Erasmus University/Erasmus MC Rotterdam
Department of Health Policy and Management
Thesis Master of Science in Health Economics

Country comparison of health inequalities

Inter-individual variation in healthy life span in a sample of OECD countries

Student:	Steef Baeten	Supervisor:	Dr. Xander Koolman
Student no.:	264886	Co-evaluators:	Dr. Teresa Bago d'Uva
Date:	July/August 2008		Dr. Tom van Ourti

Abstract

Background: Measures of health inequality have so far mainly been focused on systematic differences in morbidity or mortality between groups of society. Although, this provides very relevant information for national policy makers, it is less suitable for the comparison of health inequities in different countries. We argue that it is more appropriate for this purpose, to simultaneously measure the inter-individual variation in both morbidity and mortality. **Methods:** We used health data from the World Health Surveys (WHS) of the World Health Organization (WHO) for countries of the Organisation for Economic Co-operation and Development (OECD) (N = 73,762). We measured functioning on eight health domains through a hierarchal ordered probit analysis (HOPIT) to account for differences in frames of references between respondents. These functioning scores were combined in a measure of health utility. By bootstrapping individual health utility scores and correlating them with life expectancy, we estimated health adjusted life expectancy (HALE) for a hypothetical sample of individuals. For each country the distribution of HALE was captured in a Gini coefficient. We applied resampling techniques to construct confidence intervals. **Results:** HALE was highest among Norwegian men (79.2) and lowest among Turkish women (42.0). Health inequality was smallest in Norway (Gini = 0.110) and largest in Turkey (0.189). The large confidence intervals around the Gini coefficients (e.g. Italy 0.099-0.149) make it impossible to draw any certain conclusions. **Conclusion:** Comparable estimates of health inequalities in different countries which show the equitability of the distribution of health are feasible and can be useful in health policy making.

1. Introduction

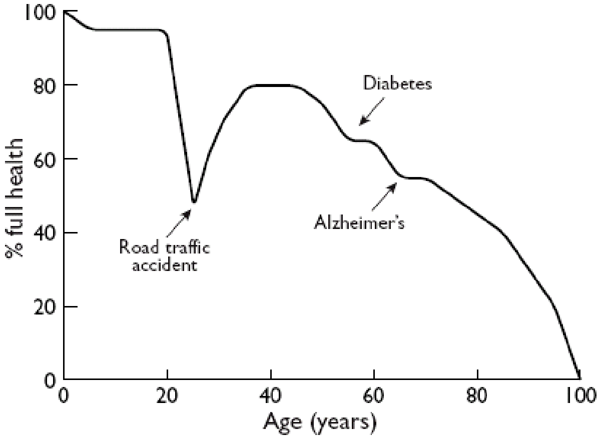
Health systems aim to generate health in the population. By providing services and generating resources through a health system, countries aim at maintaining and improving health among their citizens. However, they are not concerned with all health equally. Health decrements resulting from factors out of ones control are viewed as more unfair than those that are not. Therefore, health systems do not only try to improve health but also try to ensure an equitable distribution of health. Comparing health distributions between countries provides information about the equitability of health systems. But what constitutes health? And what health inequalities are inequitable?

Let us first have a look at the concept of health. What health comprises has been subject for debate for a long time, hitherto no consensus has been reached about a definition (Salomon, Mathers et al. 2003). We will not try to provide one, but we would like to address some characteristics of health that are important for the measurement of health inequalities. First of all, health is associated with many aspects of an individual's life. This becomes clear from the definition used by the World Health Organization (WHO) which describes health as "a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity" (World Health Organization 1946). Another important characteristic of health is, that it has a different meaning for different people. It is a subjective concept. Furthermore health varies over time. Someone that is in good health now does not have to be so tomorrow let alone in 25 years. These are important aspects of health that should be incorporated in an analysis of health inequalities. Some might say that death is the endpoint of health and others might argue that it is a concept closely related to it. Either way it is also relevant in the measurement of inequitable health inequalities.

Many authors have researched the equitability of health distributions using either cross-sectional health (Van Doorslaer, Wagstaff et al. 1997; Humphries and Van Doorslaer 2000; van Doorslaer and Koolman 2004) or mortality (Mackenbach, Stronks et al. 1989; Koskinen and Martelin 1994; Kawachi and Kennedy 1997; Daly, Duncan et al. 1998). However, there is still inequity in a society in which all people that are alive have the same level of health if some only live to the age of 15 and others live to become a hundred. This shows that measuring health inequalities using morbidity data does not give a proper insight. But, mortality alone is not enough for the measurement of health inequalities either. The distribution of health might seem equitable if everybody dies at the age of 75, but if some individuals have lived their whole lives in perfect health and others have been handicapped since birth, it is not. Therefore, a suitable measure of health for the assessment of inequitable differences in health consists of both a morbidity (health state) component and a mortality (life span) component.

Two of such measures are the healthy life span and the health adjusted life expectancy (HALE). The healthy life span and HALE are closely related, they both combine survival with weighted non-fatal health outcomes. In figure 1 the healthy life span of one individual is presented. The valuation of the life span (y-axis) is represented by the curve, which diminishes as the individual gets older. This individual survives until the age of 100 when health equals zero. The healthy life span is then given by the area under the curve (Mathers, Salomon et al. 2003). The HALE values a life in the same way as the healthy life span. However, the HALE looks at a prospective live. It looks at a risk profile of individuals to be in certain health states over their life span. In other words, the health adjusted life expectancy is the expected healthy life span at birth.

Figure 1: Healthy life span



(Gakidou, Murray et al. 2003)

To measure health inequalities we will first need to choose the most suitable quantity. We already argued that a health measure should incorporate both morbidity and mortality. A measure of health inequality is no different. We will start by explaining why the variation in healthy life span is better than previously used measures. Subsequently, we will start our calculations by estimating population health. Initially we will estimate both components of health, morbidity and mortality separately.

We will first determine the cross-sectional health utility (HU) in each country by sex and age. We used data from the World Health Surveys (WHS) of the WHO to measure different HU scores. Health utility is not an objective measure, i.e. people with the same level of morbidity experience their health differently and thus report different health utilities. In the WHS health was self-rated and therefore responses (i.e. health utilities) are not comparable across individuals. Our measurement of health will be corrected for these differences in perception, using a relatively new method: the hierarchal ordered probit analysis (HOPIT). This method does not only measure health utility itself, but also the perceptions of individuals towards health. After correction by the HOPIT analysis self-reported health question will be comparable across different individuals (Tandon, Murray et al. 2002). We will use self-reported health data about different domains of health to measure the variation in all dimensions of health. Our analysis will synthesize the corrected self-rated domain performances into health utilities using a valuation function previously developed by the WHO.

Subsequently, we will measure the uncorrected life expectancy (LE) at birth using WHO life table data. Again we will distinguish by country and gender. Crude mortality rates will be used as inputs to determine the years lived on average in different age-categories. The measured health utility by country, gender and age will then be used to value each of these years and calculate health adjusted life expectancy at birth.

This measure is very suitable for the measurement of population health (Mathers, Salomon et al. 2003). However, the aim of this article is not to measure the average health within a country, but the health of (all) individuals in a country. This means that we will need multiple estimates of both health utility and survival. For the former we will use individual responses to self-rated health questions from the WHS, for the latter we will use a hypothetical dataset in which we determine survival in different age-categories by chance. In this case we can not combine health utility and survival directly as with the measurement of population health. Both of these measures will generally be correlated with each other (Lubetkin, Jia et al. 2005), that is people who die will generally have had a lower health utility in preceding years than those who survived. This correlation will have to be accounted for because health inequalities measured using either morbidity or mortality will typically underestimate the existing inequality. The magnitude of the correlation between health utility and survival has previously been investigated and reported in a hazard ratio of lowered health utility on survival (Kaplan, Berthelot et al. 2007). We will use this hazard ratio for different individual health utilities to adjust mortality rates from the life tables and calculate individual survival. Finally, the individual healthy life spans resulting from the individual health utilities and survival will be summarized in a measure of (health) inequality, the Gini coefficient.

2. On the measurement of health inequity and inequality

Traditionally most health inequality studies have focused on systematic differences between groups determined by characteristics such as socio-economic status. See for example (Winkleby, Jatulis et al. 1992; Lillie-Blanton and Laveist

1996; Macintyre, Hunt et al. 1996; Van Doorslaer, Wagstaff et al. 1997). An alternative for measuring health inequalities through systematic differences was proposed by Gakidou et al. who suggested that the distribution of health expectancy within a country was the quantity of most interest (Murray, Gakidou et al. 1999; Gakidou, Murray et al. 2003). This point-of-view did not receive the acclaim the authors hoped for as most researchers continued the measurement of between group variation as opposed to total differences (i.e. including within group variation) (Gakidou and King 2003). See for examples (De Irala-Estevez, Groth et al. 2000; Blakely, Lochner et al. 2002; Grundy and Sloggett 2003; Zere and McIntyre 2003; Van Lenthe, Schrijvers et al. 2004). We do agree with Gakidou et al. to a large extent. We also believe that health inequalities should be reflected in the total health variation between individuals. However, Gakidou et al. suggested that health expectancy would be the best measure of health to show this variation. This is where we disagree, because we believe that the variation in health adjusted life expectancy as they suggest does not reveal all the variation in health relevant for an equitable distribution of health. We propose using variation in healthy life span instead, that is we also model the chance that results in health differences between individuals with equal health risks. In the next section we will argue why we believe this method is more suitable than both a measurement of systematic health differences and the distribution of health expectancy. We will us theoretical arguments as well as practical considerations to plead our case.

2.1. Causes of health inequalities

One might argue that a health system aiming at an equitable distribution of health would have to *try* to reduce the variation in healthy life span of all individuals to zero. To achieve total equality (i.e. no health variation) two conditions should be satisfied. First of all, everybody should have an equal health expectancy. That is, both the chances of being in a certain disease state over a life span are equal and the chances of dying are equal. In addition, it would require that all these chances are either one or zero. However, it is clear that this is not possible. Consider twins who probably face the same health expectancy. All endogenous (e.g. genes) and exogenous (e.g. education and attitude towards smoking) variables will be (almost) equal for these twins. This would probably mean that their health expectancies will be equal. However, these twins won't constantly experience the same health, nor will they die simultaneously. This is why Gakidou et al. argue that we should only measure inequalities in health expectancies (Gakidou, Murray et al. 2003). The variations in healthy life span are most probably the results of chance and therefore some will argue that they are not attributable to the health system and not unfair.

Although we disagree with Gakidou et al. on this point, it does raise an important question: Which health inequalities are considered to be fair and which are not? Before we can answer this question, we will first have to identify which kind of inequalities exist. In general there is consensus about the division of causes of health inequalities, although there are differences in point-of-view which causes are equitable and which ones are not. A first set of health inequalities can be viewed as outside of the control of both the individual and society. This can be due to chance, like the previous example, but also due to biological differences (e.g. genes). A second group of causes of health inequalities are the causes to voluntary risk seeking behavior of an individual. Some clear examples of such behavior are mountain climbing, speeding in traffic and joining the army. Health inequalities attributable to these two first sets of causes are generally considered not to be unfair, because they are out of the control of both society and the health system. Inequalities due to any other reason are often deemed to be inequitable (Whitehead 1991).

2.2. Theoretical arguments

Now that we have determined a division of causes of inequalities we can try to determine which causes lead to inequitable inequalities and which to equitable inequalities. Let us first have a look at the inequalities due to chance. Some might argue that since variation due to chance is out of the control of anyone it can never be inequitable. We disagree with this view because the risks associated with chance are pliant. Through preventive actions and cure services risks can be reduced. For example, a pandemic could take place like the bird flu or the Spanish flu at beginning of the 20th century. One could take the view that it is chance that such a disease strikes harder in one country than in another, because the chance in health variation is not due to amendable causes. However, it is very well possible that the first country has taken preventive action such as a vaccination program (if possible) only in large cities opposed to the second country that took country wide preventive action. This means that the countries health system could be held responsible for the larger health impact of the pandemic.

In addition we argue that if all health inequalities would be due to chance (i.e. none of the inequalities can be explained by any variable), but the variation would be larger in country A than in country B, the distribution of health would be more equitable in country B (Koolman 2008). This preference is revealed if someone is asked to choose a society to live in, without knowing in which position he or she would be in. This claim is supported by Rawls' theory of justice. This theory states that when choosing from behind a 'veil of ignorance' people will deem the society with the best position for the worst-off to be the most just (Rawls 1971).

We also argue that a measure of health inequality should also accommodate health inequalities that are the result of free choice. The fact that we show apprehension for them is evidence that we regard them unwanted. Even though many may believe that injuries resulting from a car crash at a speed of 80 mph (138 km/h) are the result of free choice, they probably will still be concerned with his health. Therefore services are put in place to reduce the impact (i.e. the resulting inequality) of the crash to a minimum. The person in the car crash might regret his actions afterwards and change his ways in the future. This might alleviate the views of society towards the past voluntary risk seeking behavior of this individual.

