Comparative Population Studies
Vol. 49 (2024): 181-214 (Date of release: 05.08.2024)

Mismatches in Health: A Global Analysis of Discrepancies Between Self-Reported and Tested Mobility and Cognition*

Vanessa di Lego, Sonja Spitzer, Patrick Lazarevič

Abstract: The health of individuals is frequently assessed based on self-reported information derived from surveys. However, self-reports are often inconsistent with their tested equivalents, indicating measurement issues. While discrepancies between self-reported and tested health indicators have been investigated for high-income countries in Europe, little comparative research has been conducted involving other low-income regions. This paper analyses discrepancies between self-reported and tested health limitations across 25 countries from six world regions with different income-levels, cultural backgrounds, institutional settings, and epidemiological trajectories. Using harmonised data from the Gateway to Global Aging, we match self-reported mobility and cognition with their tested equivalent to assess discrepancies at the individual level. Our results suggest that the consistency between these measures is strongly correlated with the Human Development Index, with lower scores of development showing higher discrepancies. Examining patterns by age reveals that self-reports do not accurately reflect the deterioration of health associated with aging – tested health exhibits a pronounced age gradient, whereas self-reported health varies little over the life course, particularly self-reported memory. We find no persistent gender differences in consistency. These discrepancies cast doubt on the reliability of mobility and cognitive self-reports, especially when comparing health across nations with differing development levels.

Keywords: Self-reported health · Objective health · Mobility · Cognition · Global health

* This article belongs to a special issue on "Levels and Trends of Health Expectancy: Understanding its Measurement and Estimation Sensitivity".

** This article has an Online Appendix with supplementary material URL: https://www.comparativepopulationstudies.de/index.php/CPoS/article/view/656/411

Federal Institute for Population Research 2024

URL: www.comparativepopulationstudies.de
DOI: https://doi.org/10.12765/CPoS-2024-08
URN: urn:nbn:de:bib-cpos-2024-08en3

1 Introduction

The ability of self-reported health indicators to accurately reflect the true health status of individuals has long been questioned. Research dating back to the 1950's (*Suchman et al.* 1958) and 1960's (*Maddox* 1962) investigated discrepancies between respondents' health reports and assessments of outside observers like physicians. Discrepancies have usually been attributed to biases in respondents, as judgments of others are less likely to be influenced by factors unrelated to health. This is particularly true when the assessment is done by a skilled healthcare expert. Indeed, multiple studies have consistently shown that self-reports of health are often impacted by non-health characteristics, including the respondent's gender, age, place of origin, level of education, and personality (*Barsky et al.* 1992; *Jürges* 2007; Shumway/Stoffer 2011; *Layes et al.* 2012; *Warner et al.* 2012; *Lazarevič et al.* 2018; *Lazarevič* 2023).

A commonly employed approach to assess the credibility of self-reports is the comparison of self-reported health data with performance-based measurements. Overall, such studies indicate that tested health is more reliable and sensitive to actual health changes for a series of conditions, including risk factors for cardiovascular disease, such as hypertension (*Kislaya et al.* 2021). In addition, tested health has been shown to be less sensitive to socio-cultural and psychological factors among respondents, making them more robust to comparisons across time and populations (*Guralnik et al.* 1989; *Reuben et al.* 1992; *Simpson et al.* 2004). Lastly, the reliability of self-reported or tested health data can vary by different socio-demographic factors, like gender and age group. For example, there is evidence that self-reported prevalence of hypertension and hypercholesterolemia underestimated educational inequalities among younger men (ages 25-49), but overestimated it among older individuals (ages 50-74), when compared to examined-based prevalence (*Kislaya et al.* 2019).

For the reasons mentioned above, it has been argued that self-reports in health may have limited utility in detecting clinically significant changes (*Guralnik et al.* 1989; *Fors et al.* 2006), yielding inaccuracy not only on health prevalence, but also on any indicator that is prevalence-based, such as health expectancy indicators (*Santos et al.* 2021). Most of these studies, however, assess differences between self-reported and tested health in clinical or small epidemiological population subgroups (*Wuorela et al.* 2020), which may not reflect the context of everyday life of the participants nor reflect national or population-level health (*Glass* 1998).

Another approach that has been suggested by researchers is to employ concordance or consistency measures, which compare the results of two measurements (*Bland/Altman* 1986; *Guyatt et al.* 1987). Consistency measures have been argued to be preferable to other reliability or correlation measures for contexts when the researcher would like to know to what extent two measurements yield the same results (are concordant with each other) (*Kramer/Feinstein* 1981). In the case of self-reported and tested health, the rationale is to have a counterpart test for the self-report and contrast those two results, to assess whether there are discrepancies between respondents' self-perceived health and their objectively measured health markers. Recent work has employed this approach to evaluate consistency between self-reported and tested health, finding significant disparities across various European nations (*Spitzer/Weber* 2019; *Arni et al.* 2021). These variations were more important among older individuals, while gender differences made only a small contribution to the observed discrepancies (*Spitzer/Weber* 2019).

Nonetheless, the majority of these studies were limited to wealthier nations or focused on regions characterised by relatively homogeneous cultural beliefs, health care systems, and gender norms. In this paper, our goal is to employ the consistency method devised by *Spitzer* and *Weber* (2019) and *Arni et al.* (2021) to assess consistency between population-level reported and tested health both within- and across countries with different development levels. Besides being a recommended approach for comparing two indicators that measure the same outcome (*Kramer/Feinstein* 1981), using consistency measures has the advantage of relying on population health surveys, going beyond medical settings and thus a more suitable approach for understanding population-level health. Since levels of development have been shown to correlate with health response outcomes, we also explore whether there is a correlation between levels of development and consistency between self-reported and tested health. We additionally aim to identify whether there are relevant gender- and age-specific patterns in consistency between reported and tested health across the different countries. We use The Human Development Index (HDI) as an indicator of development since it is comparable across countries and is an important summary measure of human development in key dimensions of life (*United Nations Development Programme* 2022). The HDI has also been shown to be highly correlated to health outcomes (*Chandra et al.* 2022). In order to explore broader regional heterogeneity, we also group countries according to the six global regions defined by the World Bank Development indicators and their corresponding HDI scores.

To date, studies that have performed global comparisons use less detailed health indicators and typically lacked harmonisation across them, which can yield misleading results (*Tolonen et al.* 2021; *di Lego* 2021). For our analyses, we use harmonised health data from the Gateway of Global Aging (*Lee et al.* 2021), in order to compare the prevalence of agreement between tested and reported health among women and men in China (CHARLS), India (LASI), Mexico (MHAS), Europe (SHARE), United States (HRS) and England (ELSA). We limit our analysis of health dimensions to mobility and cognition because they cover different and important dimensions of health status. Moreover, our consistency measure requires health dimensions that include both reported and tested information in all surveys. As a measure of mobility, we use walking speed test and self-reported difficulty walking 100 meters. For cognition, we use the immediate word recall test and self-reported memory.

We use this strategy since the level of consistency between self-reports and tested health does not depend on the prevalence of health conditions in the population, but only on how tested and self-reported health diverge. Similarly, because we are utilising HRS-sister surveys and they have been designed to replicate the same questions and survey methodology as the HRS, inconsistencies due to survey design are minimised. This enables us to capture age- and gender-specific patterns in consistency for specific countries and whether there is stability across different nations in the discrepancy between tested and reported health.

By performing this analysis across different countries, we shed light on how consistency between perceived and tested health may differ across different cultures and how in turn this affects the sensitivity of some health indicators. Quantifying the divergence between self-reports and tested health is important, since how individuals perceive their own health can greatly impact their health behaviours, healthcare utilisation and prevention (*Spitzer et al*. 2022; *Spitzer/Shaikh* 2022). This may additionally impact how we interpret health indicators when performing cross-country comparisons and when comparing gender differences in health.

2 Material and methods

2.1 Data and sample

2.1.1 The Gateway to Global Aging Data

The Gateway to Global Aging Data, produced by the Program on Global Aging, Health & Policy at the University of Southern California Dornsife Center for Economic and Social Research is an effort to harmonise the HRS-sister surveys (*Lee et al*. 2021). The purpose is to facilitate comparisons across countries and allow for more accurate estimates of health. The harmonised versions have followed the RAND HRS conventions of variable naming and data structure. Hence, the surveys are all harmonised having the structure and variable naming of the HRS study as a reference. For more details on how the harmonisation procedure is implemented, refer to their website (https://g2aging.org/).

Despite the Gateway to Global Aging's efforts, international comparisons are still challenging and limited to the variables and survey years that have been harmonised. In addition, even with the harmonisation, some surveys experienced variations across some waves as detailed in the harmonisation reports, like adjustments in tested health measurement, changes in self-reported response scales, and modification of rules for proxy respondent inclusion (*Kwon/Hu* 2018; *Young et al*. 2021). Furthermore, some countries like Mexico and India have limited number of waves, which – unfortunately – prevent time trend analysis and fully exploring the longitudinal feature of other surveys. Our aim is to include as many countries as possible from different regions in the world to evaluate whether country-specific patterns in consistency emerge. In addition, to employ the approach for assessing consistency between self-reported and tested health described in *Spitzer* and *Weber* (2019), we need to have a perfect match between self-reported and tested health variables for the same health dimension. Hence, we used the following criteria for selecting years and health dimensions: 1. Health dimensions that were present for the most diverse set of countries and that had the perfect match between self-reported and tested health variables for the same health dimension; 2. To address potential confounding effects arising from temporal variations in mortality and health trends within the surveys, we selected the years that are comparable or overlapping across all waves; 3. Comparable waves in the sense that there were no documented changes in either the health test, the equipment needed, proxy respondents and self-reported health, as detailed in the reports from the Gateway to Global Aging.

