(11) vs. (12)	<0.0001
	12	No	No	-6710.365	-6707.374	-6703.228	(10) vs.(12)	0.0368

Fractional polynomials								
Fractional Polynomial Degree	Power of age	Model with interactions	Model with Dispersion covariates	AIC			Models compared Based on logit link function	Likelihood Ratio Test (p value)
				Logit	Probit	Cloglog		
	0.5	(1) Yes	Yes	-6773.563	-6773.091	-6772.784	(1) vs.(2)	<0.0001
1	0.5	(2) Yes	No	-6706.393	-6708.569	-6706.393	(1) vs.(3)	<0.0001
	0.5	(3) No	Yes	-6745.117	-6742.54	-6742.535	(3) vs (4)	<0.0001
	0.5	(4) No	No	-6699.653	-6698.659	-6694.694	(2) vs. (4)	0.0014
	(0.5,3)	(5) Yes	Yes	-6776.933	-6775.307		(5) vs. (6)	<0.0001
2	(0.5,3)	(6) Yes	No	-6709.036	-6710.999		(5) vs. (7)	<0.0001
	(0.5,3)	(7) No	Yes	-6748.933	-6747.875		(7) vs (8)	<0.0001
	(0.5,3)	(8) No	No	-6700.049	-6702.674		(6) vs. (8)	0.0068
2	(0,0.5)	(5) Yes	Yes			-6775.910	(9) vs.(10)	<0.0001
	(0,0)	(6) Yes	No			-6710.113	(9) vs.(11)	<0.0001
	(0,0.5)	(7) No	Yes			-6747.094	(11) vs. (12)	<0.0001
	(0,0)	(8) No	No			-6701.665	(10) vs.(12)	0.01167
3	(2,2,3)	(9) Yes	Yes	-6798.743	-6797.754			
		(10)Yes	No	-6727.181	-6725.487			
		(11) No	Yes	-6769.662	-6768.785			
		(12) No	No	-6723.113	-6721.819			
3	(2,3,3)	(9) Yes	Yes				-6798.219	
		(10)Yes	No				-6730.897	
		(11) No	Yes				-6773.021	
		(12) No	No				-6724.304	

Tests based on both polynomial and fractional polynomial models indicated that interactions may be needed. Moreover, the tests for the need of regressors in dispersion sub models under polynomials of order 1, 2 and 3 all resulted to significant likelihood ratio tests with a p-value of <0.0001. This indicated the necessity of incorporating regressors in the dispersion sub models to account for the underlying heterogeneity. Based on AIC values, it can be seen that polynomials and fractional polynomials of order 3 and degree 3 yielded remarkable better fits than models of orders and degrees 1 and 2 ; since they had the least AIC values. Comparatively, selected best models under each link function seemed to fit the data equally well (figures 6 (a) and (b)). Predictions using best fractional polynomial models were possible only for individuals older than zero years.

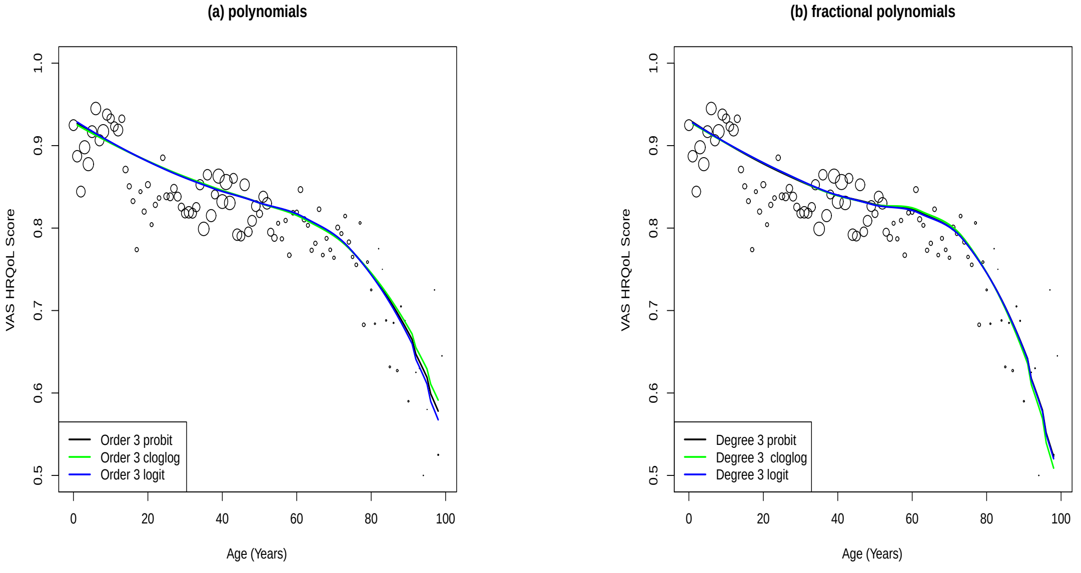

Figure 6: *Graphical representation of the best fits for polynomial and fractional polynomial models under different link functions*

Therefore, logit polynomial of order 3 (model 9: AIC=-6787.665) was selected among the fitted polynomial and fractional polynomial models, since inferences based on it would cover the whole data.

The parameter estimates and associated standard error estimates for the selected model are presented in table 4, together with the p-values based on the Wald Chi-square test.