The remaining causes of health inequalities, which are considered to be unfair, should clearly be part of any health equity measure. Summarizing, we argue that it is most appropriate to incorporate all health inequalities in a measure of health inequity, even though some of it might be considered just.

2.3. Practical arguments

In addition to the above there are also some practical arguments why variation in healthy life span is the most suitable measure to determine the equitability of health differences within countries. These arguments relate foremost to the way differences are currently measured and the biases that result from them

First of all, the groups compared in an analysis of systematic differences are to some extent arbitrary. The relevancy of groups to be compared is largely determined by the context. In a large number of western countries it might be appropriate to investigate differences between migrant workers and autochthons. However, in many other countries the number of migrant workers is so low that such a comparison would be irrelevant. Furthermore, it is not viable to select and measure all characteristics for which systematic health differences are unwanted. Even if it was possible to construct such a list of characteristics, the list itself would vary between settings, because it is highly determined by context.

A second criticism about systematic differences applies when we would be able to agree on an exhaustible and comparable set of characteristics. The way these characteristics would influence differences in health again varies by the setting. Take the example of (un-) employment, which has an entirely different effect on health in the United States than in societies with universal health coverage. In the United States people are generally insured through their employer. As a result being unemployed usually means being uninsured for health expenses. This will have a large impact on the access to health care for unemployed Americans and consequently on their health. In countries with universal health coverage everybody is covered for (basic) health care services, from which their health will benefit.

In addition, this example shows how difficult it is to measure the effect of groups. Due to its health care system the US can be divided in three groups: (i) those with Medicaid coverage (provided by the government), (ii) those with private health insurance and (iii) those without health insurance. Group (i) are usually the poorer, potentially unemployed, Americans. The second group mainly accommodates individuals with good jobs, which include health insurance or allow them to pay the high premiums of private health insurance. The last group are those who are not poor enough to qualify for Medicaid and are not able to purchase private health insurance. This results in health insurance for the low and high income groups, but no health insurance for the middle incomes. This might lead to the false conclusion that there is no effect of income on health insurance coverage.

Even if we could measure all systematic differences in a cross-setting comparable way without these kinds of measurement errors, we would still have objections to the use of systematic differences. An important reason is that the differences might in fact not be inequitable. Having a good income might allow someone to take up survival holidays or any other voluntary health risk seeking behavior. If we would only focus on these free choices we would conclude that a higher income has a negative effect on health. The true effect of income is going to be far larger than this effect and overall higher income will still be associated with better health, but the effect has been modified downward.

A fourth problem arises when we have a look at inequalities resulting from chance. Obviously inequalities due to chance, in absolute terms, are equally important in all socio-economic strata or any other kind of groups relevant in the measurement of systematic differences. This variation might possibly be explainable in the future, through the advances in medicine. If these advances in medicine would not only be able to explain health differences, but also be able to reduce them for all individuals, then health inequity would be reduced. But the total differences between groups remain and have become relatively more important (i.e. larger share of total health inequalities). This shows how the knowledge about the causes of illnesses effects conclusions about systematic inequities between groups (Koolman 2008).

Determining whether health inequalities are the result of free choice is difficult and raises a lot of discussion. The examples of these causes given earlier might be very straight forward, but there are many situations imaginable which are far harder to evaluate. For example, one might argue that heavy drinking is a known health risk and the result of free choice. However, if someone is addicted and is unable to stay away from alcohol, is it still a free choice to drink if someone is alcohol dependent? Secondly, in some circles of acquaintances drinking is seen as a part of the social process. Is it fair if someone drinks too much, when he/she is expected to drink alcohol regularly? This is an additional argument to measure total variation, because imposing a cutoff between free and unfree choices will make health inequality data incomparable to a certain extent. Preferably we would not, after all the health system can not be held accountable for ‘bad’ choices of its citizens. However, as these examples illustrate it is hard to determine whether actions are the result of free choice or not. Therefore, excluding a part of the variation in health is to some extent arbitrary and debatable.

In addition, what constitutes a free choice in one setting might not in another. For example, is walking between savage animals a freely chosen health risk? It probably is if it concerns a tourist on a safari, but it probably is not if it is a native. Another example is the equality of education on health risks. In western societies the health risks associated with smoking have been known for decades, but in other countries with a less well educated population, like China, people are less knowledgeable about the hazards associated with smoking.

What is more, in most cases it would be unfeasible to exclude certain parts of the variation in health inequality. It would require data on causes of health outcomes and the causes of those causes to differentiate between types of health inequalities. Generally, this kind of information is not available, most certainly not on the national level. Therefore, it is more convenient to use all the variation in health.

Summarizing, we believe that it is best to assess the equitability of health distributions through a measure that incorporates both morbidity and mortality, such as the healthy life span. Excluding specific kind of differences in the measurement is undesirable, as it reduces the comparability of results. In addition, we believe that inequalities between individuals are more relevant than those between groups, because inter-individual inequalities best reflect the distribution of health within a setting.

2.4. Formal arguments

In this section we will use an equation that can help clarify some of our arguments of the previous paragraphs why systematic health differences alone do not provide the total overview with regard to health inequality measurement. The equation is as follows: $y_i = \beta x_i + \varepsilon$ This describes health (y_i) by a set of covariates (x_i) with coefficient β and an unexplained part ε. Total inequality is given by the variance in health between individuals $\sigma(y)^2$. The equation partly explains this variance by the set of covariates βx_i, the part it explains is usually summarized in the R^2 measure. This reflects the systematic differences aspect of health inequality, where x indicates the group and β indicates the size of systematic difference between the groups. The percentage of variation not explained (i.e. $1 - R^2$) remains in the error term ε. We believe both the explained and unexplained variance in health should be incorporated in studies of health inequalities.

This proportion of unexplained variance is partly determined by randomness in health (i.e. unavoidable health inequalities), but also by the covariates that are excluded. That is, there will be room for improvement with regard to the incorporated covariates because the health equation can always be better formulated (e.g. age is not one of the covariates) or more importantly advancements in the understanding of health determinants will justify the incorporation of extra covariates (e.g. increased knowledge about the effects of genes). As we argued earlier we believe that health inequalities that can not (yet) been addressed due to the current capabilities of the medical sciences should also be included in a measure of health inequality. This means that we at least want to incorporate that part $(\varepsilon_{unknown})$ of the unexplained variance in health. However, it is not possible to disentangle these two parts of unexplained variance, that is we can only measure ε_{total} and not ε_{random} or $\varepsilon_{unknown}$ separately. Therefore we need to incorporate ε_{total} instead. Anyway, our argument still holds that if in country A $\sigma(y)^2$ is smaller than in country B and the entire difference is attributable to differences in ε_{random} most people will prefer country A over country B.

It is important to note that we do not argue that countries should also try to reduce $\sigma(y)^2$ to zero, because that would not be feasible. What we believe is that within ε_{total} there is a lot of variance that is not random. This shows from the mere fact that the same set of covariates will not explain the same percentage of the health variation in two different countries. There is no reason to assume that the random variation in health differs over countries, it is much more reasonable to assume that this is due to determinants not included in the health equation. This means that there is room for each country to reduce its health variation to at least the level of the best performing country. That is to try $\sigma(y\ )^2 \rightarrow \sigma(y)^2_{\min}$.

3. Methods

3.1. Selection of countries

Although health systems influence the level of population health considerably, other factors (e.g. sanitation, traffic regulations and GDP per capita) also have an effect on population health. In addition, these factors also influence the way the health system is organized. This could potentially distort the measurement of the effect of the health system on health inequalities. Therefore, only countries for which these factors are similar were incorporated in the analysis. This assumption most probably holds for OECD countries. Therefore, we only incorporated OECD countries in our analysis. Data requirements limited the number of OECD countries to 20 (Austria, Belgium, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Italy, the Netherlands, Norway, Mexico, Portugal, Spain, Slovakia, Sweden, Turkey and the United Kingdom). Because of the recent entry of Israel in the development Centre of the OECD, we also incorporated that country in our analysis.

3.2. Data

The main data source used in the analysis were the World Health Surveys (World Health Organization Evidence and Information for Policy 2002; Üstun, Chatterji et al. 2003). The WHS, a cross-sectional dataset developed by the WHO, collects information on the household as a whole and a single individual from the household. The individual questionnaire was used to assess the level and the distribution of health (World Health Organization Evidence and Information for Policy 2002). There was a long and short version of the WHS individual questionnaire. The data used in this analysis is mainly derived from the short questionnaires, which only contain information on social-demographics, health state descriptions, coverage, responsiveness and health goals. The OECD countries in which the long questionnaire was implemented are: the Czech Republic, Hungary, Mexico, Slovakia, Spain and Turkey. The Surveys in the WHS program employed a probability sampling design. This means that every adult in the country could be sampled and that the chance of being selected is known. An important advantage of this approach is that not mainly healthy individuals were selected, but also people who are in ill health or even institutionalized because of their health state. To create a representative sample, i.e. equal opportunity of selection for all respondents, the WHO applied multi-stage stratified cluster sampling. In this method individuals are not directly selected, but from a larger unit (e.g. street). For example, within a country first a province is sampled, subsequently a city is randomly selected, then ZIP-code area, a household and finally an individual. In multi-stage stratified cluster sampling it is important that each selected cluster is similar to the other clusters, but homogenous within the cluster (i.e. the diversity within a country is represented within the cluster).

3.2.1. Survey questions

The WHS contained questions about the health of the individual on eight different domains and on overall health. Questions were formulated in the following: *"Overall in the last 30 days (…)how much difficulty did you have (…): 1. None; 2. Mild; 3.*

Moderate; 4. Severe; 5. Extreme?" In addition, the survey tried to establish a frame of reference for each respondent with respect to the different health domains. This was done by means of health vignettes. Vignettes are hypothetical situations/conditions a respondent is asked to assess in addition to his own situation/condition (in this case functioning on a domain of health). To clarify, we will give an example of a vignette on the health domain mobility: *[Mary] has no problems with walking, running or using her hands, arms and legs. She jogs 4 kilometres twice a week.* The vignette questions were formulated similar to the self-assessment question, but instead of asking "*how much difficulty did you have…"* they were asked *how much difficulty did [name of person described in vignette] have …"".* When rating the vignettes the respondents were asked to imagine the subject to be of the same age and background as him or herself (World Health Organization Evidence and Information for Policy 2002).

3.2.2. The WHS sample

Respondents of whom the answers to all health questions were documented, were included in the study. In addition, age, gender and the number of education years had to be non-missing as well. In table 1 these characteristics of the population are given. The survey contained more men than women and the respondents were on average middle-aged. A large variation existed in the number of respondents per country. This is largely due to the fact that the samples were larger for the countries with the long individual questionnaire. The average number of formal education differs noteworthy between the countries as well. The respondents of the country with the most educated population (France: 13.9) on average received over two times the years of schooling as the respondents in the country with the lowest level of education (Turkey: 6.0).

3.2.3. Other data sources

The mortality in OECD countries is well documented. The countries document death occurrences well and these are publicly available. The WHO has summarized mortality rates and corresponding life expectancies for all of its member states in so called life tables. These mortality rates did not only vary by country but also by gender and age (Shibuya, Mathers et al. 2002; World Health Organization 2007). We used the WHO life tables as the inputs for the mortality component of health. For additional information about the valuation of the health vignettes, we performed a time trade-off analysis. For this purpose we took a convenience sample of 28 respondents. This sample mainly contained higher educated individuals aged 20 to

35. Every respondent was presented with all health vignettes and there were no missing values.

Table 1: Population characteristics

Country	Number	Proportion males	Age		Years of formal education	
			mean (sd)		mean (sd)	
Total	73,762	42%	43.6	(17.5)	8.3	(5.2)
Austria	989	37%	45.0	(16.1)	10.9	(2.8)
Belgium	768	43%	46.1	(17.1)	13.9	(3.5)
Czech Republic	870	44%	48.2	(18.2)	12.4	(2.7)
Denmark	988	47%	50.9	(16.9)	11.7	(4.0)
Finland	996	45%	52.7	(17.2)	11.6	(4.0)
France	847	42%	43.9	(16.5)	13.8	(4.4)
Germany	1,058	40%	51.8	(17.2)	10.9	(3.1)
Hungary	1,399	42%	49.4	(18.2)	11.6	(3.8)
Ireland	775	43%	43.9	(17.0)	12.7	(3.1)
Israel	1,188	43%	44.9	(17.6)	13.4	(3.9)
Italy	947	43%	48.4	(17.9)	11.2	(4.8)
Mexico	38,745	42%	41.0	(16.7)	7.2	(5.0)
the Netherlands	979	31%	43.9	(18.5)	13.0	(3.6)
Norway	958	50%	47.6	(18.3)	12.1	(4.2)
Portugal	884	40%	47.7	(18.3)	7.3	(4.3)
Slovakia	1,702	33%	37.5	(14.6)	13.3	(2.9)
Spain	6,113	41%	52.7	(18.4)	9.0	(5.3)
Sweden	963	42%	50.9	(18.2)	12.1	(3.7)
Turkey	10,993	43%	42.0	(16.0)	6.0	(4.4)
United Kingdom	1,159	37%	50.5	(19.4)	12.1	(3.0)

3.3. Health utility

To measure health inequalities it is first necessary to measure health. We already reasoned why using either morbidity or mortality is undesirable. In the introduction we stated that HALE is the best suited measure for this purpose, because it accommodates both morbidity and mortality. Please note, that we do not believe HALE to be best suited to estimate health inequalities.