Following these selection criteria after extensive screening, the resulting health dimensions were mobility and cognition. These were the only two health dimensions for which the exact match between self-reported and tested health were available and that had the largest variety of countries for comparable years. For mobility, the match is between self-reported difficulty walking 100 meters and the walking speed test, where respondents' timed walk over a walking course is recorded. For cognition, the match is between self-reported memory and the test of immediate word call (see details on these two health variables in section 2.1.2 below).

Because available waves for mobility and cognition were different, the samples were constructed separately. Furthermore, because we are comparing data globally from countries with very different levels of socio-economic development and epidemiological trajectories, we also grouped countries into regions as defined in the World Bank Development Indicators (*World Bank* 2022). We used this country grouping to assess the role of regional development on the disparities found between self-reported and tested health across surveys. In addition, as the Human Development Index (HDI) has been shown to be highly correlated to health outcomes (*Chandra et al.* 2022), we assign HDI scores by country as a summary measure of development (*United Nations Development Programme* 2022). Lastly, the surveys have different age eligibility for inclusion to the survey sample, varying from age 45 (CHARLS, LASI) to 51 (HRS). In some cases, the tested health variable is restricted to specific ages. To account for this, we focused on ages 65 and above. The exception is SHARE, where the tested health for mobility (walking speed) was undertaken only for participants older than 74 (Wave 2) and 75 (Wave 1) years old (*Börsch-Supan et al.* 2013; *Bergmann et al.* 2019). To assess whether the older age structure available for SHARE impact the analysis, a robustness check with models for mobility restricted to ages 75+ for all countries was also adjusted (see Appendix Table A1.1).

The waves and surveys that met our criteria for both health dimensions are described in Table 1. Any remaining inconsistencies between surveys and waves are addressed in the robustness analyses. For mobility, the available years that mostly overlapped were 2012-2015. Exceptions are the SHARE study for Europe where the walking speed test was only undertaken in waves 1 and 2 (years 2004 and 2006), and the LASI study for India that has only one wave in year 2016. In addition, data for mobility (Wave 1 and Wave 2) and cognition (Wave 4 and Wave 5) come from SHARE waves that included different sets of countries. For mobility, the countries in Wave 1 are: Germany, Austria, Belgium, Denmark, Spain, Greece, Italy, the Netherlands, Sweden, Switzerland, France and Israel. In Wave 2, the following countries are added: Ireland, Poland and the Czech Republic. For cognition, in addition to the previously mentioned, the following countries are included: Estonia, Hungary, Portugal, Slovenia and Luxembourg.

Tab. 1: Overview of the Samples for Mobility and Cognition

World Bank Region	Country	Survey	Waves	Observation Period	*N* (unweighted)**
		Mobility			
North America	USA	HRS	W11, W12	2012-2015	7,584
Europe and Central Asia / Middle East and North Africa	Europe+ Israel	SHARE	W1[1], W2[2]	2004-2007	6,629
	England	ELSA	W6, W7	2012-2015	8,796
East Asia and Pacific	China	CHARLS	W2, W3	2012-2015	8,302
Latin America and Caribbean (LAC)	Mexico	MHAS	W3	2014-2015	734
South Asia	India	LASI	W1	2016-2017	18,366
		Cognition			
North America	USA	HRS	W10, W11	2010-2013	21,702
Europe and Central Asia / Middle East and North Africa	Europe+ Israel	SHARE	W4[3], W5[4]	2010-2013	63,627
	England	ELSA	W7, W8	2014-2017	11,069
East Asia and Pacific	China	CHARLS	W1, W3	2010-2015	10,726
Latin America and Caribbean (LAC)	Mexico	MHAS	W3, W4	2012-2015	15,433

** Sample size refers to the total number of observations. Because the studies are longitudinal, participants in one wave can also be present in the other. We account for this via weights and clustered standard errors in subsequent analyses.

Source: [1] (*Börsch-Supan* 2022a) [2] (*Börsch-Supan* 2022b) [3] (*Börsch-Supan* 2022c) [4] (*Börsch-Supan* 2022d). Data for mobility (W1 and W2) and cognition (W4 and W5) come from different SHARE waves, which for these years imply different sets of countries. For mobility, we have the following countries in Wave 1: Germany, Austria, Belgium, Denmark, Spain, Greece, Italy, the Netherlands, Sweden, Switzerland, France and Israel. In Wave 2, the following countries are added: Ireland, Poland and the Czech Republic. For cognition (W4 and W5), we have the following set of countries included, in addition to the previously mentioned: Estonia, Hungary, Portugal, Slovenia and Luxembourg.

2.1.2 Measuring discrepancies between tested and self-reported limitations

To assess discrepancies between tested and self-reported limitations, we construct a measure of consistency, following the approach by *Spitzer* and *Weber* (2019) and *Arni et al.* (2021). The rationale for this approach is that, all things equal, self-reported and tested health should be consistent when considering the same health dimension within a population that is surveyed. The degree of consistency is thus defined according to how well self-reported and tested health match for the same variable. For this, we match tested health measures on mobility and cognition with their self-reported equivalent. In a first step, the tested and self-reported health measures are dichotomised so that their outcome is either "healthy" or "unhealthy".

In a second step, we assess the consistency between the tested and the self-reported measure. Consistency is reached if survey participants are healthy according to both measures, or unhealthy according to both measures. The measures are deemed inconsistent if the tested measure does not agree with the self-reported measure. We do not differentiate between negative consistency, i.e. both measures suggest health limitations, and positive consistency, i.e. both measures suggest no health limitations.

Self-reported mobility is based on the survey question: "Do you have difficulty walking 100 meters?" Participants are considered healthy if they did not report any problems with the activity, and unhealthy if they reported some difficulty. Tested mobility is based on a walking speed test, for which survey participants have to walk a distance of 2.5 meters (with the exception of Mexico, which is 3 meters). Prior to the actual test, respondents were asked if they had any problems from recent surgery, injury, or other health conditions that might prevent them from walking. Only persons aged at least 60 years, willing to do the test, and able to walk (walking aids were permitted) were asked to perform the test. Because the age eligibility for the test varied across surveys, we restricted the analysis to ages 65+. SHARE only performs the walking test on individuals older than 74 years of age, so for European countries, the analysis on mobility is restricted to ages 74-85+. To ensure that the older age structure of SHARE countries for mobility does not impact the cross-country comparisons, we provide an additional robustness check where we re-estimate all models for mobility with ages 65+ and 75+ only (refer to Table A1.1)

In line with the literature and following previous work, we estimate walking speed by dividing distance over time (m/s), with individuals considered unhealthy if their average walking speed is 0.4 meter per second or lower (*Studenski et al.* 2011; *Middleton et al.* 2015). Respondents that tried to do the test but were unable to complete it are also considered as having poor health. This is also the threshold and protocol applied by official survey reports for ELSA and SHARE (*Steel et al.* 2003). To ensure that our results are not driven by the cutoff of walking speed at lower or equal to 0.4m/s, we ran a robustness check using a higher cutoff of 0.6m/s, since the onset of mobility limitation among community dwelling residents has been shown to be associated to that threshold (*Cawthon et al.* 2021) (refer to the Appendix Table A4). In addition, as Mexico is the only country where the parameter for walking speed is different (3m-walking, instead of 2.5m), we perform an additional robustness check where we re-estimated all models excluding Mexico, to evaluate any differences across countries that could be driven by this exceptionality (refer to the Appendix Tables A4.1, A4.2).

Self-reported cognition is based on the survey question "How would you rate your memory at the present time?", with possible answers "excellent", "very good", "good", "fair", and "poor". Following previous research on the topic, individuals are considered unhealthy if they report "fair" and "poor" and healthy if they report "excellent", "very good" or "good" memory (*Gardner et al.* 2017; *Spitzer/Weber* 2019). Cognition is tested via an immediate word recall test, for which survey participants are asked to recall a list of ten words within a minute in any order. Individuals are considered unhealthy if they recall three words or less or if there is a proxy respondent (*Grodstein et al.* 2001; *Purser et al.* 2005; *Spitzer/Weber* 2019). Similar to walking speed, to ensure that our results are not being driven by the cutoff of three words, we also consider respondents that recall two words or less as having poor health, instead of 3 words or less, as a robustness check (Table A6 in the Appendix). Mexico is again an exception for tested cognition, with a word list that is composed of eight words, instead of ten. For this reason, we re-estimated the models excluding Mexico to evaluate whether there is any impact in the cross-country comparison (see Tables A6.1 and A6.2).

2.2 Empirical strategy

We first explore discrepancies between tested and self-reported health with descriptive statistics by age, gender, and country.

We then investigate country differences by regressing our consistency measure on country dummies and a set of control variables. Given the binary nature of the consistency measure, we employ a simple logistic regression model that looks as follows:

$$logit(\pi_i) = \alpha + \beta \mathrm{COUNTRY}_i + \delta X_i + \varepsilon_i \qquad (1)$$

where π_i indicates the probability of individual i achieving consistency between the tested and the self-reported measure. COUNTRY is a categorical variable that indicates the individual's country of residence with the United States as the reference country, since all surveys are considered HRS-sister studies.

