HEDG
Health, Econometrics and Data Group

THE UNIVERSITY of York

HEDG Working Paper 08/01

New prospects in the analysis of inequalities in health: a measurement of health encompassing several dimensions of health

Sandy Tubeuf
Marc Perronnin

February 2008

ISSN 1751-1976

New prospects in the analysis of inequalities in health: a measurement of health encompassing several dimensions of health

Sandy Tubeuf* and Marc Perronnin†‡

February, 2008

Abstract

This paper develops an innovative method of constructing a concrete measure of health by taking into account individual health information. Using individual survey data from the 2002 IRDES Health and Health Insurance Survey, we propose a measurement of health based on the number of diseases and their respective severity level. The construction relies on a latent variable regression model explaining self-assessed health and controlling various social and health individual characteristics. We compare this construction to other methods proposed in literature for the measurement of health. Moreover, we show how the health index allows to compare distributions of health among different populations and to evaluate inequalities in health in France by using stochastic dominance at first-order.

JEL codes : C13, C43, D63, I12

Keywords : health measurement - France - reported morbidity - stochastic dominance

---

*IDEP-GREQAM. st538@york.ac.uk. IDEP - Centre de la Vieille Charité - 2, rue de la Charité - 13002 MARSEILLE - France

†IRDES. perronnin@irdes.fr

‡We are grateful to IRDES for access to the data and numerous fruitful discussions about the paper. We also thank Fabrice Etilé, Florence Jusot, Carine Milcent, Andrée and Arié Mizrahi, Lise Rochaix, Catherine Sermet and Alain Trannoy for their helpful comments on previous versions. All remaining errors are our responsability.

1 Introduction

The majority of French studies on social inequalities in health uses mortality data. Fewer are those related to morbidity or global health indicators. In addition to their exhaustiveness, mortality data permit following the evolution of inequalities over long retrospective periods and carrying out international comparisons. Nevertheless, they do not take into account dimensions of health such as disability and pain that may have major consequences on individual well-being. Moreover, they establish only the existence of social inequalities without explaining social processes having led to these disparities in mortality. Chauvin and Lebas (2007) also point out that mortality data do not take into account the most recent social changes and are particularly sensitive to medical changes that occurred over time such as preventive behaviours, diagnosis or reimbursement of diseases' costs. Therefore, health indicators as well as surveys on living conditions and social determinants are of considerable importance to improve the understanding of social inequalities in health and also to develop public policies. Bellanger and Jourdain (2004) precisely consider that the economic approach of health in a pragmatic dimension cannot be conceived without the use of health indicators in order to evaluate the results of a health care system. As an illustration, researches were marked by the publication of the World Health Organisation book (WHO, 2002) which concerns measurements of health at both individual and collective levels. These measurements are essential tools for the analysis of inequalities in health as well as for decision makers in Member States of WHO.

Indicators measuring individual general health are particularly interesting as they provide a synthetic information on health. The measurement of an individual's health status that approximates his "true" health status is not only a crucial issue, but also one of the most interesting challenges for studies of health economics. Indeed, there are few measures of health which approach health status as a global concept whereas there is an interest to do so (Chauvin & Lebas, 2007). In addition, although some of French or European surveys on health generate a numerical score representing synthetically individual health, there is no validated and accepted general indicator, representative of the total health of individuals.

The major challenge in measuring health is that the concept of interest cannot be measured directly in its globality; it can only be measured indirectly by indicators such as self-assessed health collected by surveys, or partially, by clinical observations. These indicators are incomplete capturing only parts of the concept to be measured, and sometimes require to be aggregated.

Self-assessed health remains a widespread variable in survey on health, which includes all the physical and psychological dimensions of health. In France, few studies apprehend social inequalities in health using this variable as epidemiologists, sociologists and demographers strongly reject the subjective aspect of self-assessed health. It seems however convenient to us to use this type of variable to measure health and to analyse social inequalities in health in an economic context. First of all, self-assessed health is a measure of quality of the life related to health in the broad sense of the term. Following this assumption, Cutler and Richardson (1997) use this variable to approximate QALY. Then, longitudinal surveys show that it predicts morbidity and mortality. For the last ten years, a large amount of literature was dedicated to the construction of health indicators using self-assessed health. Methodological refinements were proposed to make self-assessed health robust to represent individual health and to provide it good properties for a use in inequality analyses. For example, Shmueli (2003) proposes to reduce individual reporting heterogeneity using MIMIC regressions or van Doorslaer and Jones (2003) introduce a method to cardinalise self-assessed health a distribution of health. Following these methodological refinements we propose an alternative approach to the measurement of health. Our motivation to construct a new measurement of health relies on two elements. Firstly, a continuous and cardinal health indicator is lacking in France[1]. Secondly, we have at our disposal a rich health survey containing information on health. In this context, we define an appropriate conceptual framework to measure health. As we cannot decide which one between the individual and the physician has the best ability to measure health, we propose to construct a concrete measure of health using both qualitative and quantitative variables from health surveys. In doing so, we suggest the construction of a health score. Our construction relies on three elements: (i) we assume that the number of diseases and their severity characteristics is the least subjective health information available in surveys; (ii) we assume that the subjective health status contains implicit general health information, and (iii) we control for individual characteristics within a latent variable model.

The second section presents these elements. The third section describes the modelling strategy of the index of health. The fourth section presents empirical results. Several

[1] There is a French version of the HUI but this health utility index is experimental and has been developed on a specific and restricted sample of about fifty children (Le Galès *et al.*, 1999).

methods have been proposed in literature to change self-assessed health into a continuum. Generally, these methods impose a scaling assumption on the ordinal categorical variable, which contrasts with our construction. In the fifth section, we compare our methodology with these approaches. In the sixth section we illustrate an utilisation of this health index for the analysis of inequalities in health. Conclusions are described in the last section.

2 Aggregating several dimensions of health to measure a general and cardinal health status

Two approaches are followed to obtain a measure of health on a unique scale from multiple indicators. The first approach relies on multidimensional analysis techniques and consists of summarising information provided by different indicators into few factors or into a unique one. This method implicitly assumes that the different dimensions of health are influenced by a common latent variable or are interacting. It is thus advised when all the indicators considered are highly correlated and it consists of a common factor analysis. However, when the indicators are relatively independent this approach induces a reduction of information. As a consequence, the second approach that relies on the aggregation of different dimensions of health might be preferred. Aggregate measures of health are generally based on assessment of individual's utilities with regard to a set of health characteristics. There are various methods used to evaluate these utilities, such as individual self-rating, standard gamble or time-trade-off. Unfortunately, standard gamble and time-trade-off heavily rely on specific questionnaires and so, are difficult to implement on a large scale with any dataset. Thus we follow an approach based on individual self-rating. We assume that if a health status provides a higher utility than another, the individual will attribute to it a higher report. In addition, we believe that a discrete health indicator or an ordinal indicator restricts empirical uses and measuring health on a continuum is preferred. Three arguments support the continuous aspect. (i) Our first argument relies on the preference for numerical indicators in empirical reasoning. Numerical health indexes are generally intended for economic analyses of outputs and for comparing results. Indeed, such indicators enable us to calculate synthetic statistics such as means or variance and to construct confidence intervals. They also permit the calculation of a health stock in the population, the graphical representation of detailed distribution or, the decomposition of indices such as concentration indices. Therefore, they permit to draw a distribution analysis. (ii) Our second argument concerns continuity as opposed to dichotomisation.

Any categorical variable can be transformed into a numerical dichotomous indicator by dividing items into two categories. Although this type of indicators is easy to interpret, it provides weak information: an individual is either ill or not. There is thus no gradation in his health status and we cannot describe the distribution of health status as asymmetric, heavy tailed, etc. The dichotomisation clearly induces a loss of information for an initial indicator described in more than two categories. Moreover, the choice of the cut-off point is not straightforward and will influence subsequent use of the health indicator. Considering self-assessed health, Wagstaff and van Doorslaer (1994) have pointed out that the lower the cut-off point, the greater is the degree of inequality. (iii) Our last argument concerns health utility differences within categories of self-assessed health. Indeed, when an individual reports a good health status equal to the category "good health", it does not mean that his health status is strictly equal to the health status of all the other respondents in the same category. Therefore there is a need for a distinction of individual health statuses within categories of self-assessed health. Ideally, this distinction would be done if individual health statuses were defined on a continuum of health statuses. Finally, we support that the technical foundation of health measurement relies on the ability to rank each individual's health status on a continuous scale.

We use data from the 2002 Health and Health Insurance Survey from IRDES (so-called *Enquête Santé, Soins et Protection Sociale*) to get an indicator measuring health on a continuum of health states in France. Considering the abundance of health information contained in this dataset, it is appropriate to rely on it in order to construct a cardinal and general health index. Run annually from 1988 to 1998 and every other year then, the IRDES-HHIS represents data on French households (except those living in overseas territory or those living in "collective housing" such as long-term care hospitals, religious communities and elderly people's homes) and covers about 20,000 individuals in 7,338 households. The IRDES-HHIS provides information on socioeconomic and demographic characteristics as well as on health status and health insurance coverage. Moreover, each household keeps a medical consumption record for one month by filling out a form. All pharmaceutical expenditures, hospital and ambulatory care consultations are also reported. A basic issue in constructing a health measure is how to choose among the large number of information that could potentially be included. We consider two types of information: medical and functional health, and subjective health.

2.1 Reported diseases count and severity induced

Diseases are a morbidity indicator that give an important information on health status. In the literature, the self-reported incidence of some ailments have already been used as less subjective than self-assessed health (Baker *et al.*, 2004). In our context, we consider that the individual number of diseases enables to identify information on health coming from self-assessed health. We exploit the fact that a stock of diseases represents a cardinal indicator. The IRDES-HHIS diseases report depends on a combination of answers to the question "*Which diseases, health difficulties or disabilities do you have at the present time?*" together with a list of disorders provided as a prompter[2]. Thus, a continuous health variable can be constructed from this dataset by summing the total number of diseases per individual. A medical team in IRDES validates the reported morbidity file by considering it as a whole and corrects glaring errors in reports.

Although a sum of pathologies would give interesting information on an individual's health status, a simple sum has important limits. Indeed, it would come to a conclusion that someone suffering from any two diseases is in worse health than someone suffering from any one disease. However, if the second individual is a terminal cancer patient and the first one has for example, diabetes and eczema, it seems essential to balance this sum of diseases. Similarly, a disease sometimes is just an event occurring in life with complete recovery afterwards; whereas it can also become a chronic part of life sometimes resulting in death. It is therefore important to incorporate a severity level to diseases. A good health indicator has to ignore illnesses with very-short term effects. In this context, we choose to measure the extent of physical limitations as well as prevalence of life risk to evaluate morbidity. We identify diseases that individuals have and evaluate the effects of these diseases on quality of life.

The IRDES-HHIS has the particularity to contain a clinical assessment of each individual file through two health indicators, namely vital risk and disability levels (Mizrahi & Mizrahi, 1985). Each of the reported health data such as diseases, daily treatment, smoking, previous surgery operations, pregnancies etc. except the self-assessed health, are considered by a doctor in order to attribute to each individual a vital risk and a disability level. The vital risk is a prognosis on life-expectancy for the respondent at the time the codification is done, this morbidity indicator would translate a quantitative aspect of life. The disability level represents a degree of difficulties in daily-life activities. This second

[2]The prompter permits limiting the under-declaration of diseases. It is an interesting detail to mention as reports of diseases have been shown biased by social characteristics.

health indicator translates a qualitative aspect of life. These individual-level indicators are ordered categorical variables. The vital risk is composed of seven categories whereas the disability level is divided into eight categories. It is assumed that other diseases from which individuals may suffer could only increase the vital risk or the disability level, but in no case to reduce them. The table 1 presents these two morbidity indicators. In order to channel doctors' assessments and to avoid large disparities in the way they assess an individual's vital risk and disability level, minima levels have been developed. Researchers from the IRDES have developed successive tests methods, in close cooperation with doctors and statisticians to generate minimal vital risk and minimal disability level for diseases (Com-Ruelle *et al.*, 1997). They have assigned a minimal vital risk and a minimal disability level to each reported disease in reference to the International Classification of the Diseases ICD-10 and without any other information. These minima levels are thus created prior to the attribution of vital risk and disability level at individual level and intervene at the end of the doctor's assessment process. If the level of one of the two indicators is lower than the minima levels of the most serious disease reported, the doctor is informed of the anomaly on the screen during the data capture. He is then free to modify the levels he has affected.

Each disease is thus positioned on a scale of six minima vital risk graduations (MVR) comprised between 0 and 5, and a scale of seven minima disability levels (MDL), comprised between 0 and 6. The table 2 describes these graduations. These minima levels provide an indication of a disease's severity feature, as both the minimal vital risk and minimal disability level respectively give information about the decrease in life expectancy and the reduction of activity caused by diseases. We are particularly interested in these minima levels as they allow diseases to be weighted according to severity. We intend to consider diseases listed in the International Classification of Diseases, whose minimal disability level and minimal vital risks have been evaluated by the IRDES researchers.