The HALE has some desirable properties for evaluating population health (Murray and Evans 2003). A health measure should get worse if a certain disease gets worse (i.e. the health utility of the health state decreases), cetris paribus. In addition, an increase of the disease’s age specific mortality rate, prevalence or incidence should lead to a reduction of the health measure. The HALE has all these properties (Murray, Salomon et al. 2000). The first step in measuring HALE was to determine health utility for each country, age and gender.

3.3.1. Hierarchal ordered probit analysis

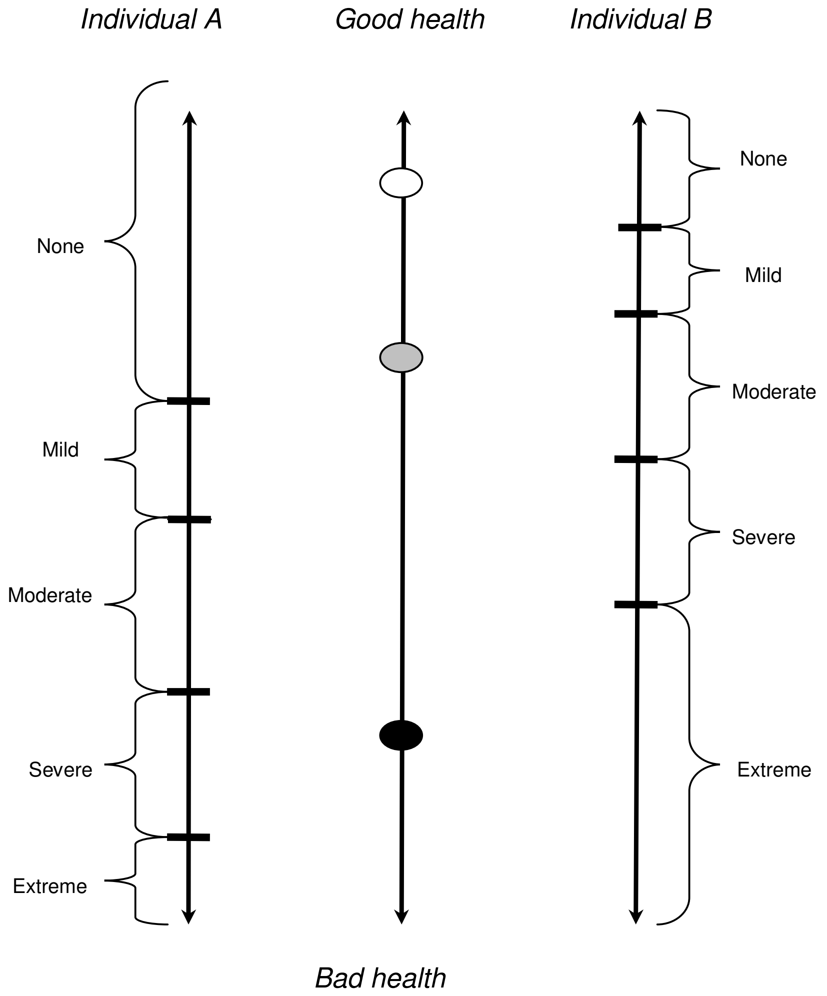
It is obvious that health does not have a limited number of values; probably everyone will consider health as a continuous concept. However, we can not objectively and directly measure health, because there is no measuring device available for it. Although it is unobservable it does determine the response given by a respondent answering a question about his/her health. Thus, the unobservable continuous concept of health determines which categorical response is given by a respondent; we also refer to this as the underlying latent scale. Values on this latent scale range from minus infinity to plus infinity and have no clear interpretation. Cut-points are values of health on the latent scale at which respondents switch from one response category to another. Generally, these cut-points are measured using an ordered probit analysis. This analysis uses these cut-points to estimate scores on the latent scale.

As mentioned earlier we used a hierarchal ordered probit model (HOPIT) and not a standard ordered probit to measure health. This fairly new method is used for analyzing categorical data in which *response heterogeneity* is present. Response heterogeneity is sometimes referred to as differential item functioning (DIF) (King, Murray et al. 2003), ‘state-dependent reporting bias’ (Kerkhofs and Lindeboom 2002), ‘scale of reference bias’ (Groot 2000) or ‘response category cut-point shift’ (Sadana, Mathers et al. 2002). Response heterogeneity occurs when responses not only differ between individuals because of differences in actual health levels, but also because they have different frames of reference. For example, think of a 20 year old male who is not able to walk two miles to the town center because of his health. This person will probably rate his own health as being bad. Now suppose an 85 year old male is in the same health, he is not able to make the walk either. How will he rate his health, when all else is assumed to be equal? He will probably not rate it as bad as the 20 year old, because he does not expect to be able to walk two miles. This difference should actually be attributed to different frames of reference, because the health of both individuals is in fact the same. Thus, response heterogeneity in a survey essentially entails the interpersonal incomparability of subjective self-reported data (King, Murray et al. 2003). Questions about a respondent’s personal health are almost always self assessed. The WHS is no exception which becomes clear from the example question in the data section. Therefore response heterogeneity is a valid concern in self assessed health in general and in health state descriptions of the WHS in particular.

Essentially, the HOPIT analysis is an extension of the standard ordered probit which is not able to handle response heterogeneity in categorical data. Both the standard ordered probit and the HOPIT, estimate these response cut-points, but unlike the HOPIT the values of these cut-points are the same for everyone in the standard model. In other words, the HOPIT lets each cut-point differ over respondents as a function of individual characteristics. Allowing cut-points to differ between respondents corrects for the fact that a response ‘good’ does not have the same meaning for all respondents. This means that response heterogeneity has been corrected for. That is why it is more suitable to analyze data in which this response heterogeneity is present with a HOPIT analysis than with a standard ordered probit analysis

The likelihood function used to estimate the HOPIT analysis has two components. The first component of the HOPIT analysis uses the responses to vignette questions to estimate the cut-points and the second component uses the responses to the self-assessment questions to estimate domain scores. The first component is used to make the self-assessment questions comparable across populations. A vignette essentially is a clear cut characterization of the level of ability on specific domain. Respondents assess this vignette on the same scale as the self-assessment question about their own health. The level of ability is fixed in the vignette description, thus the variation in ordinal responses to the vignette questions is the result of the variation in use of cut-points (i.e. response heterogeneity) (Tandon, Murray et al. 2002). This is illustrated in figure 2 in which all lines represent true latent health, i.e. equal positions on the different lines means objectively equal health. The three bullets on the middle line are three different vignettes; their position indicates what the true level of health of these vignettes is. The two outer lines show at which level of health respondents chose what response, i.e. where the respondent cut-points lie. Both individuals will consider the white vignette to represent good health and therefore respond that there are no health problems. However, they will reply differently to both other vignettes (i.e. the grey and the black dot on the line). The individual on the left will consider the grey dot as having no health problems and the black dot as having severe problems, whereas the individual on the right will regard them as having moderate problems and extreme problems respectively. The vignettes show that the left individual has a lower frame of reference with regard to health than the right individual. Because health is fixed in the vignettes, all variation in vignette responses is due to response heterogeneity (Kapteyn, Smith et al. 2007). In short, the responses to the vignette questions drive the estimation of the cut-points. These allow the HOPIT to separately identify what part of the variance in responses is due to response heterogeneity and what part is the result of actual differences in health. This latter part provides the information about individual health necessary to estimate health utility. The calculation of individual domain functioning is possible, because the vignettes drive the cut-points and set the scale, which makes it possible to determine the variance of the latent variable (Tandon, Murray et al. 2002).

Figure 2: Vignettes and response heterogeneity



The estimation of the HOPIT model was derived from Tandon et al.(Tandon, Murray et al. 2002). First of all, the vignette questions (v) allowed for the estimation of a latent variable Y_{ij}^{v*} normally distributed with mean μ_{ij}^{v} and a variance of 1. The subscripts *i* and *j* denote the different respondents and vignettes respectively.

$$(1) \qquad Y_{ij}^{v*} \sim N(\mu_{ij}^{v},1)$$

In other words, the valuation of the health described in the vignette *j* is valued by individual *i* on a continuous scale, which is unbounded and one-dimensional. The value of μ_{ij}^{v} was a function of a set of individual characteristics and the vignette $J_i^{'}$.

$$(2) \qquad \mu_{ij}^{v} = J_i^{'}\alpha$$

The reported answer of individual *i* to each of the vignette questions with *k* ordinal response categories is determined by Y_i^{v*} . This means that a higher score on the latent scale for an individual results in a higher ordinal response to the vignette question of that individual. Thus, he or she elicits a response on a five point Likert scale, which he or she deems appropriate for the health described in the vignette and what is deemed appropriate is indirectly (i.e. through the latent variable) related to the personal characteristics of the respondent. This elicitation is the result of the following mechanism:

$$y_i^v = k \text{ if } \tau_i^{k-1} \le Y_i^{v*} < \tau_i^k$$

Here y_i^v is the response (k) of individual *i* to vignette question *v*. τ_i^{k-1} is the value on the latent scale (Y^*) at which the respondent shifts from response k to k-1 and vice versa. From now on we will refer to τ as a cut-point. The cut-points are restricted by the following rule:

$$-\infty \le \tau^0 < \tau_i^1 < \tau_i^2 < \tau_i^3 < \tau_i^4 \le +\infty$$

Furthermore, the cut-points are assumed to be determined by covariates. This means that they are considered to vary systematically with individual characteristics; if they were not there would be no response heterogeneity. In equation 3 this mechanism is shown: $X_i^{'}$ is a vector of personal characteristics which all have their own effect (γ) on the value of the individual's cut-point (τ_i).

$$(3) \qquad \tau_i^k = X_i^{'}\gamma^k$$

γ does not only have a different value for each individual characteristic, but also for each cut-point, which is indicated by k . If the effect of individual characteristics on the cut-points was constant there would be no response heterogeneity, but parallel index shift (Lindeboom and van Doorslaer 2004).

Equation 3 provides the essential information to disentangle response heterogeneity and 'true health differences' from the total variation in responses between individuals. That is, it allowed us to determine a value of individual *i's* health on the latent scale using his/her ordinal response to the self-assessment question.

Because the latent scale is one-dimensional the scores of all individuals on the latent scale are comparable.

Equation 4 gives the value on the latent scale Y_i^* for the respondent's own health (s). Like with the vignette questions this value is determined by μ_i and varies, but no longer with 1 but with σ.

$$(4) \qquad Y_i^{s*} \sim N(\mu_i^s, \sigma^2)$$

This μ_i follows from equation 5 in which $Z_i^{'}$ is a set of characteristics for individual i and β is the effect these characteristics have on their 'true health' (i.e. value on the latent scale).

$$(5) \qquad \mu_i^s = Z_i^{'}\beta$$

The corresponding maximum likelihood estimation was performed in Stata MP 10.0 using an appropriately adjusted code that was previously published (Jones 2007). Next to country dummies the set of individual characteristics consisted of age and gender as demographic variables and years of formal education as an indicator of social economic status (Winkleby, Jatulis et al. 1992). This set of characteristics is identical to the one that was used for the estimation of the cut-points, which was driven by the vignettes. This estimation resulted in three important outcomes. First and foremost, this allowed for the prediction of individual performance on the different health domains. Secondly, the HOPIT analysis enabled us to estimate cut-points on the latent scale, again based on individual characteristics. The final output of the HOPIT analysis was a constant estimate of the different vignettes.

3.3.2. Aggregation of health domains

Performing a HOPIT analysis to calculate comparable health estimates based on a single self-reported health question is not feasible, because health is more complex. According to the WHO health is a multi-dimensional concept (World Health Organization 1946). So, the measurement of overall health involves individual attainment on all domains of health. In other words a person's health is determined by individual scores on different domains. In the WHS eight domains of health are distinguished. At first, in the WHO Multi-country Survey Study on Health and Responsiveness (MCSS) (Üstün, Chatterji et al. 2003), the WHO considered health to be best described by six domains (Sadana 2002). However, after this pilot study for the WHS, the WHO decided upon the following eight domains of health: mobility, affect, pain, personal relations, sleep and energy, vision, cognition and self care (Salomon, Mathers et al. 2003). To do justice to the complexity of health we used individual information on all eight domains to estimate health utility. To do so, we performed the previously described HOPIT analysis for each domain separately. Afterwards scores on the eight health domains were weighted to arrive at the overall health measure: health utility.

