GREDI
Groupe de Recherche en Économie
et Développement International

Cahier de recherche / Working Paper
17-08

A novel health metric based on biomarkers

Qing LI
Alain A. COHEN
Linda P. Fried

UNIVERSITÉ DE
SHERBROOKE

A novel health metric based on biomarkers*

Qing Li† Alan A. Cohen‡ Linda P. Fried§

Abstract

We propose a biomarker-based, objective and continuously distributed health measure which is novel to the economics literature. Using 41 commonly available biomarkers in two leading biomarker databases, we consider the individuals as points in a 41-dimensional biomarker space and measure their objective health as the Mahalanobis distance to the centroid of a reference group. In this way the centroid of the reference group represents an "ideal state" of health, and a bigger distance from this centroid indicates worse health. We validate versions of our health measure using different number of biomarkers and through the link with a commonly used measure of general health (self-reported health); we find that our health measure is positively but not perfectly linked to self-reported health. Additionally, we find that the signal of health increases with the number of biomarkers included; nonetheless, it is clearly feasible to have a signal with fewer biomarkers though the signal could be weaker. Finally we illustrate our health measure in two applications: 1) the estimation of health distribution where we find a long tail representing individuals in very bad health; 2) the concentration index where we can truly satisfy the requirement for continuously defined general health.

Keywords: Biomarkers; health measurement; Mahalanobis distance.

*We are grateful to Jean-Yves Duclos, Sabine Kröger and Owen O'Donnell for helpful and constructive comments. We aslo thank many other participants of the seminar at the Université Laval, the presentation at the 55th Congrès Annuel de la Société Canadienne de Science Économique (Montreal) and the presentation at the 11th World Congress in Health Economics (Milan), for stimulating discussions. We thank the Canadian Institutes of Health Research (CIHR) - Operating Grant 120305, for financial support.

†PRIMUS, University of Sherbrooke, 3001 12e Ave N, Sherbrooke, QC, J1H 5N4, Canada. **Corresponding author**. E-mail: qing.li@usherbrooke.ca; herrliqing@yahoo.com

‡Department of Family Medicine, University of Sherbrooke, 3001 12e Ave N, Sherbrooke, QC, J1H 5N4, Canada.

§Mailman School of Public Health, Columbia University, 722 W. 168th Street, R1408, New York, NY 10032, USA.

1. Introduction

One of the most important issues in health economics is the measurement of health. Many health measures are self-reported. Although these measures are often good predictors of other outcomes, they are often conditional on cultural or social norms making it difficult to compare across populations (Lindeboom and Van Doorslaer, 2004; Jürges, 2007). Hence it is important to have objective measures of health (Bound, 1991; Burkhauser and Cawley, 2006). Additionally, commonly used measures of general health are often discretely distributed (e.g. self-reported health).

Apart from self-reported health (SAH), widely used health measures include the Disability-Adjusted Life Year (World Bank, 1993), Disability-Adjusted Life Expectancy (Murray et al., 2002), the Health and Activity Limitation Index, (Erickson, 1998; Sondik, 2002; Stewart et al., 2005), the McMaster Health Utility Index Mark 3 (HUI3) (Feeny et al., 1995), among many others. Recently, Meijer et al. (2011) stressed international comparability and proposed a health index by interrelating self-reported health with functional limitations and a physical measure. Many continuous health measures are composed of discrete health indicators. For example, it is possible to count the number of health problems and give each individual a score, as reflected in the health measure of Jürges (2007) or the Fried frailty scale (Fried et al., 2001), among others. Single biomarkers are also used occasionally as health indicators. Generally speaking, biomarkers are objectively measured health indicators. There is usually a normal range for each biomarker; values out of the normal range, whether too low or too high, are associated with certain health problems. A way is to categorize single biomarkers one by one, for example, categorizing BMI into three categories "underweight," "normal" and "overweight"; this is because values of single biomarkers, especially those within the "normal" range, are difficult to compare: if one has BMI of 23 and another has 24, we cannot see a difference of health from their BMI.

Here we propose a continuous and objective measure of general health based on simultaneous use of multiple biomarkers. This may lead to the following benefits: first, it facilitates estimation of health distribution and inequality; second, it facilitates application of statistical tools such as OLS or IV regression whose consistency requires very weak assumptions; other techniques such as the elimination of fixed individual effect can also be easily applied. Third, it could be compared more easily across different populations. In addition to the three technical advantages, a more intrinsic difference from existing health measures is that our measure is constructed from truly continuous variables. The existing health measures, including categorized single biomarkers, can be viewed as proxies of health status. In this sense, our measure describes an exact biological state of each individual in the

biomarker/health space. Therefore, our health measure allows researchers to see the exact difference in health from one individual to another without much need of model selection. Our health measure does not need categorization of continuous biomarkers, avoiding loss of information (Barnwell-Ménard et al., 2015). Comparing with the use of single biomarkers one by one, the simultaneous use of multiple biomarkers could be more convincing at least when the objective is to measure general health. When a single biomarker changes, it is not easy to predict the exact effect on general health without knowing the biomarker's interaction with other biomarkers (see Appendix for mathematical demonstration).

In socio-economic surveys there are often less available biomarkers than in biomedical databases. While it is feasible to have a signal of general health with few biomarkers (e.g. five biomarkers, see section 3), it could be preferable to combine the biomarkers with other (discrete) indicators of health in certain cases. This could be done with, for example, the health measurement model of Meijer et al. (2011) where our health measure could serve as the internationally comparable objective health measure required by their model, in place of grip strength. We leave this for future research.

The paper is organized in the following way. In section 2 we introduce the health measure; in section 3 we validate versions of the health measure through the link with self-reported health which is another commonly used measure of general health; in section 4 we illustrate the health measure in two applications; finally we conclude in section 5.

2. Health metric

2.1. Biomarker data in socio-economic research

Biomarker data have not been systematically explored in health economics to our knowledge; nonetheless, there has been growing interest. The collection of biomarker data has also started in well-known socio-economic surveys. For example, the Health and Retirement Study (HRS) has collected two waves of biomarkers together with anthropometric measures leading to a total of fourteen objective indicators of health (Crimmins et al., 2013). Similarly, the Survey of Health, Ageing and Retirement in Europe (SHARE) collected biomarker data in wave 4 in Germany; SHARE plans to include the collection of dried blood spot samples in most of the SHARE countries in wave 6 (SHARE, 2014).

Single biomarkers have been used occasionally in economic studies. Among others the Body Mass Index (BMI) is a widely used indicator of obesity, calculated from weight and height with a normal range between 18.5 and 25 (WHO, 2000). Powdthavee (2010) used hypertension as a biomarker for health and studied the effect of education on health with data from the Health Survey for England together with the 1947 and 1973 compulsory law

changes to estimate RD and IV probit models. Some recent studies explored blood test biomarker(s). For instance, Li (2014) measured health outcome with mean corpuscular hemoglobin (MCH). Jürges et al. (2013) also used the changes in UK compulsory school law in 1947 and 1973 to study the causal effect of education on biomarkers (blood fibrinogen and C-reactive protein), as well as more traditional measures of self-reported health.