A set of $1,281$ diseases has been recorded in 2002. As few diseases have a very high minimal vital risk and/or a very high minimal disability level, we propose to collect together the two last categories for both MVR and MDL. Each square of the table 3 contains the number of diseases with the minimal disability level and the minimal vital risk considered.

We test the linear association of these two minima levels and the table 3 represents the correlation matrix. Percentages represent column and row percentages. For instance, we observe 135 diseases with a minimal disability level of 1 and a nought minimal vital risk, which represents 76.3% of diseases with a minimal disability level of 1 and 17.8% of those with a nought minimal vital risk.

We also perform the most common statistical tests to identify the relationship between these two ordinal qualitative variables (cf. table 4). The significance of Chi-square test and the high value of the Pearson correlation (almost equal to 0.7) indicate that the two variables are strongly dependant and tend to rank diseases on a similar pattern. The Gamma coefficient is based on the number of concordant and discordant pairs of observations, its value is significantly different from 0. These tests confirm the linear association of the two variables, which can be either increasing or decreasing. Tests also emphasised that an aggregation of the two variables in a unique indicator is worthwhile for two reasons. Firstly, minimal vital risk and minimal disability level are highly correlated so if they are considered individually in the same regression they would induce multicollinearity. Secondly, the dependence relation between the two variables indicates that the two minima levels assemble around the diagonal such that sets are clearly associated. Our choice is thus to construct a synthetic indicator combining the two dimensions. Moreover, the strong correlation of the two dimensions underlines that the simple sum of categories of the two indicators would not have any sense as it would produce the same calculation twice.

Considering the high correlation between the two dimensions, an aggregation in a classification of possibilities is advisable. The most adapted method is a correspondence analysis, which provides results similar to those produced by factor analysis techniques. It is based on correlation evidence between the two dimensions considered. It has been argued that correlation approaches produce results that vary according to the particular sample used in an analysis (McDowell, 2006). Nevertheless, as our sample is a set of reported diseases in a representative population survey, this use seems less reprehensible. Moreover, as regard to the small number of combinations (30) produced by the two crossed variables, it is not particularly useful to carry out a correspondence analysis, whose main objective is to simplify wide tables. In this context, we propose an analogous reading of the previous correlation table, behind a correspondence analysis and correlation evidence. We observe for each minimal vital risk the corresponding minimal disability level; more precisely, among the diseases with a given level of vital risk, we observe some levels of disability that are overrepresented. On the diagonal, five sets of minimal vital risk and minimal disability level are clearly associated and they combine similar levels of severity in the two dimensions. Assuming that $k = 1, ..., K$ represents the severity class related to a disease, we define the following severity levels:

- $k = 1$ representing the severity class for which both the minimal vital risk (MVR) and the minimal disability level (MDL) equal nought.
- $k = 2$ representing the severity class for which both the minimal vital risk (MVR) and the minimal disability level (MDL) are low.
- $k = 3$ representing the severity class for which both the minimal vital risk (MVR) and the minimal disability level (MDL) are average.
- $k = 4$ representing the severity class for which both the minimal vital risk (MVR) and the minimal disability level (MDL) are high.
- $k = 5$ representing the severity class for which both the minimal vital risk (MVR) and the minimal disability level (MDL) are very high.

We then considered the remaining sets, these are combinations a low level of minimal vital risk and a moderate or high level of minimal disability or vice-versa. Although the method seems to be done at a rough guess, we propose to make cut-out figures combining both correlation and sample size in order to avoid very small classes. With this method, we ensure that singular but interesting sets of minima levels are also emphasised. Indeed, using a programmed data analysis, these sets would have been included in the diagonal. The last four classes are thus

- $k = 6$, the minimal vital risk (MVR) is nought whereas the minimal disability level (MDL) is high.
- $k = 7$, the minimal vital risk (MVR) is average whereas the minimal disability level (MDL) is very low.
- $k = 8$, the minimal vital risk (MVR) is average whereas the minimal disability level (MDL) is high.
- $k = 9$, the minimal vital risk (MVR) is high whereas the minimal disability level (MDL) is low or average.

The table 5 gives a representation of the layout of diseases' severity classes. This severity index is thus related to diseases. For each individual, his/her number of diseases in each of these nine sets is counted. With regard to a situation where diseases would have been counted separately by level of vital risk and disability, this classification will give a more accurate estimation when included in regressions, because it permits avoiding multicollinearity. It will also enable to estimate cross-effects between vital risk and disability.

2.2 Self-assessed health

Self-assessed health indicators offer a good opportunity to capture individual preferences and thus to aggregate a wide set of health information. As supported by the philosopher Bergson (1920), each individual is able to make his assessment with regard to his global health. This variable is therefore likely to account for the main dimensions of health. For example, Liang *et al.* (1991) highlight that chronic diseases have an impact on functional heath and that both chronic diseases and functional status influence self-assessed health. Collected in surveys, this indicator has a discrete form as it is more practical to ask individuals to choose among a set of items. In the 2002 IRDES-HHIS, self-assessed health is collected using the following question: *"Could you grade your health status from 0 to 10? (with 0 being the lowest health status)"*. This scale is slightly different from most of all the other self-assessed health questions, which are usually similar to the one promoted by the European Office of WHO (2000) and consist of categories from *"very good"* to *"very poor"* [3]. In the IRDES-HHIS, respondents have no explicit reference on which they can base their evaluation, such as a comparison with people of their age or a precise time period, so they position their health according to their own scale. The representation of the distribution of self-assessed health (see figure 1) shows that a majority of individuals reports a health level higher than 7.

The distribution is highly skewed and this skewness is also manifest in the inter-category distances, much smaller between levels 7 to 10 than between 0 and 6. In view of the small number of respondents with a self-assessed health status between 0 and 4, these five categories are hereafter grouped together into a single category identified as the lowest one. We choose to use self-assessed health as an element of health but we aim to erase as far as possible its disadvantages with the number of diseases.

[3] Considering the distinctive feature of its self-assessed health question, IRDES has recently tried to be comparable with more widespread self-assessed health questions. As a consequence, the 2002 IRDES-HHIS questionnaire introduced a 5-points scale question asked to one half of the sample, along with the usual 11-points scale. A comparison of the two scales has been performed and shows that a score evaluated between 8 and 10 appears to be equivalent to categories *good* and *very good* grouped together (Jusot *et al.*, 2005).

3 A health assessment model

The first step of our analysis is to construct a health utility score. We believe that the number of diseases combined with the severity levels is a quasi-objective health indicator. We are aware that self-reported diseases can also suffer from individual response judgment. However, the IRDES surveys data have the great advantage to be well-checked by medical experts. In addition, we can also rely on the argument proposed by Jürges (2007), who suggests that diagnosed conditions and measurements are objective health indicators, because diseases are subjective information in factual matters. As a result, we use the number of diseases per severity level to adjust self-assessed health status and so, introduce them as explanatory variables. Our construction relies on Lindeboom and van Doorslaer (2004), who suggest that estimated parameters should be used as weights in their conclusions on the state dependent reporting errors in subjective health measure in their conclusions on the state dependent reporting errors in subjective health measure. Following their suggestion, we investigate an ordered Logit regression explaining the self-assessed health with several individual variables, including the quasi-objective health variables. We then use the estimated parameters to generate the health measure.

In this context, we assume that individuals assess their health considering two issues. Firstly, they score their health with regard to their diseases and the level of severity induced. Secondly, they grade their health by positioning their score into a scale whose graduations are supposed to vary according to their characteristics. In this model, the observed effect of any individual characteristics on self-assessed health is either due to its impact on the health utility score or its impact on the responses scales. These two effects cannot be separately identified in the ordered regression model. In order to solve this issue, we assume that the number of diseases combined with severity levels only influences self-assessed health through the health utility score and does not influence *ceteris paribus* the responses scale.

3.1 The model specification

We shall denote h_{ij}^{subj}, the self-assessed health of the individual i in the household j, and h_{ij}^{*}, the latent variable which represents the "true" health status according to which the individual i in the household j self-assesses his health. This latent variable is an utility measure, which allows various health dimensions to be aggregated. It is thus a continuous and unobserved variable whereas h_{ij}^{subj} is a discrete dependent variable that takes multinomial ordered values from 4 to 10 [4].

We assume that h_{ij}^{*} is explained by a vector of individual characteristics. Firstly, it depends on D_{ij}^{k} the number of reported diseases of a severity level k, with $k = 1, \ldots, 9$ and $D_{ij} = (D_{ij}^{1}, D_{ij}^{2}, ...D_{ij}^{9})$. We believe that the same illness can have a different impact on the health utility score. For example, a fractured leg would have more harmful consequences on an elderly person's health status, because of the increased risk of disability induced. Moreover, the older the person, the harder the healing is. Likewise, a same cancer may have different stages of development and cancers from one stage to another are not comparable.

Therefore, a severity index may not capture the whole "true" health. That is why h_{ij}^{*} may also depend on X_{ij}, a set of demographic, socioeconomic and health-related behaviour variables, and on an unexplained part. The vector X_{ij} is described in the following subsection. As for the unexplained part, it is composed of two residual terms u_i and ϵ_{ij}, which respectively represent household effects and individual effects taken into account by X_{ij}. This means that the "true" health status of an individual is expressed by the sum of the these two residuals terms and two linear equations, the first one concerning the number of reported-diseases by severity level and the second one containing all the other individual characteristics. This model can formally be written as

$$h_{ij}^{*} = f_1(D_{ij}, \alpha) + f_2(X_{ij}, \beta) + u_j + \epsilon_{ij} \tag{1}$$

On the other hand, we assume that the responses scale of self-assessed health varies with individual characteristics. We denote $c_{a,ij}$, the cut-off points of each category of self-assessed health. The latent health variable h_{ij}^{*} relies thus on h_{ij}^{subj} as follows.

$$\begin{aligned} h_{ij}^{subj} &= 4 \text{ if } \; -\infty < h_{ij}^{*} \leq c_{4,ij} \\ h_{ij}^{subj} &= a \text{ if } \; c_{a-1,ij} < h_{ij}^{*} \leq c_{a,ij} \text{ where } a = 5, \ldots, 9 \\ h_{ij}^{subj} &= 10 \text{ if } \; c_{9,ij} < h_{ij}^{*} \leq +\infty \end{aligned} \tag{2}$$

We assume that the cut-off points $c_{a,ij}$ vary with X_{ij} and with the two residual terms v_j^a and ω_{ij}^a on the adaptative scale g_a. We denote φ_a as a set of coefficients related to each of the covariates in the X-vector, the cut-off points of each category are defined by the

[4] Note that categories from 0 to 4 were grouped in the fourth category.

following equation.

$$c_{a,ij} \quad = \quad g_a(X_{ij};\varphi_a) + v_j^a + \omega_{ij}^a \tag{3}$$

In this context, even if individuals have identical levels of "true" health h^*_{ij}, they will assess their health status differently because of their individual characteristics. This can be written as follows.

$$\begin{aligned} & h_{ij}^{subj} = a \text{ if} \\ g_{a-1}(X_{ij};\varphi_{a-1}) + v_j^{a-1} + \omega_{ij}^{a-1} < \quad & h^*_{ij} \quad \leq g_a(X_{ij};\varphi_a) + v_j^a + \omega_{ij}^a \end{aligned} \tag{4}$$

If we introduce these assumptions into the expression 1, then our model is represented by the following reduced form.

$$\begin{aligned} & h_{ij}^{subj} = a \text{ if} \\ g_{a-1}(X_{ij};\varphi_{a-1}) + v_j^{a-1} + \omega_{ij}^{a-1} \quad & < f_1(D_{ij},\alpha) + f_2(X_{ij},\beta) + u_j + \epsilon_{ij} \leq \quad g_a(X_{ij};\varphi_a) + v_j^a + \omega_{ij}^a \end{aligned} \tag{5}$$

Assuming that each previous function is a linear combination of explanatory variables, the equation explaining h_{ij}^{subj} can be written as

$$\begin{aligned} & h_{ij}^{subj} = a \text{ if} \\ cst_{a-1} + X_{ij}.\varphi_{a-1} + v_j^{a-1} + \omega_{ij}^{a-1} \quad & < D_{ij}.\alpha + X_{ij}.\beta + u_j + \epsilon_{ij} \leq \quad cst_a + X_{ij}.\varphi_a + v_j^a + \omega_{ij}^a \end{aligned} \tag{6}$$

where cst_{a-1} and cst_a represent constant terms.