To enable the union of the eight domains the results from the HOPIT had to be rescaled to a value between zero and one. Recall that the HOPIT provided health on a single domain (i.e. a domain score) based on individual characteristics. This domain score was measured on the latent scale. Thus, its boundaries were minus and plus infinity. To rescale it was necessary to determine 'anchors' on the latent scale which corresponded with the minimum and maximum value on that domain. These values should reflect the best and worst imaginable functioning on the domain of health. In previous HOPIT studies in the health domain (Mathers, Murray et al. 2003; Salomon, Murray et al. 2003), the best and worst vignettes were used as anchors (i.e. the values zero and one on the new scale). However, it is not certain that the vignettes describe the best and worst imaginable scores on a domain. That is, the health of respondents can be better or worse than any vignette. In these studies, individual scores better than the best vignette (i.e. smaller than zero) or worse than the worst vignette (i.e. larger than one) were recoded to zero and one respectively. Consequently, this approach led to loss of variation.

To preserve this variation, we preformed a time trade-off (TTO) analysis, which essentially is a preference based health state valuation method (Bleichrodt and Johannesson 1997). In a TTO study respondents are asked to imagine themselves in a health situation. The health situation contains both a description of a health condition and the number of years the respondent would live in that condition. The respondent is then asked to indicate how many of the remaining life years he/she is prepared to surrender to be cured from this condition. Being cured from a condition would result in living in the best imaginable condition.

We asked 28 individuals to answer such kind of questions. The health conditions we described were identical to the best and worst vignette of each health domain. Hence, in total each respondent answered 16 times how many years in perfect health he/she viewed as equivalent to the concerning vignette. Subsequently, the number of surrendered years was divided by the number of years the respondent would live in the described condition. This ratio is considered to reveal the utility a respondent would derive from the condition.

The results of the TTO were used to rescale the HOPIT results so that zero and one represented the best and worst imaginable health states respectively. Equation 6 shows how the results of the HOPIT and the TTO are combined in domain scores (*D*) on a scale from zero to one.

$$D_i^d = \overbrace{t_+}^{I} + \overbrace{(t_- - t_+)}^{II} \frac{\overbrace{Y_i^s - Y_+^d}^{IV}}{\underbrace{Y_-^d - Y_+^d}_{III}} \tag{6}$$

Where I is the point on the new scale corresponding to the TTO score for the best vignette on health domain d. Part II is the distance between the TTO scores of the best (+) and worst (-) vignettes. Part III also reflects the distance between the best and worst vignette, but then on the latent scale. These values are between minus and plus infinity instead of between zero and one. Element IV indicates the disparity between the score of individual i (i.e. the score on the HOPIT corresponding to his/her characteristics) and the score of the lowest vignette on the latent scale. Accordingly, the fraction III / IV gives the relative part of latent scale covered by individual i. This fraction multiplied by II gives the difference between individual i's domain score and the best vignette on the zero to one scale. Finally, adding I results in an individual score between zero and one. For those cases in which $D_i^d < 0$ or $D_i^d > 1$ we rescaled D_i^d to 0 and 1 respectively.

Subsequently, the domains scores were combined into a single measure of health utility. Simply averaging the scores from equation 6 for all eight domains would be naïve, because clearly not all domains are valued equally. The WHO has investigated values for different health domains and examined different models which would reflect these values best (Salomon, Murray et al. 2003). We used the truncated model with only main effects, because it had the ability to valuate across the full scale (i.e. both good and worse health state); it performs better for good health states than most other models and it is fairly simple (Salomon, Murray et al. 2003). The valuation model had to be adjusted because it only incorporated valuations for the six health domains from the MCSS: affect, cognition, mobility, pain, self care and usual activities. To calculate health utility the function was transformed to facilitate current eight health domains. Again data from the WHS was used to perform this transformation. In the WHS respondents answered a question about their overall health in addition to questions about health domains. We performed a standard ordered probit analysis with this self-reported overall health response as dependent variable and all eight domain responses as determinants.

The coefficients of this analysis were not bounded between two values and their scales were not identical. In other words equal coefficients for different health domains, does not imply equal importance, because the mean and variance differ between domains. Therefore we standardized these coefficients for a unit variance in the independent variables. Standardized coefficients are not uncommon in medicine (Eggena, Barugahare et al. 2005; Kistorp, Faber et al. 2005) and social sciences (Davern, Cummins et al. 2007) but are relatively rare in health economics. The standardized $\beta' s$ are unaffected by the independent variable's underlying scale of units, which allows for better comparison of coefficients. The standardize coefficients were still not restricted by an identifiable scale, but were known to be all measures on the same scale.

Therefore, the relative importance of the domains could be determined by dividing the standardized coefficients by the sum of all standardized coefficients. This relative importance of a domain indicated which part of overall health was determined by that domain. By multiplying this relative importance of domains with the sum of all original coefficients new coefficients (β^*) for the valuation function with eight domains of health were obtained. The constant term was unaffected. The new valuation function is given by equation 7. Individual health utilities (HU_i) can easily be derived from the results of this equation.

$$(7) \qquad 100 * (1 - HU_i) = \beta^0 + \sum_{d=1}^{8} \beta^{d'} D_i^d$$

its coefficients $\beta^{d'}$ are provided in table 3. The values for D_i^d are determined by an individual's (i) result from the HOPIT analysis (see equation 6).

Subsequently, average health utility was calculated for a number of groups, which were defined by country, sex and five year age-intervals. The WHS did not include any respondent below the age of 18. Therefore, the health utility of children and adolescents was estimated using the HOPIT results of the respondents between ages 18 and 25. Very few respondents were older than 85, so also a larger age-interval grouping was used to determine the average health utility for this group of respondents. In this grouping we used the health utility of all respondents above the age of 70.

3.4. Life Expectancy

Life tables are generally used to calculate life expectancies (at birth) for populations. The life tables use the number of person years lived in (all) age-intervals (i.e. typically five years) to calculate the life expectancy. These person years are the product of the people who start an age-interval (i.e. people alive at t=0) and the part of the age interval they on average survive. In other words, the person years are a function of the number of persons, the probability of dying and the average moment of dying during the age-interval.

Annual mortality rates (M_x^{cs}) were the inputs for the life tables. These rates are based on the fraction of the number of deaths during the age-interval (D_x^{cs}) and the size of the population still alive half-way the interval (P_x^{cs}).

$$(8) \qquad M_x^{cs} = \frac{D_x^{cs}}{P_x^{cs}}$$

Here the x indicates the age-interval (Mathers, Vos et al. 2001) and c indicates the country of interest and s denotes the sex.

For the calculation of the person years it is necessary to obtain the probability of dying during an age-interval. This probability is calculated using the population at the start of an age-interval rather than the population still alive half-way the interval. First we have to make an assumption about the moment during an interval at which people on average die. Generally, we assume people to die half-way during an age-interval, in that case the proportion (a_x) of the age-interval lived by those who die equals 0.5. Only in the age-intervals at the start and end of life we assume a smaller proportion, i.e. people on average die before the half of the interval. By using a in equation 9 we can estimate the number op people within age-interval x at the start of the interval (N_x^c).

$$(9) \qquad N_x^{cs} = P_x^{cs} + (1 - a_x) * D_x^{cs}$$

Subsequently, we can estimate the probability to die during age-interval x for someone alive at the start of the interval of death (nq_x) by using the information from equations 8 and 9. The probability of dying is defined as the proportion of those alive at the start of an age-interval that dies during the interval (Mathers, Vos et al. 2001), this is represented in equation 10.

$$\begin{aligned}(10) \qquad q_x^{cs} &= \frac{D_x^{cs}}{N_x^{cs}} \\ &= \frac{D_x^{cs}}{P_x^{cs} + (1 - a_x) * D_x^{cs}} \\ &= \frac{D_x^{c} / P_x^{c}}{P_x^{cs} / P_x^{cs} + (1 - a_x) * D_x^{cs} / P_x^{cs}} \\ &= \frac{M_x^{cs}}{1 + (1 - a_x) * M_x^{cs}}\end{aligned}$$

Using q_x we can compute the number of years someone born today would on average live in each age interval. In other words we want to know how many person-years (L_x^{cs}) he or she is expected to live between the ages 0 and 100.

We first have to calculate the number of person-years someone is expected to live if he/she already survived until the start of the age-interval. A person who survives that age-interval lives the full n years of the age-interval. If someone deceases he/she will only be alive $a*n$ years. The corresponding probabilities are $1-q_x$ and q_x respectively. So, the number of person-years someone alive at the start of interval x is expected to live is given by:

$$(11) \qquad L_x^{cs} = (1 - nq_x^{cs}) * n_x + q_x^{cs} * a_x * n_x$$

These expected person-years are conditional on the survival (S_x^{cs}) of all intervals until then. This again depends on the values of the probability of dying in the age-intervals, namely q_x. That is the product (i.e. multiplication) of all values of q_x^{cs}:

$$(12) \qquad S_x^{cs} = \prod_{x=2}^{x-1} (1 - q_x^{cs})$$

where $x-1$ is the number of age-interval prior to age-interval x, for $x = 1$ the probability of survival equals 1.

By combining equations 11 and 12 for each age interval separately we can compute the average person-years lived by someone born today in each interval.

$$(13) \qquad e_x^{cs} = S_x^{cs} * L_x^{cs}.$$

Life-expectancy at birth (e_0) can then be calculated by summation over all 22 age-intervals.

$$(14) \qquad e_0^{cs} = \sum_{x=0}^{22} e_x^{cs}$$

3.5. Health expectancy

The expected number of person-years from equation 13 could also be multiplied with the health utilities corresponding to the same age-intervals. This resulted in health adjusted life expectancy (HALE) at after summation over all age-intervals. This methodology to combine health utility and mortality is referred to as Sullivan's method (Mathers, Sadana et al. 2000).

$$(15) \qquad HALE_0^{cs} = \sum_{x=0}^{22} S_x^{cs} * L_x^{cs} * HU_x^{cs}$$

3.6. Distribution of health

The above calculation of health adjusted life expectancy is based on life expectancy measured by *average* survival and health utility. However, these are unsuitable for the measurement of health inequalities (i.e. the variation in HALE), because HU^{cs} is assumed to be equal for all individuals. Secondly, the life expectancy does not vary across individuals either, although it is evident that not all individuals of a population exactly live the life expectancy.

In the next session we describe how we adjusted the calculation of $HALE_0^{cs}$, so that we could measure variation in both survival and health utility. First of all, we will introduce probabilistic sampling, based on probabilities to die, to measure the variation in mortality. Next, we will combine individual responses in the WHS sample with the HOPIT results to determine the distribution of health utility. Finally, we shall use data on the correlation between health utility and survival to combine these two new measures into individual healthy life spans.

3.6.1. Variation in survival

We constructed hypothetical samples of the population to calculate the distribution of healthy life spans. This provides more meaningful estimates than performing follow-up studies of the WHS respondents. Such a follow-up would result in different HALE estimates for all respondents and that distribution could function as a measure of inequality. However, that would not give information about inequality at this moment, but about inequality in the future. More importantly when that information would have become available it would have been information about the past. Therefore we constructed hypothetical samples of 100,000 individuals, of whom we would 'predict' their health.

Each hypothetical person in the sample was assumed to be born at the same time. All of them had the same chance to survive to the age of fifteen, namely one minus the probability of death. We chose intervals with a length of 15 years, because this allowed for the best possible combination with health utility in a later stage. In formal terms the probability to survive the interval, with length 15, was:

(16) $$q_{x+15} = 1 - \prod_{t=0}^{t=15}(1 - q_{x+t})$$

Where $1 - nq_{x+t}$ is the probability to survive a part of the fifteen year interval, for instance from the age of five to the age of ten. Thus, a proportion according to the probability of death died and the rest survived. For the former the ultimate life 'expectancy' could be calculated, for the others the life expectancy was still unknown, because it was determined by survival in the subsequent intervals. The average life expectancy of the deceased was calculated using the different $1 - q_{x+t}$ within the fifteen year interval, i.e. the probabilities of death for the five year periods constituting the fifteen year period.

