Comparative analyses of more than 50 household surveys on health status (abridged version)*

by: Ritu Sadana[1], Colin D. Mathers[1], Alan D. Lopez[1], Christopher J.L. Murray[1], Kim M. Iburg[2]

For presentation during Session 5 of the

Joint UN-ECE and WHO Expert Meeting on Health Status Measurements

hosted by Statistics Canada

Ottawa, 23-25 October 2000

Global Programme on Evidence for Health Policy
World Health Organization
Geneva

* For the full version of this paper, see www.who.int/evidence, discussion papers, #15

[1] Global Program on Evidence for Health Policy, World Health Organization, Geneva, Switzerland
[2] National Institute of Public Health, Copenhagen, Denmark

Introduction

In this year's *World Health Report*[1] the primary summary measure of population health used is Disability-Adjusted Life Expectancy, or DALE.[2-3] DALE measures the equivalent number of years of life expected to be lived in full health, or *healthy life expectancy*. As a summary measure of population health, DALE combines information on the impact of premature death and of disability and other health outcomes. In order to calculate DALE the population based prevalence of different severity levels of disability and other aspects of health are required for each country. In this paper we report on our evaluation of whether existing data from nationally representative household interview surveys conducted in a large number of countries may be used towards estimating the distribution and levels of severity of health at the population level. Two main challenges exist that prevent the meaningful comparison of household interview data across populations. The first concerns limitations in the cross-population comparability of data as surveys with only self-reported data lack external criteria to calibrate responses across countries. The second is that existing nationally representative surveys use different questions and response scales to assess health. Through out this paper, we use the term *health* or *health status* in the broadest sense to include illness, impairments, disability and other states of health spanning from the worst health state to full health for an individual alive. This is in contrast to mortality and its risk factors or causes of death.

Current empirical approaches to assess health within household surveys are based on the model depicted in Figure 1. A population's underlying *true level of health* may vary on a variety of domains, such as physical or cognitive functioning, or mental health, among others. The true level of health may be reflected by *tested health* (measured through laboratory or functional tests), *observed health* (based on professionals' clinical assessments or other ratings) and *perceived health* (based on individual's knowledge and beliefs) in related domains. No gold standard measurement technique exists to assess all aspects of health or to assess what people perceive. *Self-reported health* is what individuals report within a survey to a lay interviewer. This may differ from what an individual perceives, particularly when specific incentives or sanctions influence reporting behavior. Perceptions may differ from true health due to different definitions of health and well-being, different expectations for health and different cognitive processes. Such differences reflect variations in cultural and gender norms, knowledge and information, and other factors that shape people's perceptions.[4-8] What is observed, tested or reported is not always consistent given these variations in perceptions and differences in the sensitivity and specificity of tests, beyond measurement error.[9-13]

Differences in perceptions and subsequent reporting biases may specifically compromise the cross-population comparability of data collected within household surveys. To achieve high cross-population comparability based on self-reported data (1) respondents should interpret identically questions and response scales – such as ordinal scales with end-points of *excellent* and *very poor* and cut-points of *very good, good* and *fair*, (2) responses provided within surveys should be unbiased in relation to the true level of health and (3) for the same true level of health, responses should be identical irrespective of socio-economic or demographic characteristics.[14-16] Not surprisingly, these criteria can not be met in practice.[17-20]

More surprisingly is that users of household interview surveys often ignore these limitations and interpret self-reported data at face value. Results reported from the European Community Household Panel survey, relying on self-reported data conducted in 12 countries using identical methodologies, are illustrative. Figure 2 shows the proportion of the population in each country reporting *good* and *very good* health, in response to the question *how is your health in general?* The accompanying article notes that although questions on self-reported health status "may be sensitive to differences in language and 'culture' between Member States, it seems worth noting that for instance 'very good' health is reported by as much as 53% of the Danish and as little as 8% of the Portuguese population."[21] We believe it is unlikely that differences in the true level of health, translations or measurement error, account for such large variations between Denmark and Portugal, or other countries within the European Union. Understanding "cultural" and other factors contributing to these differences are crucial before drawing conclusions about the comparative health status across populations. Other studies have documented that sub-populations with lower expectations for health (such as the elderly) or less exposure to what

constitutes full health (such as those with lower socio-economic levels) actually report themselves in better health in comparison to tested or observed health, or other external criteria.[22-24] Such biases reduce the cross-population comparability of data. We recognize that self-reports may be most appropriate for certain domains of health, such as pain, in comparison to other measurement approaches,[25] but may still contain biases.