2.2. Our health metric

Here we view general health as a multidimensional summary measure of an individual's overall biological state. We consider the observed individuals as points in a multi-dimensional biomarker space where each biomarker is an axis of the space. Hence each individual is represented by a vector of observed biomarker values. We first define a reference group whose centroid represents the "ideal state". After this we calculate the multivariate distance with respect to the centroid for each individual in the study group. In this way a bigger distance from the centroid indicates worse health while a smaller distance represents better health.

We use the Mahalanobis distance to measure health in the biomarker space. The distance is proposed by Mahalanobis (1936) and accounts for the joint probability distribution of the variables under a supposition of multivariate normality. Formally, let $x_i \in \mathcal{X} \subset \mathbb{R}^p$ and $x_j \in \mathcal{X} \subset \mathbb{R}^p$ be two points in a p-dimensional biomarker space $\mathcal{X}$ where the points typically represent observed individuals but can also represent certain hypothetical states of health an individual might have. A distance $D(x_i, x_j)$ is a metric which defines a mapping over the biomarker space $\mathcal{X} \subset \mathbb{R}^p$ for $x_i, x_j \in \mathcal{X}$. Assuming $x_i, x_j \sim N(\mu, \Sigma)$ where N denotes multivariate normality, the Mahalanobis distance between x_i and x_j is defined as

$$D_M(x_i, x_j) := \|T(x_i) - T(x_j)\| \tag{2.1}$$

where $T(x) = \Sigma^{-\frac{1}{2}}(x - \mu)$ and $\|\cdot\|$ is the Euclidean norm. Mahalanobis explained the distance in parallels with the Galilean transformation, a standard method in physics that maps all coordinates into a frame of reference and thus ensures comparability between the objects; and indeed the distance is a composition of the Euclidean distance and the transformation $T(x)$. Therefore the Euclidean distance is weighed by the (empirically computed) covariance matrix Σ. When the biomarkers are correlated, $T(x)$ shrinks (down-weights) distances along highly correlated axes so as to avoid redundancy, and the Mahalanobis distance reduces to the Euclidean distance when Σ is the identity matrix, i.e. if the biomarkers are uncorrelated with each other. As an intuitive example, take the 2-D case where height and weight are the two biomarkers/axes. It is relatively common to see (A) people with 2m and 130kg (B) people with 1.5m and 40kg and it is relatively rare to see (C) people with 2m and

40kg or (D) 1.5m and 130 kg. The four cases (A), (B), (C) and (D) have similar Euclidean distance with respect to a centroid, say, 1.7m and 70kg, while (C) and (D) have much bigger Mahalanobis distance than (A) and (B), when taking into account the correlation between height and weight. Classical application of the distance assumes multivariate normality; nonetheless, existence and uniqueness of the Mahalanobis distance still hold when data are not normally distributed; the distance can be computed after proper transformation of the data (Eström 2011). In the application we take logarithm of the biomarkers in order to approximate multivariate normality. While the logarithm is not always the best transformation, it is often close to. Additionally, the Mahalanobis distance can become unreliable when the scales of the variables differ; we thus standardize each biomarker with respect to the mean and standard deviation of the reference group.

Biologically, the idea to use a statistical distance such as D_M to measure health status is based on the theory of physiological dysregulation (see e.g. Fried et al., 2005; Cohen et al., 2012). The theory states that organisms maintain their physiological function through a complex regulatory network that allows adjustments to changing internal and external conditions including stress, diet composition, disease, and so forth. These networks wear out over time, and this departure/dysregulation from the ideal health state is a key cause of aging and chronic diseases. Because each individual's system may become dysregulated in a unique way, we seek here to compare the different individuals using each individual's physiological dysregulation as an indicator of her general health status, which references her status in terms of degree of "normalcy" relative to a representative population.

Recent empirical results from a biomedical context also support this interpretation of D_M. The studies find that (1) D_M increases with age and predicts mortality, clinical frailty, number of comorbidities, cardiovascular disease, and diabetes (but not cancer, which could be affected by other socio-economic factors); (2) including more biomarkers in the calculation of D_M provides a clearer signal but the marginal gain from adding a marker is diminishing; (3) the signal is not qualitatively sensitive to choice of biomarkers in general; (4) using a younger, healthier reference population to estimate the centroid could boost the signal; and (5) these results are replicable across multiple populations from Europe and North America (Cohen et al., 2013; Cohen et al., 2014; Milot et al., 2014b; Cohen et al., 2015; Cohen 2015; Li et al., 2015). Milot et al. (2014a) also found similar results using biomarkers of wild birds.

2.3. Choice of reference group and centroid

An important issue is how to define a reference group. There are widely accepted normal ranges of single biomarkers, but there is no consensus on a point-wise multivariate centroid.

For this study we compare two choices of reference group. Since our study group consists uniquely of aged women, we first use a group of younger women as reference. However, there could be some physiological adaptation with age, hence part of the distances from the centroid of the younger reference group may not only represent a deterioration in health; additionally, physiology of the same population may change from generation to generation with changes in life style. For this reason we also use the aged women themselves as their own reference group, and we compare the two results.

It is likely that neither of the two choices are optimal and we admit this as a limitation. Possible improvements include selecting a healthy reference group with specific criteria, which may provide a better centroid and weighting matrix. Concerning the two datasets we use, there is no disease information in NHANES. WHAS contains self-reported information on chronic diseases, but we did not include such information with the purpose of keeping the health measure completely objective. In other words, we assume that the diseases are reflected in the biomarkers. For instance, even a psychological problem can be reflected in iron deficiency, particularly observed in women, and this is associated with apathy, depression and rapid fatigue when exercising (Bentona and Donohoe, 1999). Although the biomarkers we use might not provide enough information on certain aspects of health (e.g. healed broken bones that are more fragile than normal bones), we believe that they provide a valuable proxy for general health.

3. Validation of the health measure through self-reported health

3.1. Method

If D_M in the biomarker space is a valid measure of general health, we should expect a positive link between D_M and self-reported health because the latter is also a measure of general health. We also expect that the biomarker-based signal of general health should be strengthened when more biomarkers are included in the calculation of D_M. We validate this through an ordered Logit regression of self-reported health on logarithm of D_M controlling for disability and socio-economic status. We include in the regression a linear term and a quadratic term of D_M, similar to what Lindeboom and Van Doorslaer (2004) did with self-reported health and the HUI, although their final goal was to test reporting bias in self-reported health.

Formally, we define for individual i

$$H_i^O := \frac{\log(D_{M,i}) - \overline{\log(D_M)}}{sd(D_M)}; \tag{3.1}$$

where $D_{M,i}$ is the individual's biomarker-based health. Taking the logarithm eliminates possible influence of extreme values.