The first step in the calculation was to determine the number of years lived by the people who died in the first five year interval, which is equal to $x + a_x * 5$, where x is the number of years lived until the start of the fifteen year interval and a is the average point of dying during the interval (see equation 9). The chance this will be someone's life expectancy is q_x, the probability of dying. The average number of years lived by someone dying during the second five year interval is given by $x + 5 + a_{x+5} * 5$, the plus five is added here because this person already survived the first five year interval. The corresponding chance of dying in that interval is conditional on surviving the first five year interval (i.e. $1 - q_x$) and dying in the second interval (q_{x+5}). The calculation of the years lived by a person in the third (and last) five year interval is the same and given by $x + 10 + a_{x+10} * 5$, the chance to actually die in that interval is given by $(1 - q_x) * (1 - q_{x+5}) * q_{x+10}$. After division by the total probability to die in the fifteen year interval (q_{x+15}), the average life expectancy of someone who dies between the ages of x and x+15 is estimated, formally,

(17) $$(LE^{cs}_{x+15} \middle| S^{cs}_x = 1 \middle| S^{cs}_{x+15} = 0) = \frac{1}{q_{x+15}} * ((x + a_x * 5) * q_x + (x + 5 + a_{x+5} * 5) * (1 - q_x) \\ * q_{x+5} + (x + 10 + a_{x+10} * 5) * (1 - q_x) * (1 - q_{x+5}) * q_{x+10})$$

To clarify, suppose that the probability of death between the age of zero and fifteen is 0.006. Then if there are 100,000 newborns (i.e. the size of the hypothetical sample) about 600 (0.006) of them will not survive to the age of fifteen. Those children were assumed to all have died at the age calculated as described above. Because, survival was sampled (i.e. a random number determined whether someone survived or not) not exactly 600 out of 100,000 would die, but for example 596 or 603.

Subsequently, approximately 99,400 survivors would reach the start of the next age-interval (i.e. 15-29) and again faced a chance of dying and a chance of surviving. Once more we sampled survival for this age-interval and determined the life ‘expectancy’ for the ones who died, in this case 15 (i.e. the length of the previous age-intervals plus average years lived in the current age-interval before dying). This procedure was repeated for all age-intervals until the age of 105 was reached, when all remaining survivors were assumed to die.

This method of sampling survival among hypothetical newborns is an application of probabilistic sampling. The survival (based on a probability) in the first sample determines the size of the second sample, and so on. Life ‘expectancy’ for 100,000 newborns was estimated in this way for all countries in our subset of the WHS. Because the probabilities of death differ between these countries the composition of the life ‘expectancies’ in the hypothetical sample also differ. This compositional difference partly determines which country performs relatively well on the distribution of health and which does not.

3.6.2. Variation in health utility

However, this method only takes variation in mortality into account and in the earlier we argued that we were interested in the distribution of mortality and health utility combined. When we tried to estimate the variation in health utility between individuals, we were faced with a drawback of the standard application of the HOPIT analysis. In the HOPIT analysis individual scores on a health domain were determined by personal characteristics. However, this meant that individuals with the same characteristics all had the same domain scores and because the same characteristics were used for all health domains, they also all had the same health utility. In reality there will be variation in health utility between individuals with the same characteristics. Applying the average health utilities in the probabilistic sampling process would lead to an underestimation of health inequality.

A good indicator of individual performance on a health domain was the individual response to the domain question. A value on the latent scale corresponding to this individual response was not available, because an ordinal response only provides information about an interval in which the actual individual score lies. However, the boundaries of this interval are known from the HOPIT analysis (i.e. the cut-points estimated using the vignette responses). For the response 1 (no problems) we used the mid-point between the first cut-point and the best imaginable health state (equation 6) and for response 5 (extreme problems) we used the mid-point between the fourth cut-point and the worst imaginable health state. In those cases where either the first cut-point was smaller than the best imaginable health state or the fourth cut-point was larger than the worst imaginable health state, we rescaled the value back to the best and worst imaginable health state respectively. An individual score was then determined by using the mid-point value between these boundaries. After repeating this for all domains of health, we first rescaled the scores back between zero and one using equation 6 and then used equation 7 to calculate individual health utility scores.

3.6.3. Adding variation in health utility to variation in mortality

Subsequently, we attached these health utilities to years lived in the different age-intervals by the hypothetical sample described earlier. We did this randomly by sampling a single *individual* health utility from the WHS dataset and attaching this to a person in the hypothetical sample. We only matched for country, age and gender. For example, in the hypothetical sample for French males there was a person still alive at the start of the age-interval from 30 to 45, we then selected an *individual* health utility from the WHS belonging to a men from France between the ages of 30 and 75. If the hypothetical person survived the age-interval fifteen times (i.e. the length of the age-interval) the health utility was added fifteen times to his healthy life span; if he died the health utility was added fewer times. The number of times the health utility was added depended on the results of equation 17.

3.6.4. Correlating mortality and health utility

This method combined variation in both survival and health utility. Yet, one issue remained. We earlier reasoned that mortality and health utility should be combined in a measure of health inequality, because the two are correlated. The approach described above does not yet facilitate this correlation.The causal relationship is clear in this case: lower health utility presents itself before death. Therefore, we first sampled the health utility from the WHS dataset. Secondly, we adjusted the probabilities of death in the hypothetical sample based on this sampled health utility for each individual. This adjustment was performed through a hazard ratio (HR=0.47 for a decrease of one unit health utility) that relates health utility to all-cause mortality (Kaplan, Berthelot et al. 2007). First individual hazard ratios were calculated using the individual health utility, average health utility and the overall hazard ratio,

$$(18) \qquad HR_i^{cs} = EXP(LN(HR) * (HU_i^{cs} - \overline{HU}^{cs}))$$

Secondly, the average probability ($\hat{q}_x$) to die in a specific (x) fifteen year interval was transformed to an average hazard to die in that interval ($\hat{h}_x$).

$$(19) \qquad h_x^{cs} = -LN\left(1 - q_x^{cs}\right)$$

After multiplication of equations 18 and 19, the individual hazard was recalculated to an individual probability to die in the next fifteen years.

$$(20) \qquad q_{x,i}^{cs} = 1 - EXP(-HR_i^{cs} * h_x^{cs})$$

The individuals in the hypothetical sample now no longer had the same probability of death, but an adjusted probability based on their sampled health utility. Those who got a health utility attached that was higher than the average had a lower chance of death and vice versa. Finally, we used the adjusted probabilities of death to determine the survival per age-interval; we attached the sampled health utility to the survived years in the corresponding age-interval and summed these over all age-interval to calculate health adjusted life expectancy.

Using this method for each country a hypothetical sample was constructed with 100,000 different healthy life spans. Sex ratio at birth was used to determine the number of men and women in the hypothetical samples (Norberg 2004). The distribution of the resulting healthy life spans reflected the true distribution of health in that country.

To show the relative equality of these distributions, the objective of this analysis, we calculated the Gini coefficient (Atkinson 1970) for each country separately. The Gini coefficient first compares each (hypothetical) individual's health with that of all other individual's in the country (sample). Next, the absolute difference between each possible pair of individual health scores is summated. Finally, this quantity is adjusted to the size of the country (sample) and the average level of health. Equation 21 shows the formal definition of the Gini coefficient (Gakidou, Murray et al. 2003).

$$(21) \qquad G^c = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \left| h_i^c - h_j^c \right|}{2 {n^c}^2 \overline{h}^c}$$

In this equation *h* denotes the health of individuals *i* and *j* who are part of the population from country *c* with size *n*. Gini coefficients were only calculated for countries as a whole, there was no division by gender because these inequalities are also relevant and unwanted. Calculating the Gini coefficient using its formal definition is computationally very demanding. Therefore we rewrote equation 21 which yields the exact same results if the dataset is sorted in ascending order (i.e. $h_i^c \geq h_j^c$). This adjusted formula is given in equation 22. In which the expression R_i^c is equal to the rank (in order of worst to best) of individual i's (Damgaard and Weiner 2000).

$$(22) \qquad G^c = \frac{\sum_{i=1}^{n}(h_i^c * (2R_i^c - n^c - 1))}{n^{c^2}\overline{h}^c}$$

3.7. Confidence intervals

3.7.1. Uncertainties in the analysis

There were a lot of inputs involved in the analysis of both the level of health and the distribution of health. The results of the analysis depended heavily on the values of these inputs. However, the values of many inputs were not definite, because estimations cohere with uncertainty. In the confidence intervals around the point estimates of health expectancy and the Gini coefficients these uncertainties should be incorporated (Salomon, Mathers et al. 2001). Inputs that led to uncertainty around the point estimate were: the WHS sample taken by the WHO and all estimates derived from it; the valuation of the worst and the best vignette in the time trade-off analysis; the coefficients of the original valuation function; and the drawn hypothetical samples. Conversely, the life expectancies do not cause any additional uncertainty, as the mortality rates are based on observed deaths in the entire population.

3.7.2. Resampling

It was not feasible to estimate the uncertainty associated with all these inputs analytically. If it would have been possible to do so, we could have estimated the distribution (*F*) of both HALE and the Gini coefficients and could have determined the range from the point estimates to the boundaries of the confidence interval. In the bootstrap procedure (Efron 1979) we performed the entire calculation from start to end multiple times. In each *run* of the analysis different values for all the inputs were chosen randomly, yielding different outcomes every time. Al these different outcomes together formed an empirical estimation of the distribution of HALE and the Gini coefficient. In other words, we took B samples (x_c^*) with size N (i.e. the number of respondents from the WHS dataset). In each sample different values for the other inputs were selected based on their reported distributions. All B samples ($x_c^{*1}, x_c^{*2}, x_c^{*3},..., x_c^{*B}$) resulted in B different values for the HALE

($HALE_c^{*1}, HALE_c^{*2}, HALE_c^{*3},..., HALE_c^{*B}$) and the Gini coefficient ($G_c^{*1}, G_c^{*2}, G_c^{*3},..., G_c^{*B}$). This last set of values together resulted in an empirical distribution ($\hat{F}_c$) of the Gini coefficient for each country (c).

Ideally we would have constructed an empirical distribution using a large number of bootstrap samples (B ≥ 10,000). In that case we could have dropped the 2.5% lowest and highest results. That would have given good indication of the uncertainty associated with our findings. However, the entire analysis was computationally very demanding, preventing the bootstrap to produce sufficient replications. Time constraints limited us to a total number of 50 replications. Therefore, the assumption was made that the empirical distribution of outcomes (both HALE and Gini coefficient) was normally shaped (Efron 1985), formally

$$(23) \qquad \frac{G_c^b - G_c}{se_c^B} \sim N(0,1)$$

in which G_c^b is the health adjusted life expectancy calculated for country c in bootstrap sample b; G_c is the point estimate of health expectancy in country c; and se_c^B is the standard error of health expectancy in country c over the B bootstraps. After performing the bootstrap we checked whether this assumption in fact did hold.

From the empirical distribution we first calculated the standard error of the HALE and the Gini coefficient for each country. We used the formula for the standard deviation provided by Efron and Tibishirani (Efron and Tibshirani 1986) to estimate this standard error. In equation 24 the example is given for the standard error of the Gini coefficient, but the calculation of the standard error of the HALE is identical, in that case G_c^b is interchanged for $HALE_c^b$

$$(24) \qquad se_c^B(G) = \sqrt{\sum_{b=1}^{B}\left[G_c^b(b) - \sum_{b=1}^{B}\frac{G_c^b(b)}{B}\right]^2 /(B-1)}$$

In this equation G_c^b is the Gini coefficient calculated for country c using sample b. The total number of bootstrap samples is given by B. Equation 24 can be rewritten to equation 25 because $\sum_{b=1}^{B}\frac{G_c^b(b)}{B}$ is equal to the average value of the gini coefficients in all bootstrap samples.

$$(25) \qquad se_c^B(G) = \sqrt{\frac{\sum_{b=1}^{B}\left[G_c^b - \overline{G}_c\right]^2}{B-1}}$$

Equation 25 shows that the estimation of the standard error is similar to the calculation of a standard deviation in an arbitrary sample.

If the assumption of normality (equation 23) holds, the calculation of the standard error allows for the calculation of the 95% confidence interval for each country through equation 26. All parameters have been explained above, except $z^{(0.025)}$ and $z^{(0.975)}$ which are values from the standardized normal distribution.

$$(26) \qquad CI_{95\%} = \left[G_c - z^{(0.025)} * se_c^B ; G_c + z^{(0.975)} * se_c^B\right]$$

4. Results

Figure 3a and 3b depict the relative shift of the cut-points for each country. These figures show that response heterogeneity is present in the health questionnaire of the WHS.