Most current efforts to enhance comparability of survey data concentrate on improving translation protocols before data collection[26] or judging the equivalence of question wording across different national surveys.[27-28] Many have estimated the psychometric properties of the same instrument in different languages and countries.[29-31] As no gold standard test exists to measure health, several weaker forms of validity are estimated for groups of questions hypothesized to assess the same underlying construct.[32] If these weaker estimates of validity reach acceptable levels within each population, then an instrument is judged comparable across those populations. More recently, several specifically use Item Response Theory (IRT) to estimate internal relationships between an individual's response on any one item in a survey and the underlying construct of the domain or scale.[33-34] Proponents of IRT analysis argue that if these relationships are similar in each population tested, then an item acts homogeneously across these populations and provides comparable data on the same scale.[35] However, one assumption in order to use IRT models is often violated: that is for individuals with the same level of health, the probability of answering any item in the same direction should be unrelated to the probability of answering any other item in the same direction.[16,36] Similar to classical psychometric approaches, IRT analysis does not use external criteria to establish the equivalence of scales across populations. Without external criteria, we believe that current efforts provide insufficient evidence to establish the cross-population comparability of data.

We therefore prefer to base our analysis on nationally representative surveys that incorporate methods to calibrate self-reported responses using the same external criteria across populations. However none exist. Only a handful of countries conduct nationally representative surveys that combine interviews and examinations.[37] These may provide estimates of the within-population comparability of data.[12-13] Available for secondary analysis from a large number of countries are different types of household interview surveys that include different questions addressing health. No standard methodologies exist that facilitate the comparability of data from different surveys using different questions. This study represents an initial effort to address the two main challenges that limit the meaningful comparison of household interview data across populations by developing and testing a simple post-harmonization method using only self-reported data.

Methods

Data. Unit level data obtained from 64 recent household interview surveys with nationally representative non-institutionalized civilian population samples were collected from 46 countries. Countries from each of the WHO Regions were included (*African:* Ghana, South Africa, United Republic of Tanzania, Côte d'Ivoire; *Pan-American:* Brazil, Guyana, Jamaica, Panama, Paraguay, Peru, USA; *Eastern Mediterranean:* Bahrain, Egypt, Jordan, Morocco, Pakistan, Tunisia; *European*: Austria, Belgium, Bulgaria, Denmark, France, Germany, Greece, Ireland, Italy, Kyrgyzstan, Luxembourg, Netherlands, Portugal, Russian Federation, Spain, United Kingdom; *South East-Asian:* Bangladesh, Democratic People's Republic of Korea, India, Indonesia, Myanmar, Nepal, Sri Lanka, Thailand; *Western Pacific*: China, Fiji, Malaysia, Philippines, Republic of Korea.) Details concerning the surveys, samples and questions are found elsewhere.[38] Questions assessing the level of physical or mental health, physical or cognitive functioning, other disability, as well as the degree and duration of activity or role limitations as a proximate measure of severity, were included. The number of questions meeting the inclusion criteria varied considerably across surveys (not shown).

Extraction of health information. We utilized common factor analysis[32] as an initial attempt to extract information on health status at the individual level from each survey. An exploratory factor analysis on several surveys provided evidence that most items meeting the inclusion criteria loaded on one factor. We hypothesized that all questions describing health tap a general health construct, H, and that H is equivalent across populations. Although we believe that health is a multi-dimensional construct, that an

underlying factor ties together these domains is similar to other findings.[30] We estimated model parameters for each survey and sex using the Mplus statistical analysis program[39] which uses specialized estimation techniques for categorical and non-normally distributed data. We calculated a variant of the intraclass correlation coefficient[40] to estimate the internal consistency of all variables as a measure of *H* by survey and sex. Seventy-eight percent of the models reached .7 or higher, the arbitrary cut-off on a zero to one scale suggesting that the variables are tapping a similar construct.[32] This is not, however, an evaluation of the cross-population comparability of data.

Estimation of the level of health. We estimated scores for each individual's level of health utilizing survey and sex specific multiple regression models.[32] In general, the factor score for individual *i* for health, *H*, can represented as:

$$\hat{H}_i = \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \ldots \hat{\beta}_p x_{ip}$$

where $\hat{H}_i$ is the estimated factor score (severity level) of health for individual *i* , $\hat{\beta}_p$ is the sex specific estimated factor score coefficient (factor loading) for variable *p*, and x_{ip} is the *p*th variable included within the survey for individual *i*. The Mplus statistical analysis program[39] calculates biserial and tetrachoric correlation estimates to produce correlation matrices within factor analyses. Given the current constraints of this program, all ordered categorical variables have been dichotomized within this analysis.

Fixing of endpoints. We attempted to equalize the scale of $\hat{H}$ across populations, while maintaining the relative differences in the level of health and distribution of severity within each population. We first transformed the lowest and highest estimated scores of $\hat{H}$ to zero (e.g., worst living state) and 100 (e.g., full health) respectively. Scores between these values represent the percentage of the highest estimated score. We then adjusted for differences in the end-points of the scale for $\hat{H}$, given no possibility to calibrate end-points or intermediary scores through an external process. We fixed all scores to zero that represent the bottom 0.1% of the cumulative distribution of $\hat{H}$. Although arbitrary, 0.1% maintains a small absolute number of individuals in the worst state across surveys. Fixing the top end-point was more problematic given the highly skewed distribution of $\hat{H}$ towards better levels of health. In one extreme case,[38] 75-80 percent of males and females within the Pakistan Living Standards Measurement Study Survey of 1991 (5 questions addressing health status) have a transformed score of 99 or 100. We believe that the distribution at the top end is truncated to some degree in all surveys. Several factors contribute to high ceiling effects, including the insensitivity of questions to mild deviations from full health, the relatively few number of questions and reporting biases, among other factors.[32] Given the large variations in ceiling effects (not shown) and no basis to judge what is the true proportion of individuals in each population at full health, we simply fixed 100 to the highest estimated score of $\hat{H}$.