The vector of control variables X includes the individual's age to address country-differences in the age structure as well as period dummies to account for different survey waves. It also includes a gender dummy, except for when the model is estimated separately for women and men to obtain gender-specific coefficients for each country.

All computations are weighted to adjust for representation error and to account for the impact of complex sample survey design on standard errors. In addition, we are not performing longitudinal analysis, but we use more than one wave for some countries, which means that some participants are included in the sample more than once, because we pool different survey waves. To account for this, we estimated standard errors that are clustered at the individual level using the HC3 estimator, which has been shown to outperform the HCO, HC1 and HC2 estimators, since it avoids underestimating the true variance when there are observations with extreme values of the predictor variables (*Long/Ervin* 2000). This approach is adopted to mitigate the potential influence of artificially inflating the sample size due to measures that come from the same individual when treating the data as cross-sectional (*White* 1980; *Long/Ervin* 2000; *Abadie et al.* 2022).

Since levels of development have been shown to be an important mediator of health differentials, we further explore the relationship between regional development and consistency between tested and reported health. For this, we first group countries into regions as defined in the World Bank Development Indicators.

Subsequently, Spearman correlation coefficients are employed to assess the link between the odds of consistency, derived from the regression models, and the Human Development Index (HDI) scores. The HDI scores are for year 2013, the mid-point year for most of the survey period selected. Far from aiming at a causal analysis and given possible omitted variable bias, this exploration is a correlation exercise to elucidate regional disparities, provide context to these countries and put our comparison in perspective.

Lastly, for illustrative purposes and prompt discussion, we rank countries from low to high prevalence according to self-reported and tested health. The goal is to elucidate how countries can fare very differently if one chooses one indicator over the other. Specifically for the ranking analysis, as we are rating absolute prevalence, we restrict the analysis to age 75+ to avoid biasing the results by performing comparisons that mix prevalence from different cut-off ages.

3 Results

3.1 Mobility

3.1.1 Descriptive results

There are large variations in consistency across countries as shown in Table 2. Consistency is as high as 89 percent in England but only 66 percent in India. Among countries that are in the European region and part of the SHARE survey, there is great heterogeneity. While consistency is 88 percent in Sweden, 86 percent in the Netherlands and 86 percent in Switzerland, it reaches only 64 percent in Greece, 66 percent in Spain, and 69 percent in Poland. China and Mexico have 85 percent and 81 percent of consistency, respectively. Women and men have different patterns of consistency. Women show lower shares of consistency for all countries except Ireland and the Netherlands. For men, consistency reaches its highest in Switzerland, with almost 92 percent and followed by England (90 percent), Sweden (90 percent), and the U.S. (88 percent). While the high share of consistency holds for English women as well (88 percent), it is much lower for women in the U.S. (84 percent). Greek and Indian women show particularly low levels of consistency – the self-reports agree with the tests for less than two thirds of the survey participants in those countries (60 percent).

Importantly, consistency substantially declines with age, regardless of gender, in all countries. China has the steepest age gradient, where consistency drops from a high of 90 percent in the youngest age group (65-70 years old) to 66 percent for those ages 85 and over, a decline of more than 26 percent between those ages. Across European countries, the age gradient is also substantial, despite the youngest age group available being 75-80. Poland, Austria and Greece experience the most dramatic declines in consistency, at 27 percent, 26 percent and 21 percent, respectively. By the age of 85 and over, only half of respondents in Poland experience consistency between their self-reported and tested mobility.

Tab. 2: Descriptive summary of %consistency between tested and reported mobility, by survey, country, gender and age

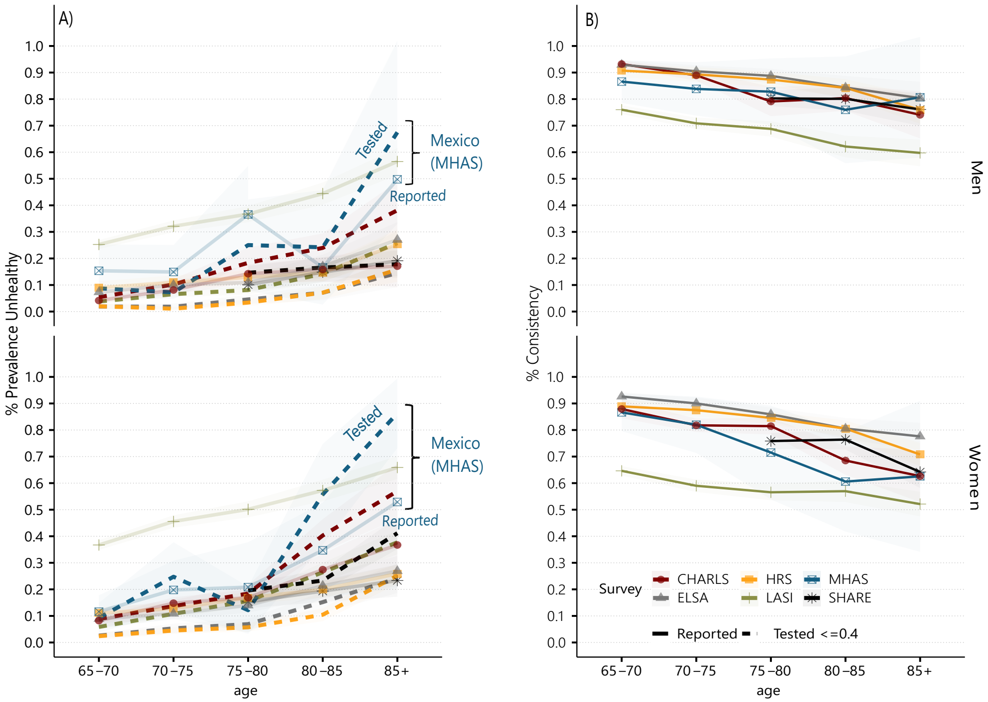
Survey/ Country	Total %	Gender Women %	Men	Age 65-70 %	70-75	75-80	80-85	85+
HRS								
USA	85.8	84.4	87.6	89.7	88.3	85.8	82.1	72.7
SHARE								
Europe	76.0	73.8	79.7	-	-	77.7	77.7	67.7
Austria	72.4	70.5	75.2	-	-	77.3	72.7	57.3
Belgium	80.1	77.2	84.7	-	-	84.4	77.4	72.5
Czechia	81.2	79.4	83.8	-	-	83.5	80.8	69.1
Denmark	84.2	82.4	87.0	-	-	88.2	81.6	80.1
France	76.2	74.0	80.1	-	-	78.3	76.9	69.5
Germany	80.0	78.0	83.6	-	-	80.9	83.0	68.8
Greece	64.2	60.4	68.6	-	-	67.5	66.2	53.1
Ireland	72.4	76.1	66.7	-	-	71.2	74.3	70.5
Israel	77.9	75.3	80.8	-	-	77.9	75.3	82.6
Italy	73.8	70.3	80.0	-	-	74.2	79.2	59.2
Netherlands	85.9	87.0	84.3	-	-	87.7	87.9	78.3
Poland	69.2	61.9	79.1	-	-	69.9	75.6	51.3
Spain	65.7	63.7	68.5	-	-	69.0	61.3	64.5
Sweden	87.6	86.3	89.6	-	-	90.3	90.0	79.5
Switzerland	86.2	83.2	91.9	-	-	89.0	87.4	79.6
ELSA								
England	88.5	87.6	89.5	92.8	90.2	87.2	82.3	78.7
CHARLS								
China	84.6	81.7	87.4	90.6	85.5	80.2	74.0	66.6
MHAS								
Mexico	80.7	78.2	83.2	86.6	82.9	77.1	69.7	69.6
LASI								
India	66.0	60.4	71.5	70.4	65.0	62.5	59.5	55.6

Note: %Consistency is reached if survey participants are healthy according to both tested and reported measures, or unhealthy according to both of the measures. For mobility, this is agreement between reporting some difficulty in walking 100m and having an average walking speed that is 0.4 m/s or lower. We do not differentiate between negative consistency, i.e. both measures suggest health limitations, and positive consistency, i.e. both measures suggest no health limitations. All values consider complex survey design and are weighed accordingly.

Source: Own calculations based on the Gateway to Global Aging data.

As shown in Figure 1, the age-related decline in consistency (Panel B) can be attributed to the distinct age-related patterns of tested health (dashed lines Panel A) and the self-reported health measure (solid lines Panel A) (See Appendix Fig. A1-A2 for country-specific figures). The dashed lines (tested health) present a sharper increase by age compared with the solid lines (self-reported health). This suggests that mobility-losses over the life cycle are more severe according to the tested measure than to the self-reported measure. This discrepancy is particularly

Fig. 1: %Prevalence (Panel A) and %consistency (Panel B) between reported and tested mobility, by age, gender and survey



Note: %Prevalence (Panel A) is weighted and considers complex survey design for each survey. %Consistency (Panel B) is reached if survey participants are healthy according to both tested and reported measures, or unhealthy according to both of the measures. We do not differentiate between negative consistency, i.e. both measures suggest health limitations, and positive consistency, i.e. both measures suggest no health limitations.