We finally assume that the two residual terms v_j^a and ω_{ij}^a are identical for each grade a. We can then estimate the model through a generalised linear latent model. Our analysis relies on a vector of individual characteristics as well as specific modelling assumptions, which are described in the following subsections.

3.2 A set of demographic, socioeconomic and health-related behaviour variables

The model considers some individual characteristics independent of the aggregated health information, namely health-related variables and socioeconomic variables.

3.2.1 Health-related behaviours

Following the point of view of Cutler and Richardson (1997), we assume that health-related behaviours are information of both current and future health status because of their negative effect on health. For instance, they interact with chronic as well as mental diseases. Their omission may bias the effect of diseases on self-assessed health. Therefore we include in the model, three risk factors, which are available in the dataset: body mass index, tobacco and alcohol consumption[5]. Body mass index reflects health status when low as well as when high, and it is associated with elevated risks of mortality and morbidity[6]. Body mass index values[7] can thus be included as a determinant of the health utility score. Tobacco consumption has a long-lasting effect on health related to the quantity and the length of consumption. In IRDES-HHIS, individuals are first asked if they smoke, and if so, they are then asked how many cigarettes they smoke per day, how many years they smoked, whether they smoke at home, whether they are trying to stop smoking and whether they smoked before[8]. As for the alcohol consumption, questions are asked on the frequency and the quantity of drinking habits. Another question concerns the frequency with which individuals drink more than six glasses at the same time in a month[9].

[5]The categories of these three risk factors are constructed behind the questionnaire, they rely on medical assessment (Com-Ruelle *et al.*, 2006; Dauphinot *et al.*, 2006).

[6]In order to avoid multicollinearity among regressors, we have excluded obesity and other diseases related to weight from the reported diseases count used to construct the health index. Indeed, these pathologies were not consequences of overweight or obesity on health status but a direct observation of a state of fact. On the contrary, cardiovascular diseases or diabetes are consequences of obesity and overweight so they have been kept in reported-diseases.

[7]Body mass index is generated with individual height and weight; respondents are classified accordingly, using international references such as underweight ($BMI < 18.5$), normal weight ($18.5 \leq BMI < 25$), overweight ($25 \leq BMI < 30$) and obesity ($BMI \geq 30$). A fifth category is included for missing values.

[8]Tobacco consumption is divided into four categories: heavy smoker (more than ten cigarettes or five cigars), low (less than ten cigarettes or five cigars), former and non-smoker. A fifth category is introduced for missing values.

[9]Alcohol consumption is also divided into four categories (slight, moderate, heavy and non consumer) and a fifth one for missing values.

3.2.2 Sociodemographic variables

Van Doorslaer and Jones (2003) emphasise the importance to consider a vector of individual characteristics in order to get greater individual-level variations in the health measure. We also believe that individual characteristics when included have a valuable contribution to the control of for reporting bias on reported health.

In addition to health information, the IRDES-HHIS gives detailed social and demographic variables at individual level that we include in our vector of individual characteristics. The table 6 describes variables introduced in the analysis. Concerning demographic variables, 10 age-gender categories are created for men and women. Three levels of education are considered. The main occupational activity variable has six modalities: employed, unemployed, inactive, homemaker, retired and student. Professional activity is also included. Considering that some individuals (about 17%), do not have an occupational class, for example students and homemaker, who has never worked, the occupational class of the household head is assigned to them. In addition, in the survey, individuals are asked to report their income in full and/or using an interval scale. When the exact income is missing, the median of the bracket is used. We use the OECD scale[10] to compute the equivalent household income. Besides income, education, labour market status and activity status, several health insurance variables are collected, indicating whether the person is covered by private voluntary supplementary health insurance[11] or by a means tested public scheme (Rochaix & Hartmann, 2005). As in 2000, the poorest subgroups of the French population have been granted a limited coverage through the so-called *Couverture Maladie Universelle* (CMU), information also includes whether the individual is covered by a private health insurance beyond compulsory insurance or the CMU complementary insurance in 2002.

The analysis is also restricted to those in a position to respond to the self-assessed health status question, i.e. those aged 16 and above. Finally, individuals with incomplete health questionnaires and those who did not answer some of the sociodemographic questions were also excluded. In the end, the sample contains 8,635 individuals for 2002. The omitted reference in the analysis is a young man, in employment, highly educated,

[10] The OECD scale gives a weight of 1 to the first adult, 0.5 to the second and subsequent adults and 0.3 to each dependent.

[11] In France, public health insurance is compulsory and universal. It covers about 75% of health expenditures. To finance the remaining part, individual can subscribe a supplementary health insurance, which can be provided through their workplace (being sometimes mandatory) or individually.

non-smoker, with a normal weight, who drinks with moderation and has private health insurance.

3.3 Using individual characteristics to correct the drawbacks of self-assessed health

Considering that individuals which grant the same utility to health status are likely to report different self-assessed health according to their personal characteristics such as age, gender, socioeconomic status and health conditions. We assume that a good health measure should disentangle the health utility score from personal response bias. Therefore, we propose a correction at two levels. The first level is to consider the reporting variations in the thresholds of self-assessed health categories according to individual's characteristics. The second level relies on a random effect, according to which people of the same household are likely to report a similar self-assessed health.

3.3.1 Considering individual variability in self-assessed responses scale

The correction for individual report variability is supposed to allow our indicator to approximate more precisely the health utility score. Our testing strategy is in two phases.

<u>Phase 1:</u> **Ordered Logit model without varying thresholds**

In the first phase, we suppose that the vector of individual characteristics has the same effect on each threshold. In this context, the responses scale is changing through only one translation from one individual and the gap between categories stays the same:

$$\varphi_a = \varphi \tag{7}$$

Nevertheless, constant terms still vary with categories a. As a result, we write the following reduced form.

$$\begin{gathered} h_{ij}^{subj} = a \text{ if} \\ cst_{a-1} + X_{ij}.\varphi + v_j^{a-1} + \omega_{ij}^{a-1} < D_{ij}.\alpha + X_{ij}.\beta + u_j + \epsilon_{ij} \leq cst_a + X_{ij}.\varphi + v_j^a + \omega_{ij}^a \\ \text{i.e. if } \; cst_{a-1} < D_{ij}.\alpha + X_{ij}.(\beta - \varphi) - v_j^{a-1} - \omega_{ij}^{a-1} + u_j + \epsilon_{ij} \\ \text{and } \; D_{ij}.\alpha + X_{ij}.(\beta - \varphi) - v_j^a - \omega_{ij}^a + u_j + \epsilon_{ij} \leq cst_a \end{gathered} \tag{8}$$

It is important to remind that in this model, β and φ cannot be identified and their respective effects on h_{ij}^{subj} cannot be distinguished either. Indeed, the effects of covariates X_{ij} both on h_{ij}^{*} and on the adaptative scale g_a cannot be separately estimated. Thus, the coefficients may integrate two types of effect, an effect on "true" health and an effect on the responses scale of self-assessed health.

Phase 2: Ordered Logit model with varying thresholds

In the second phase, we allow the thresholds to vary with covariates. Gaps between thresholds are thus supposed to vary from one individual to another. The figure 2 explains the reporting process of self-assessed health for two individuals A and B, whose "true" health is respectively represented by H_A^* and H_B^*.

They report their health status according to their own responses scales, which are respectively represented by $C_A^4, ..., C_A^9$ and $C_B^4, ..., C_B^9$. From one individual to the other, the position of the thresholds is varying. This means that each individual positions his "true" health on his own responses scale and reports his health level according to this position. As a result, individual A evaluates his "true" health status h_A^* between C_A^8 and C_A^9 , and reports a self-assessed health equal to 9; whereas individual B evaluates his "true" health status h_B^* between C_B^5 and C_B^6 and reports then a self-assessed health equal to 6. We notice that if individual B had the same responses scale as individual A, he would report a self-assessed health equal to 7.

Thresholds can be estimated with linear or loglinear specifications. We assume a linear specification[12], which allows us to interpret coefficients in the model, easily whereas the linear specification does not ensure that thresholds are well-ordered i.e.

$$g_{a-1}(X_{ij}; \varphi_{a-1}) < g_a(X_{ij}; \varphi_a)$$

In order to avoid significant calculation time, we assume that there is only one covariate that greatly influences the thresholds. We test one by one the effects of each of the covariates on thresholds using an ordered Logit with shifting cut-off points[13]. The likelihood ratio test allows us to select the individual characteristic on which the thresholds vary the most, the lowest log-likelihood. The table 7 recapitulates the log-likelihood val-

[12] Other specifications are conceivable, for instance an exponential link for differences in thresholds or sequential models, which would be used to estimate $p(SAH \geq k)$ instead of $p(SAH = k)$.

[13] We assume that the introduction of the cluster effect hypothesis in all these regressions is not changing the covariate that greatly influences thresholds. Consequently, we ignore cluster effects in this ordered Logit.

ues of each of these models. Among all the covariates, the occupational activity being the variable, which have the highest impact on reporting bias, the log likelihood associated to the model equals $-12,700.837$, whereas it equals $-12,735.764$ for income. In other words, occupational activity is now excluded from explaining variables. We now include cluster effects within the ordered Logit model for health.

3.3.2 Correcting for cluster effect

Unobserved heterogeneity may have several well-known negative consequences on the estimation if it is ignored (Allison, 1999). Indeed, a bias in standard error of estimated parameters leads to an overestimation of the accuracy of statistical test, a lack of efficiency, a heterogeneity shrinkage and a spuriousness bias[14]. We choose to account for this unobserved heterogeneity through a random effect.

Our specification allows to avoid all the previous issues except the spuriousness bias because we use an ordered Logit regression considering random effects[15]. Our motivation to provide for cluster effect relies on the common occurrence in households to report the same self-assessed health for all the members. As shown in figure 3, in our sample, more than one quarter of individuals belongs to a household[16] where all the members are reporting the same self-assessed health. As a result, a similar way of reporting health is operated in about 29% of households of more than one individual. It is necessary to highlight that when respondents of same households are not reporting exactly the same self-assessed health status, a quarter of them report a level of health status which differs of one category, only. This cluster effect would be explained either by a similar "true" health status itself, such as genetic endowment, exposition to similar risks for health, similar preferences for health, or similar reporting behaviour, due to cultural factors or similar

[14]The heterogeneity shrinkage means that the variance generated by unobserved heterogeneity attenuates regression coefficients. Spuriousness bias is due to the correlation between household effects and individual effect which bias estimations of the coefficients.

[15]The spuriousness bias could have been corrected by a mixed model, but much more covariates would have been required, leading to unreasonable time calculation. Alternatively, we could have used a fixed effect model to avoid the restrictions on u_i. In particular, unobserved heterogeneity is allowed to be correlated to the covariates, and we thus correct the spuriousness bias due to this correlation. Whereas this type of model is difficult to generalize in non linear cases, an ordered Logit with fixed effect is developed by Ferrer-i-Carbonell and Frijters (2004). Nevertheless, this method presents limitations in our case. Firstly, it discards a considerable proportion of data as it excludes households with no variation in SAH. This exclusion increases standard error since in our sample 30% of individuals are in households with a same SAH level for all the members. Secondly, it does not provide an estimation of variables that are fixed within households, like income by consumption units, which makes our model less informative and more difficult to interpret.

[16]We considered all the households composed of more than one individual.

perception of pain, for instance. Members of a same household are likely to assess their health statuses in a similar way because of common unobservable factors that are not taken into account by the socioeconomic and health variables.

3.4 Construction of the health index

The construction relies on the use of the estimated coefficients of each severity level to weight the number of diseases. These coefficients allow us to give a weight, which is not biased by individual responses heterogeneity. The continuous health measure is generated using the combination of diseases by severity level, multiplied by its estimated effect $\hat{\alpha}$ on the latent health variable. For the sake of interpretation, we propose to normalise this continuous health measure in two steps.

In a first step, we choose to normalise each coefficient by $\hat{\alpha}$, which is the estimated coefficient associated to the lowest severity level. The direct use of estimated coefficients $\hat{\alpha}$ as weights would generate arbitrary values as in ordinal regressions, parameters are estimated up to scale[17]. The weight given to a disease of severity level k is thus equal to

$$w_k = \frac{\hat{\alpha}_k}{\hat{\alpha}_1} \tag{9}$$

The interpretation of such quantity is straightforward; it represents the number of diseases with the lowest severity level which is needed to produce the same effect on self-assessed than a disease with a severity level k. The health measure can then be written as the sum of all the diseases weighted by the severity level associated with it.

$$I_{ij}^{raw} = \sum_{k=1}^{9} \frac{\hat{\alpha}_k}{\hat{\alpha}_1} D_{ij}^{(Sev_k)} \tag{10}$$

This health measure can be compared to a health index as it summarises health into a single number. Our measurement of health combines the medical health and the subjective health controlled by various social dimensions in one instrument. In economic evaluation, these measurements are variously termed "general health status measure" or "measures of health related quality of life". However, we would say that quality of life is broader than

[17]In particular, their value is sensitive to the distributional assumption for residuals. For example, if we assume that residuals are following a normal law instead of a logistic law, coefficients would be divided by 1,64. In effect, standard normal distribution has a standard error equal to 1 whereas standard Logit distribution has a standard error equal to $\frac{\pi}{\sqrt{3}}$.

our construction. For example, other topics such as daily activities are also considered in the EQ-5D, or such as work and role performance in the SF36.