Validity tests. Given the lack of external criteria, we conducted a series of internal validity and reliability checks.[38] The comparison of two different national surveys from Denmark conducted in 1994 serve as one test of convergent validity. Figure 3 shows that the Danish Health and Morbidity Survey (DHMS) (36 questions) provides a relatively smoother distribution of the cumulative frequency of different levels of health than the shorter European Community Household Panel (ECHP) survey (4 questions), as expected. Both surveys have high ceiling effects. Yet the similarity of these distributions provide some evidence that our methods provide a reasonable approach to compare different surveys conducted in the same population.

Results

Figure 4 shows estimates for the level of health, by 10 year age intervals and sex, for selected countries in Africa, South East Asia and Mediterranean Europe, with additional results provided elsewhere.[38] A score of 100 is the best level of health, or full health, whereas a score of zero is the worst level of health, excluding death. The level within age groups and trends across age groups are generally more similar within regions than across regions, except for the Mediterranean European countries. Greater variation exists at the older age groups than in younger age groups, partially reflecting that at younger age groups

little deviation from full health is reported. Furthermore, females generally report worse levels of health in comparison to males, but this is not always the case across all age groups.

Figure 5 summarizes age-standardized[41] estimates for health status across all 46 countries included, for people aged 65 years and older (e.g., the age group common across all surveys). Considerable variation exists across countries on the aggregate level of health based on self-reported health status. For both males and females, Indonesia and China report the best levels of health (almost at full health), Germany and Cote D'Ivoire report mid range levels of health, while France and Kyrgyzstan report the lowest levels of health (less than half of full health).

Discussion

Information content. The evidence is rather mixed concerning the information content of surveys within countries. Some surveys clearly appear to meet basic criteria, (1) that a range of health states exist at the population level, and (2) that health status declines with age. Even if differences between age groups provide only a weak test of criterion validity,[32] many surveys do not pass this necessary test. Almost no differences are found across age groups in surveys from Bangladesh, Bulgaria, China, Indonesia, Paraguay, Peru, Morocco, Nepal and South Africa. Self-reported health status is almost at full health in these countries, except for Bulgaria. Not surprisingly, these are the same countries with the highest aggregated health level for the 65 and older population (Figure 5). Given that different types of surveys have been utilized, some with extensive questions, such as in Indonesia and Bangladesh, and some with more limited questions, such as in Peru or South Africa, we do not conclude that these results only reflect differences in methods. In countries where we would expect worse levels of health for the 0-4 age interval in comparison to the 5-9 age interval, interview reported data do not always conform to expectations, such as in Morocco or Nepal.[38] In other countries, such as Belgium, Guyana and Panama, the level of self-reported health status is better in the oldest age group in comparison to the second oldest. Such patterns reduce our confidence in the information content of surveys. Overall, we hypothesize that different norms and expectations by age and potentially other factors contribute to an under-reporting of decrements to full health, particularly at older age groups. Females report more severe levels of health than males, conforming to expected patterns of self-reporting of health status. Whether this is a reflection of reporting biases including different norms and expectations,[7] or differences in the true level of health due to biological or gender based factors, can not be addressed given that we only have self-reported data. We do not assume that reporting biases only exist in surveys that fail the weak test of expected differences across age groups.

Cross-population comparison. We now consider whether the results may be compared across countries within the same region. In the African region (Figure 4), very little variation is noted across age groups, as discussed, except for Tanzania, where self-reported health levels drop in the oldest population groups, especially for females. The survey conducted in Tanzania (at the sub-national level) included a broader range and depth of items addressing health status than those conducted in the other three African countries. Differences in survey methodologies may therefore account for some of the differences within this region. However, given that almost no difference in the level of health status was reported across age groups in South Africa, we suggest that some of the differences are due to reporting biases. We do not conclude that South African females over the age of 55 are significantly healthier than Tanzanian women of the same age group. Similar limitations exist concerning the comparability of data within other regions.[38]

Within Europe, we have the possibility to eliminate differences due to different methodologies, given that the same survey was conducted in 13 European countries. For both sexes, the average levels of health vary considerably across age groups, with Greece and Ireland (not shown) at the highest levels across most age groups, and France at the lowest levels across all age groups (Figures 4). We find it improbable that France has significantly lower levels of health, starting with the 15-24 age interval, in comparison with the other European countries included in the same survey. Instead, we believe that reporting biases are a major factor contributing to these results, beyond variations in the true level of health or measurement error.