In calculating D_M, we do *not* give any special weight to any of the biomarkers over and above the weights from the (empirical) covariance matrix; the reason is that certain biomarkers are well-known to be important for certain diseases or physiological systems, such as glucose for diabetes, but there is no consensus that one biomarker is more important for general health than another, to our knowledge; and thus subjective weighting of individual variables could bias the measure.

The number of available biomarkers may change from one database to another. We wonder how the signal of health could change with the number of biomarkers and whether it is possible to have a signal of health with a smaller number biomarkers. We validate this by varying the number of biomarkers included in the calculation of D_M: in each different case we validate the signal of health through the link of D_M with self-reported health. Hence the standardization in the definition of H_i^O in (3.1) is necessary because the scale of D_M changes with the number of biomarkers included.

3.2. Datasets

We use two datasets for the studies in this paper, the Women's Health and Aging Study (WHAS) and the National Health and Nutrition Examination Survey (NHANES). WHAS is a population-based study of community-dwelling women in Baltimore (Guralnik et al.,1995); we only use the first wave because self-reported health was reported only once. NHANES is a cross-sectional study based on representative samples of the US population and conducted in various waves since the 1970s. Different individuals were surveyed in each different wave with information on hundreds of biomarkers. We combine six waves of data collected from 1999 to 2010 with a total sample size of 56,528 individuals.

Both WHAS and NHANES contain information on many biomarkers, however, each biomaker has some missing values. We take all biomarkers that are available in both datasets and that can be retained while keeping the sample size of the study group (WHAS) equal to 1,000 with no missing data. The need for a common list across two datasets results in a final list of 41 biomarkers, namely Albumin, Alkaline phosphatase, Alpha-carotene, ALT, AST, Beta-cryptoxanthin, Blood urea nitrogen, C-reactive protein, Calcium, Chloride, Cholesterol, Diastolic blood pressure, Ferritin, Folate, GGT, Globulin, Glucose, HDL, Heart rate, Hematocrit, Hemoglobin, Iron, Iron saturation, LDH, MCV, Mean corpuscular hemoglobin, Mean corpuscular hemoglobin concentration, Phosphorous, Potassium, Red blood cell count, Red blood distribution width, Sodium, Systolic blood pressure, Total bilirubin, Total protein, Transferrin, Triglycerides, Uric acid, Vitamin B12, Vitamin D-25 total and White

blood cell count. Because WHAS is all female, we take the female individuals in NHANES aged between 20 and 40 as the reference group, which leads to a sample size of 1,277 for the reference group.

Table I shows the variables and summary statistics. The subjective health H^S is categorical and objective health H^O (as defined previously) is continuous. The objective health H^O_N is with respect to the centroid of the younger reference group from NHANES (superscript "O" for objective health, subscript "N" for NHANES); the objective health H^O_W was calculated using WHAS as its own reference group (subscript "W" for WHAS). Years of education is treated as a continuous variable; the mean is slightly above secondary school (9 years).

3.3. Results

Table II and Table III show the coefficients from the ordered logistic regressions for the whole sample and various subsamples; the results in Table II use H^O_N as the objective health measure and those in Table III use H^O_W; in both cases all 41 biomarkers are used to construct the objective health measure.

As expected, there is always a positive coefficient for the linear part of H^O. The coefficient of the quadratic part is negative, indicating a concave relationship. The coefficient of the linear part is positive even in case we omit the quadratic part. Note that the coefficients of the control variables are not directly interpretable because they could include 1) an effect on health due to the fact that H^O may not perfectly control for true health and 2) the manners in reporting one's health that change with socio-economic status (Lindeboom and Van Doorslaer, 2004). In addition, the regression is cross-sectional and there is certainly a doubt for endogeneity due to omitted variables, which we admit as a limitation. Meanwhile, the study group WHAS can be viewed as a relatively homogeneous subsample obtained after stratification, which could help reduce the problem of omitted variable.

The four panels in Figure 1 show the changes in odds ratios associated with H^O_N and $(H^O_N)^2$ (first row) and those associated with H^O_W and $(H^O_W)^2$ (second row). We choose to report the odds ratios because the magnitude of the coefficients from an ordered logistic regression are not directly interpretable; another way is to report the partial effect because it is possible to estimate quantitative effects of the covariates on the probabilities of reporting each health category, for each individual (Jones et al., 2013, chapter 4.3); however, in our case there would be too many results if we report different probabilities with respect to different categories; so we use instead the odds ratio defined as the exponential $\exp(\beta)$ of a coefficient β. The odds ratio is the odds for H^S to be in a category that is greater than j versus less than or equal to j (McCullagh, 1980). Hence in our case there are five categories

of H^S and the ordered logit regression can be viewed as equivalent to a proportional odds model (POM) of four logits, one for each of the following cut-points: (A) 1 vs. 2,3,4,5; (B) 1,2 vs. 3,4,5; (C) 1,2,3 vs. 4,5 and (D) 1,2,3,4 vs. 5; the various logits in this POM are constrained to be equal and for this reason the POM is also referred to as the constrained cumulative logit model (Hosmer and Lemeshow, 2000). As a result, there is only one odds ratio associated with each regressor; the odds ratio bigger than 1 corresponds to a positive original coefficient and a value smaller than 1 corresponds to a negative original coefficient.

The number of biomarkers used to construct H_N^O and H_W^O varied from 2 to 39. In each of the 38 cases (2 to 39 biomarkers)[1] we randomly chose 200 biomarker combinations from the C_{41}^l ($l = 2, \cdots, 39$) possibilities. The red curve represents the median of the 200 replications and the blue curves 2.5% and 97.5% quantiles. From the odds ratios associated with the linear part of objective health (the two panels on the left) we observe an increasing signal with the number of biomarkers; the increase slows down but has not stopped even with 39 biomarkers. Nevertheless, it is clearly possible to have a signal with fewer biomarkers, although the signal could be weaker. For the quadratic terms, there also appears to be a negative and increasingly stable signal with increasing biomarker number.

The results we obtain here are, of course, sample specific. Validation of a health measure cannot be done with a simple exercise and the validation in this section is rather illustrative. Studies using data from different populations/countries are necessary to further validate our health measure. We also refer interested readers to the biomedical empirical results mentioned in section 2.2.