Figure 3a: Response heterogeneity by country, domain and cut-point, first set of domains

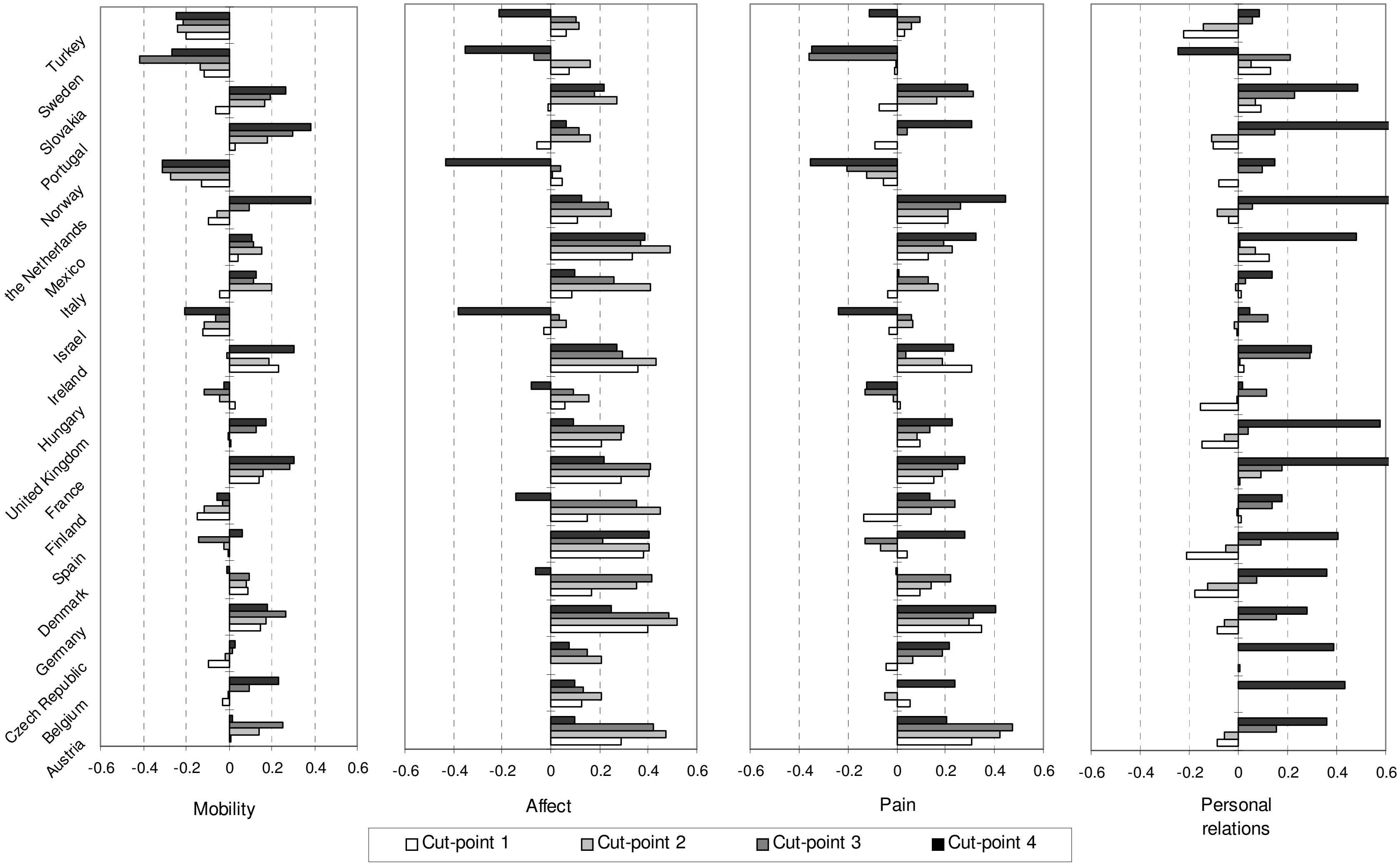


Figure 3b: Response heterogeneity by country, domain and cut-point, second set of domains

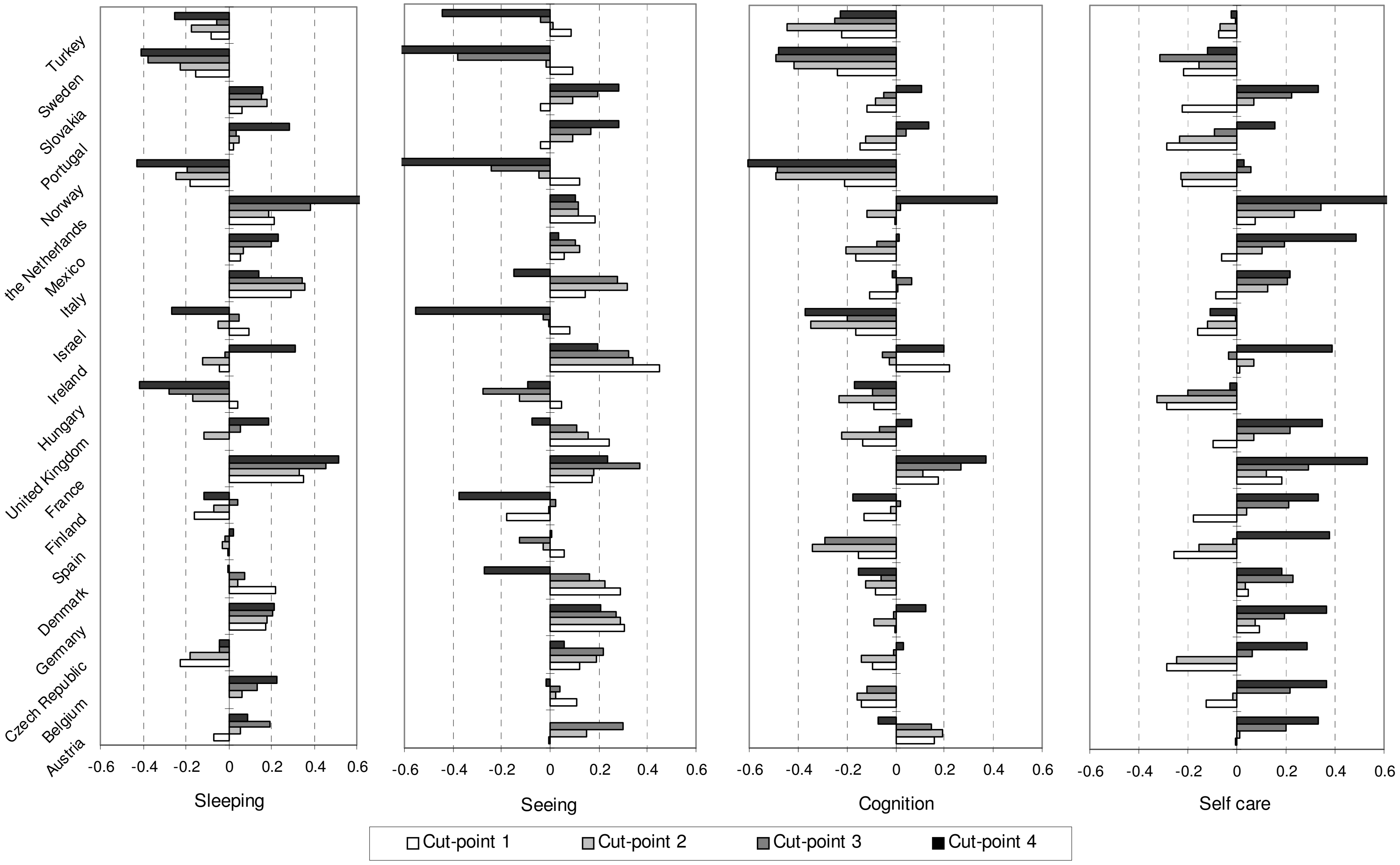


Table 2 shows the results of the HOPIT analyses of the functioning on health domains. These are values on the latent variable and have not yet been rescaled to a zero to one scale. The absolute values of the figure have no intuitive interpretation. The sign of the values show the direction of the country effect. That is, a negative value indicates that a country performs better than the reference country (Greece) on that specific domain of health; conversely a positive sign indicates worse performance. This can be generalized as follows higher values indicate less health.

Table 2: Results Hierarchal Ordered Probit Analysis (HOPIT)

Domain of Health	Mobility	Affect	Pain	Interpersonal realtionshi	Vision	Sleep and Energy	Cognition	Self Care
	χ^2 5,985 P< 0.001	χ^2 3,043 P< 0.001	χ^2 4,215 P< 0.001	χ^2 1,792 P< 0.001	χ^2 3,766 P< 0.001	χ^2 3,763 P< 0.001	χ^2 3,475 P< 0.001	χ^2 3,601 P< 0.001
Covariate	Coefficient	Coefficient	Coefficient	Coefficient	Coefficient	Coefficient	Coefficient	Coefficient
Gender	-0.247**	-0.407**	-0.282**	-0.146**	-0.205**	-0.299**	-0.299**	-0.162**
Age	0.028**	0.009**	0.016**	0.015**	0.024**	0.015**	0.017**	0.030**
Education years	-0.042**	-0.024**	-0.030**	-0.034**	-0.023**	-0.017**	-0.028**	-0.057**
Austria	0.122	-0.605**	0.061	-0.167	-0.413**	-0.183	-0.059	0.038
Belgium	0.343**	-0.172	0.133	0.367**	-0.131	0.348**	0.252*	0.236
Czech Republic	0.692**	-0.070	0.480**	0.258	-0.127	0.397**	0.428**	0.556**
Germany	0.656**	-0.172	0.284**	0.083	-0.013	0.439**	0.078	0.359*
Denmark	0.245	-0.146	0.291**	-0.242	-0.494**	0.193	0.288**	-0.286
Spain	-0.041	0.007	-0.095	-0.122	-0.056	-0.098	-0.007	-0.091
Finland	0.124	-0.374**	0.140	-0.012	-0.800**	-0.021	0.208*	-0.245
France	0.302**	0.034	0.194*	0.305	0.093	0.489**	0.706**	0.468**
United Kingdom	0.465**	-0.162	0.300**	-0.049	-0.489**	0.372**	0.179	0.307*
Hungary	0.514**	-0.267**	0.039	-0.025	-0.262*	0.040	0.209*	0.084
Ireland	-0.067	-0.482**	-0.247*	-0.325*	-0.382**	-0.047	-0.001	-0.055
Isreal	0.649**	-0.108	0.305**	0.639**	0.111	0.425**	0.414**	0.611**
Italy	0.341**	-0.010	0.078	0.148	0.236*	0.320**	0.353**	0.223
Mexico	0.124	-0.106	0.007	-0.280*	0.251*	-0.333**	0.053	0.169
the Netherlands	0.531**	-0.082	0.485**	0.363**	-0.357**	0.296**	0.499**	-0.051
Norway	-0.150	-0.578**	-0.150	-0.143	-0.954**	-0.237*	-0.118	-0.901**
Portugal	0.443**	-0.165	-0.067	-0.067	0.142	0.157	0.405**	0.054
Slovakia	0.971**	-0.184	0.259**	0.653**	0.278*	0.152	0.305**	0.330*
Sweden	0.396**	-0.169	0.076	0.178	-0.544**	0.304**	0.365**	-0.095
Turkey	0.519**	0.395**	0.483**	-0.062	0.203*	0.183*	0.365**	0.583**

** significant at 1% Level

* significant at 5% Level

Table 2 does not allow for a comparison of the domains as the scales have no interpretation and the relative importance of each domain is unknown. In Table 3 the relative importance is stated as it shows the results of the ordered probit analysis of individual domain responses on overall self reported health. The results show that the most important determinants of health utility are mobility and pain.

Table 3: Adjustment of valuation function

Domain	Original coefficient	Relative importance*	New coefficient
Affect	26.65	13%	12.49
Cognition	12.36	11%	10.77
Mobility	15.72	28%	27.15
Pain	9.01	30%	28.85
Self care	22.85	4%	3.48
Usual activities	10.89	N/A	N/A
Interpersonal relations	N/A	1%	0.59
Vision	N/A	4%	4.37
Sleep and energy	N/A	10%	9.78
Constant	2.43	N/A	2.43

* The domain coefficient from the ordered probit analysis divided by the sum of all coefficients

Table 4 provides estimates of health, including a separate morbidity and life expectancy component. The health utility scores are averages of the WHS dataset and are the result of the HOPIT analysis and the valuation function. Men have higher health utility than women and Ireland performs best on this component of health, Irish men and women score on average 96.9 and 93.8 respectively. Turkey on the other hand clearly has the lowest health utility of all participating countries with scores of 76.6 and 58.5 on average for men and women. The life expectancy estimates show that women on average live longer than men do. Mortality is lowest among French women; they are expected to live for 83.9 years. Swedish men have a higher life expectancy at birth than men from any other OECD countries (78.7). Again Turkey scores lowest with life expectancy of 73.6 for women and 69.0 for men. The last two columns present the health adjusted life expectancy by country and gender. Again Ireland and Turkey are the best and worst presenting country with HALE scores of 75.3 and 74.4 for Irish women and men respectively and 42.0 and 54.3 for Turkish women and men.