Interpretation. Figure 5 raises several issues concerning the meaningful interpretation and use of self-reported data. Can one conclude that the level of health status is much better in Nepal or Bangladesh, than it is in Spain or France, based on self-reported data? We do not. Nor have we included these results *as is* within estimates of DALE for these 46 countries.[3] We recommend that reporting biases should be estimated and self-reported data potentially adjusted. This is not to imply that one set of norms or expectations is superior. Rather, we argue that comparisons of health status across populations should compare health and not differences in norms and expectations. For example, Figure 6 is a scatter plot of the per capita health expenditures[42] and the average level of self-reported health for the over 65 population (males and females combined), for all 46 countries. *Higher* levels of per capita health expenditures are correlated with *lower* average levels of health, an improbable result. Although not conclusive, this suggests that with more information, resources, and exposure to health services, population norms and expectations differ and that these differences are associated with the self-reported level of health across countries. This correlation, even if weak, is consistent with other findings[6,22-24,38] in that sub-populations or countries that are wealthier and spend more resources on health, also self-report worse levels of health in comparison to those with fewer resources. However, the contrary is more likely true. A meta-analyses of available epidemiological data within the Global Burden of Disease Study[43] documented higher levels of disability in regions that are less wealthy and spend fewer resources on health, than those regions with greater resources.

We conclude that the cross-population comparability of existing data from household interview surveys is limited, even where the data collection approaches are standardized. This conclusion is disappointing given WHO's goal to include health, not only mortality levels, in its estimates and comparisons of the level of health status across populations. Given that a growing number of countries implement nationally representative household interview surveys addressing health topics, we believe further investment to improve survey methods is urgently warranted.

Next steps. WHO plans a comprehensive strategy to enhance the comparability of data from household interview surveys. Two specific areas of work are currently being pursued through international consensus and research with Member States and our collaborators.[38] These are (1) identifying a core set of domains to describe health and appropriate survey questions for household interviews, and (2) developing methods to estimate cross-population comparability of data collected and to calibrate self-reported response categories across populations to enhance comparability. In the first phase, we plan to design and test in nationally representative surveys across some 20 populations a range of methods, such as:

- Comparing self-reported health and tested health (performance tests) in selected domains, such as vision, cognition and mobility.
- Selecting and testing the equivalence of end-points and cut-points for different domains of health, based on external criteria.
- Estimating differences in thresholds for response categories across countries, using a series of vignettes to describe health status in specific domains of health such as mobility, pain, affect and cognition.
- Estimating the reporting biases of selected homogeneous sub-populations across countries, based on health and non-health selection criteria: a comparison of the reporting biases of different groups may provide a defensible means to adjust self-reported data.
- Developing standardized heuristics, such as clearly defined anchor points in the framing of questions and response scales, to reduce biases associated with forming self-assessments of health status.

In subsequent phases, we plan to facilitate the use of methods developed in a broader range of countries and surveys. These strategies, among others, are to be used in addition to current approaches to estimate the validity and reliability of questions and domains and may be considered as a valuable dimension in cross-population research in this area.

Acknowledgments

The authors thank the following individuals for their contributions: Brodie Ferguson for his perseverance in obtaining data sets; David Cutler for previous collaborative work that contributed to the introductory section; Luis Prieto, Bedirhan Üstün and Omar Ahmad for their comments and suggestions; and Nankhonde Kasonde, Sue Piccolo and Marie-Claude von Rulach for text processing assistance.