Table 4: Model Coefficients, Standard Errors and Significance Tests

		Selected Model			Final Model	
Parameter	Coefficients	s.e	p value	Coefficients	s.e	p value
Location submodel						
β_0 (Intercept)	2.7500	0.1933	< 0.0001	2.6970	0.1026	<0.0001
β_1 (A1)	-0.0582	0.0174	0.0008	-0.0502	0.0076	<0.0001
β_2 (A2)	0.0013	0.0004	0.0005	0.0009	0.0002	<0.0001
β_3 (A3)	-0.000009	0.000003	0.0047	-0.000006	0.000002	<0.0001
β_4 (G-Male)	-0.1847	0.2586	0.4750	-0.0071	0.0603	0.9070
β_5 (D-Yes)	-0.9660	0.4200	0.0214	-0.6649	0.0695	<0.0001
β_6 (A1*G)	0.0181	0.2191	0.4066	-	-	-
β_7 (A1*D)	0.0083	0.2886	0.7732	-	-	-
β_8 (A2*G)	-0.0006	0.0005	0.2909	-	-	-
β_9 (A2*D)	-0.0002	0.0006	0.7354	-	-	-
β_{10} (A3*G)	0.000004	0.000004	0.2475	-	-	-
β_{11} (A3*D)	0.000002	0.000004	0.6538	-	-	-
β_{12} (G*D)	0.3086	0.1450	0.0333	-	-	-
Precision submodel						
d_0 (Intercept)	0.8594	0.1985	<0.0001	0.6111	0.0820	<0.0001
d_1 (A1)	-0.0149	0.0185	0.4208	0.0107	0.0016	<0.0001
d_2 (A2)	0.0009	0.0005	0.0646	-	-	-
d_3 (A3)	0.000009	0.000004	0.0138	-	-	-
d_4(G-Male)	-0.2706	0.2667	0.3104	0.1445	0.0725	0.0461
d_5(D-Yes)	0.4919	0.4516	0.2761	0.4517	0.1150	<0.0001
d_6 (A1*G)	0.0268	0.0243	0.2679	-	-	-
d_7 (A1*D)	-0.4004	0.0336	0.2344	-	-	-
d_8(A2*G)	-0.0006	0.6279	0.3181	-	-	-
d_9 (A2*D)	0.0009	0.0007	0.2196	-	-	-
d_{10} (A3*G)	0.000005	0.000004	0.2444	-	-	-
d_{11}(A3*D)	-0.000004	0.000005	0.4293	-	-	-
d_{12} (G*D)	-0.5715	0.1838	0.0018	-0.5490	0.1507	0.0003

Clearly the covariate capturing whether a person had suffered a severe illness before as well as age had significant effects at $\alpha = 5\%$. This observation is noted both in the location and precision sub models. However, gender is significant in the precision sub model and not significant in the location sub model. Even though it does not have an effect in the location sub model, it contributes in accounting for variability in the mean of VAS HRQoL scores.

Figures 7 (a) and (b) show VAS HRQoL predictions based on the final model. A difference in health scores is noted between individuals who had suffered severe illnesses and those who had not, and the difference is larger for older individuals. While there is no notable difference in HRQoL scores between males and females. The prediction confidence band shown in figure 7 (c) is narrower at the beginning and wider at the end.

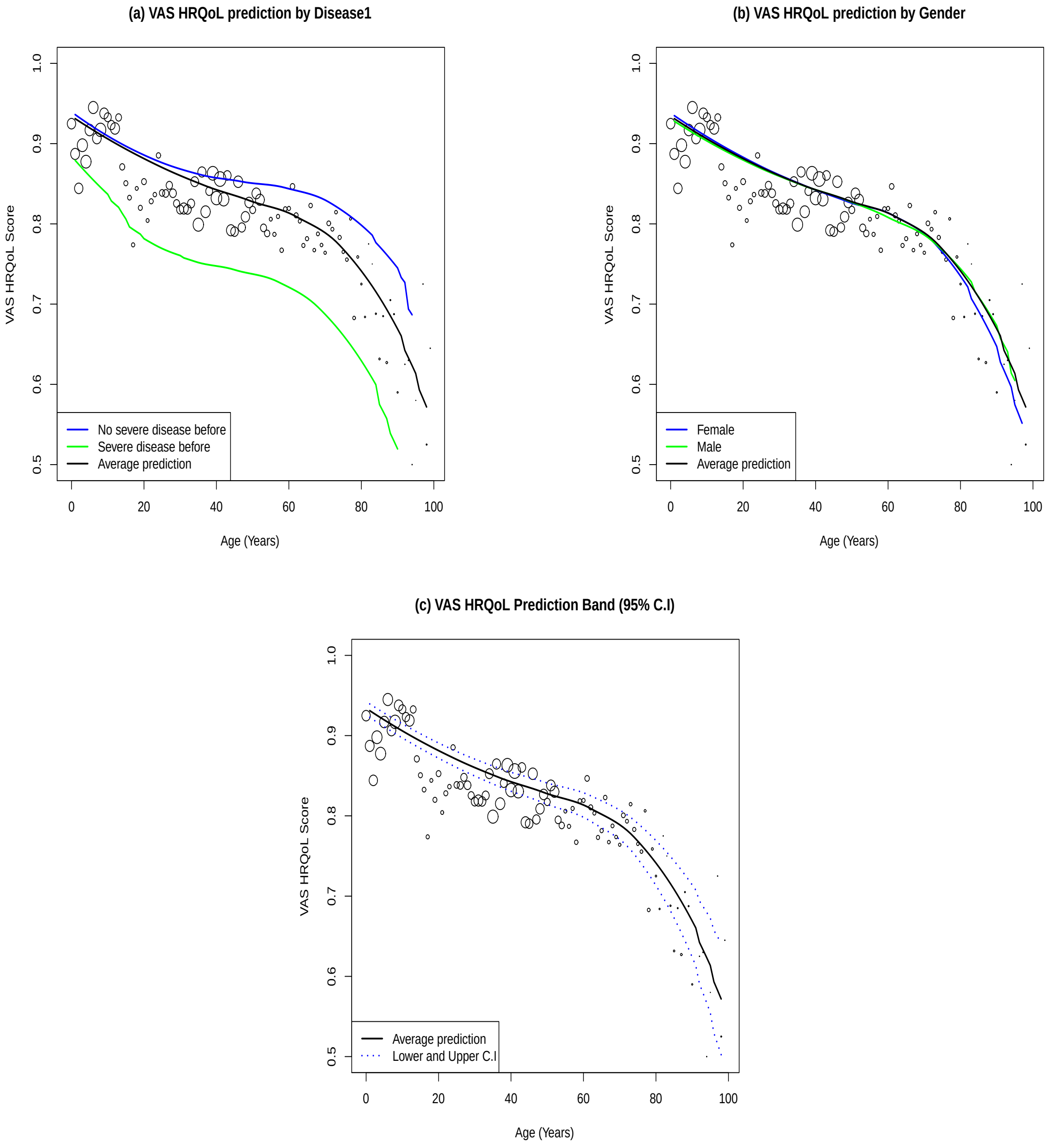

Figure 7: *Final model VAS HRQoL predictions by Disease1 and gender , and 95% prediction Confidence interval*

Bias corrected , bias reduced and non-parametric bootstrap estimates based on the selected model are shown in table 5 below.