In the second step, we change the health measure into a health index described in the interval $[0; 1]$ so as to compare it to other general health status measures such as Health Utility Index or the two summary measures on physical health and mental health from the SF36. In order to do so, we calculate the gap to the highest value it can reach and divide it by the range of its values. This health index can thus be generated using the equation 11.

$$I_{ij} = \frac{I_{\max}^{raw} - I_{ij}^{raw}}{I_{\max}^{raw}} \tag{11}$$

This health index can be used in its current form in different analyses. The health index is based both on a medical approach as the number of diseases are taken into account, and on the subjective approach as self-assessed health is considered. Our approach is conservative as we do not include the effects of X_{ij} on h_{ij}^*. As a matter of fact, we cannot distinguish between effects of individual characteristics on “true” health and effects on the scale of self-assessed health. The coefficients integrate two types of effects: an effect on the health utility score and an effect on the reporting bias. We assume that a substantial part of the socioeconomic variations in self-assessed health is attributed to reporting bias. Furthermore, we do not account for h_{ij}^* in its entirety.

The question of the incorporation of risk factors in the health index is tricky. As mentioned earlier, these variables reflect health status but they are also changing overtime and their effects on health are mediated by other health indicators, such as medical or functional ones (Manderbacka *et al.*, 1999). In our case, their consequences on health are taken into account through reported diseases. In addition, risk factors may capture other aspects than medical well-being, as for example a lower individual care granted to health. Therefore, they may influence the reference scale as well as “true” health. It is the reason why we choose not to use these factors within the construction of the health index. Nevertheless, other information could be taken into account to describe all the dimensions of “true” health, for instance, functional characteristics.

The generalised linear latent and mixed model is carried out for equivalent health status, the same diseases and the same severity induced levels.

4 Empirical results

4.1 Ordered logit models with or without cluster effects

4.1.1 The importance of cluster effects

In a first stage, we estimate an ordered Logit model without variation of thresholds and without cluster effects, and in a second stage we take into account the cluster effect due to the ability to self-assess a similar health status in the same household. The table 8 recapitulates the results of these two models.

If we compare results of the two models, we notice that health-related are the parameters whose effects on health are changing the most. For example, overweight and obesity do not have any significant effect on health in the regression with cluster effects whereas these two same variables do have a significant effect on health in the regression model which does not consider cluster effects. A similar pattern is observed for light smokers. Whilst having a high consumption of alcohol has an impact on health in the first model, it does not have such an impact in the second one. The Khi square statistic of the cluster effects parameter equals 210 with one degree of freedom which indicates that inter cluster variance is significantly different from zero. Therefore, it suggests that some unobserved household characteristics have a strong effect on the global health or on the scale. It is thus relevant to introduce cluster effects in the model as taking into account this unobserved heterogeneity substantially modifies coefficients and their significance. In particular, coefficients associated with the numbers of diseases by severity level are changing. Our decision to take into account cluster effects was motivated by these results.

The following part outlines relevant results concerning the ordered Logit regression with cluster effects and observed for individuals with the same health status.

4.1.2 The impact of health variables on self-assessed health

Regardless of the severity level, for each class of severity self-assessed health is decreasing when the number of diseases increases.

The effect on self-assessed health is stronger when the severity level is high. Being a heavy smoker has a significant and negative impact on self-assessed health. This result is inconsistent with the hypothesis we could have formulated saying that smokers enjoy smoking and increase their well-being by doing so, and that they would self-assess a good health status. It is either that smokers have got bad habits but are conscious of smoking bad consequences on their life expectancy, or they are unconscious that smoking is the cause of their bad health but they suffer from health conditions such as respiratory problems or cardiovascular diseases. For same pathologies, smoking degrades more self-assessed health status.

Not consuming alcohol has a significant and negative impact on self-assessed health. This impact is explained by individuals, who cannot drink alcohol because of medical prescriptions. The fact that individuals do not drink alcohol often stems from a constraint due to health status. Indeed, data do not separate those who do not consume from those who consumed alcohol in the past. Heavy drinkers are likely to report a poor self-assessed health, but this result is not significant.

The impact of the body mass index on the self-assessment of health status is relevant for overweighted and obese people. The higher the BMI, the worse is the self-assessed health. As for smoking habits, individuals who are suffering from overweight must be conscious of the reduction of their ability in daily life, or they suffer from diseases that are consequential to their high weight.

4.1.3 The impact of demographic variables on self-assessed health

Self-assessed health decreases as age increases. Even if results are controlled according to health, the effect of age can be explained by a more pessimistic assessment in older age categories or by an impact of health status which would not be caught entirely. In effect, the same disease can have worse consequences on an elderly person than on a younger person. Considering gender, young women assess a significantly worse health status than young men, and inversely in older ages. These results are consistent with previous studies (van Doorslaer & Jones, 2003), particularly those concerning elderly people (Groot, 2000) which were explained in terms of life expectancy. Before self-assessing their health status, men would compare themselves to other men of their age and would observe that mortality among men is higher than among women. Thus, they would give a lower assessment of their own life expectancy and of their health status.

4.1.4 The impact of social variables on self-assessed health

Household equivalent income plays a positive and significant role on self-assessed health; the higher the income level, the better is self-assessed health. Intuitively, as expected, the richest have a better access to the health care system and benefit from a higher quality of cares when they are ill.

Education level has a non-significant impact on self-assessed health whatever the level of education considered.

Concerning the main occupational activity status, being a student has an effect on self-assessed health, which can be compared to the one of age. As age classes are large (16-35 years old), student effect could be explained by a hidden age effect or the absence of particular diseases, such as those due to work conditions. Inactivity, which excludes homemakers, has a negative impact on self-assessed health. That can be explained by both a direct and an indirect health effect. Indeed, in a direct way, individuals out of the labour market at working ages are likely to be excluded because of their health status. The indirect health effect relies on the fact that an individual in precarious conditions often has a poor health. Finally, unemployment, retirement as well as being homemaker have a non-significant impact on self-assessed health.

Farmers and unskilled workers are likely to assess a worse health status than employees. The common explanation comes from working conditions. Inversely, executives assess a better health status. As we consider individuals having the same health status, an explanation can be found in respect of executives, who may have less health problems because of their higher social status.

Following this idea, having no supplementary health insurance plays a negative role on self-assessed health. That counters to the self selection hypothesis. However, two theories explain this impact on health. Firstly, although people with a lower self-assessed health would have a greater propensity to ask both for care and for supplementary insurance, premiums of this supplementary insurance are more expensive and so, would lead to higher health care expenditures. Secondly, people who cannot afford a supplementary health insurance could be sicker because they cannot have a good access to health care they need, which worsens their health. This first analysis supports the importance of cluster effects. This is why the third model, which considers varying thresholds, includes clusters effects.

4.2 Ordered logit model with cluster effects and varying thresholds

As described in the previous method, we choose to make thresholds varying with a unique variable. According to the log-likelihood value of various regression models, occupation status has appeared to be the most relevant. The results of this last model are presented in table 9.

These results are similar to those of the previous model with cluster effects but without varying thresholds. However, if we represent the effects of occupational status on the thresholds of self-assessed health, we notice the importance of taking into account varying thresholds.

Figure 4 represents the distance from one self-assessed health category to another according to occupational status. It allows us to understand that according to the occupational status, individuals have different levels of health expectations. For instance, the interval of self-assessed health comprised between 9 and 10 is the largest for active individuals, which means that they have a higher probability to self-assess a health status of this level than individuals with other occupational status.

Conversely, retired and unemployed people have lower expectations of good health and are less likely to report a self-assessed health higher than 9. This hypothesis of varying thresholds implies a strict analysis of their effects on health.

4.3 The continuous health indicator

The regression coefficients $\hat{\alpha}$ are used as an unbiased weight to construct the health indicator. In a first step, we normalise each estimated coefficient by the one associated to the lowest severity level. The table 10 gives weights that are attributed to each severity level according to the modelling concepts and corresponds to the values of the coefficients normalised to the lowest one. This table can be analysed as "equivalent number of diseases of the lowest severity level": a disease with a severity level of 5, is equivalent to 3.9 diseases with a severity level of 1 in the model with cluster effects and varying thresholds, respectively 3.66 in the second model and 3.57 in the first one. If we represent the distribution of these severity weights according to the model specification, we observe the same pattern whatever the model. However, by comparison to the simplest model, we can see that the correction for cluster effects as well as the consideration of varying thresholds emphasise weights. When the severity is the highest (i.e $k = 5$), the associated weight is the strongest and the model relies thus on varying thresholds and cluster effects. The severity level estimates are particularly different in the model specification when there is an existent level of vital risk. Indeed, there are light differences between severity levels for which $k = 2, 3, 6, 7, 8$ according to the model specifications. For the other values of k, we confirm previous results according to which the cluster effect influences values of coefficients, even when they are normalised by the coefficient associated to the lowest severity level in order to drop the shrinkage effect.

Our hypotheses of cluster effects and varying thresholds are thus directly relevant to the health measure, which will be constructed. They emphasise the weight of diseases' severity levels in the indicator and so, the weight of objective health. The raw continuous health indicator can then be generated using equations 11 and estimated coefficients of these diseases severity levels. Nevertheless, which estimated coefficients are preferred within the construction of the health index?

Our model specification in three steps has emphasised the importance of cluster effects.

As for the effect of varying thresholds, even if it exists, its implementation is time-consuming and the choice of the covariate on which it is based, depends on the sample considered. In this context, we prefer to construct our health measure using estimated coefficients from the ordered Logit with cluster effects and without varying thresholds. The distribution of the constructed continuous health indicator is represented in the figure 5, and is compared to the one of the self-assessed health variable. The health index reports an average health equal to 0.89. Generally speaking the distribution of the indicator is concentrated among good health statuses and is spread among bad health. This health index is synthetic and allows comparisons between different populations. Its continuous aspect enables us to make a distributional analysis, in particular to calculate standard error or confidence intervals. We consider it to be another measurement of health for analysis of inequalities in health.

5 Analysis of inequalities in health as measured by the health index

The cumulative distribution function for the health index is drawn in figure 6 for the full sample. The inverted L-shape of the empirical distribution function emphasises that there is a long left-hand tail which represents relatively few individuals in very bad health. Many people are concentrated in the right-hand tail and so have a higher health index. The vertical line at the right-end of the distribution shows a large proportion of individuals having a health status equal to 1. We shall now understand how health, as measured by the health index is unequally distributed over some individual characteristics such as age, income, education and economic status. Empirically, we rely on a graphical representation of cumulative distribution functions and on tests of stochastic dominance at first order as described in Lefranc *et al.* (2004). To do so, we use unilateral *Kolmogorov-Smirnov* tests of equality of distribution.

5.1 Distribution of health over age classes

We consider the health index according to age classes. The empirical distribution of health shifts to the right as income increases as described in figure 7. It emphasises that health status worsens with age. We carry out dominance tests based on a conjunction of *Kolmogorov-Smirnov* unilateral tests to compare distributions of health over age classes. They confirm that people aged 16-25 years old are significantly in better health than all the other age classes (cf. table 11) and that each age class is always dominating the upper age classes.

5.2 Distribution of health over socioeconomic statuses

We consider the distribution of the health index according to activity statuses. It emphasise in figure 8 that "students" and "employed" experience a better health than "retired", "inactives" or "homemakers". In other words, younger age classes and having a job have a better health. We supplement the graphical analysis by unilateral tests whose P-values are presented in table 12. The distribution of health of "students" significantly dominates the distribution of health of all the other activity statuses. "Unemployed", are significantly in worse health than "employed" people. This result has already been shown in other empirical studies (Khlat & Sermet, 2004). Distributions of health of "retired" people and "inactives" people are significantly dominated by the distribution of health of all the other activity statuses, which is respectively explained by the strong link between health and age and inactivity due to health status. Moreover, the distribution of health of "inactives" dominates significantly the distribution of health of "retired".

As for education level represented in figure 9, the distribution of health of poorly educated individuals (i.e those having no diploma) is situated on the left of the distributions of health of the two higher education levels. The unilateral tests emphasise that the distribution of health of individuals having at least A-level significantly dominates the distribution of health of individuals having either no diploma or a diploma of primary or secondary level.