4. Illustrative applications

4.1. Distribution of health

We first illustrate our health measure with the estimation of health distribution. General health may be heavy-tailed in a way similar to income or cost data. Our health measure describes how bad one's health is, and a long tail in the distribution corresponds to individuals in very bad health. As reviewed by Jones et al. (2015) in their study of healthcare costs, the mean of a distribution is clearly an important feature, nonetheless, it is generally not the only aspect of interest to policymakers (Vanness and Mullahy, 2012). There is a growing literature in econometrics that has been developping techniques to model the entire distribution, thus 'going beyond the mean' (Fortin et al., 2011). Analysis based solely on the mean misses out potentially important information in other parts of the distribution (Bitler

[1] The number of included biomarkers is from 2 to 39 because $C_{41}^2 = C_{41}^{39} = 820$ allows us to have 200 replications but $C_{41}^{40} = 41$ does not.

et al., 2006). In health economics, there is a particular emphasis on identifying individuals or characteristics of individuals that lead to very large costs, and there is a demand for empirical strategies to 'target the high-end parameters of particular interest' including tail probabilities (Mullahy, 2009). The existence of heavy tail has important influence on model selection. For example, cost data are often modeled with generalized linear models (GLMs) (Blough et al., 1999); while the GLMs have attractive properties for researchers concerned only with the conditional mean of the distribution, they have been found to perform badly with heavy-tailed data (Manning and Mullahy, 2001). For instance, whatever distribution is adopted, the skewness is directly proportional to the coefficient of variation, and the kurtosis is linearly related to the square of the coefficient of variation (Holly, 2009). Our health measure provides a way to identify tails.

We estimate the distribution of health for the individuals in NHANES and WHAS. We divide the respondents in NHANES into three age groups namely 20-40, 41-64 and 65-85 in order to represent "young," "mid-aged" and "old". The choice of these thresholds are arbitrary but the results are not sensitive to these particular choices. In order to compare with WHAS, we only include females in NHANES; nonetheless, the results are qualitatively similar if the males in NHANES are included. We estimate as well the distribution putting all respondents in NHANES together (the group "all"). Concerning choice of reference group, we first compute D_M for the respondents in each group, each time using the group as its own reference population (hence five different centroids for five different groups). In order to compare, we also compute D_M taking the young group (20-40) from NHANES as reference for all groups (hence the same centroid for the five groups).

Table IV shows minimums, maximums and different percentiles in the health distributions; the first part of the table shows the results where D_M is computed using the young reference group while the second part shows the results with D_M computed using each group as its own reference group. Figure 2 and Figure 3 show the distributions visually. There is a long tail in the distribution of health. When referenced by the young group, the mean of health is higher for older group; the highest is for WHAS which includes women aged between 65 and 100 comparing to the aged group from NHANES where the oldest age is 85. In contrast, we do not see clear difference in the mean level of health when each group is referenced by itself. Hence, the health values are not directly comparable across different groups when D_M is computed for each group using its own centroid, although the distribution is clear within the group. We do not go into more detailed discussions here as the application is demonstrative; the important thing is that a continuous measure of health makes it easier to spot the individuals on the tail and choose proper statistical tools to deal with the issue.

4.2. Concentration index

Next, we illustrate the health measure through the estimation of health concentration index with respect to income. The concentration index (Kakwani 1977, 1980) quantifies the degree of the inequality (Kakwani et al., 1997; Wagstaff et al., 1989) and has become the standard measure to quantify income-related inequalities in health economics (Wagstaff and Van Doorslaer, 2000). Here we give a short summary of some results and issues while a detailed review can be found in O'Donnell (2008, Chapter 8). A convenient formula for the concentration index defines it in terms of the covariance between the health variable and the fractional rank in the living standards distribution (Jenkins 1988; Kakwani 1980; Lerman and Yitzhaki 1989),

$$C = \frac{2}{\mu} cov(h_i, r_i) \tag{4.1}$$

where μ is the mean of the health variable. Standard error of a concentration index can be computed with the formula derived in Kakwani et al. (1997) with

$$var(\widehat{C}) = \frac{1}{n}[\frac{1}{n}\sum_{i=1}^{n} a_i^2 - (1+C)^2], \tag{4.2}$$

where

$$a_i = \frac{h_i}{\mu}(2r_i - 1 - C) + 2 - q_{i-1} - q_i, \tag{4.3}$$

and

$$q_i = \frac{1}{\mu n}\sum_{j=1}^{i} h_j \tag{4.4}$$

is the ordinate of the concentration curve with $q_0 = 0$.

Strictly speaking, the concentration index is an appropriate measure of socioeconomic-related health inequality when health is measured on a ratio scale with a true zero (which in our case is the state of "ideal health"). The values of the health variable should be nonnegative. The concentration index cannot be computed directly from categorical data (e.g. self-reported health). Although the ordinal data can be transformed into some cardinal measure which allows the computation of the concentration index, (Van Doorslaer and Jones 2003; Wagstaff and Van Doorslaer 1994), the value of the index will depend on the transformation chosen (Erreygers 2005). In cross-country comparisons, even if all countries adopt the same transformation, their ranking by the concentration index could be sensitive to differences in the means of health that are used in the transformation. A partial solution to this problem would be to dichotomize the categorical health measure; unfortunately, this introduces another problem. Wagstaff (2005) has demonstrated that the bounds of the

concentration index for a dichotomous variable are not –1 and 1 but depend on the mean of the variable. Our health measure has a truly continuous support; hence we can largely avoid these problems. Additionally, many existing continuous measures of health are constructed (partly or entirely) from self-reported information on health. As is well documented, there is often a reporting bias in self-reported health. Therefore, for example, when we see bad health is more concentrated among poor people, we may not be sure to what extent the result is due to self-perception rather than true health. By using biomarkers, in contrast, we can guarantee the objectivity of the results.

We assign the individuals to the same age groups as we did in the application to estimate the distributions of health. For NHANES, we use the poverty income ratio (PIR) as the income variable. The variable was calculated by dividing family income by the poverty guidelines, specific to family size, as well as the appropriate year and state. If family income was reported as a range value, the midpoint of the range was used to compute the variable. Values at or above 5.00 were coded as 5.00 or more because of disclosure concerns. As for WHAS, the variable income is obtained from respondents' answers to a question on their total household income before taxes from all sources, including Social Security, retirement income, job earnings, public assistance, help from relatives, rent from property, and any other income (Guralnik et al., 1995). The value ranges from 0 to 300000. Some respondents refused to answer the question and this is a main reason why there are more missing values in income than in other variables (and this is the reason why we did not include it previously in the validation study, although inclusion does not change the results qualitatively).

Table V shows the results. To facilitate comparison we also provide information on sample size as well as mean and standard error of health and family income for each group. We see that the concentration index is negative in most cases. Because the Mahalanobis distance measures how bad one's health is, a negative concentration index means bad health is more concentrated in poor people. In the young group there is evidence of inequality and this inequality exacerbates in the mid-aged group. This is perhaps because between the age of 40 and 64 income difference widens from one person to another and in the meantime people start to have more health problems during this period, both leading to bigger inequality in health with respect to income. As for the two aged groups from NHANES and WHAS, no result is significant at 5% level. When each group is referenced by itself, the results follow the same line. We infer that inequality might still exist in the aged group, but if it exists, it is smaller than in the young and mid-aged groups. Generally speaking, it is often observed that health gaps narrow after the age of 65, which could be due to differential survival and safety net programs such as Medicare and Social Security that begin at that age (Evans et al., 2012).