Table 5: Bootstrap, Bias corrected and Reduced model estimates

Parameter	Bias Corrected Estimates	Bias Reduced Estimates	Non-parametric Bootstrap Estimates
Location submodel			
β_0 (Intercept)	2.6970(0.1026)	2.6980(0.1026)	2.6967(0.1091)
β_1 (A1)	-0.0504(0.0076)	-0.0504(0.0076)	-0.0502(0.0129)
β_2 (A2)	0.0009(0.0001)	0.0009(0.0001)	0.0009(0.0004)
β_3 (A3)	-0.000006(0.000001)	-0.000006(0.000001)	-0.000006(0.000002)
β_4 (G-Male)	-0.0073(0.0605)	-0.0078(0.0605)	-0.0071(0.0530)
β_5 (D-Yes)	-0.6649(0.0697)	-0.6651(0.0698)	-0.6648(0.0738)
Precision submodel			
d_0 (Intercept)	0.6093(0.0820)	0.6102(0.0820)	0.6111(0.0959)
d_1 (A1)	0.0107(0.0017)	0.0107(0.0017)	0.0108(0.0021)
d_4 (G-Male)	0.1440(0.0725)	0.1434(0.0605)	0.1445(0.0928)
d_5 (D-Yes)	0.4433(0.1149)	0.4428(0.1149)	0.4517(0.1808)
d_{12} (G*D)	-0.5479(0.1505)	-0.5481(0.1505)	-0.5490(0.2490)

The obtained bias corrected and bias reduced estimates were similar to the proposed model estimates. This could have resulted due to the considerable large sample size used for the analysis. Stability was further confirmed by the non-parametric bootstrap estimates (based on 1500 bootstrap samples) which were quite similar to selected model estimates presented in table 4.

2.4.2 Statistical Modelling of EQ5D HRQoL outcomes:Results

Table 6 below presents AIC results for models fitted under polynomials and fractional polynomials.

Table 6: Model comparisons based on AIC and Likelihood Ratio Tests

polynomials								
Polynomial Order	Model	Model with interactions	Model with Dispersion covariates		AIC		Models compared Based on logit link function	Likelihood Ratio Test (p value)
				Logit	Probit	Cloglog		
1	1	Yes	Yes	1496.067	1497.503	1503.868	(1) vs.(2)	<0.0001
	2	Yes	No	1602.195	1603.950	1611.873	(1) vs.(3)	0.0009
	3	No	Yes	1506.058	1508.327	1517.693	(3) vs (4)	<0.0001
	4	No	No	1605.632	1608.179	1618.567	(2) vs. (4)	0.0017
2	5	Yes	Yes	1402.973	1402.278	1404.214	(5) vs. (6)	<0.0001
	6	Yes	No	1545.833	1545.462	1552.039	(5) vs. (7)	0.0675
	7	No	Yes	1396.832	1396.127	1399.689	(7) vs (8)	<0.0001
	8	No	No	1537.864	1538.469	-1545.164	(6) vs. (8)	0.2830
3	9	Yes	Yes	1406.404	1405.659	1410.018	(9) vs.(10)	<0.0001
	10	Yes	No	1530.520	1529.846	1537.266	(9) vs.(11)	<0.0798
	11	No	Yes	1395.047	1395.424	1537.266	(11) vs. (12)	<0.0001
	12	No	No	1525.601	1526.041	1531.611	(10) vs.(12)	0.0718

Fractional polynomials								
Fractional Polynomial Degree	Power of age	Model with interactions	Model with Dispersion covariates	AIC			Models compared Based on logit link function	Likelihood Ratio Test (p value)
				Logit	Probit	Cloglog		
	3	(1) Yes	Yes	1433.095	1432.939	1436.032	(1) vs.(2)	<0.0001
1	3	(2) Yes	No	1559.465	1559.638	1559.951	(1) vs.(3)	<0.0105
	3	(3) No	Yes	1431.686	1431.909	1434.186	(3) vs (4)	<0.0001
	3	(4) No	No	1557.053	1557.587	1562.357	(2) vs. (4)	0.0171
	(3,3)	(5) Yes	Yes	1402.283	1401.554	1404.024	(5) vs. (6)	<0.0001
2	(3,3)	(6) Yes	No	1529.075	1528.885	1533.715	(5) vs. (7)	<0.0700
	(3,3)	(7) No	Yes	1396.032	1396.120	1399.136	(7) vs (8)	<0.0001
	(3,3)	(8) No	No	1526.715	1526.945	1532.198	(6) vs. (8)	0.2756
3	(-1,3,3)	(5) Yes	Yes	1395.739	1396.141		(9) vs.(10)	<0.0001
	(-1,3,3)	(6) Yes	No	1524.349	1524.804		(9) vs.(11)	<0.0798
	(-1,3,3)	(7) No	Yes	1393.913	1394.773		(11) vs. (12)	<0.0001
	(-1,3,3)	(8) No	No	1521.817	1522.730		(10) vs.(12)	0.2542
3	(-2,3,3)	(9) Yes	Yes			1397.900		
	(-2,3,3)	(10)Yes	No			1529.356		
	(-2,3,3)	(11) No	Yes			1396.302		
	(-2,3,3)	(12) No	No			1527.740		

Tests for interactions based on order 1 polynomial and degree 1 fractional polynomial (p values = 0.0017 and 0.0171) indicated the need for interactions, while tests in second and third order polynomials had non-significant p values indicating that interactions were possibly not useful (p values = 0.2830 and 0.0718). This deduction was similarly observed for fractional polynomial models. Besides, significant likelihood ratio tests showed that inclusion of regressors was important in the dispersion sub models.

Predictions based on best models under each link function in table 6 were shown in figure 8 (a and b). Since there was existence of two sub groups for EQ5D response, another modeling technique that could be considered is non-linear piecewise regression but this was alternatively modeled non parametrically using cubic splines which resulted to equally a good fit to the data (figure 8 c). These parametric and non-parametric techniques resulted to models which gave almost equivalent fits to the data. It is important to note that prediction based on best fractional polynomial model was only possible for individuals whose ages were greater than zero years. Even though fractional polynomial model was the best in terms of AIC (1393.913), inferences based on it could not offer full coverage to the data.