Table: 4 Health Utility, Life Expectancy and HALE by country and gender

	Health utility scores		Life expectancy		Health adjusted life expectancy			
	females	males	females	males	females		males	
					mean	(95% CI)	mean	(95% CI)
Austria	88.3	94.9	82.2	76.7	71.4	(69.0; 73.7)	72.2	(71.1; 73.3)
Belgium	82.4	92.0	81.5	75.6	66.9	(63.6; 70.3)	70.0	(68.6; 71.5)
Czech Republic	64.3	81.9	79.3	72.9	55.5	(51.3; 59.6)	62.9	(60.6; 65.2)
Germany	67.2	82.4	82.0	76.5	61.0	(56.4; 65.5)	67.0	(64.5; 69.5)
Denmark	72.7	86.2	80.4	75.6	63.0	(59.2; 66.7)	68.0	(66.1; 69.8)
Spain	77.7	90.4	83.6	76.9	69.4	(66.9; 71.9)	71.8	(70.7; 72.8)
Finland	78.4	90.1	82.4	75.7	69.1	(66.3; 72.0)	70.2	(69.0; 71.4)
France	76.9	90.1	83.9	76.8	62.9	(58.3; 67.4)	68.0	(65.6; 70.4)
United Kingdom	72.8	81.9	81.1	76.6	62.4	(58.9; 66.0)	68.0	(66.1; 69.9)
Hungary	79.4	91.8	77.1	68.6	65.4	(62.8; 67.9)	64.8	(63.9; 65.7)
Ireland	93.8	96.9	81.3	76.9	75.3	(73.4; 77.1)	74.4	(73.5; 75.3)
Israel	74.4	89.3	82.2	78.1	61.8	(58.0; 65.7)	68.6	(66.4; 70.9)
Italy	72.9	90.5	83.8	77.9	64.7	(61.2; 68.1)	70.7	(68.9; 72.4)
Mexico	84.3	92.5	76.9	71.8	62.0	(59.5; 64.5)	65.7	(64.6; 66.7)
the Netherlands	69.3	87.7	81.3	76.9	58.4	(53.2; 63.6)	66.5	(63.5; 69.4)
Norway	91.2	95.9	82.4	77.5	71.5	(69.8; 73.1)	79.2	(78.4; 80.1)
Portugal	74.2	88.0	81.5	74.9	63.4	(60.1; 66.7)	67.6	(65.9; 69.2)
Slovakia	83.8	94.6	78.0	70.1	60.9	(57.9; 63.9)	63.9	(62.7; 65.1)
Sweden	76.6	88.3	83.0	78.7	67.6	(64.4; 70.8)	71.7	(70.0; 73.4)
Turkey	58.5	76.6	73.6	69.0	42.0	(37.6; 46.3)	54.3	(51.7; 56.9)

In table 5 Gini coefficients are given for all participating countries. Again scores are presented based on three different methods of measurement: (i) using individual health utility scores from the WHS dataset, (ii) the results of probalistic sampling of life expectancy, and (iii) the probalistic sampling of healthy life spans. The first method of measurement leads to the largest variation within countries (i.e. the highest Gini coefficients) and between countries. Norway is the most equal country with regard to the distribution of health utility. Turkey has the highest Gini coefficient (0.305), i.e. the largest differences in health utility. The differences in the distribution of life expectancy are smaller. The highest Gini coefficients measured in that way can be found in Mexicio (0.122) and Turkey (0.124), the lowest inequalities are found in Sweden (0.077). In all other countries except Hungary the Gini coefficient lies between 0.082 and 0.099. The Gini coefficients based on the distribution of health adjusted life expectancy are generally higher than those based on life expectancy, but smaller than those based on health utility. The lowest inequality is found in Norway (0.110) and the highest in Turkey (0.189).

Table 5: Gini scores by type of health differences

	Health Utility		Life Expectancy		Health Expectancy	
	Mean (95% CI)		Mean (95% CI)		Mean (95% CI)	
Austria	0.198	(0.175; 0.221)	0.085	(0.084; 0.085)	0.121	(0.116; 0.126)
Belgium	0.214	(0.172; 0.255)	0.087	(0.085; 0.089)	0.126	(0.117; 0.136)
Czech Republic	0.290	(0.237; 0.342)	0.091	(0.089; 0.092)	0.139	(0.120; 0.158)
Germany	0.267	(0.214; 0.321)	0.085	(0.084; 0.087)	0.132	(0.124; 0.141)
Denmark	0.215	(0.142; 0.288)	0.089	(0.088; 0.090)	0.128	(0.120; 0.135)
Spain	0.247	(0.213; 0.281)	0.085	(0.083; 0.086)	0.122	(0.115; 0.128)
Finland	0.214	(0.161; 0.267)	0.090	(0.089; 0.091)	0.126	(0.120; 0.133)
France	0.218	(0.180; 0.257)	0.092	(0.091; 0.093)	0.125	(0.115; 0.136)
United Kingdom	0.264	(0.209; 0.319)	0.087	(0.086; 0.087)	0.138	(0.129; 0.146)
Hungary	0.242	(0.199; 0.284)	0.109	(0.108; 0.110)	0.140	(0.133; 0.147)
Ireland	0.161	(0.141; 0.180)	0.082	(0.081; 0.083)	0.117	(0.111; 0.124)
Israel	0.248	(0.204; 0.291)	0.083	(0.082; 0.084)	0.134	(0.126; 0.143)
Italy	0.240	(0.188; 0.292)	0.083	(0.082; 0.084)	0.124	(0.099; 0.149)
Mexico	0.208	(0.188; 0.229)	0.122	(0.120; 0.125)	0.156	(0.152; 0.161)
the Netherlands	0.231	(0.132; 0.329)	0.082	(0.081; 0.083)	0.130	(0.118; 0.142)
Norway	0.152	(0.119; 0.184)	0.082	(0.081; 0.083)	0.110	(0.104; 0.116)
Portugal	0.274	(0.233; 0.316)	0.091	(0.090; 0.092)	0.134	(0.124; 0.144)
Slovakia	0.222	(0.200; 0.245)	0.099	(0.098; 0.101)	0.138	(0.132; 0.144)
Sweden	0.232	(0.159; 0.304)	0.077	(0.076; 0.078)	0.126	(0.116; 0.136)
Turkey	0.305	(0.249; 0.361)	0.124	(0.122; 0.126)	0.189	(0.176; 0.203)

5. Discussion

This analysis provides estimates of health inequality based on good data for the selected countries and recently developed appropriate methods. The used approach assures comparability of results and incorporates all uncertainty resulting from the different inputs. Our method of combining health and mortality estimates is unique and allows for better assessment of true inequalities than previous studies. The hierarchal ordered probit models allowed for valid comparisons between countries with regard to health utility, even though there was concern for response heterogeneity. The resulting estimates were combined with mortality estimates in accordance with available data on the correlation between these two components of health.

In the introduction we argued which health inequalities should be measured for a meaningful comparison of health equity between countries. In the remainder of the article we have shown that this type of measurement is also possible. Our method uses a different theoretical foundation, with which we aimed at measuring a larger part of the variation relevant in health than traditional approaches do. In addition we have tried to address a number of methodological issues in preceding studies.

First of all, this analysis is the first to use comparable health estimates from a HOPIT analysis in any measure of health inequality. As mentioned earlier many previous studies used self-rated health to determine individual health. Such a measure raises questions about the comparability of data, especially in an international comparison. By correcting for response heterogeneity through a HOPIT analysis our results ensure comparability of health, within and between countries.

Furthermore, previous HOPIT analysis that involved rescaling of the latent variable used the best and worst vignettes as end points (Mathers, Murray et al. 2003). Because all scores that are respectively lower or higher than these vignettes get the same score on the new scale, there is a loss of variation. By using a TTO analysis to put the best and worst vignettes somewhere between the endpoints of the new scale, this variation was better preserved in our analysis.

While our findings may be valid, reliability is problematic. Our results show relatively large confidence intervals compared to the point estimates of health inequality. Considering that the point estimates of most countries are very close, it is impossible to do any enunciation about relative performance of most countries. The large confidence intervals are the result of the many different inputs of our analysis. Each of these inputs has its own uncertainty. All these uncertainties together amount to the large confidence intervals around the Gini coefficients. This shows that there is a lot of room for improvement concerning the reliability of health inequalities in different countries. However, it does not mean that our method is biased. Continuing research into the different inputs will decrease the uncertainty and will result in smaller confidence intervals.

5.1. Interpretation of results

The health utility estimates for men and women are remarkably far apart. However, the estimates from table 4 are not the best indication of cross-sectional health in the participating countries. These estimates are averages of the health utility over all age groups. However, the WHS did not take representative samples, for example all children below the age of 18 were excluded. We know that this group forms a substantial part of any society and that its health utility is unequal to that of WHS average (table 1). However, our estimates of healthy life expectancy have been corrected for the absence of this age group in the WHS by using the country life table to show the future population structure.

These estimates show small differences in HALE between men and women, which is the result of the higher health utility for men and the longer life expectancy for women. Even though, we did not perform statistical tests to assess the significance of these intervals, the overlapping confidence intervals show that females and males are about in equal health. These large intervals also make it impossible to draw any conclusions about the relative rankings of the OECD countries with respect to health, although Norway (with regard to men) and Turkey seem to stand out as most and least healthy country respectively.

Our estimates with regard to health inequality provide insight in the characteristics of the three measures of health underpinning the Gini coefficients. The variation is largest in the Gini-coefficients based on health utility. This is a result of using a relatively small dataset which is not a representative sample of the entire population, some parts of the population (e.g. children and adolescents) were not (fully) represented. In addition, the HOPIT analysis is not very suitable as a measure of *individual* health utility, because its calculation is based on group characteristics and not on individual scores.

On the other hand, measuring health inequality through mortality rates derived from life tables leads to the lowest estimates. A possible explanation is that this method assumes individuals to be equal that are not. That is, if all people were assumed to die at the same age it would assume that there are not health inequalities. While in fact there can still be inequalities due to variation in health utility.

Even though these two measures solely are not able to capture the variation in health within a country, combined they do provide a good measure of health inequality. In our HALE inequalities we resolved the problem of a small dataset by creating a hypothetical sample in which we sampled health utility for all ages, so that this sample was representative, even for children. In addition it allowed for variation in health utility over the life span. Therefore, our measure of health inequality provides us with a better insight in the true distribution of health in the OECD countries than measures based solely on either health utility or mortality.

5.2. Assumptions

The HOPIT analysis is a fairly new method that solves problems associated with the measurement of self reported health. However, for the HOPIT analysis to give valid results two measurement assumptions have to be made. First of all, there should be response consistency, which states that respondents should use the response categories in the vignette questions in the same way as in the self-assessment question. For example, the results would be biased if a respondent would rate the health described in a vignette as having no health problems, whereas he would rate it as having mild problems if he would be in the same health state himself. This would imply that the frame of reference in the vignettes is different from that in the self-assessment question. In that case cut-points estimated through the vignettes would not be applicable for the persons own health.

Secondly, vignette equivalence is required to perform a HOPIT analysis; this in fact is an assumption of unidimensionality. So, differences in responses to the vignette questions should be the result of different ratings of the one latent scale (e.g.

mobility) and not be influenced by anything else (e.g. age or another health domain) (King, Murray et al. 2003).

To add the scores of all individual domains into one single measure of health a global valuation function was used. This implies that the weight of each domain is the same for everybody irrespective of age, sex, nationality, etc. However, it is questionable whether this is true in all cases. In addition it is not only uncertain how important each health domain is, but also if these domains capture all aspects of health. It might be the case that the current set of health domains does not include all domains considered to be important in terms of societal health goals (Salomon, Mathers et al. 2003).

Due the computational complexity of the analysis it was not possible to perform more then 47 successful replications. Therefore, the assumption that both the HALE estimates and the gini coefficients were normally distributed was made to determine the confidence intervals. However, it is possible that the underlying distribution of these two outcomes is not normal. The tests for normality did not provide clear cut conclusions about this issue. In three of the twenty samples of gini coefficients based on HALE the assumption of normality did not hold.

5.3. Limitations

Remarks have to be made about some aspects of the analysis. First of all, the average height and weight by age in Austria were higher for women than for men. We believed this to be a coding error in the WHS data which we corrected. However, we could not verify this, because we did not have the original questionnaires to our disposal.

Secondly, the only available valuation functions for the construction of an overall measure of personal health were not suited for the set of health domains in the WHS. The adjustment through an ordered probit of the WHS health domains on overall health seems to be a good method to solve this problem. However, the fact that the proportions of the valuation function and the standardized coefficients of the ordered probit regression differ substantially, is reason for concern. In addition, the valuation function had a large impact on the outcomes. Therefore, it would have been desirable to estimate a new valuation functions with the new set of health domains from the WHS.

Thirdly, the TTO analysis did allow for a more accurate measurement of individual attainment on a health domain. However, due to time and costs constraints it was not possible to perform this analysis in a large representative sample; it only contained 28 higher educated respondents. We also have to address the limited number of respondents in some of the countries. The WHO stated that it would require at least 1,000 respondents per country (Üstun, Chatterji et al. 2003). Table 1 shows that, even though our sub sample of the WHS contained 73,762 respondents from 21 countries, in most countries the sample was smaller than 1,000. By using sub-optimally large groupings within countries we managed to generate estimates for almost all countries. Unfortunately the sample in Greece was so small (N=441) that the results took on extreme values. Therefore, we decided to expunge the estimates for Greece from all output tables and figures.