References

1. WHO 2000. *World Health Report 2000*. Geneva, World Health Organization.
2. Murray CJL, Salomon J, Mathers C 1999. *A critical review of summary measures of population health.* Geneva, World Health Organization, 2000 (GPE Discussion Paper No. 2).
3. Mathers C, Sadana R, Salomon J, Murray CJL, Lopez AD (2000). *Estimates of DALE for 191 countries: methods and results.* Geneva, World Health Organization, 2000 (GPE Discussion Paper No. 16).
4. Kleinman A, Eisenberg L and Good B 1978. Culture, illness and care: clinical lessons from anthropologic and cross-cultural research. *Annals of Internal Medicine* 88: 251-258
5. Kahneman D, Sovic P, Tversky A (eds.). 1982. *Judgment under uncertainty: heuristics and biases*, Cambridge University Press, Cambridge
6. Johansson SR 1991. The health transition: the cultural inflation of morbidity during the decline of mortality. *Health transition review* 1(1):39-68.
7. Annas J 1993. Women and quality of life: two norms or one? In *The quality of life* Nussbaum M and Sen A (eds.) Clarendon Press, Oxford, pp.279-296.
8. Sen A 1994. Objectivity and position: assessment of health and well-being. In *Health and social change in international perspective*, LC Chen, A Kleinman, NC Ware (eds.). Harvard series on population and international health, Harvard School of Public Health, Boston, pp. 115-128.
9. Zurayk H, Khattab H, Younis N et al. 1995. Comparing women's reports with medical diagnosis of reproductive morbidity conditions in rural Egypt. *Studies in Family Planning* 26(1):14-21.
10. Kempen GI, van Heuvelen MJ, van den Brink RH et al. 1996. Factors affecting constrasting results between self-reported and performance-based levels of physical limitations. *Age & Ageing* 25(6):458-64
11. Kivinen, P, Halonen P, Eronen M, Nissinen A 1998. Self-rated health, physician-rated health and associated factors among elderly men: the Finnish cohorts of the Seven Countries Study, *Age & Ageing*, 27:41-47.
12. Iburg KM, Murray CJL, Salomon J (2000). *Examining the relationships between self-reported, physician-assessed and observed measures of health*. Geneva, World Health Organization, 2000 (GPE Discussion Paper No.14).
13. Thomas D and Frankenberg E. 2000. The measurement and interpretation of health in social surveys. In CJL Murray, CD Mathers, AD Lopez, J Salomon, R Lozano (eds). *Summary measures of population health*, Geneva, World Health Organization (under preparation).
14. Tourangeau R 1984. Cognitive sciences and survey methods. In *Cognitive aspects of survey methodology: building a bridge between disciplines.* T Jabine, M Straf, J Tanur and R Tourangeau (eds). Pp. 73-100. National Academy Press, Washington D.C.
15. Krosnick JA 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology* 5:213-236
16. Streiner DL and Norman GR 1995. *Health Measurement Scales: a practical guide to their development and use.* Second edition. Oxford University Press.
17. Kleinman A 1994. An anthropological perspective on objectivity: observation, categorization, and the assessment of suffering. In *Health and social change in international perspective*, LC Chen, A Kleinman, NC Ware (eds.). Harvard series on population and international health, Harvard School of Public Health, Boston, pp. 129-138.
18. Idler E.L., Hudson S.V., and Leventhal H. 1999 The Meanings of Self-Ratings of Health: A Qualitative and Quantitative Approach. *Research in Ageing*
19. Groot W 2000. Adaptation and scale of reference bias in self-assessments of quality of life. *J Health Economics* 19:403-420.
20. Weir C and Seacrest M 2000. Developmental differences in understanding of balance scales in the United States and Zimbabwe. *J Genetic Psychology* 161(1):5-21
21. Eurostat 1997. Self-reported health in the European Community. *Statistics in Focus, Population and social conditions*. ISSN 1024-4352
22. Kroeger A, Zurita A, Perez-Samaniego C, Berg H, 1988. Illness perception and use of health services in North-East Argentina. *Health Policy and Planning* 3: 141-151.
23. Murray CJL 1996. Epidemiology and morbidity transitions in India. In *Health, Poverty and Development in India*, eds. DasGupta M, Chen LC and Krishnan TN, Oxford University Press, Delhi, 122-147.
24. Mackenbach J.P., Looman C.W.N. van der Meer J.B.W. 1996. Differences in the Misreporting of Chronic Conditions, by Level of Education: The Effect on Inequalities in Prevalence Rates. *American Journal of Public Health* 86(5):706-711.
25. Murray CJL and Chen LC 1992. Understanding Morbidity Change. *Population and Development Review*, 18(3):481-503

26. Ware JE, Keller SD, Gandek B, Brazier JE, Sullivan M et al. 1995. Evaluating translations of health status questionnaires: methods for the IQOLA project. *Int J Technol Assess Health Care* 11(3):525-551
27. Rasmussen N, Gudex C and Christensen S 1999. *Survey Data on Disability*, Eurostat Working Papers, Population and Social Conditions 3/1999/E/n 20, European Commission, Luxembourg.
28. Robine JM, Jagger C and Egidi V 2000. *Selection of a Coherent Set of Health Indicators, Final draft* A First Step Towards A User's Guide to Health Expectancies for the European Union Montpellier (France), Euro-REVES
29. Anderson RT, Aaronson NT and Wilkin D 1993. Critical review of the international assessments of health-related quality of life. *Quality of Life Research* 2:369-395.
30. Keller SD, Ware JE, Bentler PM et al. 1998. Use of Structural Equation Modeling to Test the Construct Validity of the SF-36 Health Survey in Ten Countries: Results from the IQOLA Project. *J Clin Epi*, 51(11):1179-1188
31. WHOQOL Group 1998. The World Health Organization Quality of Life Assessment (WHOQOL): development and general psychometric properties. *Social Science & Medicine*. 46(12):1569-85
32. Nunnally JC and Bernstein IR 1994. *Psychometric Theory*. Third edition. McGraw Hill, New York.
33. Raczek AE, Ware JE, Bjorner JB, Gandek B, et al. 1998. Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: results from the IQOLA project. *J Clin Epidemiology* 51(11):1203-1214
34. Cooke DJ and Michie C 1999. Psychopathy across cultures: North America and Scotland compared. *Journal of Abnormal Psychology*. 108(1):58-68
35. Hambleton RK and Swaminathan H 1985. *Item response theory: principles and applications*. Kulwer Nijhoff, Boston.
36. Hambleton RK, Swaminathan H and Rogers HJ 1991. *Fundamentals of item response theory*. Sage, Newbury Park, CA.
37. Fisher G, Pappas G and Limb M 1996. Prospects, problems and prerequisites for national health examination surveys in developing countries. *Soc Sci Med* 42(12):1639-1650.
38. Sadana R, Mathers C, Lopez AD, Murray CJL, Iburg KM (2000). *Comparative analyses of more than 50 household surveys on health status.* Geneva, World Health Organization, 2000 (GPE Discussion Paper No. 15).
39. Muthén LK and Muthén BO 1998. *Mplus: the comprehensive modeling program for applied researchers, User's Guide, Version 1.0*. www.StatModel.com, Los Angeles.
40. Werts CE, Linn RL and Joreskog KG 1974. Intraclass reliability estimates: testing structural assumptions. *Educational and Psychological Measurement* 34:25-33.
41. Ahmad O, Boschi-Pinto C, Lopez AD, Murray CJL, Lozano R, Inoue M. 2000. Age standardization: the new WHO world standard population. Geneva, World Health Organization, (GPE Discussion Paper No. 31).
42. Poullier JP and Hernández P 2000. Estimates of national health accounts. Aggregates for 191 countries in 1997. Geneva, World Health Organization, 2000 (GPE Discussion Paper No. 27).
43. Murray CJL and Lopez AD, eds. 1996. *The Global Burden of Disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2030,* Global Burden of Disease and Injury Series, Vol.1, Harvard University Press, Cambridge