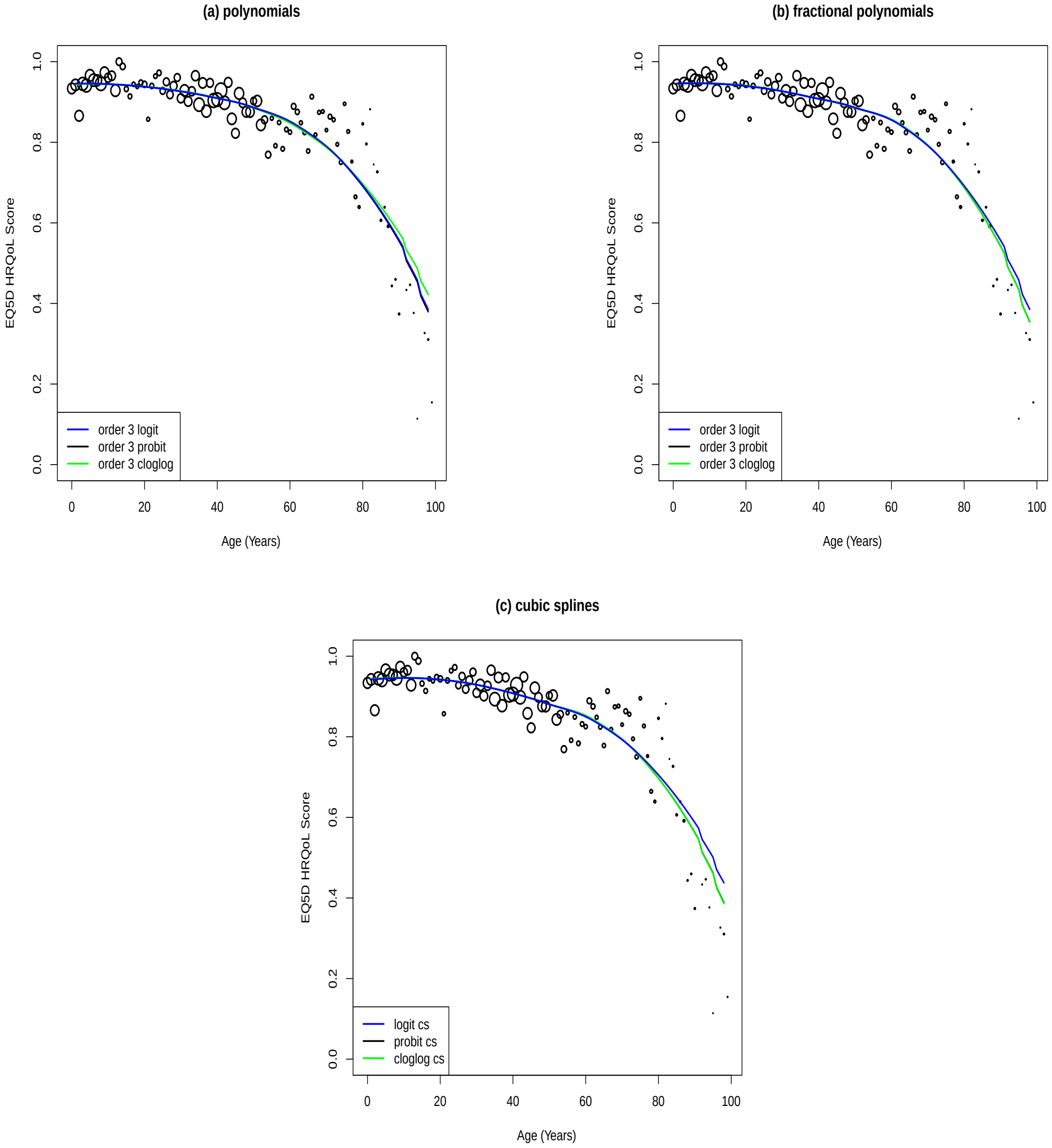

Figure 8: *Graphical representation of the best fits for polynomial, fractional polynomial and cubic spline models under different link functions*

Therefore, order three logit polynomial (AIC=1395.047) was selected whose predictions and inferences would cover the whole data. The non-significant parameters were systematically eliminated from the model by backward selection. As such the resulting final model was a polynomial which is a special case of degree 2 fractional polynomial with powers 1 and 3. Thus table 7 presents the final model estimates.

Table 7: Model Coefficients, Standard Errors and Significance Tests

		Selected Model			Final Model	
Parameter	Coefficients	s.e	p value	Coefficients	s.e	p value
Location submodel						
β_0 (Intercept)	1.8310	0.2020	<0.0001	1.6810	0.1468	<0.0001
β_1 (A1)	-0.0344	0.0201	0.0878	-0.0128	0.0049	0.0096
β_2 (A2)	0.0006	0.0005	0.2707	-	-	-
β_3 (A3)	0.000007	0.000004	0.0957	-0.000003	0.0000008	0.0006
β_4 (G-Male)	0.1780	0.1068	0.0957	0.1783	0.1068	0.0953
β_5 (D-Yes)	-1.1970	0.1392	<0.0001	-1.1870	0.1388	<0.0001
Beta Regression submodel						
β_0(Intercept)	0.7770	0.1074	<0.0001	0.6979	0.0749	<0.0001
β_1 (A1)	0.0002	0.0091	0.9846	0.00928	0.0023	<0.0001
β_2	0.0002	0.0002	0.2944	-	-	-
β_4 (A3)	-0.000004	0.000001	0.3545	-0.000002	0.0000003	<0.0001
β_4(G-Male)	-0.0568	0.0473	0.2303	-0.0541	0.0472	0.2522
β_5 (D-Yes)	-0.3881	0.0604	<0.0001	-0.3868	0.0604	<0.0001
Dispersion submodel						
d_0 (Intercept)	2.3940	0.2275	<0.0001	2.426	0.1845	<0.0001
d_0 (A1)	0.0445	0.0188	0.0181	0.0410	0.0076	<0.0001
d_2 (A2)	-0.0006	0.0004	0.1761	-0.00053	0.00007	<0.0001
d_3 (A3)	0.0000007	0.000003	0.8245	-	-	-
d_4 (G-Male)	-0.3534	0.1091	0.0012	-0.3463	0.1086	0.0015
d_5 (D-Yes)	-0.9429	0.1180	<0.0001	-0.9379	0.1179	<0.0001

The logistic sub model showed that a unit difference in age resulted to a decrease in odds for HRQoL by 0.98 times when considering younger to older individuals. Cubic function of age was significant in both of the mean models and denoted a decrease in HRQoL of life with age. Additionally, logistic regression resulted to a notable decrease in odds for HRQoL by 0.305 times for individuals who had previously suffered severe illnesses compared to individuals who had not suffered any severe illnesses. Predicted mean scores were weighted averages from these two models. Considering the dispersion sub model, quadratic function of age, gender and disease status explained reduction in variability for the mean.