5. Conclusion

The main objective of this paper is to propose a continuous and objective health measure. Despite the enormously rich information on objective health, biomarkers have not to our knowledge been widely explored, and most existing works used a few single biomarkers as health indicators. Because changes in single biomarkers are complex and there could be billions of different ways for human health to "go wrong", we propose a measure that describes the exact biological state of each individual in the biomarker space - a method being able to reflect any possible status of the biomarkers as a whole. While there remain issues such as the search for a better centroid, our distance-based framework provides a novel way to measure general health. Human health is high-dimensional, and in this sense our health measure can be viewed as a projection of the true health onto a subspace of available biomarkers.

The choice of health measure in empirical studies obviously depends on the objectives of any specific study. For instance, if the objective is to predict costs, an good measure of general health might not be the best predictor; bone fragility might be a small component of general health in a sense, but a disproportionately large contributor to costs. For this reason we do *not* claim that our health measure is optimal; nonetheless, our measure reflects advances in biological research and technically it also provides a convenient tool for health economists. Finally, our health measure could be further combined with discrete health indicators, which we leave for future research.

References

Barnwell-Ménard J-L, Li Q, Cohen AA. 2015. Effects of categorization method, regression type, and variable distribution on the inflation of Type-I error rate when categorizing a confounding variable. *Statistics in Medicine* 34(6): 936-949.

Bentona D, Donohoe RT. 1999. The effects of nutrients on mood. *Public Health Nutrition*, 2: 403-409.

Blough DK, Madden CW, Hornbrook MC. 1999. Modeling risk using generalized linear models. *Journal of Health Economics* 18(2): 153–171.

Bitler MP, Gelbach JB, Hoynes HW. 2006. What mean impacts miss: distributional effects of welfare reform experiments. *American Economic Review* 96(4): 988–1012.

Bound J. 1991. Self-reported versus objective measures of health in retirement models. *Journal of Human Resources* 26: 106–138.

Burkhauser RV, Cawley J. 2006. The importance of objective health measures in predicting early receipt of social security benefits: the case of fatness. *Working Paper WP 2006-148*, Michigan Retirement Research Center, University of Michigan, Ann Arbor, MI.

Cohen AA, Martin LB, Wingfield JC, McWilliams SR, Dunne JA. 2012. Physiological regulatory networks: ecological roles and evolutionary constraints. *Trends in Ecology & Evolution*, 27: 428-435.

Cohen AA, Milot E, Yong J, Seplaki CL, Fülöp T, Bandeen-Roche K, Fried LP. 2013. A novel statistical approach shows evidence for multi-system physiological dysregulation during aging. *Mechanisms of Ageing and Development*, 134: 110-117.

Cohen AA, Milot E, Li Q, Legault V, Fried LP, Ferrucci L. 2014. Cross-population validation of statistical distance as a measure of physiological dysregulation during aging. *Experimental Gerontology*, 57: 203–210.

Cohen AA, Li Q, Milot E, Leroux M, Faucher S, Morissette-Thomas V, Legault V, Fried LP, Ferrucci L. 2015. Statistical distance measures of physiological dysregulation are largely robust to variations in calculation method. *PLoS One*, 10.1371/journal.pone.0122541.

Cohen AA. 2016. Complex systems dynamics in aging: new evidence, continuing questions. *Biogerontology*, 17(1): 205-20.

Crimins E, Faul J, Kim JK, Guyer H, Langa K, Ofstedal MB, Sonnga A, Wallace R, Weir D. 2013. Documentation of Biomarkers in the 2006 and 2008 Health and Retirement Study. *HRS Documentation Report*.

Ekström J. 2011. Mahalanobis' distance beyond normal distributions. *UCLA Statistics Preprints*, 624.

Erreygers G. 2005. Beyond the Health Concentration Index: An Atkinson Alternative to the Measurement of Socioeconomic Inequality of Health. University of Antwerp. Antwerp, Belgium.

Erickson P. 1998. Evaluation of a population-based measure of quality of life: the health and activity limitation index (HALex). *Quality of Life Research*, 7: 101–114.

Evans W, Wolfe B, Adler N. 2012. The income-health gradient. Article based on the chapter one in Wolfe B, Evans W, Seeman TE, (eds) *The Biological Consequences of Socioeconomic Inequalities*, CUP Services .

Feeny, D., Furlong, W., Boyle, M., Torrance, G., 1995. Multi-attribute health status classification systems: health utilities index. *Pharmacoeconomics* 7 (6), 490–502.

Fortin N, Lemieux T, Firpo S. 2011. Decomposition methods in economics, In Part A of *Handbook of Labor Economics*, Ashenfelter O, Card D (eds). Vol. 4. Elsevier: Amsterdam; 1–102.

Fried LP, Tangen CM, Walston J, Newman AB, Hirsch C, Gottdiener J, et al. 2001.

Frailty in older adults: evidence for a phenotype. *Journal of Gerontology*, A Biol Sci Med Sci, 56: M146-M156.

Fried LP, Hadley EC, Walston JD, Newman AB, Guralnik JM, Studenski S, Harris TB, Ershler WB, Ferrucci L. 2005. From bedside to bench: Research agenda for frailty. *Science of Aging Knowledge Environment*, 31: 24.

Guralnik JM, Fried LP, Simonsick EM, Kasper JD, Lafferty ME. 1995. *The Women's Health and Aging Study: Health and Social Characteristics of Older Women with Disability.* National Institute on Aging, NIH Pub. No. 95-4009.

Holly A. 2009. Modeling risk using fourth order pseudo maximum likelihood methods, Institute of Health Economics and Management (IEMS), University of Lausanne: Switzerland.

Hosmer DW, Lemeshow S. 2000. *Applied Logistic Regression* (2nd ed.), New York: Wiley.

Jenkins S. 1988. Calculating income distribution indices from microdata. *National Tax Journal*, 61: 139–42.

Jones AM, Rice N, d'Uva TB, Balia S. 2013. *Applied Health Economics*, 2nd Edition.

Jones AM, Lomas J, Rice N. 2015. Healthcare cost regressions: going beyond the mean to estimate the full distribution. *Health Economics*, 10.1002/hec.3178

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

Jürges H, Kruk E, Reinhold S. 2013. The effect of compulsory schooling on health—evidence from biomarkers. *Journal of Population Economics*, 26: 645-672.

Kakwani NC. 1977. Measurement of tax progressivity: an International comparison. *Economic Journal,* 87(345): 71–80.

Kakwani NC. 1980. *Income Inequality and Poverty: Methods of Estimation and Policy Applications.* New York: Oxford University Press.

Kakwani NC, Wagstaff A, Van Doorslaer E. 1997. *Socioeconomic Inequalities in Health: Measurement, Computation and Statistical Inference.* Journal of Econometrics, 77(1): 87–104.

Lerman RI, Yitzhaki S. 1989. Improving the accuracy of estimates of Gini coefficients. *Journal of Econometrics*, 42(1): 43–47.

Li Q. 2014. Identifiability of mean-reverting measurement error with instrumental variable. *Statistica Neerlandica*, 68: 118–129.