Figure 1. Empirical Model for the Assessment of Health Status

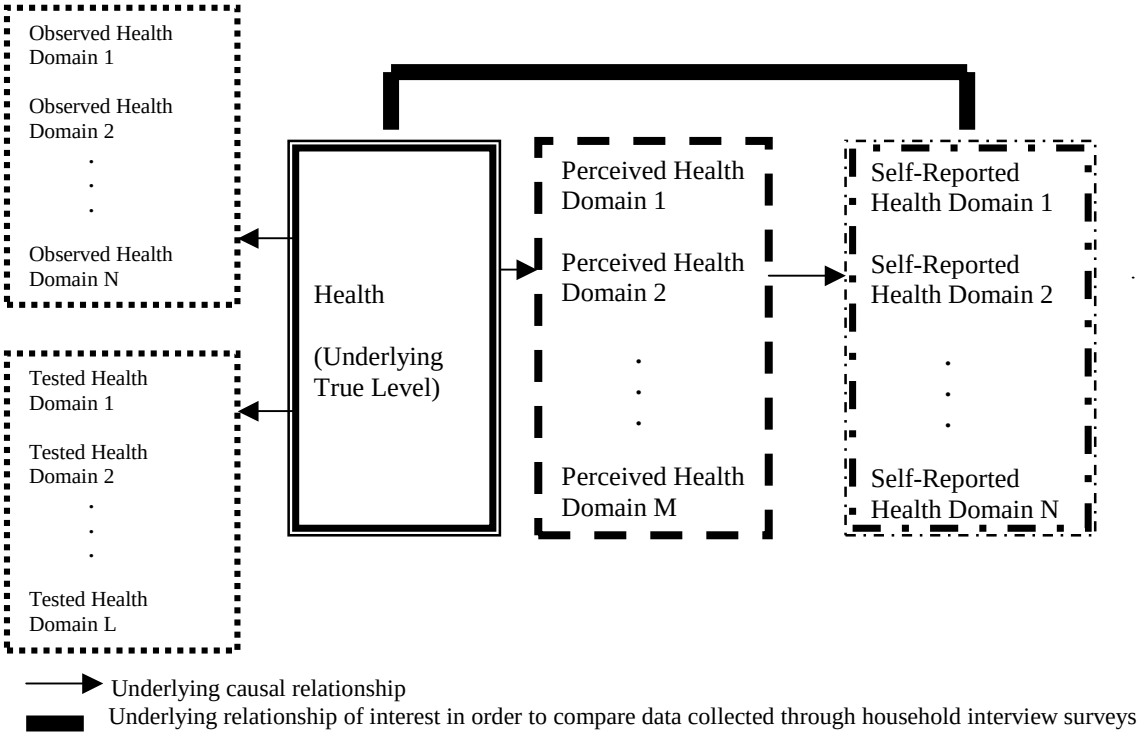

Figure 2. Proportion of population ≥ 16 years of age, reporting very good and good general health, 12 European countries, Eurostat 1997
(European standard population age-standardization)