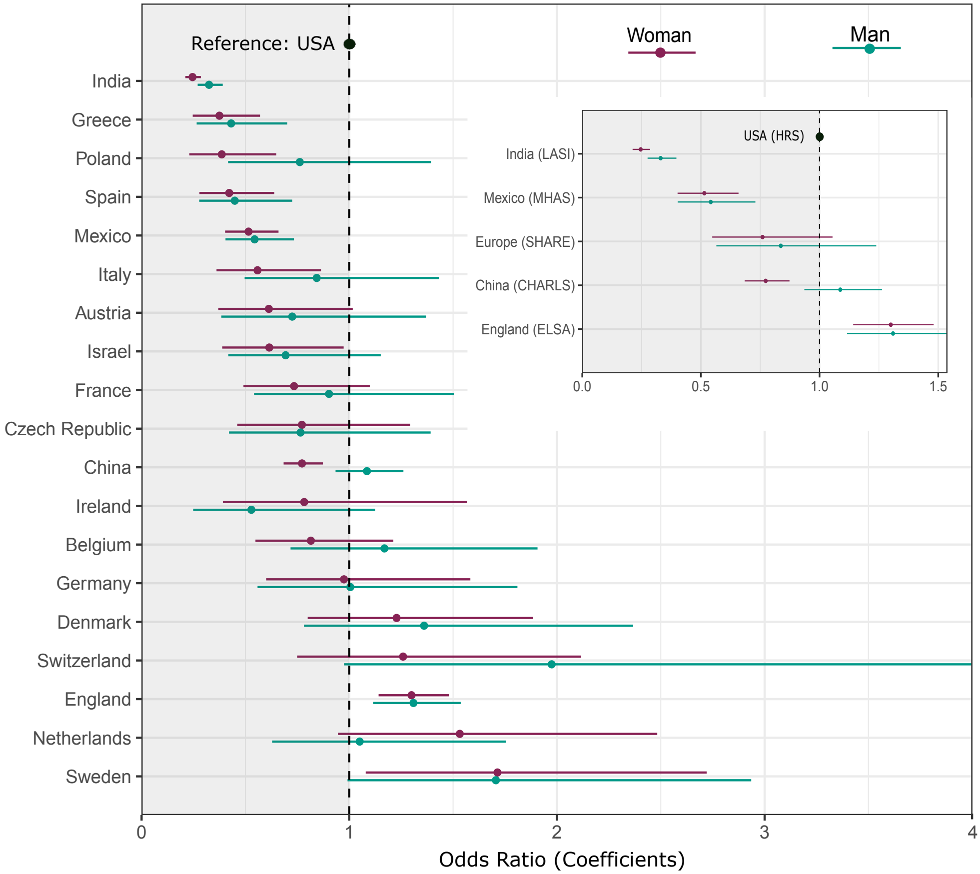
Source: Own calculations based on the Gateway to Global Aging data.

pronounced in Mexico – highlighted in the figure – and China, but also in some European countries like Greece and Italy (see Fig. A1 in the Appendix for all countries). India stands out as having the highest prevalence of self-reported health until the age group of 80-85 and the lowest consistency across all ages, compared to all other surveys.

3.1.2 Regression results

The above results show a clear decline in consistency with age. This age gradient is likely driving country differences in consistency and possibly between women and men. It is thus important to perform comparisons across countries while controlling for difference in the age composition. We thus fit regression models controlling for age and survey period. This enables us to examine the relationship between country and consistency in mobility while keeping age constant. Figure 2 presents the estimated odds ratio for consistency, categorised by country and stratified by

Fig. 2: Odds ratio coefficients from regression models for mobility, by gender, country and survey



Note: Models adjusted for women and men separately. Reference value for the odds ratio is the United States (HRS) and the countries are shown in ascending order based on the estimates for women. The threshold for unhealthy in the walking speed test is 0.4m/s or less. We do not differentiate between negative consistency, i.e. both measures suggest health limitations, and positive consistency, i.e. both measures suggest no health limitations.

Source: Own calculations based on the Gateway to Global Aging data.

gender (See regression table with all coefficients in Table A1 in the Appendix). The countries are shown in ascending order based on the estimates for women. We use the United States as a reference since the HRS is the basis for all other sister-studies.

Supporting the descriptive analysis presented in Table 1, India has significantly lower odds of consistency for both women and men when compared to the United States. Nevertheless, this is closely followed by countries in Europe. Greece, Spain, and Poland (women only) had lower consistency odds than the United States (61 percent, 57 percent, and 48 percent, respectively). In contrast, Sweden has the highest level of consistency among all countries, regardless of gender. Specifically, the odds of Swedes displaying consistency is 68 percent higher than those of North Americans. This finding supports the significant variation in consistency among European nations in the SHARE survey. Moreover, this heterogeneity persists even after controlling for age composition and survey year. However, these results must be interpreted with caution, since for mobility SHARE countries only have available data for ages 75+.

For the model adjusted for both women and men together and controlling for age and survey year, women have 34 percent lower odds of being consistent for mobility when compared to men, most likely an effect driven by the aforementioned countries (see Table A1 in the Appendix).

Depicted in the top plot located in the upper right section of the figure are the odds for the whole surveys. India (LASI) and Mexico (MHAS) are the countries/surveys with the lowest odds of consistency when compared to the United States (HRS). Although the odds of consistency are comparatively lower in the SHARE survey for Europe when compared to the HRS survey for both women and men, the difference is not significant. When analysing the odds by survey only, ELSA has the highest odds of consistency across all surveys for mobility, when compared to the HRS study, while the lower-income regions have a much lower probability of achieving consistency.

3.1.3 Robustness analyses

To ensure that the results are not driven by the cutoff of walking speed at lower or equal to 0.4m/s, we ran the analysis using a higher cutoff of 0.6m/s. When using a higher cutoff mark, it is expected that a larger share of respondents will be reclassified as unhealthy, when compared to the 0.4m/s cutoff. We investigate whether this impacts the odds of consistency. Tables A3-A4 in the Appendix show that our main conclusions do not change when using different walking speed thresholds. The share that achieves consistency is lower overall for the 0.6m/s, but the direction of the effect by age is the same as with 0.4m/s (consistency declines with age). Gender remains significant in the overall regression model and does not change direction, i.e., women have lower odds of consistency when compared to men. In addition, country-specific effects do not change in ranking and direction, but only the level of consistency.

3.2 Cognition

3.2.1 Descriptive results

For cognition, there is lower consistency between tested and self-reported measure relative to mobility for all countries. Keep in mind, however, that the overall level of consistency depends on the thresholds set for dichotomising the tested and the self-reported outcome (see Table A3 of the Appendix for a comparison). In addition, the set of countries and waves used for mobility and cognition are different, due to previously mentioned data availability issues. Hence, comparing the level of consistency in cognition with that of mobility is not informative. By contrast, country and gender differences as well as age patterns of consistency can be compared between the two health dimensions. As shown in Table 3, consistency is particularly low for the countries with the lowest income, namely Mexico (48 percent) and

Tab. 3: Descriptive summary of %consistency between tested and reported cognition, by survey, country, gender and age

Survey/ Country	Total %	Gender %		Age %				
		Women	Men	65-70	70-75	75-80	80-85	85+
HRS								
USA	69.0	69.7	68.0	73.4	70.0	67.0	64.9	63.1
SHARE								
Europe	66.5	66.1	66.9	71.4	66.0	62.8	64.4	65.0
Austria	74.8	75.6	73.6	80.6	77.0	71.6	69.6	67.4
Belgium	67.4	66.0	69.2	74.6	69.8	63.2	63.1	60.1
Czechia	66.8	64.4	70.2	72.3	69.1	63.5	55.5	61.1
Denmark	75.6	75.7	75.6	84.1	77.0	68.8	66.8	66.5
Estonia	51.3	50.8	52.5	54.8	48.9	46.2	49.6	64.3
France	63.1	61.5	65.5	69.2	64.6	60.6	58.6	58.7
Germany	70.4	71.6	69.0	75.9	71.2	66.9	69.4	65.0
Hungary	58.9	55.5	64.5	62.8	53.9	52.4	59.8	70.6
Israel	71.4	73.8	68.6	75.8	66.4	67.9	68.7	76.4
Italy	64.0	63.9	64.2	68.1	60.5	58.3	64.9	71.0
Luxembourg	73.4	72.0	75.0	77.6	78.2	76.8	66.7	57.7
Netherlands	74.2	75.5	72.7	81.8	73.3	70.5	71.7	62.4
Poland	60.4	58.3	63.9	66.7	53.8	64.9	56.4	57.7
Portugal	65.7	63.5	68.7	67.0	75.2	48.1	68.7	72.3
Slovenia	66.1	66.7	65.2	72.5	62.7	69.0	60.2	58.6
Spain	62.9	63.0	62.7	62.7	58.7	60.5	66.2	70.6
Sweden	65.1	65.0	65.4	72.2	64.1	61.5	57.0	62.2
Switzerland	76.1	75.1	77.4	85.3	78.5	72.4	66.0	66.2
ELSA								
England	63.2	64.4	61.8	67.5	62.6	61.5	58.0	61.1
CHARLS								
China	59.8	63.4	56.2	51.6	57.1	65.5	75.2	86.7
MHAS								
Mexico	48.0	46.0	50.1	43.2	43.9	47.5	56.3	69.2

Note: %Consistency is reached if survey participants are healthy according to both tested and reported measures, or unhealthy according to both of the measures. For cognition, this is agreement between reporting “poor” and “fair” memory and recalling 3 words or less in the immediate word recall test. We do not differentiate between negative consistency, i.e. both measures suggest health limitations, and positive consistency, i.e. both measures suggest no health limitations. All values consider complex survey design and are weighed accordingly.