Li Q, Wang S, Milot E, Bergeron P, Ferrucci L, Fried L, Cohen A. 2015. Homeostatic dysregulation proceeds in parallel in multiple physiological systems. *Aging Cell*, 14(6), 1103–1112.

Lindeboom M, Van Doorslaer E. 2004. Cut-point shift and index shift in self-reported health. *Journal of Health Economics,* 23, 1083–1099.

Mahalanobis PC. 1936. Mahalanobis distance. *Proceedings National Institute of Science of India*, 49: 234–256.

Manning WG, Mullahy J. 2001. Estimating log models: to transform or not to transform? *Journal of Health Economics* 20(4): 461–494.

McCullagh P. 1980. Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Series B*, 42(2): 109-142.

Meijer E, Kapteyn A and Andreyeva T. 2011. Internationally comparable health indices. *Health Economics*, 20: 600–619.

Milot E, Cohen AA, Vézina F, Buehler DM, Matson KD, Piersma T. 2014a. A novel integrative method for measuring body condition in ecological studies based on physiological dysregulation. *Methods in Ecology and Evolution*, 5: 146-155.

Milot E, Morissette-Thomas V, Li Q, Fried L, Ferrucci L, Cohen A. 2014b. Trajectories of physiological dysregulation predict mortality and health outcomes in a consistent manner across three populations, *Mechanisms of Ageing and Development*, 141–142: 56–63.

Mullahy J. 2009. Econometric modeling of health care costs and expenditures: a survey of analytical issues and related policy considerations. *Medical Care* 47: S104–S108.

Murray CJL, Salomon JA, Mathers CD, Lopez AD (eds). 2002. *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications.* World Health Organization: Geneva, Switzerland.

O'Donnell O, Van Doorslaer E, Wagstaff A, Lindelow M. 2008. *Analyzing Health Equity Using Household Survey Data: a Guide to Techniques and Their Implementation.* World Bank.

Powdthavee N. 2010. Does education reduce the risk of hypertension? Estimating the biomarker effect of compulsory schooling in England. *Journal of Human Capital*, 4(2): 173–202.

SHARE website. 2014. http://www.share-project.org/methodological-research/biomarker-project.html

Sondik E. 2002. Summary measures of population health: applications and issues in the United States. In *Summary Measures of Population Health: Concepts, Ethics, Measurement and Applications*, Murray CJL, Salomon JA, Mathers CD, Lopez AD (eds). World Health Organization: Geneva, Switzerland; 75–81.

Stewart ST, Woodward RM, Rosen AB, Cutler DM. 2005. A proposed method for monitoring U.S. population health: linking symptoms, impairments, and health ratings. *Working Paper 11358*, National Bureau of Economic Research, Cambridge, MA, revised

2007.

Van Doorslaer E, Jones AM. 2003. Inequalities in self-reported health: validation of a new approach to measurement. *Journal of Health Economics*, 22(1): 61-87.

Vanness DJ, Mullahy J. 2012. Perspectives on mean-based evaluation of health care, In *The Elgar Companion to Health Economics* (2nd edn), Jones AM (ed). Elgar Original Reference: Cheltenham; 563–575.

Wagstaff A, Van Doorslaer E, Paci P. 1989. Equity in the finance and delivery of health care: some tentative cross-country comparisons. *Oxford Review of Economic Policy*, 5(1): 89–112.

Wagstaff A, Van Doorslaer E. 1994. measuring inequalities in health in the presence of multiple-category morbidity indicators. *Health Economics*, 3: 281–91.

Wagstaff A, Van Doorslaer E. 2000. Equity in health care finance and delivery. In: Culyer A, Newhouse J (eds), *Handbook of Health Economics*. North Holland: Amsterdam, 1804–1862.

Wagstaff A. 2005. The bounds of the concentration index when the variable of interest Is binary, with an application to immunization inequality. *Health Economics*, 14(4): 429–32.

Wansbeek T, Meijer E. 2000. *Measurement Error and Latent Variables in Econometrics*. Amsterdam: Elsevier.

WHO. 2000. Obesity: Preventing and managing the global epidemic, report of a WHO consultation. *WHO Technical Report Series 894*. Geneva: World Health Organization.

World Bank. 1993. *World Development Report: Investing in Health*. Oxford University Press: New York.

Table I: Summary statistics

	Mean	Sd	Min	Max
$H^S = 1$	0.08	0.275	0	1
$H^S = 2$	0.21	0.410	0	1
$H^S = 3$	0.35	0.477	0	1
$H^S = 4$	0.26	0.436	0	1
$H^S = 5$	0.10	0.303	0	1
D_M^N	10.77	6.033	6.37	191.76
D_M^W	6.20	1.589	2.61	31.33
Disabled	0.63	0.484	0	1
White	0.75	0.432	0	1
Education	10.72	3.807	1	18
Married	0.28	0.451	0	1

H^S: 1 = Excellent, 2 = Very good, 3 = Good, 4 = Fair, 5 = Poor

D_M^N: The Mahalanobis distance from the centroid of the NHANES reference group

D_M^W: The Mahalanobis distance from the centroid of WHAS as its own reference group

The scripts "N" and "W" indicate NHANES as reference group and WHAS as reference group, respectively.

Table II: Results from ordered logistic regressions - young women from NHANES as reference group

	All	Disabled	Non-disabled	White	Non-White	Low Edu.	High Edu.	Married	Unmarried
H_N^O	0.24 *	0.22 *	0.38 *	0.27 *	0.19	0.32 *	0.21 *	0.58 *	0.14
	(0.071)	(0.080)	(0.151)	(0.085)	(0.141)	(0.114)	(0.094)	(0.144)	(0.082)
$(H_N^O)^2$	−0.02*	−0.12*	−0.02	−0.02*	−0.03	−0.07	−0.01	−0.04	−0.01
	(0.008)	(0.049)	(0.012)	(0.009)	(0.066)	(0.054)	(0.009)	(0.119)	(0.009)
Disabled	1.62 *	-	-	1.71 *	1.36 *	1.15 *	1.90 *	2.46 *	1.35 *
	(0.139)			(0.161)	(0.286)	(0.242)	(0.177)	(0.294)	(0.162)
White	−0.56*	−0.57*	−0.65*	-	-	−0.70*	−0.37	−0.40	−0.61*
	(0.141)	(0.168)	(0.269)			(0.203)	(0.198)	(0.316)	(0.158)
Education	−0.09*	−0.07*	−0.15*	−0.09*	−0.09*	−0.04	−0.15*	−0.15*	−0.08*
	(0.017)	(0.021)	(0.031)	(0.020)	(0.034)	(0.051)	(0.032)	(0.038)	(0.019)
Married	0.33 *	0.55 *	0.05	0.35 *	0.25	0.55 *	0.24	-	-
	(0.130)	(0.175)	(0.199)	(0.145)	(0.299)	(0.253)	(0.155)		
1|2	−3.13*	−4.35*	−4.13*	−2.53*	−3.26*	−3.02*	−3.80*	−4.07*	−3.04*
	(0.266)	(0.313)	(0.460)	(0.295)	(0.527)	(0.499)	(0.489)	(0.607)	(0.301)
2|3	−1.39*	−2.64*	−2.36*	−0.81*	−1.42*	−1.28*	−2.04*	−2.11*	−1.38*
	(0.246)	(0.253)	(0.428)	(0.275)	(0.444)	(0.451)	(0.469)	(0.569)	(0.276)
3|4	0.41	−1.08*	−0.01	1.07 *	0.12	0.30	−0.05	0.15	0.29
	(0.244)	(0.234)	(0.412)	(0.278)	(0.435)	(0.445)	(0.464)	(0.568)	(0.273)
4|5	2.23 *	0.72 *	2.25 *	2.92 *	1.90 *	1.99 *	1.96 *	2.20 *	2.08 *
	(0.254)	(0.234)	(0.576)	(0.299)	(0.450)	(0.456)	(0.481)	(0.577)	(0.285)