This stochastic dominance analysis confirms the existence of social inequalities in health. Considering that different statistical methods have been proposed to transform the ordinal categorical self-assessed health into a cardinal measure, it is interesting to compare our construction with this literature.

6 Discussion

In the literature, three solutions highlight the scope of methods proposed to transform an ordered categorical indicator into a continuous one. They assume that the categorical ordinal variable reflects a continuous latent variable that measures global health and then estimate this latent variable. The first method assuming that self-assessed health follows a lognormal distribution (Wagstaff & van Doorslaer, 1994), the second one using an ordered Probit model and several different dimensions of health to estimate a "health capital" (Cutler & Richardson, 1997) and the last method introducing the use of a health distribution (van Doorslaer & Jones, 2003). In the following subsections, we describe these methods and in the last subsection we discuss the features of our indicator as compared to these three methods.

6.1 Getting continuity from an "arbitrary" distribution

When there are no other information on the actual distribution of health, a health measure can be generated by imposing a functional form for its distribution, which relies on empirical observations of the distribution. Wagstaff and van Doorslaer (1994) propose to assume that the observed health distribution over a self-assessed health composed of A categories is generated by a latent unobservable and continuous variable with a standard normal density function. In the course of their analysis, the choice of an inverse lognormal distribution is preferred as regard to the skewed distribution of most of health indicators. Typically persons suffering from serious ill-health are in minority and a large proportion of any general sample population report good health[18]. Indeed, health distributions are strongly concentrated among good health statuses whereas they are spread among lower health statuses, which are more graded. Economists often model the distribution of income or wealth using a lognormal distribution (Cowell, 2000). The lognormality has some convenient properties, such as its simple relationship to the normal distribution, the preservation under loglinear transformations as well as the advantage of allowing for skewness. This last point is particularly important for the underlying distribution of health.

[18]The choice of an inverse or a standard lognormal distribution is explained by the skewness of the distribution. If this skewness is observed on the right (respectively left) then an inverse (a standard) lognormal would be preferred.

The cardinalisation process considers the frequency of each category and calculates thresholds by fitting quantiles from the ordinal categorical variable, notably the cumulated frequencies of categories of self-assessed health, with those of the inverse lognormal distribution. Category scores are obtained as the expected values within each of the intervals defined by the cut points. If an individual reports a health status a, his continuous health status is defined by the theoretical average value of the latent health variable between the thresholds c_a and c_{a+1}.

Gerdtham *et al.* (1999) validate this approach. They compare the direct assessment of health status using either the rating-scale method[19] or the time-trade-off[20] method. The main advantage of the Wagstaff and van Doorslaer's approach by comparison to the time-trade-off or to the rating scale, is that categorical information on health status is available in most of the population surveys because this indicator is much easier to collect. However, even if the latent health variable is assumed to be continuous, it is still inherently categorical and therefore it could not be used as a continuous variable in an ordinary least square regression. Its use would produce non normal and heteroscedastic residuals leading to inefficient estimates of coefficients and biased estimates of their standard error. Moreover, intra-categorical differences are not considered. The time-trade-off and the rating scale directly yield a continuous health measure whereas the third method requires an assumption of the shape. This assumption relies rather on arbitrary than obvious feature of the distribution. In particular, it assumes the same distribution of health, whatever the population considered, which may lead to biased estimates of concentration index.

As regard to these critics, a cardinalisation of the self-assessed health using health information in order to overcome the arbitrary aspect.

6.2 Getting continuity by combining different health dimensions

Cutler and Richardson (1997) discuss a theoretical framework for measuring health capital of the population. They aim to estimate quality-adjusted life years (QALY), that are weights reflecting the quality of life that somebody attaches to each of his remaining years of life taking into consideration his health conditions during these years. An individual's quality of life is scaled on a 0 to 1 basis, where 0 is equivalent to death and 1 is

[19] The rating-scale method uses a visual-analogic scale from 0 to 100 with labeled anchors from "death" to "full health".

[20] Individuals are asked to evaluate on a scale of 20, the number of years in full health that they think is of equal value to 20 years in their current health status.

equivalent to perfect health. Cutler and Richardson (1997) advocate a health measure, which relies not only on a physical measure of morbidity but which accounts also for mental and physical functioning as well as risk factors. Therefore, they choose to estimate QALY by weighting the fact of living with major chronic diseases and functional impairments. This means that suffering from a disease attributes to the individual a quality of life comprised between 0 and 1 (both excluded). Considering the possibility of using time-trade-off methods for the assessment of QALY weights, they reject this approach and argue that "there is no consensus in the literature about the disutility associated with various conditions or the change in these disutilities over time". However, they include a discount rate $\frac{1}{(1+r)^k}$ to take into account individual preference for present.

In this context, using the American National Health Interview Survey, each functional limitation is weighted measuring the extent to which a disease influences self-assessed health. Their method is to assume that people have a latent measure of health, related to their diseases, demographic characteristics and to estimate such a model using an ordered Probit model. The estimated coefficient of the diseases vector is used as a measure of health. The ordered Probit model allows to estimate all the cut-off points of the self-assessed health categories. As a QALY is scaled on $[0; 1]$, the estimated coefficient (usually range from $-\infty$ to ∞) has to be normalised. It is therefore divided by the differences between the estimated coefficients of the highest and the lowest categories of self-assessed health. The estimated coefficient of the diseases vector is interpreted as a reduction in quality of life associated with each chronic condition.

A peculiar aspect is that the QALY loss to a chronic disease is not conditioned by other variables, such as income and standard of living. Indeed, the estimated coefficient of a particular chronic condition informs how that condition changes along the scale of self-assessed health, holding constant demographic characteristics and other reported health conditions. However, a chronic disease has a different impact on an unskilled worker than on a manager, and these aspects are not considered. Indeed, a good utility function must take into account individual preferences in a given context of perfect information, as it is in Grossman (1972) as well as in a given context of uncertainty.

Nevertheless, the validity of this method has not been shown (van Doorslaer & Jones, 2003). Moreover, there is a misspecification of the quality of life; when an individual rates his health as very poor, QALY equals 0, which implies "death" according to preliminary hypotheses whereas the individual is not obviously dead. This construction could lead to give individuals predicted values of health status lower than 0 or greater than 1. Van

Doorslaer and Jones (2003) highlight this limitation and offer to overcome it with two alternatives. Firstly, they propose to rescale to the $[0; 1]$ interval, using the largest and the lowest prediction. Secondly, under the assumption that a continuous health distribution is available for the sample considered, the range of average values of this distribution for age groups could be used as an explained variable. The minimum and maximum predictions from this new model would then define the observable range of the distribution conditional on the set of regressors.

As regard to critics formulated against these two first methods to describe the latent health variable, a third method proposes to consider a health distribution, in some cases external within an interval regression.

6.3 Getting continuity using external information

The third solution relies on the creation of scattering within categories of self-assessed health by considering a health distribution. An appropriate econometric procedure to do so has been proposed by Stewart (1983). It uses a likelihood function for the application at hand. The likelihood function is a modification of that used in the estimation of the standard ordered Probit model and replaces the unknown threshold values by the set of known thresholds that delineate the intervals. The responses on the dependent variable are grouped. In the literature this type of model is referred to as a grouped dependent variable model or as interval regression model. As self-assessed health is an ordinal variable in nature but interval coded, this interval nature is exploited within an interval regression model.

In order to understand how the model is implemented, responses of self-assessed health are coded 1, 2,..., 5 to capture the five distinct health status categories. We shall denote y_i the observe self-assessed health and y_i^* an underlying variable that captures the health status of the i^{th} individual. This can be expressed as a linear function of a vector of explanatory variables X_i using the following relationship. The exact knowledge of the thresholds allows the likelihood function to be specified in a fairly straightforward manner. The variable y_i^* is best interpreted not as a latent measure but a measure with a quantitative interpretation. The interval regression provides a good alternative to ordered Probit model when the limits of the intervals of the parameter of interest are known. Interval regression has been specifically recommended as an appropriate method for analysing results from contingent valuation studies (Donaldson *et al.*, 1998). It has also been successfully applied by van Doorslaer and Jones (2003) on Canadian data, using a health distribution derived from the Canadian National Population Health Survey (CNPHS), namely Health Utility Index (HUI), to rescale the Canadian self-assessed health available in the same survey[21]. The cumulative distribution function of HUI is used as the benchmark, from which the thresholds defining HUI intervals of each self-assessed health level are derived. In concrete terms, the q^{th} quantile of the distribution of HUI corresponds to the q^{th} quantile of the self-assessed health, which is analogous to the previous inverse lognormal rescaling. In a first step, the cumulative frequency of observations for each category is computed. The second step is then to find the quantiles of the cumulative density function of HUI. Each interval is thus limited by a couple $[c_{a-1}; c_a]$, from which an interval regression can be conducted.

The interval regression thus measures individual probabilities to self-assess a health status between $[c_{a-1}; c_a]$ dependent on a vector of demographic and socioeconomic characteristics. It provides efficient estimated parameters, an identifiable variance of the error term and a definition of the scale of the latent health variable. The values of indicator can be interpreted in terms of health utility because they are obtained by rescaling the latent variable with the distribution of HUI, which is a utility-based measure obtained by a Von Newman-Morgenstern procedure.

This method relies on having a dataset that includes both self-assessed health and a cardinal index of health[22]: in their case the Canadian National Population Health Survey (NPHS), which includes self-assessed and the McMaster health utility index (HUI). This is used to construct a mapping from HUI to self-assessed health on the assumption that there is a systematic relationship between the two measures of health, such that those at the bottom of the distribution of self assessed health will also be those at the bottom of the distribution of health utility. This method cannot be replicated to the French context as we do not have at our disposal a dataset containing both self-assessed health and the questionnaire of the Health Utility Index.

6.4 Some elements of discussion

From the three previous methods, two aspects appear essential for an appropriate measurement of health status.

[21]The self-assessed health question is "*In general, how would you say your health is?*" and the five response categories are *excellent, very good, good, fair* and *poor*.

[22]Van Doorslaer and Jones (2003) go further in their conclusions and propose to use these HUI predicted thresholds to compute an interval regression on self-assessed health, even if the survey does not contain any generic health distribution.

Firstly, it is advisable to reach the continuous aspect by using several health factors. For instance, Cutler and Richardson (1997) include physical morbidity, mental and physical functioning and risk factors. Similarly van Doorslaer and Jones (2003) rely on an index of health utility along with self-assessed health. Secondly, it is important to consider the strong links between health and individual characteristics as it is done in ordered Probit as well as in the interval regression, which includes various individual characteristics. Our construction encompasses these two elements.

Nevertheless, it is noteworthy that the main difference between our procedure and the constructions proposed in Wagstaff and van Doorslaer (1994) or van Doorslaer and Jones (2003) is the initial element of the measurement of health. These methods rely firstly on self-assessed health whereas our initial element is the reported diseases count that we assume more objective as we correct it using a severity index. Then, these methods use a distribution (arbitrary or representing health) assumed more objective to correct the subjective health whereas we rely on self-assessed health to weight the number of diseases. In simple terms, we could say that these methods generate a subjective indicator of health corrected with objective health information. On the contrary, we generate an objective indicator of health corrected with subjective health information. As a result, we all propose a mixed indicator of health but with different initial assumptions.

By comparison to the measure of health proposed by Cutler and Richardson (1997), our indicator is more informative than an indicator that would be based on the occurrence of the disease, because it takes into consideration the fact that some diseases affect the length of life as well as its quality. Moreover, we can underline that our indicator could also easily involve a parameter of preference for present or preference for certainty as proposed in Cutler and Richardson (1997).

7 Conclusion

In view of the multidimensional nature of health status and the need to take into account reporting biases, we have considered the construction of a health status variable encompassing the three main dimensions of health described by Blaxter (1990), namely medical, functional and subjective, while offering a cardinal health indicator. Firstly, the medical and functional dimensions are translated into the number of diseases and their respective severity level medically evaluated. Secondly, the subjective dimension is approached by self-assessed health level. Despite the fact that diseases are self-declared data and so, can suffer from individual reporting bias, this health information seems to be less biased than self-assessed health because of the use of diseases' severity level. These severity levels allow checking for coherency between severity and number of diseases.

The new measurement of health takes into account the multidimensionality of health and offers a measurement of health, halfway between subjective health and more "objective" health. This index differs with other health measurement tools available at national and international levels. Indeed, its construction does not depend on a particular questionnaire and simply relies on survey data. This indicator does not claim to be universal; but its construction method can easily be replicated with other control variables and other samples on condition that they provide diseases report to which the severity index can be applied.

This method gives a simple way to construct a continuous indicator with variables classically collected in health surveys. Moreover, this method could be replicated on previous versions of the survey and it would enable us to study changes over time. It could also be applied with minor adaptations to other surveys as the severity index that we propose is related to the International Classification of Diseases ICD-10. This aggregation and bias correction method could also easily be used with other sociodemographic, health and health related behaviour variables. The main strength of this method is to use retrospective information from health surveys.