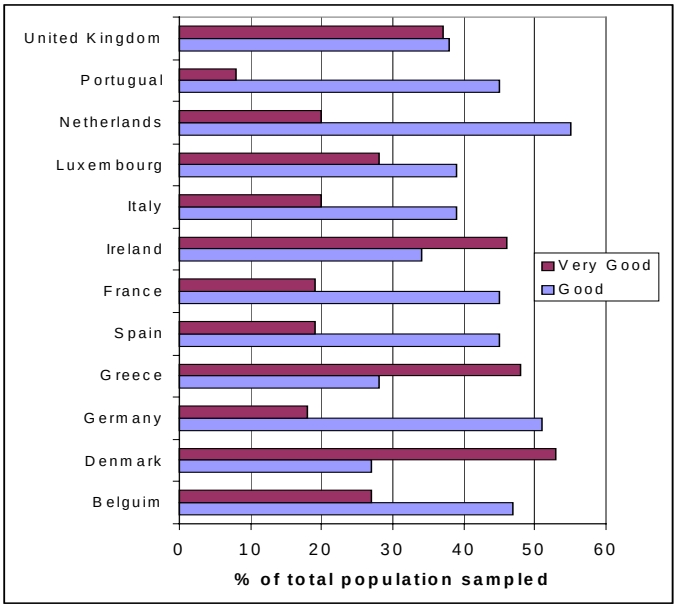

Figure 3. Distribution of Levels of Health Status, Comparison of Two Surveys from Denmark, Males and Females

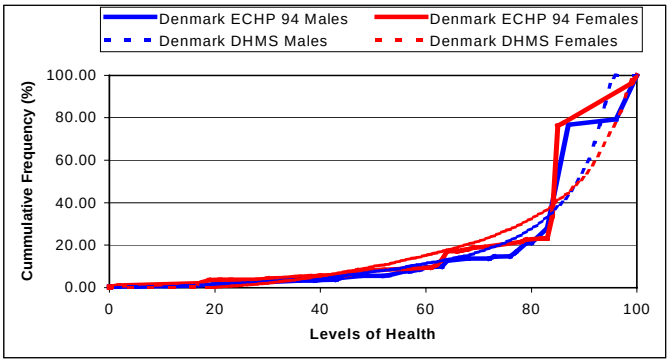

Figure 4: Self-Reported Health Status, by Age and Sex groups, Selected Countries and Regions