Note: Each column denotes a different regression. The dependent variable is self-reported health. The subscript "N" in H_N^O means that NHANES was used as the reference group. Standard errors are in parentheses. Coefficients with a star are significant at 5% level.

Table III: Results from ordered logistic regressions - WHAS as its own reference group

	All	Disabled	Non-disabled	White	Non-White	Low Edu.	High Edu.	Married	Unmarried
H_W^O	0.20 *	0.16	0.33 *	0.22 *	0.11	0.29 *	0.16	0.44 *	0.13
	(0.068)	(0.084)	(0.118)	(0.075)	(0.174)	(0.113)	(0.086)	(0.134)	(0.079)
$(H_W^O)^2$	−0.04*	−0.07*	−0.04	−0.05*	−0.01	−0.07	−0.03	−0.06	−0.03
	(0.020)	(0.034)	(0.026)	(0.024)	(0.047)	(0.038)	(0.024)	(0.070)	(0.021)
Disabled	1.48 *	-	-	1.55 *	1.27 *	1.03 *	1.76 *	2.04 *	1.28 *
	(0.136)			(0.156)	(0.279)	(0.240)	(0.171)	(0.273)	(0.159)
White	−0.55	−0.55	−0.62	-	-	−0.65	−0.38	−0.48	−0.59
	(0.142)	(0.169)	(0.269)			(0.206)	(0.198)	(0.313)	(0.160)
Education	−0.09	−0.07	−0.14	−0.09	−0.09	−0.02	−0.15	−0.16	−0.07
	(0.017)	(0.021)	(0.031)	(0.020)	(0.034)	(0.051)	(0.032)	(0.038)	(0.019)
Married	0.32	0.51	0.06	0.33	0.28	0.55	0.22	-	-
	(0.130)	(0.175)	(0.199)	(0.145)	(0.301)	(0.254)	(0.155)		
1|2	−3.24	−4.29	−4.23	−2.70	−3.27	−2.96	−3.93	−4.51	−3.10
	(0.266)	(0.313)	(0.461)	(0.298)	(0.530)	(0.497)	(0.487)	(0.607)	(0.301)
2|3	−1.50	−2.59	−2.45	−0.97	−1.43	−1.22	−2.16	−2.55	−1.43
	(0.245)	(0.254)	(0.428)	(0.277)	(0.448)	(0.449)	(0.466)	(0.563)	(0.276)
3|4	0.29	−1.04	−0.09	0.90	0.10	0.34	−0.17	−0.33	0.24
	(0.242)	(0.236)	(0.410)	(0.279)	(0.439)	(0.445)	(0.460)	(0.555)	(0.272)
4|5	2.11	0.75	2.16	2.75	1.88	2.03	1.83	1.66	2.03
	(0.252)	(0.236)	(0.575)	(0.299)	(0.453)	(0.456)	(0.477)	(0.556)	(0.284)

Note: Each column denotes a different regression. The dependent variable is self-reported health. The subscript "W" in H_W^O means that WHAS was used as the reference group for itself. Standard errors are in parentheses. Coefficients with a star are significant at 5% level.

Table IV: Quantiles of health - different age groups

Health calculated with young women (20-40) in NHANES as reference group					
Group	NHANES All	NHANES 20-40	NHANES 41-64	NHANES 65-85	WHAS 65-100
Min.	3.87	3.82	4.29	5.22	6.37
5%	5.00	4.75	5.21	6.31	8.09
25%	5.83	5.44	6.08	7.45	9.48
Mean	6.58	6.02	6.89	8.51	10.58
75%	7.68	6.77	7.95	9.85	11.60
95%	10.43	8.58	10.64	12.03	13.91
Max.	20.53	17.11	20.53	15.93	191.76

Health calculated with each group as its own reference group					
Group	NHANES All	NHANES 20-40	NHANES 41-64	NHANES 65-85	WHAS 65-100
Min.	3.61	3.82	3.83	4.01	2.58
5%	4.65	4.75	4.62	4.81	4.64
25%	5.36	5.44	5.35	5.54	5.37
Mean	5.98	6.02	6.02	6.12	5.98
75%	6.80	6.77	6.78	6.88	6.67
95%	8.67	8.58	8.41	8.42	8.54
Max.	16.95	17.11	15.09	10.65	25.03

Table V: Concentration Index - different age groups

Health calculated with young women (20-40) in NHANES as reference group					
Group	NHANES All	NHANES 20-40	NHANES 41-64	NHANES 65-85	WHAS 65-100
Observations	2076	1221	605	250	635
Health (Mean & S.D.)	7.001 (1.7894)	6.265 (1.3099)	7.287 (1.9024)	8.7635 (1.8700)	10.983 (7.4452)
Family income (Mean & S.D.)	2.605 (1.6380)	2.454 (1.6211)	3.017 (1.6570)	2.350 (1.4946)	19449.39 (23993.49)
Concentration Index	-0.015^{*} (0.0034)	-0.015^{*} (0.0038)	-0.025^{*} (0.0064)	-0.011 (0.0104)	0.026 (0.0226)

Health calculated with each group as its own reference group					
Group	NHANES All	NHANES 20-40	NHANES 41-64	NHANES 65-85	WHAS 65-100
Observations	2076	1221	605	250	635
Health (Mean & S.D.)	6.249 (1.3899)	6.265 (1.3099)	6.240 (1.4148)	6.296 (1.0953)	6.208 (1.5506)
Family income (Mean & S.D.)	2.605 (1.6380)	2.454 (1.6211)	3.017 (1.6570)	2.350 (1.4946)	19449.39 (23993.49)
Concentration Index	-0.016^{*} (0.0030)	-0.015^{*} (0.0038)	-0.023^{*} (0.0060)	-0.013 (0.0093)	-0.009 (0.0065)

Note: A star indicates that the Concentration Index is significant at 5% level.