Our model uses both an ordered Probit and new explanatory variables. As a result, the measurement of health we propose is cardinal as it initially relies on a cardinal numeral determinant: the individual number of diseases.

Another important result of our study is the significance of the cluster effect due to unobserved heterogeneity among households. It means that important common unobserved factors among households affect either the general health status or the scale itself. We have chosen to use a random effect model to correct this bias. In the process, we found evidence of instability in the value of the coefficients and their standard errors, which reduces their significance. Although the use of a random effect model rather than a fixed effect model is debatable, it is important to stress that if we do not take into account this household effect, it may generate biases, reduce the accuracy of estimates and make coefficients less comparable among populations because of shrinkage. As this household effect is significant for French data, it might also be observed in other countries. However, as far as we know, no studies have considered this household effect in health reports so far.

As to shifting thresholds, their introduction does not substantially modify the values of coefficients associated to the degree of severity, except for the highest one. A model with varying thresholds is more informative; however, in our study such a model does not involve a significant improvement of the estimation and is also costly in terms of time calculation. The model with fixed thresholds is preferred because our main purpose is to use the estimated parameters as weights of the number of diseases to construct a health indicator. In addition, the empirical illustration shows that this index allows health status comparisons between different populations and distributions analyses. Therefore, it offers news prospects of analyses such as inequality analysis using stochastic dominance.

To our opinion, the health index could also offer other prospects of analyses than those proposed in the dissertation. For instance, this indicator could be used within an analysis of health care consumption according to care need which would be defined from several morbidity indicators. In this context, the index would be a good solution to avoid autocorrelation.

8 References

Allison, P. D. 1999. *Logistic regression using the SAS System: theory and application.* Cary: SAS Institute Press.

Baker, M., Stabile, M., & Deri, C. 2004. What do self-reported, objective, measures of health measure? *Journal of Human Resources*, 39(4): 1067-1093.

Bellanger, M. & Jourdain, A. 2004. Tackling regional inequalities in France by resource allocation: a case for complementary instrumental and process-based approaches. *Applied Health Econometrics and Health Policy*, 3(4): 243-250.

Bergson, H. 1920. *L'évolution créatrice.* Paris: Librairie Félix Alcan.

Blaxter, M. 1990. *Health and lifestyles.* London and New York: Tavistock/Routledge Publication.

Chauvin, P. & Lebas, J. 2007. Inégalités et disparités sociales de santé en France, In: *Etat de santé et principales pathologies*, Ch. 42, 331-341.

Com-Ruelle, L., Lecomte, T., Le Fur, P., Mizrahi, An., Mizrahi, Ar., & Sermet, C. 1997. Un indicateur de gravité des maladies: enquête sur la santé et les soins médicaux, France 1991-92. 1189. IRDES Report.

Cowell, F. A. 2000. Measurement of inequality. In T. Atkinson & F. Bourguignon (Eds.), *Handbook of Income Distribution*: 89-169. London: Elsevier Science.

Cutler, D. M. & Richardson, E. 1997. Measuring the health of the United States population. *Microeconomics*, Brookings Paper in Economic Activity: 217-271.

Dauphinot, V., Naudin, F., Guéguen, R., Perronnin, M., & Sermet, C. 2006. Ecarts entre morbidité déclarée et morbidité diagnostiquée - L'exemple de l'obésité, de l'hypertension artérielle et de l'hypercholestérolémie. Q*uestions D'Économie De La Santé*, IRDES,(114).

Donaldson, C., Jones, A. M., Mapp, T., & Olson, J. A. 1998. Limited dependent variables in willingness to pay studies: applications in health care. *Applied Econometrics*, 30: 667-677.

Ferrer-I-Carbonell, A. & Frijters, P. 2004. How important is methodology for the estimates of the determinants of happiness? *Economic Journal*, 114(497): 641-659.

Gerdtham, U. G., Johannesson, M., Lundberg, L., & Isacson, D. 1999. A note on validating Wagstaff and van Doorslaer's health measure inequalities in health. *Journal of Health Economics*, 18(1): 117-124.

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

Jürges, H. 2007. True health vs response styles: exploring cross-country differences in self-reported health. *Health Economics*, 16(2): 163-178.

Jusot, F., Khlat, M., Rochereau, T. & Sermet, C. 2005. Les itinéraires professionnels en relation avec la santé: une exploitation longitudinale de l'enquête ESPS de l'IRDES. DARES, DREES, collection MiRe, and La Poste. Intermediate Report.

Khlat, M. & Sermet, C. 2004. La santé des chômeurs en France: revue de la litterature. *Revue D'Epidémiologique Et De Santé Publique*, 52(3): 465-474.

Lefranc, A., Pistolesi, N., & Trannoy, A. 2004. Le revenu selon l'origine sociale. *Economie Et Statistique*, 371: 49-82.

Liang, J., Bennett, J. M., Whitelaw, N. A., & Maeda, D. 1991. The structure of self-reported physical health among the aged in the United States and Japan. *Medical Care*, 29(12): 1161-1180.

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

Manderbacka, K., Lundberg, O., & Martikainen, P. 1999. Do risk factors and health behaviours contribute to self-ratings of health? Social Science & Medicine, 48: 1713-1720.

McDowell, I. 2006. *Measuring health - a guide to rating scales and questionnaires.* Oxford.

Mizrahi, A. & Mizrahi, A. 1985. Indicateurs de morbidité et facteurs sociodémographiques. Rapport IRDES, 657.

Rochaix, L. & Hartmann, L. 2005. The French public-private mix. In A. K. Maynard (Ed.), *The public-private mix for health care.*

Shmueli, A. 2003. Socio-economic and demographic variation in health and in its measures: the issue of reporting heterogeneity. Social Science & Medicine, 57(1): 125-134.

Stewart, M. 1983. On least squares estimation when the dependent variable is grouped. Review of Economic Studies, L: 737-753.

van Doorslaer, E. & Jones, A. 2003. The determinants of self-reported health inequalities: validation of a new approach to measurement. *Journal of Health Economics*, 22: 61-87.

Wagstaff, A. & van Doorslaer, E. 1994. Measuring inequalities in health in the presence of multiple-category morbidity indicators. *Health Economics*, 3: 281-291.

World Health Organization 2002. *Summary measures of population health: concepts, ethics, measurement and applications.* WHO Books. Geneva.

9 Appendix

Vital risk	**Disability level**
0 No vital risk	0 No discomfort
1 Prognosis very weakly pejorative	1 Very weakly hampered
2 Prognosis weakly pejorative	2 Moderately hampered
3 Possible risk on vital conditions	3 Hampered but normal life
4 Prognosis probably bad	4 Limited professional/domestic activity
5 Prognosis certainly bad	5 Highly hampered
6 Undetermined or deceased during the survey	6 No autonomy for domestic activities
	7 Confinement to bed
	8 Undetermined or deceased during the survey

Table 1: Two morbidity indicators in IRDES-HHIS: vital risk and disability level (Irdes, *Enquête Santé et Protection Sociale.*)

Minimal vital risk (MVR)	**Minimal disability level (MDL)**
0 No vital risk	0 No discomfort
1 Prognosis very weakly pejorative	1 Very weakly hampered
2 Prognosis weakly pejorative	2 Moderately hampered
3 Possible risk on vital conditions	3 Hampered but normal life
4 Prognosis probably bad	4 Limited professional/domestic activity
5 Prognosis certainly bad	5 Highly hampered
	6 No autonomy for domestic activities

Table 2: Minimal vital risk and minimal disability level (IRDES, Com-Ruelle *et al.*, 1997.)

	MDL=0	MDL=1	MDL=2	MDL=3	MDL=4	MDL=5	Total by row
MVR=0	351 46,2% 90,0%	135 17,8% 76,3%	164 21,6% 56,9%	78 10,3% 39,0%	28 3,7% 16,6%	4 0,5% 7,0%	760 59,3%
MVR=1	1 20,5% 8,5%	33 21,7% 19,8%	61 37,9% 21,2%	20 12,4% 10%	11 6,8% 6,5%	1 0,6% 1,8%	161 12,6%
MVR=2	5 4,62% 1,3%	4 3,6% 2,3%	40 36,4% 13,9%	38 34,6% 19%	19 17,3% 11,2%	4 3,6% 7%	110 8,6%
MVR=3	1 0,6% 0,3%	3 1,9% 1,7%	23 14,7% 8%	60 38,5% 30%	56 35,9% 33,1%	13 8,3% 22,8%	156 12,2%
MVR=4	0 0% 0%	0 0% 0%	0 0% 0%	4 4,3% 2%	55 58,5% 32,5%	35 37,2% 61,4%	94 7,3%
Total by column	390 30,4%	177 13,8%	288 22,5%	200 15,6%	169 13,2%	57 4,4%	1281

Table 3: Correlation between minimal vital risk and minimal disability level

Statistic	**DF**	**Value**	**Prob**
Chi-Square	20	900,4817	<.0001
Likelihood Ratio Chi-Square	20	812,2337	<.0001
Mantel-Haenszel Chi-Square	1	591,3311	<.0001
Phi Coefficient		0,8384	
Contingency Coefficient		0,6425	
Statistic		**Value**	**ASE**
Gamma		0,7411	0,0183
Kendall's Tau-b		0,5547	0,0171
Pearson Correlation		0,6797	0,0162
Spearman Correlation		0,6318	0,0188

Table 4: Summary statistics for minimal vital risk by minimal disability level

	MDL=0	MDL=1	MDL=2	MDL=3	MDL=4	MDL=5
MVR=0	k=1	k=1	k=6	k=6	k=6	k=6
MVR=1	k=7	k=2	k=2	k=8	k=8	k=8
MVR=2	k=7	k=7	k=3	k=3	k=8	k=8
MVR=3	k=9	k=9	k=9	k=4	k=4	k=4
MVR=4	k=9	k=9	k=9	k=9	k=5	k=5

Table 5: Definition of nine possible severity levels for a disease

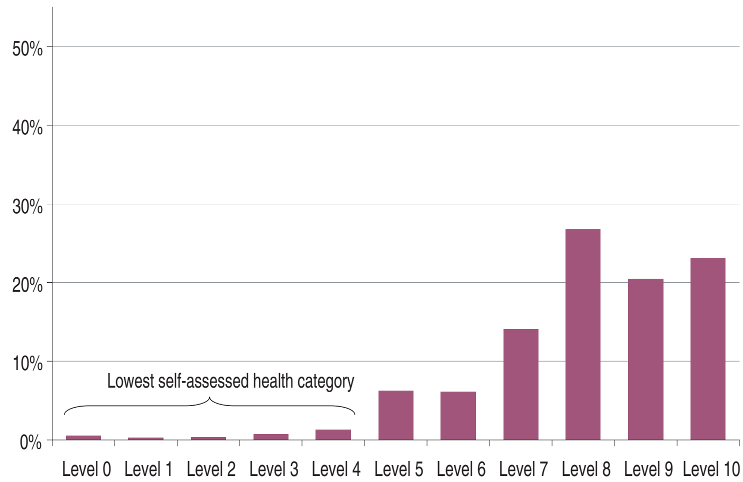

Figure 1: Distribution of self-assessed health (*2002 IRDES-HHIS*)

Variables	Mean	Proportion
Age	43.4	
Income (monthly)	1 381.16	
Education level		
Higher education	2,492	28.86%
High school	1,823	21.11%
Secondary education	4,320	50.03%
Professional activity		
Farmer	351	4.06%
Craftsmen retailer	434	5.03%
Executive	1,151	13.33%
Technician	1,926	22.30%
Other employees	2,256	26.13%
Skilled worker	1,667	19.31%
Unskilled worker	850	9.84%
Current activity		
Active	4,986	57.74%
Student	977	11.31%
Unemployed	458	5.30%
Retired	1,541	17.85%
Homemaker	492	5.70%
Inactive	181	2.10%
Social health insurance		
Private	7,766	89.94%
Cmu	291	3.37%
No supplemental insurance	578	6.69%

Table 6: Descriptive statistics of sociodemographic variables (*2002 IRDES-HHIS*)