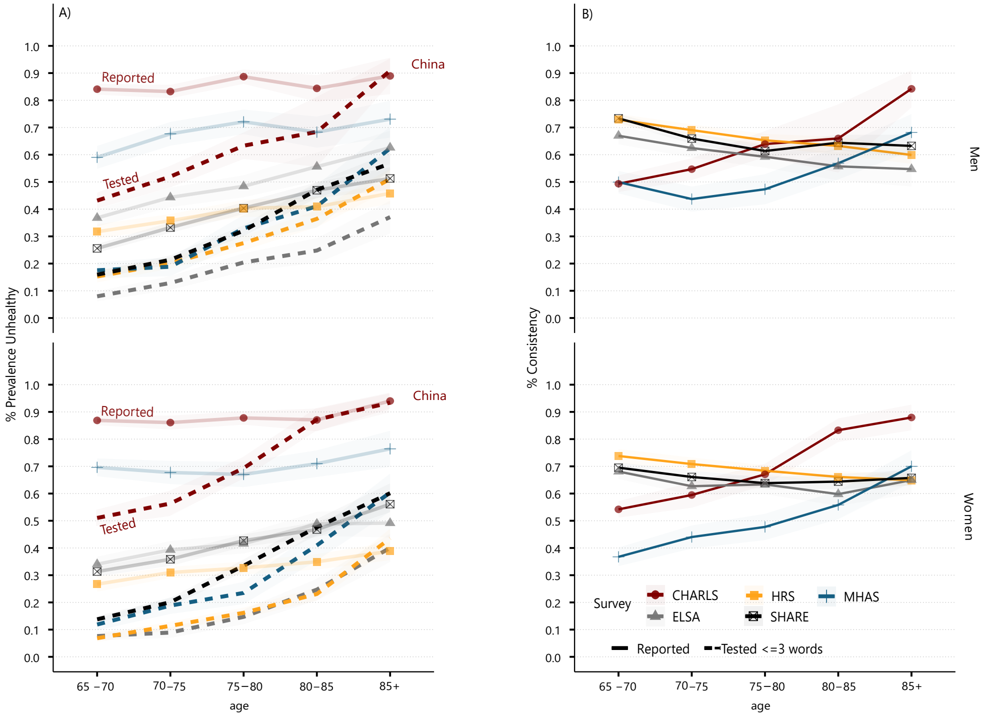
Source: Own calculations based on the Gateway to Global Aging data.

China (59 percent). Among European countries, Estonia (51 percent) has the lowest consistency, followed by Hungary (58 percent).

As indicated by the dashed lines in Panel A of Figure 3, the age gradient for tested health is steeper than the age gradient for reported health in all countries and regardless of gender, and is even steeper than the age gradient for mobility.

In general, consistency also declines with age, but not as strongly as what is observed with mobility. In addition, there are two major exceptions: China and Mexico. In these two countries, consistency increases with age, as shown in Panel B, Figure 3. There is a convergence between self-reported memory and immediate word recall test as individuals age. This convergence is particularly evident in China, where a distinct "wishbone" pattern is observed. The differential in pace by age between the self-reported and the tested measure is so great that it leads to an increase in consistency over the life course, because the tested measure slowly catches up

Fig. 3: %Prevalence (Panel A) and %consistency (Panel B) between reported and tested cognition, by age, gender and survey



Note: %Prevalence (Panel A) is weighted and considers complex survey design for each survey. %Consistency (Panel B) is reached if survey participants are healthy according to both tested and reported measures, or unhealthy according to both of the measures. The threshold for unhealthy in the immediate word recall test is three words or less. We do not differentiate between negative consistency, i.e. both measures suggest health limitations, and positive consistency, i.e. both measures suggest no health limitations.

Source: Own calculations based on the Gateway to Global Aging data.

with the self-reported measure. It should be noted again that the prevalence of unhealthy and thus the level of the self-reported line in Panel A of Figure 3 depends on the exact cutoff. Hence, the main take-away message here is not the decline or increase in consistency with age, but instead the substantial differences in the age pattern between the tested and the self-reported measure. Particularly, the steeper age gradient in the tested indicator compared to the self-reported indicator.

3.2.2 Regression results

Similar to mobility, we fit regression models controlling for age and survey period to account for differences in age structure across countries and perform more robust comparisons. It is important to note that the set of countries is different for cognition. Waves 4 and 5 of the SHARE surveys in the European region include more countries than Waves 1 and 2 used for mobility. Consequently, we also include Slovenia, Estonia, Hungary, Portugal and Luxembourg. However, neither Greece nor Ireland are included. Due to the absence of self-reported memory in the LASI survey, India could not be included in the analysis.

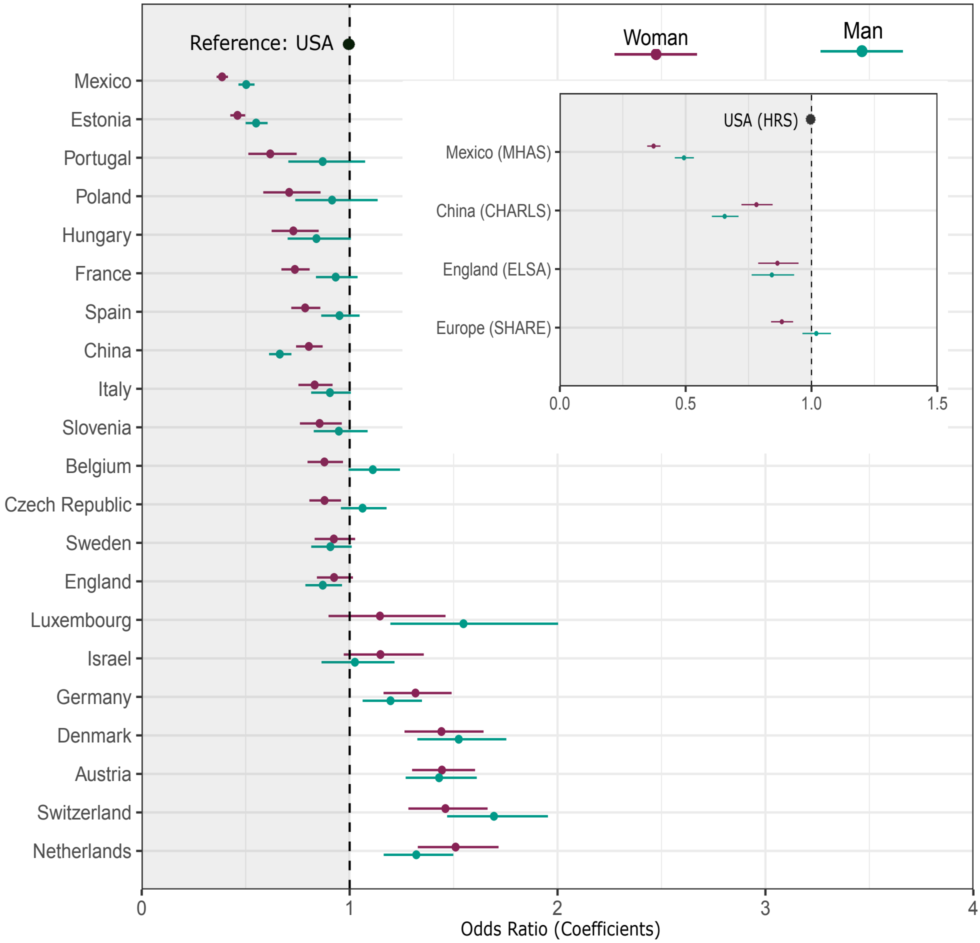
When grouping all SHARE countries (see top right plot in Figure 4), it is again the lower-income countries that show the lowest consistency. However, when looking at all European countries separately, the picture is less clear. Of all countries, women and men in Mexico have the lowest consistency in cognition when compared to women and men in the United States. This figure is closely followed by women and men in Estonia. Interestingly, the heterogeneity across European countries is smaller than for mobility. The odds ratio for each country is closer in value than what was observed for mobility. In addition, contrary to mobility, England has lower consistency odds than the United States, while other countries in Europe like Germany, Denmark, Austria, Switzerland and the Netherlands have significant higher odds for both genders.

In addition, the model that incorporates gender as a control variable together with age and survey year shows no significant difference in consistency between women and men, possibly due to this gender variability between countries (Table A2 in the Appendix).

3.2.3 Robustness analyses

For additional analysis regarding tested cognition, we apply a different threshold of recalling two words or less, instead of 3 words or less. Tables A5-A6 summarise the results for the descriptive findings and the regression results. Despite the expected changes in consistency levels, the direction of our main conclusions still holds. For the majority of countries, women have higher consistency between reported and tested health. However, the regression model controlling for country, gender, age and survey year shows no significant effect for gender. The age pattern in consistency is J-shaped, where consistency increases at very old ages (85+). China and Mexico are particular cases where consistency increases dramatically with age, regardless of the threshold for tested cognition.