The Gini coefficient based on morbidity were relatively high, because they were calculated by combining individual responses and the results of the HOPIT analysis. Recall that we tried to minimize response heterogeneity as much as possible by using the cut-points to determine individual scores on the latent scale. The average health utility resulting from this method diverged importantly from that in the standard HOPIT analysis. Therefore, we used the ratio of the two health utility measures (i.e. the point estimate and the health utility with variation) to rescale individual health utility. This ratio was so large that it occasionally resulted in unrealistically small (< -0.1) and large (> 1) health utility scores. For the measurement of the distribution of HALE this was not a large problem, because on health utility score only formed a small part of an individual's total health expectancy and more importantly that distribution contained much more (hypothetical) individuals than the WHS sample. This correction influenced the measurement of the distribution of health utility greatly.

5.4. Comparison with the World Health Report 2000

A similar attempt to measure health inequality in a large number of countries has to our knowledge only been performed by the WHO in the World Health Report 2000 (World Health Organization 2000). Those estimates were not only criticized for the lack of high-quality data (Williams 2000; Almeida, Braveman et al. 2001), but also because they were based on differences in under five mortality rates (Anand, Ammar et al. 2003). This is not a suitable measure for measuring health inequalities in developed countries, like the OECD countries. The variation in life expectancy is hardly affected by differences in mortality at the very young ages, differences between the ages of five and forty have greater explanatory power (Murray, Kulkarni et al. 2005). Nonetheless, we will try to compare our results with those of the World Health Report 2000. Their results also showed Turkey performed much worse than all other participants. Turkey ranked 109th of a total 191 countries, all other countries ranked within the first forty countries. Of the participating countries in this study the United Kingdom performed best and Norway second best. In fact, Norway scored best in our analysis. However, this seems more of a coincidence as most other rankings are dissimilar (World Health Organization 2000). The lack of discriminatory power in these countries of under five mortality rates can well be a cause of that.

References

Almeida, C., P. Braveman, et al. (2001). "Methodological concerns and recommendations on policy consequences of the World Health Report 2000." The Lancet **357**(9269): 1692-1697.

Anand, S., W. Ammar, et al. (2003). Report of the Scientific Peer Review Group on Health Systems Performance Assessment. Health Systems Performance Assessment: Debates, Methods and Empiricism. C. J. L. Murray and D. B. Evans. Geneva, World Health Organization**:** 839-916.

Atkinson, A. B. (1970). "On the measurement of inequality." Journal of Economic Theory **2**(3): 244-263.

Blakely, T. A., K. Lochner, et al. (2002). "Metropolitan area income inequality and self-rated health - A multi-level study." Social Science and Medicine **54**(1): 65-77.

Bleichrodt, H. and M. Johannesson (1997). "Standard gamble, time trade-off and rating scale: Experimental results on the ranking properties of QALYs." Journal of Health Economics **16**(2): 155-175.

Daly, M. C., G. J. Duncan, et al. (1998). "Macro-to-Micro Links in the Relation between Income Inequality and Mortality." Milbank Quarterly **76**(3): 315-338.

Damgaard, C. and J. Weiner (2000). "Describing Inequality in Plant Size or Fecundity." Ecology **81**(4): 1139-1142.

Davern, M. T., R. A. Cummins, et al. (2007). "Subjective Wellbeing as an Affective-Cognitive Construct." Journal of Happiness Studies **8**(4): 429-449.

De Irala-Estevez, J., M. Groth, et al. (2000). "A systematic review of socio-economic differences in food habits in Europe: Consumption of fruit and vegetables." European Journal of Clinical Nutrition **54**(9): 706-714.

Efron, B. (1979). "Bootstrap Methods: Another Look at the Jackknife." The Annals of Statistics **7**(1): 1-26.

Efron, B. (1985). "Bootstrap Confidence Intervals for a Class of Parametric Problems." Biometrika **72**(1): 45-58.

Efron, B. and R. Tibshirani (1986). "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy." Statistical Science **1**(1): 54-75.

Eggena, M. P., B. Barugahare, et al. (2005). "Depletion of Regulatory T Cells in HIV Infection Is Associated with Immune Activation." J Immunol **174**(7): 4407-4414.

Gakidou, E. and G. King (2003). Measuring Total Health Inequality: Adding Individual Variation to Group-Level Differences. Health Systems Performance Assessment: Debates, Methods and Empiricism. C. J. L. Murray and D. B. Evans. Geneva, World Health Organization**:** 483-496.

Gakidou, E., C. J. L. Murray, et al. (2003). A Framework for Measuring Health Inequality. Health Systems Performance Assessment: Debates, Methods and Empiricism. C. J. L. Murray and D. B. Evans. Geneva, World Health Organization**:** 471-484.

Groot, W. (2000). "Adaptation and scale of reference bias in self-assessments of quality of life." Journal of Health Economics **19**(3): 403-420.

Grundy, E. and A. Sloggett (2003). "Health inequalities in the older population: The role of personal capital, social resources and socio-economic circumstances." Social Science and Medicine **56**(5): 935-947.

Humphries, K. H. and E. Van Doorslaer (2000). "Income-related health inequality in Canada." Social Science and Medicine **50**(5): 663-671.

Jones, A. M. (2007). Applied health economics. Routledge advanced texts in economics and finance. Abingdon & New York, Routledge**:** 335.

Kaplan, M., J.-M. Berthelot, et al. (2007). "The predictive validity of health-related quality of life measures: mortality in a longitudinal population-based study." Quality of Life Research **16**(9): 1539-1546.

Kapteyn, A., J. P. Smith, et al. (2007). "Vignettes and Self-Reports of Work Disability in the United States and the Netherlands." The American Economic Review **97**: 461-473.

Kawachi, I. and B. P. Kennedy (1997). "The relationship of income inequality to mortality: Does the choice of indicator matter?" Social Science and Medicine **45**(7): 1121-1127.

Kerkhofs, M. and M. Lindeboom (2002). "Subjective Health Measures and State-Dependent Reporting Errors." Econometric Analysis of Health Data.

King, G., C. J. L. Murray, et al. (2003). "Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research." The American Political Science Review **97**(4): 567-583.

Kistorp, C., J. Faber, et al. (2005). "Plasma Adiponectin, Body Mass Index, and Mortality in Patients With Chronic Heart Failure." Circulation **112**(12): 1756-1762.

Koolman, X. (2008). Onrechtvaardige verschillen in gezondheid en gezondheidszorggebruik [Inequitable Differences in Health and Health Care usage], Draft.

Koskinen, S. and T. Martelin (1994). "Why are socioeconomic mortality differences smaller among women than among men?" Social Science and Medicine **38**(10): 1385-1396.

Lillie-Blanton, M. and T. Laveist (1996). "Race/ethnicity, the social environment, and health." Social Science and Medicine **43**(1): 83-91.

Lindeboom, M. and E. van Doorslaer (2004). "Cut-point shift and index shift in self-reported health." Journal of Health Economics **23**(6): 1083-1099.

Lubetkin, E. I., H. Jia, et al. (2005). "Relationship Among Sociodemographic Factors, Clinical Conditions, and Health-related Quality of Life: Examining the EQ-5D in the US General Population." Quality of Life Research **14**(10): 2187-2196.

Macintyre, S., K. Hunt, et al. (1996). "Gender differences in health: Are things really as simple as they seem?" Social Science and Medicine **42**(4): 617-624.

Mackenbach, J. P., K. Stronks, et al. (1989). "The contribution of medical care to inequalities in health: Differences between socio-economic groups in decline of mortality from conditions amenable to medical intervention." Social Science and Medicine **29**(3): 369-376.

Mathers, C., A. Salomon, et al. (2003). Alternative Summary Measures of Average Population Health. Health Systems Peformance Assessment: Debates, Methods and Empiricism. C. J. Murray and D. B. Evans. Geneva, World Health Organization**:** 319-334.

Mathers, C. D., C. J. L. Murray, et al. (2003). "Healthy life expectancy: comparison of OECD countries in 2001." Aust N Z J Public Health **27**(1): 5-11.

Mathers, C. D., R. Sadana, et al. (2000). "Estimates of DALE for 191 countries: methods and results." World Health Organization: GPE discussion paper series **16**.

Mathers, C. D., T. Vos, et al. (2001). "National Burden of Disease Studies: A Practical Guide. Edition 1.0. 2001." World Health Organization, Global Program on Evidence for Health Policy, Geneva.

Murray, C. J., E. E. Gakidou, et al. (1999). "Health inequalities and social group differences: what should we measure?" Bull World Health Organ **77**(7): 537-43.

Murray, C. J. L. and D. B. Evans (2003). Health Systems Perfomance Assessment: Goals, Framework and Overview. Healt Systems Performance Assessment: Debates, Methods and Empiricism. J. L. Murray and D. B. Evans. Geneva, World Health Organisation.

Murray, C. J. L., S. Kulkarni, et al. (2005). "Eight Americas: New Perspectives on U.S. Health Disparities." American Journal of Preventive Medicine **29**(5, Supplement 1): 4-10.

Murray, C. J. L., J. A. Salomon, et al. (2000). "A critical examination of summary measures of population health." Bulletin of the World Health Organization **78**: 981-994.

Norberg, K. (2004). "Partnership status and the human sex ratio at birth." Proceedings of the Royal Society B: Biological Sciences **271**(1555): 2403-2410.

Rawls, J. (1971). "A Theory ofJustice." Cambridge, Mass **1**: 33.

Sadana, R. (2002). Development of standardized health state descriptions. Summary measures of population health: concepts, ethics, measurement and applications. Geneva: World Health Organization. C. J. L. Murray, A. Salomon, C. D. Mathers and A. D. Lopez. Geneva, World Health Organization**:** 315-328.

Sadana, R., C. D. Mathers, et al. (2002). "Comparative Analyses of more than 50 Household Surveys on Health Status." Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications.

Salomon, A., C. D. Mathers, et al. (2003). Quantifying individual levels of health: Definitions, concepts, and measurement issues. Health System Performance Assessment: Debates, Methods and Empiricism. C. J. L. Murray and D. B. Evans. Geneva, World Health Organization**:** 301-318.

Salomon, J. A., C. D. Mathers, et al. (2001). "Methods for life expectancy and healthy life expectancy uncertainty analysis." World Health.

Salomon, J. A., C. J. L. Murray, et al. (2003). Health State Valuations in Summary Measures of Population Health. Health Systems Performance Assessment: Debates, Methods and Empiricism. C. J. L. Murray and D. B. Evans. Geneva, World Health Organization**:** 409-437.

Shibuya, K., C. Mathers, et al. (2002). "Global and regional estimates of cancer mortality and incidence by site: II. results for the global burden of disease 2000." BMC Cancer **2**(1): 37.

Tandon, A., C. J. L. Murray, et al. (2002). "Statistical Models for Enhancing Cross-Population Comparability." World Health Organization: GPE discussion paper series **42**.

Üstun, T. B., S. Chatterji, et al. (2003). The World Health Surveys. Health System Performance Assessment: Debates, Methods and Empiricism. C. J. Murray and D. B. Evans. Geneva, World Health Organization**:** 797-808.

Üstün, T. B., S. Chatterji, et al. (2003). WHO multi-country survey study on health and responsiveness, 2000–2001. Health systems performance assessment: debates, methods and empiricism. . C. J. L. Murray and D. B. Evans. Geneva., World Health Organization**:** 761-796.

van Doorslaer, E. and X. Koolman (2004). "Explaining the differences in income-related health inequalities across European countries." Health Economics **13**(7): 609-628.

Van Doorslaer, E., A. Wagstaff, et al. (1997). "Income-related inequalities in health: Some international comparisons." Journal of Health Economics **16**(1): 93-112.

Van Lenthe, F. J., C. T. M. Schrijvers, et al. (2004). "Investigating explanations of socio-economic inequalities in health: The Dutch GLOBE study." European Journal of Public Health **14**(1): 63-70.

Whitehead, M. (1991). "The concepts and principles of equity and health." Health Promotion International **6**(3): 217-228.

Williams, A. (2000). "Science or marketing at WHO? A commentary on 'World Health 2000'." Health Economics **10**(2): 93-100.

Winkleby, M. A., D. E. Jatulis, et al. (1992). "Socioeconomic status and health: how education, income, and occupation contribute to risk factors for cardiovascular disease." Am J Public Health **82**(6): 816-820.

World Health Organization (1946). Constituiton of the World Health Organization. Internation Health Conference, New York.

World Health Organization (2000). World Health Report 2000. Geneva, World Health Organization.

World Health Organization. (2007, 08-28-2007). "Life Tables for WHO Member States."   Retrieved August 27th, 2007, from http://www.who.int/whosis/database/life_tables/life_tables.cfm.

World Health Organization Evidence and Information for Policy (2002). Individual Questionnaire. World Health Survey. World Health Organization.

Zere, E. and D. McIntyre (2003). "Inequities in under-five child malnutrition in South Africa." International Journal for Equity in Health **2**: 1-10.