Figure 10 shows weighted mean predictions for HRQoL by disease status and gender using the final fitted model shown in table 7. A clear difference in HRQoL is shown between those who had suffered severe illnesses before and those who had not. While no serious notable difference in HRQoL for males and females, if any then the little non-significant difference observed for individuals of 80 years old and above. The male and female HRQoL predictions coincided with the average prediction for the population.

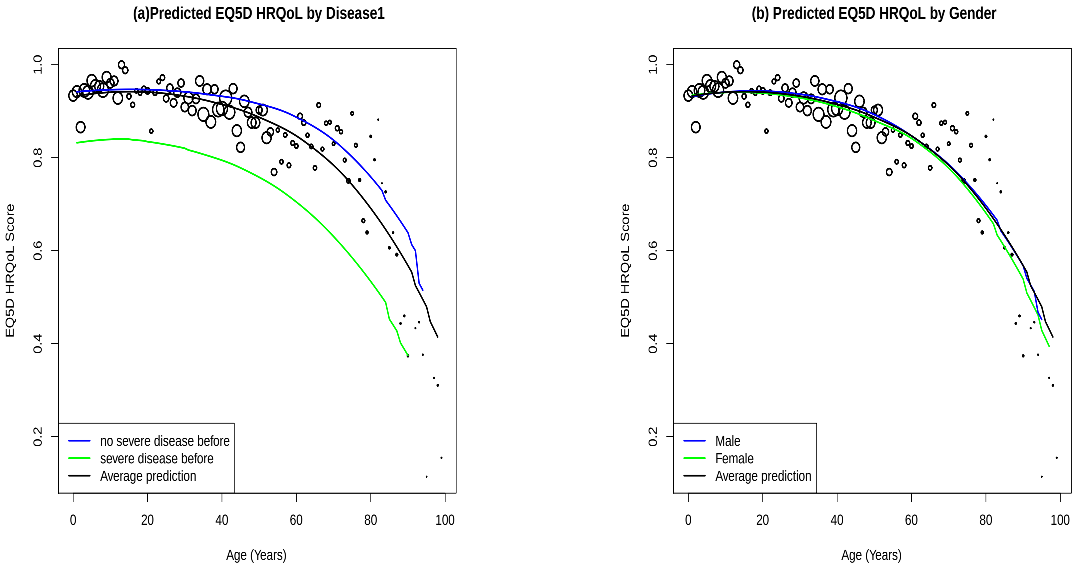

Figure 9: *Final model weighted mean EQ5D HRQoL predictions by Disease1 and gender*

The beta regression sub model showed positive and negative significant effects of age and cubic function of age respectively on HRQoL scores. Therefore, it was important to draw meaningful interpretations from the prediction plots for the individuals who had EQ5D HRQoL scores less than one.

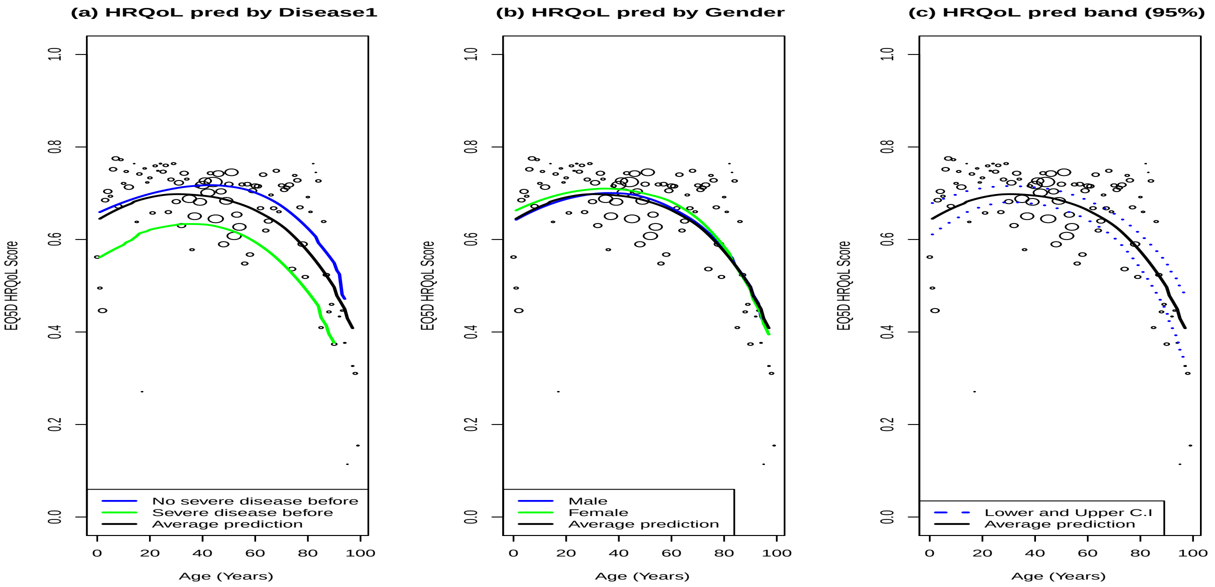

Figure 10: *Mean EQ5D HRQoL predictions by Disease1 and gender based on beta regression sub model, and 95% prediction band.*

A positive effect is observed for a unit difference in age for individuals between 0 and 38 years, while negative effect noted for individuals of 38 years and above. The confidence band in figure 10(c) is a bit wider at the beginning and at the end. This is observed due to the few individuals at these points.