Fig. 4: Odds ratio coefficients from regression models for cognition, by gender, country and survey



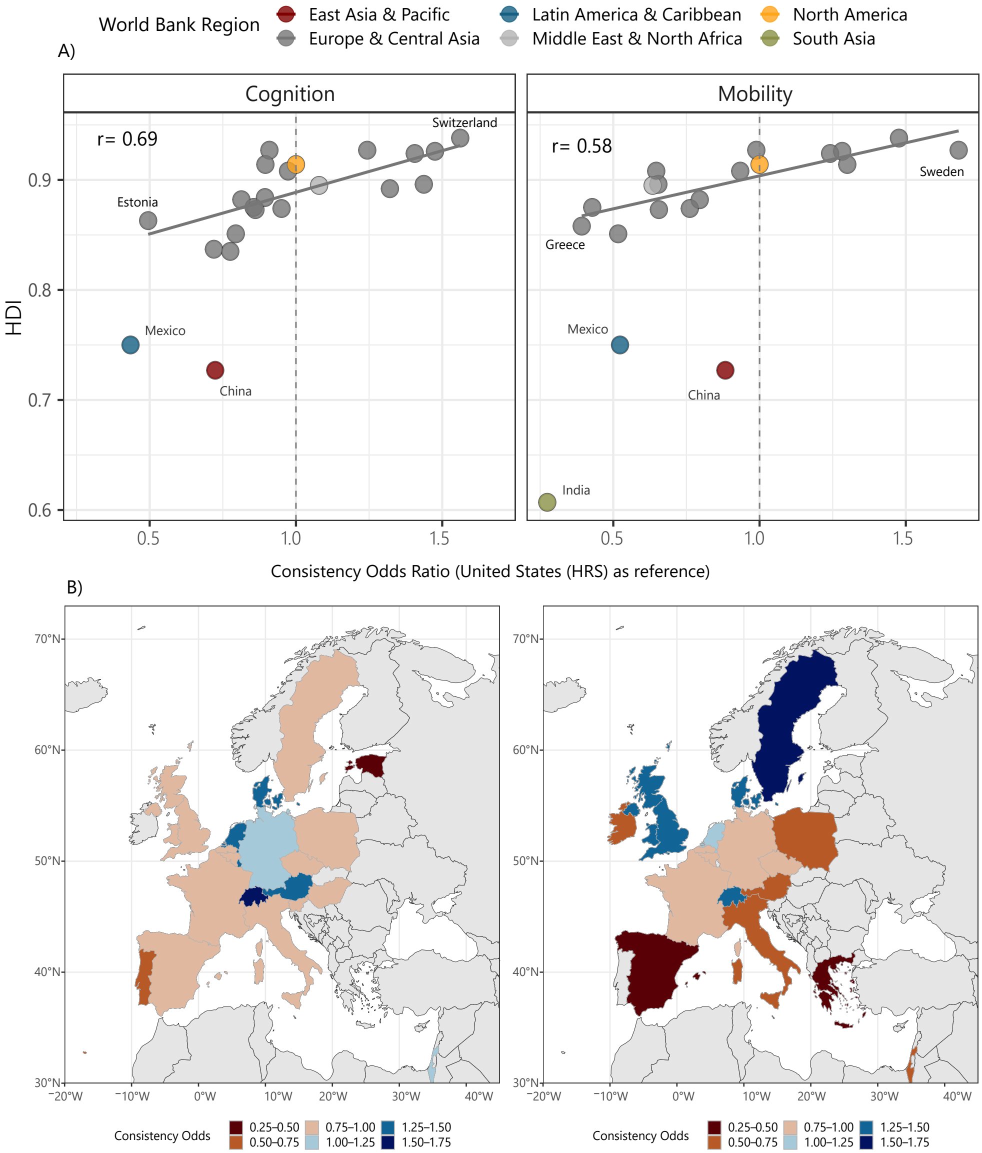
Note: Models are adjusted for women and men separately, controlling for age and survey year. Coefficients are exponentiated so this is odds of consistency between tested and reported health. The country of reference is the United States (HRS survey). The threshold for unhealthy in the immediate word recall test is three words or less. We do not differentiate between negative consistency, i.e. both measures suggest health limitations, and positive consistency, i.e. both measures suggest no health limitations.

Source: Own calculations based on the Gateway to Global Aging data.

3.3 Regional heterogeneity and association with the Human Development Index (HDI)

As demonstrated in the preceding sections, there is an overall pattern of low levels of consistency for countries that are part of less affluent regions of the globe, such as Mexico, India and China, when compared to the United States and Europe.

Fig. 5: Association between consistency odds and the Human Development Index (HDI) by World Bank Regions (Panel A) and consistency odds in Europe (Panel B)



Note: Consistency odds ratio retrieved from models controlling of age, survey period and gender. The country of reference is the United States (HRS survey). The regression line in Panel A was generated for SHARE countries only. Middle East & North Africa is represented by Israel. R refers to the Spearman correlation. The threshold for unhealthy in the immediate word recall test is three words or less. For mobility, it is walking speed equal to 0.4m/s or lower. We do not differentiate between negative consistency, i.e. both measures suggest health limitations, and positive consistency, i.e. both measures suggest no health limitations. ELSA and SHARE countries are grouped within Europe and Central Asia region. Israel is part of SHARE but categorised as a Middle East and North Africa in the world bank regions. In the map, Panel (B), we categorised England is shaded as United Kingdom, but the estimates refer only to England. This was done for visualisation purposes only.

Source: Own calculations based on the Gateway to Global Aging data. United Nations Development Programme, Human Development Report. 2022. "Human Development Index." Retrieved from http://hdr.undp.org/en/composite/HDI. World Bank Annual Report 2022.

Nonetheless, it was also evident that the European region is rather heterogeneous when it comes to consistency, especially for mobility. In order to further investigate this aspect, we grouped countries into the World Bank Regions and evaluated their association with the Human Development Index (HDI), a summary indicator of development. Due to the absence of discernible gender disparities in consistency, we used the estimated odds ratio for each country derived from the model that accounts for age, gender, and survey year, but that is not estimated separately for women and men, as shown in Figure 5.

We find that there is a positive and significant association between Human Development Index scores and the odds of consistency for both cognition (r=0.69) and mobility (r=58), as shown in Figure 5 (Panel A). In general, greater consistency odds are positively related with higher HDI scores. The regions of South Asia (India), and Latin America and Caribbean (Mexico) are distinguished from the rest of the regions with lower HDI ratings and odds of consistency when compared to the United States. The East Asia & Pacific (China) region has lower HDI but not necessarily lower consistency odds, especially for mobility, when compared with other countries. Despite the high correlation between the Human Development Index (HDI) and consistency odds, it also serves to underscore the significant variation when looking within the regions of Europe and Central Asia (ELSA and SHARE countries) and Middle East and North Africa (Israel, part of SHARE). For cognition, there is less heterogeneity in consistency across European countries, as seen by the overall similarity in color shades throughout these countries on the map on Panel B of Figure 5 (odds [0.75-1.00]). For mobility, however, heterogeneity in consistency is more pronounced, as indicated by the presence of more contrasting colors (ranging from [0.25-0.50]) to [1.50-1.75]). Switzerland, Denmark, Austria and Germany are among the countries with the highest HDI scores and that also have higher odds of consistency in their tested and reported cognition when compared to the United States.

4 Tested versus reported health: How do discrepancies impact country rankings?

As shown in the above analyses, self-reported mobility and cognition appear do deviate substantially from their tested equivalent. Most importantly, self-reports do not reflect the expected age pattern of increasing health decline over the life course. Differences between the self-reports and the tests are particularly pronounced for countries with lower scores of Human Development Index (HDI), which are indicative of lower levels of socio-economic development. This indicates that self-reports for mobility and cognition are not the most reliable indicators to be used as stand-alone tools for cross-country comparisons in health. To illustrate this issue in more detail, we provide a further overview of how differences between tested and self-reported health impact country rankings for ages 75+. For this, we estimated the %prevalence unhealthy for the whole population and for women and men for each survey and health dimension considered, as shown in Table 4. We then compute country rankings based on the total %self-reported (R) and %tested (T) health prevalence for the total population and for women and men. In addition, because for the case of mobility European countries from the SHARE sample only have available data above age 75, we have restricted the comparisons to those ages only, in order to prevent bias due to considering a younger age structure for the other countries.

Tab. 4: Discrepancies between %self-reported and %tested health prevalence, total population, survey/country and health indicator

For ages 75+ for all countries

Survey	Country	Mobility		Cognition	
		Self-Reported	Tested	Self-Reported	Tested
		%		%	
HRS	USA	18.8	10.5	37.9	30.6
SHARE	Europe	16.3	21.5	46.6	44.4
ELSA	England	17.8	10.7	49.4	25.4
CHARLS	China	19.2	26.5	88.2	74.2
MHAS	Mexico	31.9	37.5	70.7	39.5
LASI	India	49.4	18.6	-	-

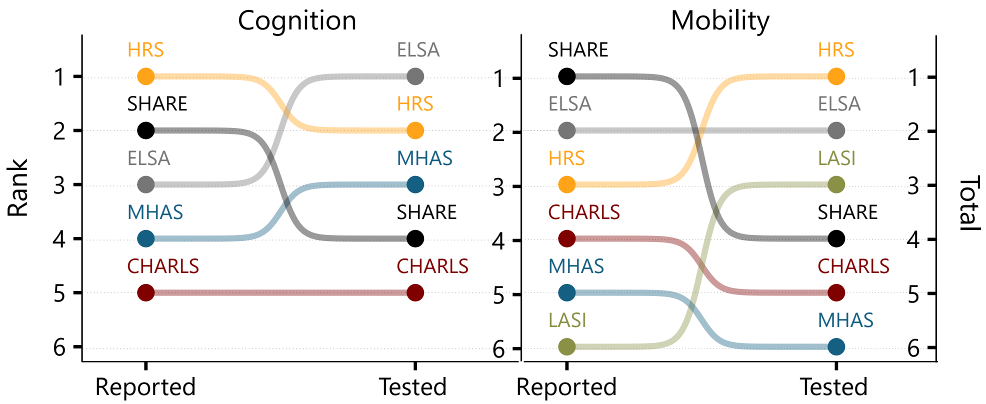
Note: %Prevalence unhealthy for self-reported cognition is reporting "poor" and "fair" memory; for tested cognition is recalling 3 words or less in the immediate word recall test. %Prevalence unhealthy for self-reported mobility is reporting some difficulty walking 100m; for tested mobility is having a walking speed that is 0.4m/s or lower. All values consider complex survey design and are weighed accordingly.