Figure 1: Odds ratios related to H_N^O and H_W^O ("N" means referenced by NHANES; "W" means referenced by WHAS itself) and the quadratic terms of H_N^O and H_W^O. The vertical axis shows the odds ratio (exponential of the coefficient from the regression); the horizontal axis shows the number of biomarkers included in the computation of the health measure. For each number of biomarkers from 2 to 39, we report the median of 200 replications (red curve) and the 2.5% and 97.5% quantiles (blue curves).

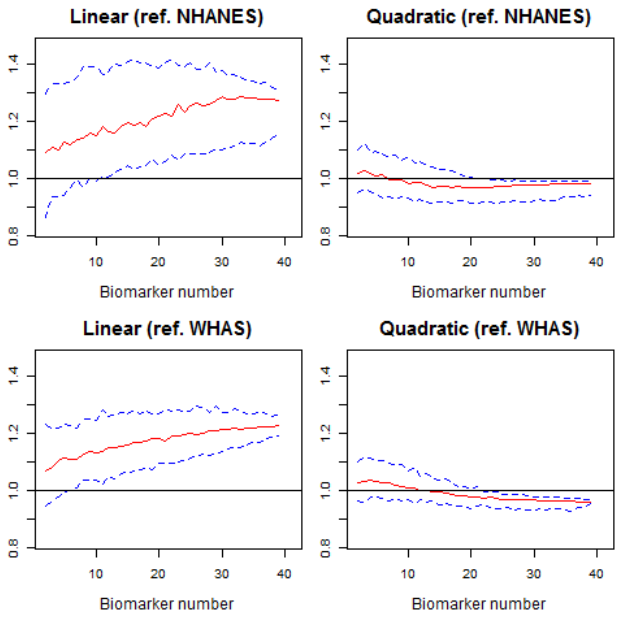


Figure 2: Health distributions of different age groups. The dotted vertical lines show the medians. The women aged 20-40 in NHANES are used as reference group.

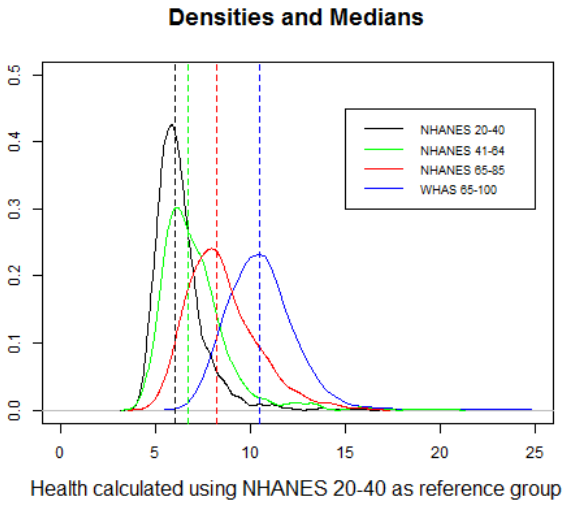
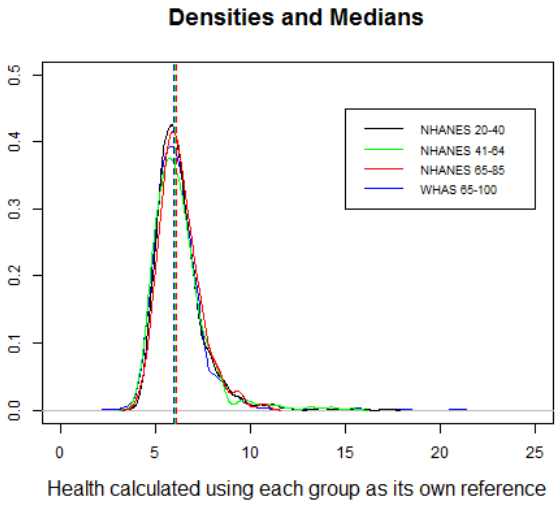


Figure 3: Health distributions of different age groups. The dotted vertical lines show the medians. Each age group is used as its own reference group.



Appendix: Effect of changes in single biomarkers

Here we derive mathematically the effect of changes in single biomarkers on the Mahalanobis distance. This can be seen with some matrix algebra (see Wansbeek and Meijer, 2000, Appendix A). By definition the Mahalanobis distance of respondent i to the centroid μ is

$$\begin{aligned} D_M(x_i,\mu) &= \sqrt{(x_i-\mu)'\Sigma^{-1}(x_i-\mu)} \\ &= \sqrt{tr\Sigma^{-1}D} \end{aligned} \tag{5.1}$$

where $D := (x_i-\mu)(x_i-\mu)^{'}$ is a square matrix having the same dimension as Σ. Letting

$$\sigma = vec\Sigma \tag{5.2}$$

be the vectorized version of Σ, it follows from the chain rule that

$$\begin{aligned} \frac{\partial D_M(x_i,\mu)}{\partial\sigma} &= \frac{1}{2}(tr\Sigma^{-1}D)^{-\frac{1}{2}}\frac{\partial(tr\Sigma^{-1}D)}{\partial\sigma} \\ &= -\frac{vec(\Sigma^{-1}D\Sigma^{-1})}{2\sqrt{tr\Sigma^{-1}D}}. \end{aligned} \tag{5.3}$$

Letting θ be a parameter in Σ, i.e. θ is a variance when diagonal or a covariance when off-diagonal, and

$$\Sigma_\theta := \frac{\partial\Sigma}{\partial\theta} \tag{5.4}$$

be the (zero-one) matrix obtained from deriving Σ with respect to θ, it follows straightforwardly upon applying the chain rule once more, that

$$\frac{\partial D_M(x_i,\mu)}{\partial\sigma'}\frac{\partial\sigma}{\partial\theta} = \frac{(vec(\Sigma^{-1}D\Sigma^{-1}))^{'}vec\Sigma_\theta}{-2\sqrt{tr\Sigma^{-1}D}}. \tag{5.5}$$

The denominator $-2\sqrt{tr\Sigma^{-1}D}$ is clearly negative and the *diagonal* elements of the square matrix $\Sigma^{-1}D\Sigma^{-1}$ are positive; hence when θ is a variance the vector $vec\Sigma_\theta$ has an element equal to one and all other elements equal to zero. As a result the product of the two vectors in the numerator of (5.5) is positive making the whole derivative negative. This means the Mahalanobis distance decreases when a variance in Σ increases; and this is why the Mahalanobis is sensitive to scales of the variables. In our case we normalized the (logarithm of) biomarkers so the diagonal elements in Σ are all equal to one. When θ is a covariance, the vector $vec\Sigma_\theta$ has two elements equal to one and all other elements equal

to zero; the two elements capture the corresponding off-diagonal elements that appear in symmetric positions in $\Sigma^{-1}D\Sigma^{-1}$. Therefore, the vector product in the numerator of (5.5) keeps the same sign as the corresponding off-diagonal element in $\Sigma^{-1}D\Sigma^{-1}$ but doubles its absolute value. Empirically this means the change in the Mahalanobis distance with respect to a change in a certain covariance cannot be predicted *a priori* without inspecting other elements in Σ^{-1}, i.e. without knowing the whole structure of Σ.