Appendix figure 14, shows predictions for the whole data using logistic regression sub model. Generally a decrease in HRQoL is observed for a unit increase in age.

3 Discussion

It was of interest to determine and explain quality of health in the general population in Flanders. It is important to note that EQ5D instrument used in this region has gained widespread acceptance as a tool for measuring HRQoL in clinical trials, analyses leading to decisions in health economics and many population surveys. Its main strength is reliability since it takes into account the underlying uncertainties of quality of health in terms of dimensions. On the other hand, VAS is perceived as a latent measurement instrument and its reliability is gradually gaining acceptance (Brazier, Ratcliffe, Salomon & Tsuchiya, 2007).

The methods discussed under data exploratory; regression trees, random forest and lasso regression have an underlying assumption of normality. However, the data did not conform to this assumption thus were used as mere exploratory tools.

Polynomials of up to order 3 and fractional polynomials of up to degree 3 were investigated. Models for the latter turned out to be the best fitting models for EQ5D and VAS outcomes based on compared AIC values even though both still achieved considerably better fits for the two outcome measures. Predictions based on polynomials covered the whole range of the data while predictions based on fractional polynomials neglected individuals at age zero. Therefore, best polynomial models were considered suitable for the data other than fractional polynomials which had poor predictions for a section of the data even though they had relatively lower AIC values.

Statistical analyses both for VAS and EQ5D HRQoL outcomes showed significant effects at 5% for age and covariate indicating whether an individual had suffered a severe disease previously or not. VAS outcomes showed that there was a rather gradual decrease in HRQoL for individuals between 0 to 20 years, and also between 60 to 100 years. However, EQ5D outcomes showed that individuals between 0 to 5 years old had almost similar HRQoL scores and a gradual decrease observed in the scores for individuals beyond 5 years to 100 years. Individuals who had suffered severe illnesses before had lower predicted quality of health scores compared to corresponding individuals of the same age who had not suffered any severe disease previously. Gender as a factor was not significant but a little difference though not significant was observed for males and females older than 80 years. Despite its non-significance, it was retained in the model since it was considered as a confounder.

Analysis was performed at the population level only, and possible extension can be considered for sub groups present in the data (children, adults and the elderly). Other modelling techniques such as non-linear modelling and p splines can be explored for comparable fits.

4 Appendix

Table 8: EQ5D Health States

Mobility	*Score*	*Self-care*	*Score*
I have no problem walking around	1	I have no problems to take care of myself	1
I have some problems walking around	2	I have some problems to take care of myself to wash or dress	2
I am bed ridden	3	I myself am unable to wash or dress	3
Daily Activities	*Score*	*Pain/Symptoms*	*Score*
I have no problem with my daily activities	1	I have no pain or other symptoms	1
I have some problems with my daily activities	2	I have moderate pain or other symptoms	2
I am unable to perform my daily activities	3	I have very severe pain or other symptoms	3
Anxiety/Depression	*Score*		
I am not anxious or depressed	1		
I am moderately anxious or depressed	2		
I am very anxious or depressed	3		

Table 9: Possible sets of models for modelling VAS and EQ5D HRQoL outcomes

Polynomial Order	Model	Location Model Covariates	Dispersioon Model Covariates
1	1	A1,G, D,A1*G, A1*D,G*D	A1,G, D,A1*G, A1*D,G*D
	2	A1,G, D,A1*G, A1*D,G*D	-
	3	A1,G,D	A1,G, D
	4	A1,G, D	-
2	5	A1,A2,G,D,A1*G,A1*D,A2*G,A2*D,G*D	A1,A2,G,D,A1*G,A1*D,A2*G,A2*D,G*D
	6	A1,A2,G,D,A1*G,A1*D,A2*G,A2*D,G*D	
	7	A1,A2,G,D	A1,A2,G,D
	8	A1,A2,G,D	
3	9	A1,A2,A3,G,D,	A1,A2,A3,G,D,
		A1*G,A1*D,A2*G,A2*D,A3*G,A3*D,G*D	A1*G,A1*D,A2*G,A2*D,A3*G,A3*D,G*D
	10	A1,A2,A3,G,D,	-
		A1*G,A1*D,A2*G,A2*D,A3*G,A3*D,G*D	-
	11	A1,A2,A3,G,D	A1,A2,A3,G,D
	12	A1,A2,A3,G,D	-

(A1-Age; A2-A1**2 ;A3=A1**3; G-Gender; D-Disease1)

Table 10: Variable Description

Population level Variables (Child, Adult and Elderly shared variables) : n=2204				
Variable	Levels	%	Type	Remarks
Gender	Female	46.71	categorical	Gender of respondent
	Male	53.29		
	missing	-		
Nationality	Belgian	96.64	categorical	
	From Another EU	2.31	(nominal)	
	Non-european	0.86		
	Missing	0.19		
Province	Vlaams-Brabant	14.86	Categorical (nominal)	Regions
	Antwerpen	27.63		
	Limburg	15.40		
	West-Vlaanderen	18.29		
	Oost-Vlaanderen	23.27		
	Missing	0.87		
Animal	Yes	62.03	categorical	Shows whether a family
	No	37.38		kept animals or not
	Missing	0.59		
Normalday	Normal	76.45	categorical	Shows Normal
	Not normal	23.05		day or non-normal day due to Sickness or other reason
	Missing	0.50		
Agecat /Age	Child	22.38	Agecat (categorical)	Age category of a person
	Adult	61.17	Age (continuous)	Age in years
	Elderly 60+ yrs	16.45		
	Missing	-		

Householdsize	1	13.91	Categorical (ordinal)	Size of the household
	2	22.02		
	3	20.66		
	4	29.18		
	5	10.78		
	6	2.49		
	7	0.41		
	8	0.13		
	9	0.04		
	11	0.04		
	12	0.04		
	Missing	0.27		
HouseholdID	Observed	100		Household unique ID
	Missing	-		
Parents	1	19.93	Categorical	Indicates whether the number of Parents in a family is 1 and 2
	2	76.16		
	Missing	3.89		
Disease1	Yes	15.04	categorical	Ever been confronted with severe disease of yourself
	No	78.93		
	Missing	6.03		
Disease2	Yes	45.58	categorical	Ever confronted with severe disease of someone in the family
	No	44.77		
	Missing	9.65		