Source: Own calculations based on the Gateway to Global Aging data.

Figure 6 shows the rankings for countries from low to high prevalence according to reported and tested health for the total population. It is important to note that these are absolute health prevalences computed for the total population aged 75 and over. This figure is thus impacted by the age structure of each country. The ranking is merely to illustrate the discrepancies across and within surveys if one looks only at either self-reported health or tested health.

For cognition, HRS (-1 ranking from reported to tested), followed by MHAS (+1 ranking from reported to tested) and CHARLS (no changes in ranking) were the most stable surveys regarding their rankings for reported and tested health. SHARE had a change of 2 positions in the ranking from reported to tested health (from 2nd to 4th). Both for self-reported and tested health, regions with higher development

Fig. 6: Ranking of countries (from low to high prevalence) according to reported and tested health for the total population, ages 75+



Note: %Prevalence unhealthy for self-reported cognition is reporting "poor" and "fair" memory; for tested cognition is recalling 3 words or less in the immediate word recall test. %Prevalence unhealthy for self-reported mobility is reporting some difficulty walking 100m; for tested mobility is having a walking speed that is 0.4m/s or lower. All values consider complex survey design and are weighed accordingly.

Source: Own calculations based on the Gateway to Global Aging data.

levels according to their HDI scores are always ranked in the first two positions (HRS/SHARE- ELSA/HRS).

For mobility, ELSA was the survey that was more stable, with no change in rankings for the total population. The change in reported to tested health is stronger for SHARE (from 1st to 4th place) and LASI (rises three places from 6th to 3rd).

Figure 6 allows us to derive some important messages both across and within surveys. First, within surveys, higher rankings in reported health are not necessarily followed by higher rankings in tested health, which reinforces the discrepancies between these two measures. Second, across surveys, it matters whether one uses reported or tested health to estimate prevalence of unhealthy and compare performance. While SHARE countries fare better than most countries with regards to reported health, this figure changes when one looks at tested health. This discrepancy raises concerns about which indicator is best for comparative purposes.

5 Discussion

We expanded earlier work that evaluated several health surveys using the consistency between reported and tested health to lower-income regions (*Cleary* 1997; *Fors et al.* 2006; *Spitzer/Weber* 2019; *Cislaghi/Cislaghi* 2019; *Huang/Maurer* 2019). A key result is that consistency decreases with age for mobility and has an overall j-shape for cognition, but no significant patterns are found by gender. Tested health has a sharper age gradient than reported health, increasing at a faster pace with age. The inconsistencies found reflect that self-reported health does not adequately capture the deterioration in health that is expected with aging (*di Lego/Sauerberg* 2023), which is in line with earlier research that found strong intertemporal stability in health reports despite objective health decline (*Galenkamp et al.* 2012; *Spuling et al.* 2017). The lack of gender differences also correspond to previous work that found no significant differences between women and men in the response behaviour regarding self-rated health, even after adjusting for reporting behaviour, but did find significant differences by age (*Oksuzyan et al.* 2019; *Spitzer/Weber* 2019; *Lazarevič/Brandt* 2020; *Lazarevič/Quesnel-Vallée* 2023). Because we are comparing countries with very different gender norms and epidemiological trajectories, this indicates that inconsistencies between reported and tested health usually go in the same direction, at least for the health dimensions we analysed.

Interestingly, we found a positive and significant association between Human Development Index scores and the odds of consistency for both cognition and mobility. In general, greater consistency odds were positively related with higher HDI scores, but important differences within the European region remained, with a large variation in consistency between reported and tested health. This is in line with previous work that documented large discrepancies in Europe between reported and tested health and merits further investigation (*Spitzer/Weber* 2019). Overall, this may be reflecting different welfare state systems and access to health care, as research has shown that differences in welfare state regimes within Europe accounted for approximately half of the national-level variation of health inequalities between European countries (*Eikemo et al.* 2008).

Most likely, this is driven by the profiles we observed in self-reported health, which challenge the validity of this indicator for performing cross-country comparisons. Indeed, research has repeatedly shown that reporting behaviour across European nations is so heterogeneous that cross-country comparisons based on self-rated health indicators are questionable (*Pfarr et al.* 2012; *Molina* 2016; *Picavet* 2017; *Luy et al.* 2023). However, self-reports of health information may provide insight into other environmental factors. For the case of mobility, for instance, the extremely high levels of self-reported health limitations for India may be due to adverse environmental conditions, security concerns, lack of accessibility and cultural factors that discourage individuals from going outside (*Gallagher et al.* 2016; *Almeida Bentes et al.* 2017; *Cantuaria et al.* 2023). Since both women and men report a higher prevalence compared to other countries, this may reflect the condition of the pavement and the ability to circulate in the streets rather than gender norms. Such aspects have long been discussed under such terms as "extra-individual factors" (*Verbrugge/Jette* 1994) as important influences on the link between objective health impairments and consequent functional limitations. As these extra-individual factors can be expected to vary widely and in relation to, e.g., the economic welfare of a country, tested health would enable a better measurement of health conditions/clinical aspects since the test is taken in a standardised, controlled environment (*Fors et al.* 2006). As a consequence, comparisons of specific health problems, such as in functioning or cognition, might not only reflect objective differences in prevalence, but also differences in welfare regimes or accessibility.

Other research that focused on comparisons of self-reported disability measures in six low- and middle-income countries (LMICs) that included China, India and Mexico, showed that self-reported data varied greatly due to differences in the inability to perform an activity because of one's environment (*Capistrant et al.* 2014). This is why it has also been suggested that tested health based on performance measures and self-reported measures are complementary measures and that they do not measure the same construct (*Hoeymans et al.* 1996).

Our results on cognition are consistent with country-specific studies that also analyse self-reported and tested memory inconsistencies, suggesting that self-reported memory may not adequately represent memory issues (*Sohel et al.* 2016). A study that focused on China strongly cautioned against relying on self-rated memory measures for performing memory assessments in primary care or survey research, since self-rated memory was only modestly correlated with performance on memory tests (*Huang/Maurer* 2019). All these results could also be interpreted as a "recalibration" of older respondents' standards for reporting a health limitation, meaning that more severe limitations could be seen as less serious in older age (*Schwartz/Sprangers* 1999; *Sprangers/Schwartz* 1999).

Gender differences on memory tests were also in line with what has been shown in previous research, with women performing better than men on memory tests in some contexts like England (*Huppert/Whittington* 1993; *Steel et al.* 2003; *Salthouse* 2016).

This study has some limitations, many of which were addressed in robustness analyses. First, as our goal was to have the most diverse set of countries as possible, the dimensions of health were restricted to the variables available across surveys. To ensure an exact match between self-reported health information and tested health, we also had to choose particular metrics within each of the health dimensions. Our findings are therefore limited to mobility (as assessed by the 100-meter walking test difficulty and the walking speed test) and cognition (as assessed by the self-reported memory and the word recall test) and cannot necessarily be generalised beyond that. However, difficulty walking short to medium distances is not only an important marker for disability, but also a key predictor for mortality, cardiovascular risk and health care utilisation (*Lan et al.* 2002; *Newman et al.* 2006; *Hardy et al.* 2011). Likewise, gait speed is a simple yet important clinical marker of current health and well-being and a powerful predictor of mortality in older adults, often being referred to as a "vital sign" of overall health and well-being (*Studenski et al.* 2011; *Peel et al.* 2013; *White et al.* 2013). Hence, despite our analysis being restricted to this aspect of mobility, it is likely a good candidate for capturing physical limitation.

Another limitation is the fact that SHARE countries only have ages 75+ for mobility, which means that SHARE has an older age structure compared to other countries. To mitigate this limitation, in the case of country rankings where we analyse absolute prevalence, we restricted the analysis to ages 75+. In the case of regression models, we performed robustness checks with models for mobility with ages 75+ only. Overall, the impact is not as important as it might have seemed due to the strong age gradient in health and the fact that the prevalence for both reported and tested health in mobility and cognition increase, while consistency declines with age.

The fact that we did not find important gender differences may also be reflective of these particular health dimensions in our analyses, since other work has shown that women rate for chronic health conditions in a less consistent way than men (*Deeg/Kriegsman* 2003; *Idler* 2003), and that important differences are found when different dimensions of health are considered (*Luy/Di Giulio* 2006; *di Lego et al.* 2020a/b).