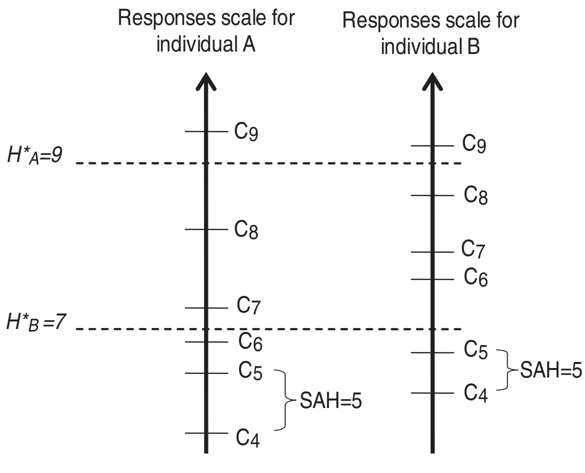

Figure 2: Process of self-assessment for health for 2 individuals A and B

Covariates	**Log likelihood**
Demographic variables	-12,716.96
Education Level	-12,725.35
Occupational activity	-12,700.82
Labor market status	-12,716.76
Household income	-12,735.76
Health insurance	-12,744.16
Smoking	-12,743.63
Alcohol consumption	-12,718.84
Body mass index	-12,740.34

Table 7: Effects of covariates on varying thresholds

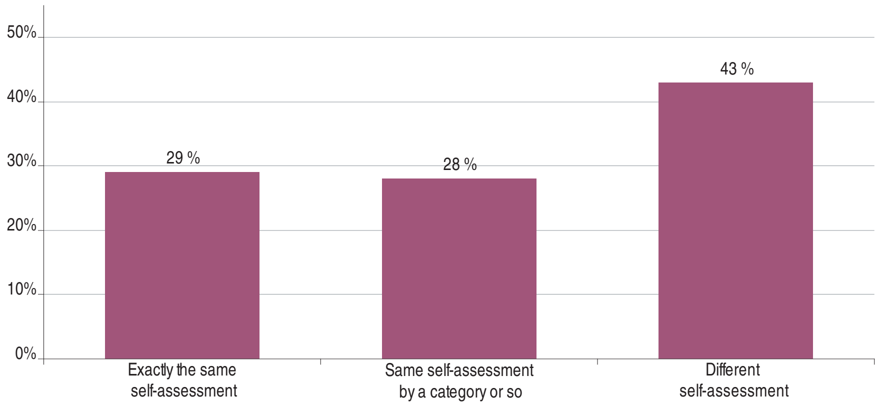

Figure 3: Variations in individual self-assessed health within the same household (*2002 IRDES-HHIS*)

Ordered Logit regression without cluster effects					Ordered Logit regression with cluster effects				
Log likelihood = -12760,638 Pseudo R2=0,1537					Log likelihood = -12488,537				
Variables	**Coef.**	**S.E**	**P>z**	**[Conf. Int.]**	**Variables**	**Coef.**	**S.E**	**P>z**	**[Conf. Int.]**
Cross product of vital risk by disability									
k=1	-0.338***	0,026	0	[-0,388; -0,287]	k=1	-0,407***	0,032	0	[-0,469; -0,345]
k=2	-0.379***	0,045	0	[-0,467; -0,291]	k=2	-0,462***	0,054	0	[-0,568; -0,356]
k=3	-0.521***	0,032	0	[-0,585; -0,458]	k=3	-0,668***	0,041	0	[-0,748; -0,588]
k=4	-0.792***	0,065	0	[-0,919; -0,666]	k=4	-1,011***	0,077	0	[-1,161; -0,861]
k=5	-1.208***	0,144	0	[-1,490; -0,927]	k=5	-1,488***	0,173	0	[-1,827; -1,150]
k=6	-0.440***	0,019	0	[-0,478; -0,402]	k=6	-0,539***	0,024	0	[-0,586; -0,491]
k=7	-0.327**	0,138	0,018	[-0,598; -0,056]	k=7	-0,301	0,230	0,191	[-0,753; -0,151]
k=8	-0.715***	0,102	0	[-0,916; -0,515]	k=8	-0,917***	0,124	0	[-1,160; -0,675]
k=9	-0.692***	0,173	0	[-1,032; -0,352]	k=9	-0,953***	0,206	0	[-1,356; -0,550]
Tobacco consumption									
No smoker	ref.				No smoker	ref.			
Former smoker	-0,048	0,053	0,365	[-0,151; 0,056]	Former smoker	-0,065	0,065	0,319	[-0,193; 0,063]
Light smoker	-0,097	0,063	0,125	[-0,220; 0,027]	Light smoker	-0,187**	0,079	0,018	[-0,342; -0,033]
Heavy smoker	-0.392***	0,067	0	[-0,524; -0,260]	Heavy smoker	-0,482***	0,086	0	[-0,651; -0,314]
Unknown	0,041	0,081	0,613	[-0,118; 0,200]	Unknown	0,042	0,103	0,685	[-0,159; 0,243]
Alcohol consumption									
No cons.	-0.143**	0,055	0,009	[-0,250; -0,035]	No cons.	-0,148**	0,068	0,030	[-0,282; -0,015]
Light cons.	ref.				Light cons.	ref.			
Medium cons.	-0,083	0,054	0,125	[-0,189; 0,023]	Medium cons.	-0,045	0,068	0,507	[-0,179; 0,089]
Heavy cons.	-0.204**	0,085	0,016	[-0,371; -0,038]	Heavy cons.	-0,172	0,105	0,102	[-0,377; 0,034]
Unknown	-0,067	0,094	0,48	[-0,251; 0,118]	Unknown	-0,077	0,116	0,508	[-0,304; 0,151]
Body mass index									
Underweight	0,258	0,157	0,102	[-0,051; 0,566]	Underweight	0,198	0,232	0,393	[-0,257; 0,654]
Normal	ref.				Normal	ref.			
Overweight	0,122	0,141	0,388	[-0,155; 0,398]	Overweight	-0,234***	0,059	0	[-0,350; -0,119]
Obesity	0,100	0,083	0,233	[-0,064; 0,263]	Obesity	-0,575***	0,090	0	[-0,752; -0,398]
Unknown	0,011	0,137	0,937	[-0,257; 0,279]	Unknown	-0,026	0,171	0,881	[-0,361; 0,309]
Log of inc.	0.163***	0,041	0	[0,082; 0,243]	**Log of inc.**	0,231***	0,061	0	[0,111; 0,350]
Professional activity									
Farmer	-0.311	0,109	0,004	[-0,525; -0,097]	Farmer	-0,423***	0,151	0,005	[-0,718; -0,128]
Craftsmen	0.246**	0,101	0,015	[0,047; 0,445]	Craftsmen	0,257*	0,130	0,048	[0,002; 0,512]
Executive	0.276***	0,077	0	[0,125; 0,428]	Executive	0,245**	0,099	0,013	[0,052; 0,439]
Technician	0.143**	0,061	0,02	[0,023; 0,263]	Technician	0,130	0,077	0,092	[-0,021; 0,281]
Employees	ref.				Employees	ref.			
Skilled worker	0,091	0,065	0,161	[-0,036; 0,218]	Skilled worker	0,047	0,081	0,564	[-0,113; 0,207]
Unskilled worker	-0.208**	0,077	0,007	[-0,358; -0,058]	Unskilled worker	-0,300***	0,096	0,002	[-0,489; -0,112]
Education									
Education 3	ref.				Education 3	ref.			
Education 2	0,044	0,059	0,458	[-0,072; 0,159]	Education 2	0,035	0,073	0,636	[-0,109; 0,179]
Education less	-0,009	0,060	0,876	[-0,126; 0,108]	Education less	-0,063	0,076	0,403	[-0,212; 0,085]
Age crossed with gender									
Male 16-34	ref.				Male 16-34	ref.			
Male 35-44	-0.377***	0,083	0	[-0,540; -0,215]	Male 35-44	-0,617***	0,104	0	[-0,820; -0,413]
Male 45-54	-0.879***	0,080	0	[-1,035; -0,723]	Male 45-54	-1,193***	0,099	0	[-1,387; -1,000]
Male 55-74	-0.996***	0,145	0	[-1,281; -0,711]	Male 55-74	-1,287***	0,176	0	[-1,633; -0,942]
Male=>75	-1.262***	0,176	0	[-1,608; -0,917]	Male=>75	-1,660***	0,219	0	[-2,089; -1,232]
Fem. 16-34	-0.206**	0,070	0,003	[-0,343; -0,069]	Fem. 16-34	-0,278***	0,082	0,001	[-0,438; -0,117]
Fem. 35-44	-0.371***	0,084	0	[-0,537; -0,206]	Fem. 35-44	-0,600***	0,105	0	[-0,806; -0,395]
Fem. 45-54	-0.830***	0,083	0	[-0,993; -0,667]	Fem. 45-54	-1,151***	0,103	0	[-1,353; -0,950]
Fem. 55-74	-0.950***	0,141	0	[-1,226; -0,675]	Fem. 55-74	-1,294***	0,174	0	[-1,635; -0,953]
Fem.=>75	-1.226***	0,163	0	[-1,546; -0,906]	Fem.=>75	-1,582***	0,202	0	[-1,979; -1,186]
Current activity									
Active	ref.				Active	ref.			
Student	0.475***	0,077	0	[0,323; 0,627]	Student	0,570***	0,099	0	[0,377; 0,763]
Unemployed	0,014	0,094	0,878	[-0,170; 0,199]	Unemployed	-0,014	0,114	0,904	[-0,238; 0,21]
Retired	0,027	0,091	0,767	[-0,151; 0,205]	Retired	0,068	0,114	0,551	[-0,155; 0,291]
Homemaker	-0,151	0,094	0,108	[-0,335; 0,033]	Homemaker	-0,091	0,111	0,433	[-0,305; 0,131]
Inactive	-0.904***	0,151	0	[-1,200; -0,608]	Inactive	-1,068***	0,180	0	[-1 424; -0,719]
Social health insurance									
Private	ref.				Private	ref.			
CMU	-0,164	0,124	0,184	[-0,407; 0,078]	CMU	-0,182	0,169	0,281	[-0,513; 0,149]
No insurance	-0.336***	0,082	0	[-0,497; -0,175]	No insurance	-0,375***	0,112	0,001	[-0,595; -0,156]
Cut-off point estimates									
Cut1	-5,329	0,318			Cut11	-6,532***	0,464	0	[-7,441; -5,622]
Cut2	-3,746	0,312			Cut12	-4,659***	0,457	0	[-5,555; -3,764]
Cut3	-2,896	0,311			Cut13	-3,617***	0,455	0	[-4,508; -2,725]
Cut4	-1,649	0,310			Cut14	-2,053***	0,453	0	[-2,954; -1,166]
Cut5	-0,025	0,309			Cut15	0,039	0,451	0,932	[-0,846; 0,923]
Cut6	1,193	0,309			Cut16	1,647***	0,452	0	[0,762; 2,533]
Significance of parameters *<0,10, **<0,05, ***<0,01					Intra cluster	1,874	0,1291		

Table 8: Results of the ordered Logit regressions without and with clusters effects.

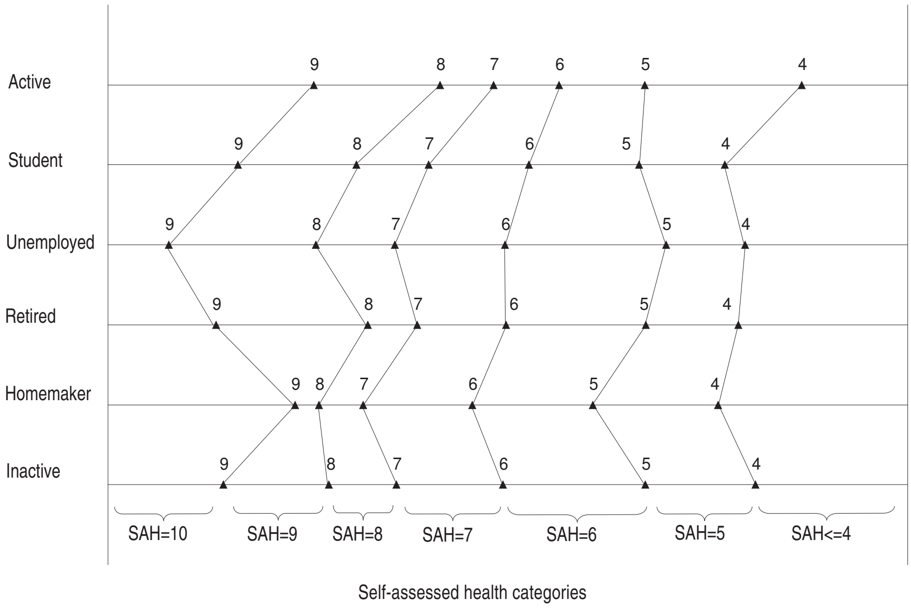

Figure 4: Effects of occupational status on the thresholds of self-assessed health.