Child level additional variables : n=494				
mumEducation	None	0.61	categorical (nominal)	Education level for a childs Mother
	Primary	0.81		
	Vocational	10.73		
	Lower technical	2.23		
	Lower Secondary	2.43		
	Upper technical	8.30		
	Upper Secondary	13.97		
	Non-university higher education	42.91		
	Graduate/postgraduate	18.02		
	Missing	-		
Adults and elderly shared variables: n=1710				
Smokestatus	Smoker	16.46	categorical (nominal)	Indicates respondent's smoke status
	Quit smoker	22.01		
	Non-smoker	60.89		
	Missing	0.64		
WorkedinHCare	Yes	22.24	categorical	Indicates whether an individual works/had worked in a health care facility or not
	No	76.94		
	Missing	0.82		
Disease3	Yes	8.48	Categorical	Ever confronted with severe disease because caring for someone else
	No	63.69		
	missing	27.83		

Profession	craftsman with no employees	3.62	categorical	Respondents profession
	craftsman with less than 5 employees	1.17	(nominal)	
	business leader with 6 or more employees	0.99		
	professional clerk	2.86		
	senior member of general management	2.45		
	middle, not part of the general management	13.66		
	other employee	32.81		
	worker with vocational training	8.69		
	worker without vocational training	6.13		
	housewife/househusband	6.13		
	disabled	2.10		
	retired	1.57		
	student	9.52		
	unemployed	2.81		
	rentier	0.12		
	missing	5.37		Missingness include persons with profession>1
Elderly level additional variables: n=360				
Working	Yes	5.51	Categorical	elderly work status
	No	92.28		
	missing	2.21		
Freq1	daily	14.84	Categorical (nominal)	Children visit freq
	a few times a week	38.37		
	a few times a month	25.45		
	once a month	2.24		
	a few times a year	4.76		
	once a year	0.56		
	less than once a year	1.68		
	missing	12.10		

Freq2	daily	6.44	Categorical (nominal)	grandchildren visit freq
	a few times a week	26.89		
	a few times a month	26.61		
	once a month	7.56		
	a few times a year	9.80		
	once a year	1.40		
	less than once a year	2.52		
	missing	18.78		
Freq3	1 or 2 glasses daily	16.25	Categorical (nominal)	Freq with which elderly Consumed alcohol
	more than 2 glasses daily	4.76		
	1 or 2 glasses a few times a week	17.93		
	More than 2 glasses a week	5.04		
	a few times a month	16.25		
	a few times a year	8.41		
	missing	31.37		
		Response variables		
VAS	Observed	95.84	Continuous	Outcome measures by VAS
	Missing	4.16		
CLMPT_EQ5D	Observed	98.28	Continuous	Outcome measures by Cleemput
	Missing	1.72		EQ5D