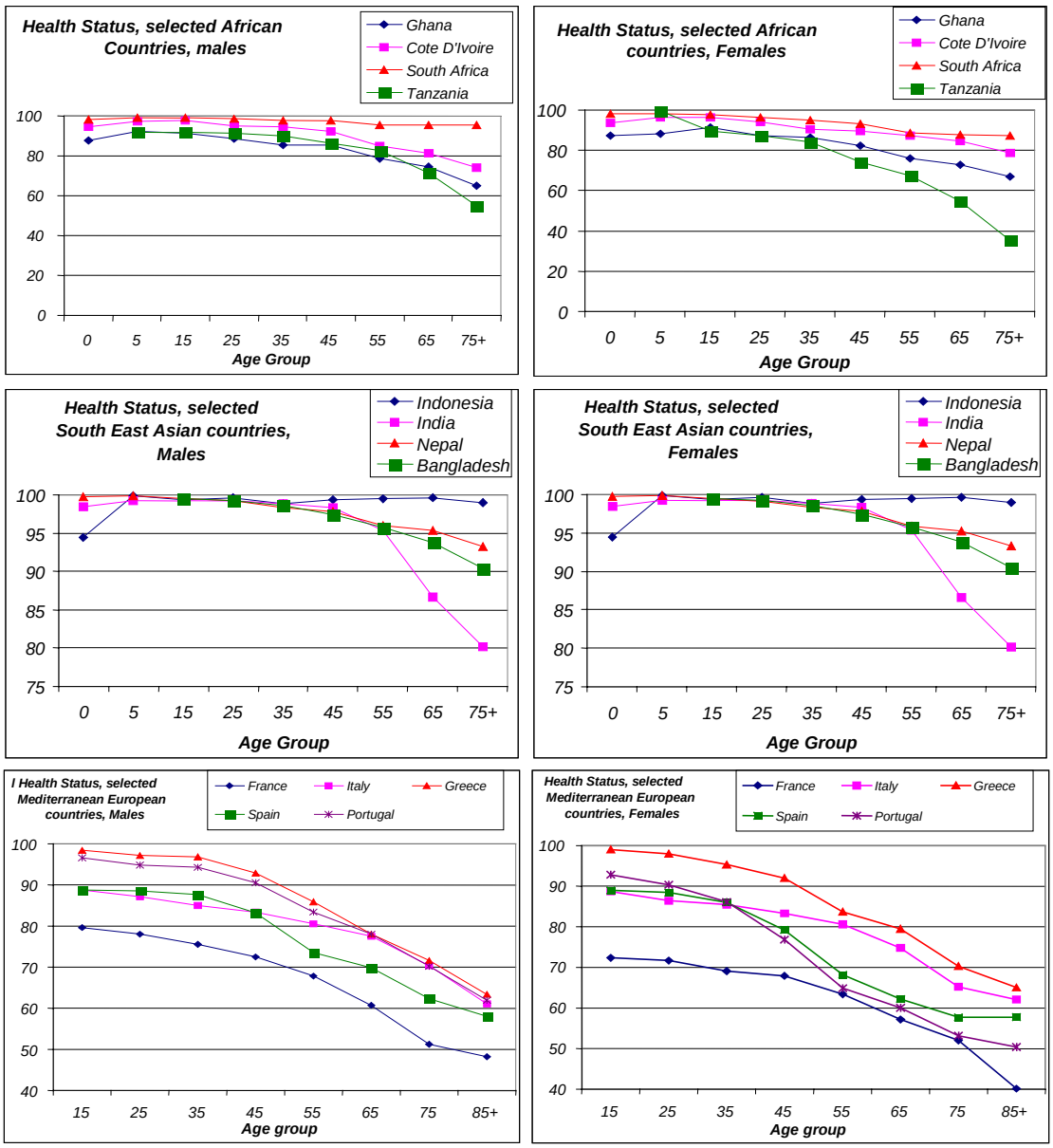

Figure 5. Comparison of average level of self-reported health, 46 countries, 65 years and older, age standardized, by sex

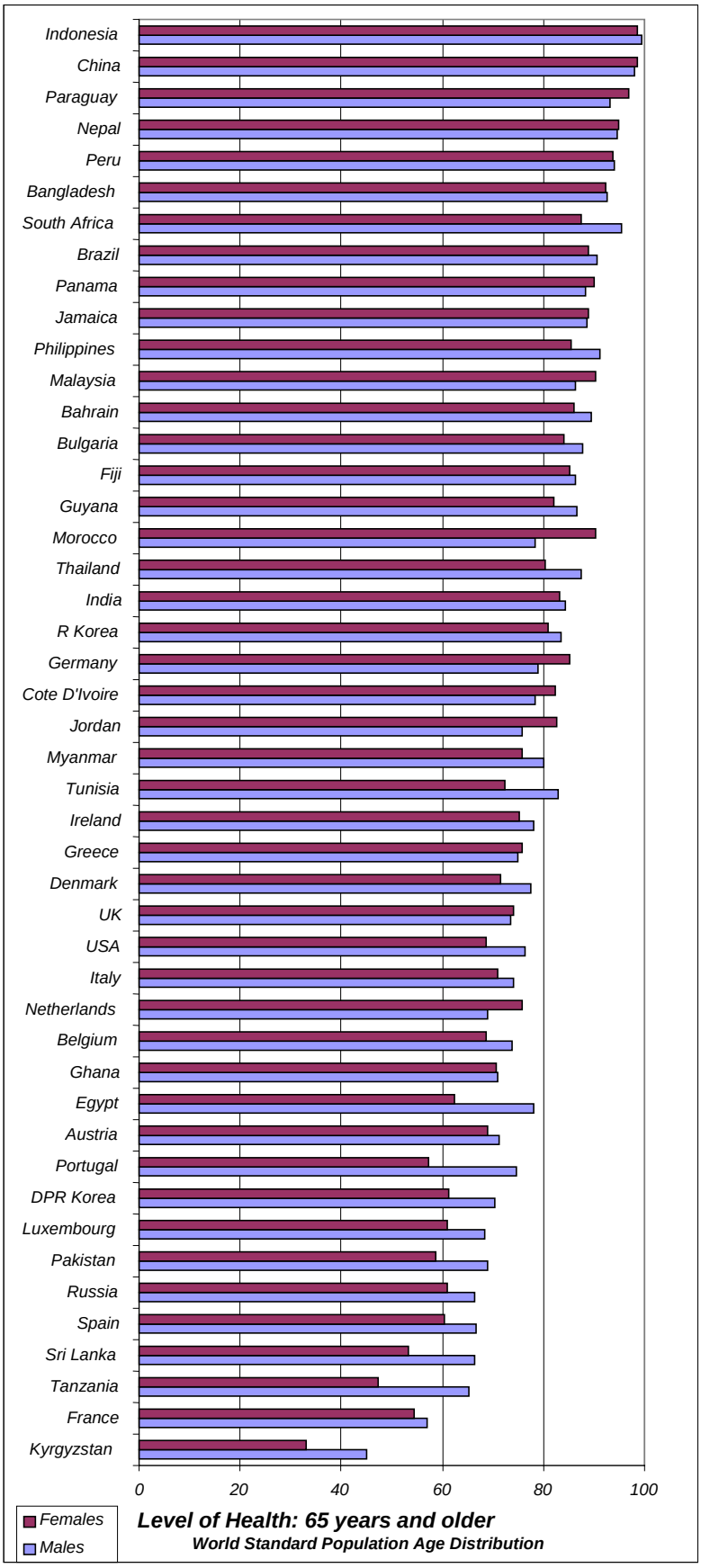

Figure 6. Per Capita Health Expenditures vs. Average level of self-reported health, 46 countries, age standardized