Disease severity level	**Without cluster effect without varying thresholds**	**With cluster effects without varying thresholds**	**With cluster effect varying thresholds**
k=1	1	1	1
k=2	1,12	1,14	1,14
k=3	1,54	1,64	1,67
k=4	2,34	2,48	2,56
k=5	3,57	3,66	3,90
k=6	1,30	1,32	1,35
k=7	0,97	0,74	0,76
k=8	2,12	2,25	2,34
k=9	2,05	2,34	2,40

Table 10: Values of weights according to the model specification

Ordered Logit regression with cluster effects and occupation varying thresholds									
Log likelihood = -12419.813									
Condition Number = 269.97443									
Variables	**Coef.**	**S.E**	**P>z**	**[Conf. Int.]**	**Variables**	**Coef.**	**S.E**	**P>z**	**[Conf. Int.]**
Cross product of vital risk by disability					**Social health insurance**				
k=1	-0,404***	0,032	0	[-0,467; -0,342]	Private	ref.			
k=2	-0,461***	0,055	0	[-0,568; -0,354]	CMU	-0,175	0,171	0,307	[-0,510; 0,160]
k=3	-0,673***	0,041	0	[-0,753; -0,593]	No supp. ins.	-0,401***	0,113	0	[-0,622; -0,180]
k=4	-1,034***	0,078	0	[-1,186; -0,882]	**Cut-off point estimates**				
k=5	-1,575***	0,175	0	[-1,918; -1,231]	**Cut11**				
k=6	-0,544***	0,024	0	[-0,592; -0,496]	Active	ref.			
k=7	-0,308	0,232	0,185	[-0,763; 0,147]	Student	1,076***	0,409	0,008	[0,275; 1,878]
k=8	-0,946***	0,125	0	[-1,191; -0,702]	Unemployed	-0,111	0,401	0,781	[-0,896; 0,674]
k=9	-0,970***	0,207	0	[-1 376; -0,564]	Retired	-0,817***	0,220	0	[-1,249; -0,385]
Tobacco consumption					Homemaker	0,221	0,309	0,474	[-0,384; 0,827]
No smoker	ref.				Inactive	1,353***	0,313	0	[0,740; 1,966]
Former smoker	-0,069	0,066	0,294	[-0,198; 0,060]	Cons	-6,268***	0,482	0	[-7,212; -5,325]
Light smoker	-0,192**	0,079	0,015	[-0,348; -0,037]	**Cut12**				
Heavy smoker	-0,474***	0,086	0	[-0,644; -0,305]	Active	ref.			
Unknown	0,037	0,104	0,722	[-0,166; 0,240]	Student	-0,145	0,344	0,673	[-0,819; 0,529]
Alcohol consumption					Unemployed	0,585***	0,220	0,008	[0,155; 1,016]
No cons.	-0,146***	0,069	0,033	[-0,281; -0,012]	Retired	-0,195	0,157	0,214	[-0,503; 0,113]
Light cons.	ref.				Homemaker	0,416**	0,203	0,040	[0,018; 0,814]
Medium cons.	-0,050	0,069	0,469	[-0,185; 0,085]	Inactive	1,667***	0,250	0	[1,177; 2,157]
Heavy cons.	-0,170	0,106	0,108	[-0,378; 0,038]	Cons	-4,686***	0,467	0	[-5,601; -3,770]
Unknown	-0,086	0,117	0,463	[-0,315; 0,143]	**Cut13**				
Body mass index					Active	ref.			
Underweight	0,210	0,234	0,369	[-0,249; 0,669]	Student	-0,503*	0,269	0,062	[-1,031; 0,025]
Normal weight	ref.				Unemployed	0,306	0,188	0,103	[-0,061; 0,674]
Overweight	-0,232***	0,059	0	[-0,348; -0,117]	Retired	-0,023	0,140	0,867	[-0,298; 0,251]
Obesity	-0,574***	0,091	0	[-0,752; -0,396]	Homemaker	0,481***	0,169	0,004	[0,151; 0,812]
Unknown	-0,021	0,172	0,901	[-0,359; 0,316]	Inactive	1,456***	0,239	0	[0,987; 1 925]
Log of income					Cons	-3,670***	0,463	0	[-4,578; -2,762]
Log of income	0,233***	0,062	0	[0,112; 0,354]	**Cut14**				
Professional activity					Active	ref.			
Farmer	-0,449***	0,154	0,003	[-0,751; -0,148]	Student	-0,461***	0,165	0,005	[-0,784; -0,138]
Craftsmen	0,278**	0,131	0,034	[0,021; 0,534]	Unemployed	0,044	0,154	0,776	[-0,258; 0,345]
Executive	0,253**	0,099	0,011	[0,058; 0,448]	Retired	0,033	0,128	0,795	[-0,218; 0,285]
Technician	0,136*	0,078	0,080	[-0,016; 0,288]	Homemaker	0,392***	0,144	0,006	[0,110; 0,674]
Other employees	ref.				Inactive	0,846***	0,240	0	[0,376; 1,316]
Skilled worker	0,047	0,082	0,565	[-0,113; 0,207]	Cons	-2,074***	0,459	0	[-2,974; -1,174]
Unskilled worker	-0,307***	0,097	0,002	[-0,497; -0,118]	**Cut15**				
Education					Active	ref.			
Education 3	ref				Student	-0,794***	0,119	0	[-1,027; -0,561]
Education 2	0,029	0,074	0,690	[-0,115; 0,174]	Unemployed	0,005	0,143	0,973	[-0,276; 0,285]
Education less	-0,063	0,076	0,407	[-0,212; 0,086]	Retired	0,310**	0,138	0,025	[0,039; 0,581]
Age crossed with gender					Homemaker	-0,100	0,141	0,480	[-0,377; 0,177]
Male 16-34	ref				Inactive	-0,008	0,263	0,976	[-0,524; 0,508]
Male 35-44	-0,594***	0,104	0	[-0,798; -0,390]	Cons	0,066	0,457	0,886	[-0,831; 0,962]
Male 45-54	-1,156***	0,099	0	[-1,351; -0,961]	**Cut16**				
Male 55-74	-1,255***	0,179	0	[-1,605; -0,905]	Active	ref.			
Male=>75	-1,694***	0,222	0	[-2,129; -1,258]	Student	-0,564***	0,110	0	[-0,779; -0,348]
Fem. 16-34	-0,288***	0,082	0	[-0,450; -0,127]	Unemployed	-0,261	0,160	0,102	[-0,575; 0,052]
Fem. 35-44	-0,597***	0,105	0	[-0,803; -0,390]	Retired	-0,161	0,163	0,324	[-0,480; 0,159]
Fem. 45-54	-1,108***	0,103	0	[-1,311; -0,905]	Homemaker	-0,459***	0,159	0,004	[-0,771; -0,148]
Fem. 55-74	-1,260***	0,176	0	[-1,605; -0,916]	Inactive	0,689*	0,367	0,060	[-0,030; 1,408]
Fem.=>75	-1,604***	0,205	0	[-2,006; -1,202]	Cons	1,718***	0,458	0	[0,820; 2 616]
Intra cluster variance						1,871	0,129		
Significance of parameters *<0.10, **<0.05, ***<0.01									

Table 9: Results of the ordered Logit regression with clusters effects and varying thresholds due to occupation status.

Figure 5: Comparison of the distributions of the health index and self-assessed health

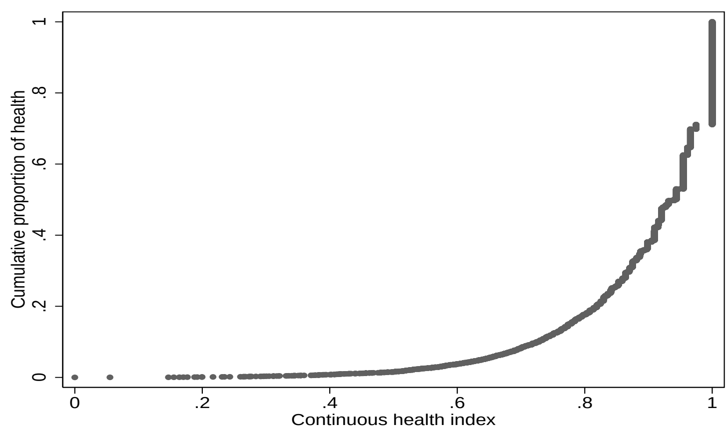

Figure 6: Empirical distribution function of the health index.

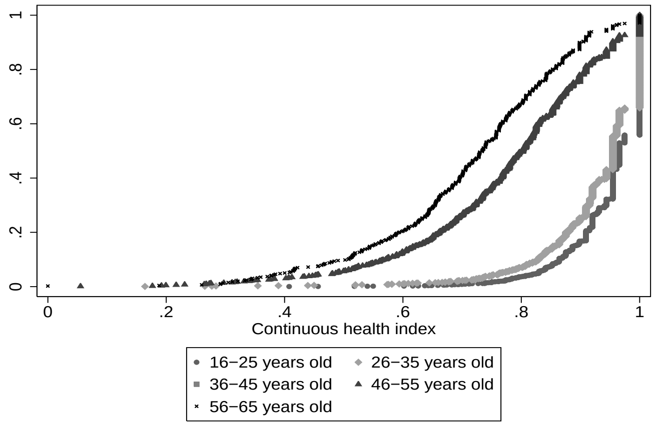

Figure 7: Empirical distribution function of the health index per age classes.

	16-25 y.o	26-35 y.o	36-45 y.o	46-55 y.o	56-65 y.o
16-25 y.o		<0,0001***	<0,0001***	<0,0001***	<0,0001***
26-35 y.o	1		<0,0001***	<0,0001***	<0,0001***
36-45 y.o	1	1		<0,0001***	<0,0001***
46-55 y.o	1	1	1		<0,0001***
56-65 y.o	1	1	1	0.977	

Significance levels: * (10%), ** (5%) and *** (1%)

Table 11: P-value of *Kolmogorov-Smirnov* test related to health according to age classes.

Explanation of the table: the result of the unilateral *Kolmogorov-Smirnov* test is read in row. The distribution of health of an individual aged 26-35 years old significantly dominates the distribution of health of individuals aged 36-45 years old, 46-55 years old, 56-65 years old as p-value<0,0001.

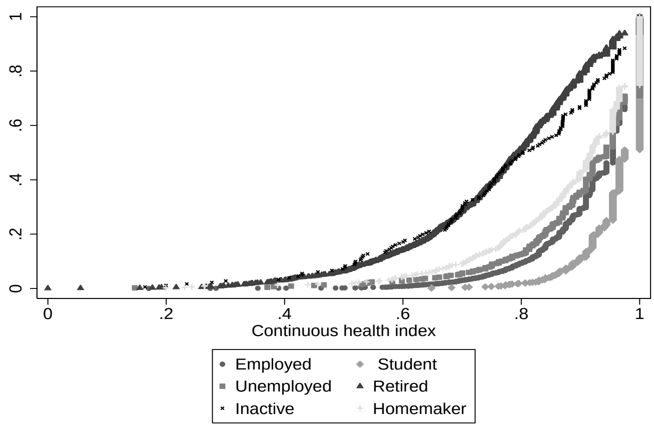

Figure 8: Empirical distribution function of the health index per socioeconomic statuses.

	Employed	Student	Unemployed	Retired	Inactivity	Homemakers
Employed		1	0,014**	<0,0001***	<0,0001***	<0,0001***
Student	<0,0001***		<0,0001***	<0,0001***	<0,0001***	<0,0001***
Unemployed	1	1		<0,0001***	<0,0001***	0,011**
Retired	1	1	0.999		0,674	0.982
Inactive	1	1	0,999	<0,003		1
Homemakers	1	1	0,998	<0,0001***	<0,0001***	

Significance levels: * (10%), ** (5%) and *** (1%)

Table 12: P-value of *Kolmogorov-Smirnov* test related to health according to socioeconomic statuses

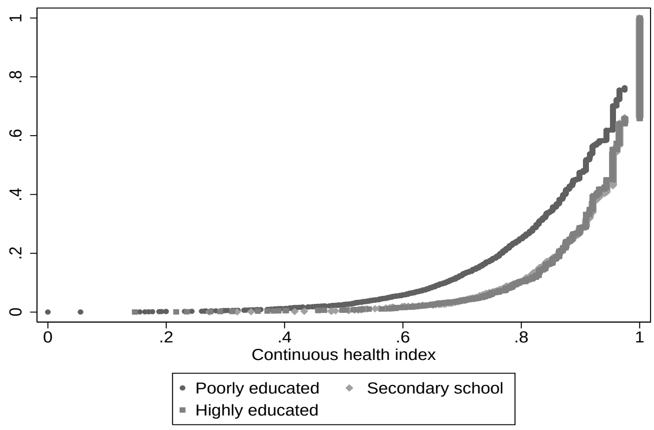

Figure 9: Empirical distribution function of the health index per education levels.

	Education 3	Education 2	Education less
Education 3		0,575	<0,0001***
Education 2	0,419		<0,0001***
Education less	1	1	

Significance levels: * (10%), ** (5%) and *** (1%)

Table 13: P-value of *Kolmogorov-Smirnov* test related to health according to education levels