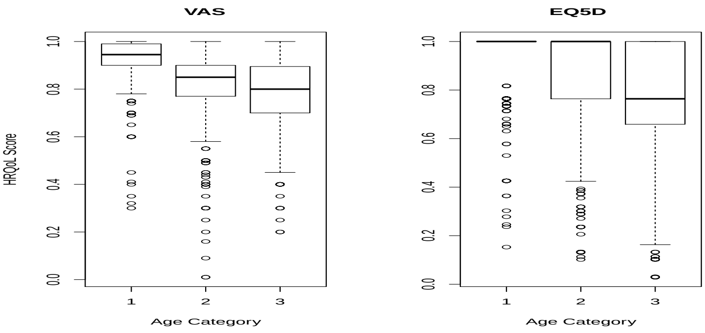

Figure 11: *Boxplots for HRQoL by age categories*

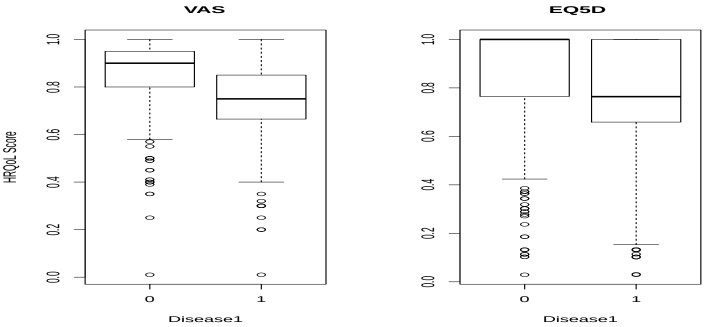

Figure 12: *Boxplots for HRQoL by disease1*

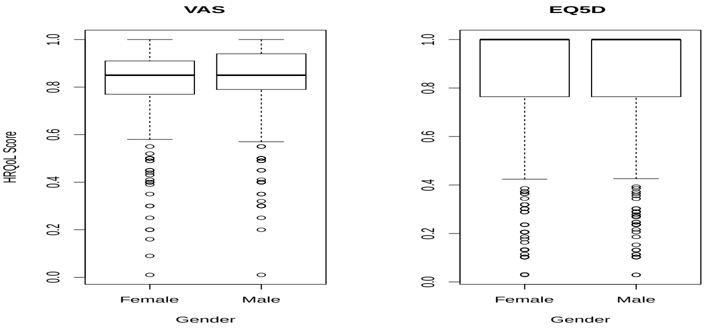

Figure 13: *Boxplots for HRQoL by gender*

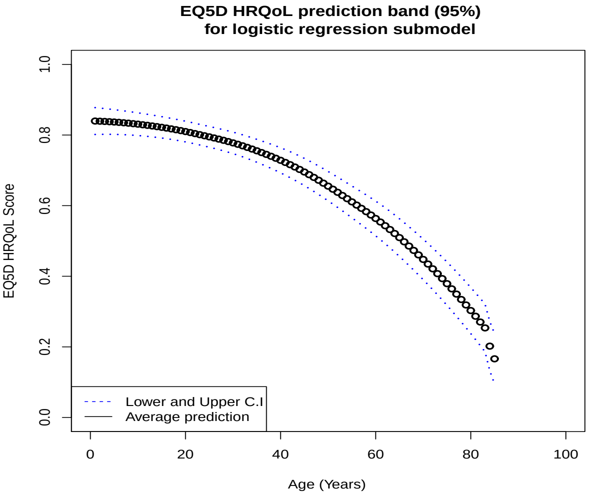

Figure 14: *EQ5D HRQoL prediction by logistic submodel*

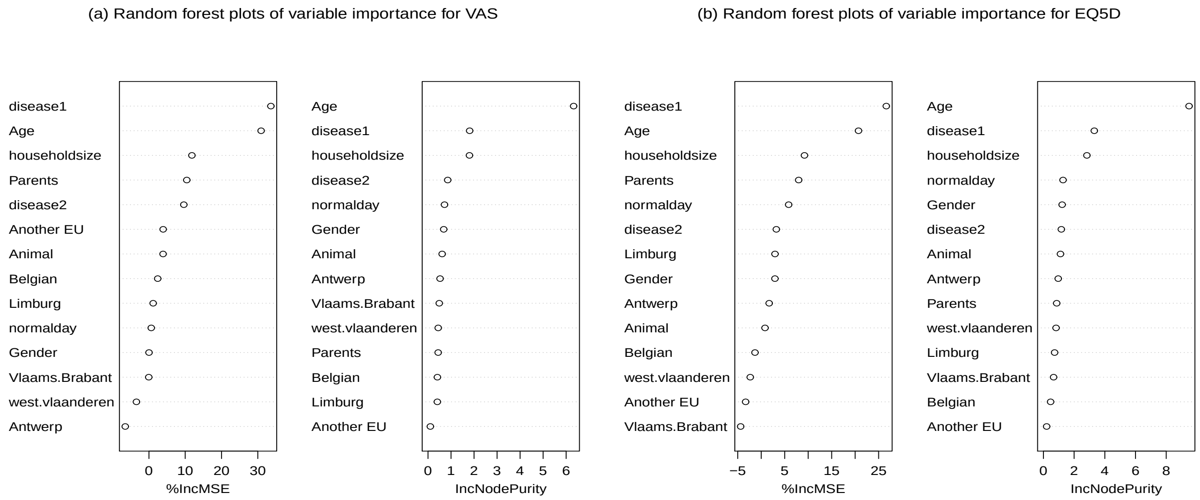

Figure 15: *Random forests for VAS and EQ5D HRQoL Scores respectively*

References

Brazier, J., Ratcliffe, J., Salomon, J. A., & Tsuchiya, A. (2007). *Measuring and Valuing Health Benefits for Economic Evaluation.* New York: Oxford University Press Inc.

Breiman, L. (2001) . *Random forests. Machine Learning*, 45, 5-32.

CDC, title : *Health Related Quality of Life*; accessed on 13th July, 2012, from the website [http://www.cdc.gov/hrqol/]

Cleemput, I. (2010). *A social preference valuation set for EQ5D health states in Flanders, Belgium.*

Cribari-Neto, F., Vasconcellos, K.L.P. (2002). Nearly unbiased maximum likelihood estimation for the beta distribution. *J. Statist. Comput. Simul.* 72, 107-118.

Diaz-Uriarte, R., & Andres, S. (2005). *Variable selection from random forests: application to gene expression.* Spanish National Cancer Center (CNIO), Madrid.

Elith, J., Leathwick, J. R.,& Hastle, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, 802-813.

EuroQol Group (1990). EuroQol *A new facility for the measurement of health-related quality of life.* Health Policy 16: 199-208

Frayback, D. G. (2010). *Measuring Health Related Quality of Life.* University of Wisconsin-Madison.

Guyatt, G. H., Feeny, D. H.,& Patrick, D. L. (1993). *Measuring Health Related quality of life.* McMaster University, Hamilton, Canada.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (Second ed.). Springer New York Inc .

Kirchengast, S., & Haslinger, B. (2008). Gender differences in health related quality of life among healthy aged and old aged austrians: Cross-sectional analysis. *Gender Medicine: The Journal for the Study of Sex& Gender Differences*, 5(3), 270-278.

KUTNER, M.H., NETER, J.,(2005). Applied Linear Statistical Models, fifth edition,McGraw-Hill/Irwin, New York.

Ospina, R., & Ferrari, S. (2007). *Inflated Beta Distributions.* Universidade de Sao Paulo, Departmento de Estatistica?IME-USP, Brazil.

Paolino. P. (2001).Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis*; 9:325-346.

WHO(1948) .Preamble to the Constitution of the World Health Organization as adopted by the International Health Conference, New York, 19-22 June, 1946; signed on 22 July 1946 by the representatives of 61 States (*Official Records of the World Health Organization, no. 2, p. 100*) and entered into force on 7 April 1948.

Royston, P., & Sauerbrei, W. (2008). *Multivariate Model Building.* John Wiley& Sons, Ltd.

Tibshirani et al.(1996) . *Regression shrinkage and selection via the lasso.* J.R.Statist. Soc. B(1996), 58,No.1, 267-288

Shmueli, A. (2005). *The Visual Analog rating scale of health related quality of life: an examination of end-digit preferences.* Hebrew University, Health Management.

Simas, A. B., Souza, W. B., & Rocha, A. V. (2008). *Improved estimators for a general class of Beta Regression Models.*Computational Statistics & Data Analysis, 54(2), 348366.

Simas AB, Barreto-Souza W, Rocha AV (2010). *Improved Estimators for a General Class of of beta regression models.* Associacao Instituto Nacional de Matematica Pura e Aplicada, Departamento de Estatstica.

Singh, R., & Dixit, S. (2010). Health- Related Quality of Life and Health Management. *Journal of Health Management.*

Smithson, M., & Verkuilen, J. (2005). *Beta Regression: Practical Issues in Estimation.* The Australian National University and University of Illinois at Urbana-Champaign.

Swearing, C. J., Castro, M. S., & Bursac, Z. (2011). *Modelling Percentage Outcomes: The % Beta_Regression Macro.* University of Arkansas for Medical Sciences, Little Rock, AR, Department of Pediatrics.

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Assessing the health related quality of life in the general population

Richting: **Master of Statistics-Biostatistics**
Jaar: **2012**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Malla, Lucas

Datum: **14/09/2012**