Further, there were also some more specific deviations in the procedures in individual surveys that reduced their comparability. For one of the countries, India, there was only one wave available. The Mexican survey also used slightly different procedures, compared to the other surveys, both for walking speed test and cognition. For cognition, the immediate word recall list is composed of eight words instead of ten words. Likewise, the walking speed test is performed over a 3m-walking course in Mexico, instead of 2.5m like the other countries. However, as we are concerned with the health of impaired individuals and the lower bound – i.e., those who are most impaired – of both walking speed and memory, these differences should not impact our analyses. In this case, the cutoff point we use, which is the same for all countries, can be interpreted as a lower bound for measuring functional limitations and cognition. What could indeed impact the results are the thresholds in the proportion that is considered unhealthy. For that reason, to ensure that the results are not driven by the cutoff of walking speed at lower or equal to 0.4m/s, we ran the analysis using a higher cutoff of 0.6m/s. When using a higher cutoff mark, it is expected that a larger share of respondents will be reclassified as unhealthy, when compared to the 0.4m/s cutoff. The conclusions were not changed with different cutoff marks. Lastly, when dropping Mexico from the analyses the odds ratios and significance were very close and in the same direction, indicating that the cross-country comparisons are not affected. Unfortunately, it is not possible to assess whether this difference in scale could impact within-country differences in Mexico. As we mentioned, since we focus on impairment for this paper, this can mitigate the difference in scale. With all the robustness checks we performed, it is unlikely that the exceptionality in Mexico's tested health scale explains the steep age gradient observed for walking speed.

Finally, the discrepancies we find between self-reported and tested health may raise the question to researchers and policy makers alike: which indicator should I use or is more reliable to capture health? Health is multidimensional and complex and that is why it is so hard to have a single indicator that can accurately capture health (*Mathers et al.* 1994; *Robine/Jagger* 2003; *Costa et al.* 2019; *Santos et al.* 2021). Indeed, the World Health Organization (WHO) has a global reference list of 100 core health indicators they categorise as: 1. health status indicators, 2. risk factors indicators, 3. Service coverage indicators and 4. health systems indicators (*World Health Organization* 2018). The WHO defines health as "a state of complete physical, mental, and social well-being, and not merely the absence of disease or infirmity" (*World Health Organization* 2005). If we consider this definition, then self-reported health can also be capturing aspects related to social and mental well-being that are not restricted to the health condition itself, but to the environment one lives in. Indeed, reported health has been shown to give clues as to the environmental limitations, cultural aspects and overall well-being (*Bombak* 2013; *Gallagher et al.* 2016).

Similarly, people-centred health systems, such as the Patient-Reported Indicators Surveys (PaRIS) launched in 2017 by the Organization for Economic Cooperation and Development (OECD), require indicators centred on people's well-being. The OECD Health Ministers developed the PaRIS in order to systematically collect data on what matters most to patients, and not only their tested health, since not all health conditions are equally felt by individuals (*OECD* 2019). In this case, using indicators that also reflect overall well-being, such as reported measures, are more appropriate, besides being inexpensive to include in national surveys (*Strawbridge/Wallhagen* 1999; *Bombak* 2013).

In other health contexts, such as critical phases of the COVID-19 pandemic, it was crucial to have tested health data and objective metrics to accurately assess infection rates and make proper decisions (*di Lego et al.* 2022). Likewise, in contexts of hospital care demand planning and forecasting hospital beds, it may also be important to have performance-based measurement (*Ravaghi et al.* 2020). In other words, there is no single best health indicator, but the most appropriate depending on the purpose of a policy, and whether it is a monitoring, preventive, targeted or action-based policy. The important thing is to be aware of how sensitive some indicators are to make better decisions on which indicator to use, instead of taking them for granted. The goal of this paper was to elucidate this sensitivity when performing cross-country comparisons, where usually the aim is to rank countries and assess which country is healthier. In this scenario, tested health is a more robust choice, particularly when comparisons are performed across countries from different levels of socio-economic development.

6 Conclusion

Health surveys are important tools used by policy makers to monitor health and establish health goals and national strategies. Self-reported health, usually derived from health surveys, is an important source of information that researchers use to assess health across populations (*Wong et al.* 2005; *Jürges* 2007; *Layes et al.* 2012; *Boerma et al.* 2016; *Galenkamp et al.* 2020). However, the inconsistency we find indicates that self-reports, at least when considering mobility and cognition, are not the most reliable to be used alone for those assessments, especially when performing cross-country comparisons.

Tested health, on the other hand, seems to be more realistically capturing the deterioration of health that happens with age in a standardised, comparable way. They are more expensive and complex to incorporate in health surveys, but there are performance tests such as gait speed, word recalls or chair stand, which can be employed as cost-effective measures for high quality survey research (*Christensen et al.* 2013; *Peel et al.* 2013).

While there are no perfect measures for health, tested health seems like a better candidate to base cross-country comparisons on, especially when comparing a highly diverse set of countries as in this study.

Acknowledgements
This work is supported within the EU Framework Programme for Research and Innovation Horizon 2020, ERC Grant Agreement No. 725187 (LETHE).

This paper uses data from SHARE Waves 1, 2, 4 and 5 (DOIs: 10.6103/SHARE.w1.800, 10.6103/SHARE.w2.800,10.6103/SHARE.w4.800,10.6103/SHARE.w5.800)see (*Börsch-Supan et al.* 2013) for methodological details. The SHARE data collection has been funded by the European Commission, DG RTD through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812), FP7 (SHARE-PREP: GA N°211909, SHARE-LEAP: GA N°227822, SHARE M4: GA N°261982, DASISH: GA N°283646) and Horizon 2020 (SHARE-DEV3: GA N°676536, SHARE-COHESION: GA N°870628, SERISS: GA N°654221, SSHOC: GA N°823782, SHARE-COVID19: GA N°101015924) and by DG Employment, Social Affairs & Inclusion through VS 2015/0195, VS 2016/0135, VS 2018/0285, VS 2019/0332, and VS 2020/0313. Additional funding from the German Ministry of Education and Research, the Max Planck Society for the Advancement of Science, the U.S. National Institute on Aging (U01_AG09740-13S2, P01_AG005842, P01_AG08291, P30_AG12815, R21_AG025169, Y1-AG-4553-01, IAG_BSR06-11, OGHA_04-064, HHSN271201300071C, RAG052527A) and from various national funding sources is gratefully acknowledged (see www.share-project.org).

This analysis uses data or information from the following Harmonized datasets: LASI dataset and Codebook, Version A.2 as of October 2021, developed by the Gateway to Global Aging Data (DOI: https://doi.org/10.25549/h-lasi). The development of the Harmonized LASI was funded by the National Institute on Aging (R01 AG042778, 2R01 AG030153, 2R01 AG051125). CHARLS dataset and Codebook, Version D as of June 2021 developed by the Gateway to Global Aging Data. The development of the Harmonized CHARLS was funded by the National Institute on Aging (R01 AG030153, RC2 AG036619, R03 AG043052). ELSA dataset and Codebook, Version G.2 as of July 2021 developed by the Gateway to Global Aging Data. The development of the Harmonized ELSA was funded by the National Institute on Aging (R01 AG030153, RC2 AG036619, R03 AG043052). SHARE dataset and Codebook, Version F as of June 2022 developed by the Gateway to Global Aging Data. The development of the Harmonized SHARE was funded by the National Institute on Aging (R01 AG030153, RC2 AG036619, R03 AG043052). MHAS dataset and Codebook, Version B.4 as of February 2022 developed by the Gateway to Global Aging Data in collaboration with the MHAS research team. The development of the Harmonized MHAS was funded by the National Institute on Aging (R01 AG030153). The Harmonized MHAS data files and documentation are public use and available at www.MHASweb.org. The MHAS (Mexican Health and Aging Study) receives support from the National Institutes of Health/National Institute on Aging (R01 AG018016) in the United States and the Instituto Nacional de Estadística y Geografía (INEGI) in Mexico. HRS dataset and

Codebook, Version C as of January 2022 developed by the Gateway to Global Aging Data and randhrs1992-2018v1. The development of the Harmonized HRS was funded by the National Institute on Aging (R01 AG030153, RC2 AG036619, 1R03AG043052). For more information about the Harmonization project, please refer to https://g2aging.org/.

References

Abadie, Alberto et al. 2022: When Should You Adjust Standard Errors for Clustering? In: The Quarterly Journal of Economics 138,1: 1-35. https://doi.org/10.1093/qje/qjac038

Almeida, Bentes Aline et al. 2017: Self-rated health and perceived violence in the neighborhood is heterogeneous between young women and men. In: BMC Public Health 17,967. https://doi.org/10.1186/s12889-017-4969-1

Arni, Patrick et al. 2021: Biased health perceptions and risky health behaviors – Theory and evidence. In: Journal of Health Economics 76,102425. https://doi.org/10.1016/j.jhealeco.2021.102425

Barsky ,Arthut J.; Cleary, Paul D.; Klerman, Gerald L. 1992: Determinants of perceived health status of medical outpatients. In: Social Science & Medicine 34,10: 1147-1154. https://doi.org/10.1016/0277-9536(92)90288-2

Bergmann, Michael et al. 2019: Survey participation in the Survey of Health, Ageing and Retirement in Europe (SHARE), Wave 1-7. Based on Release 7.0.0. SHARE Working Paper Series 41-2019. Munich. SHARE-ERIC.

Bland, J. Martin; Altman, Douglas G. 1986: Statistical methods for assessing agreement between two methods of clinical measurement. In: The Lancet 327,8476: 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8

Boerma, Ties et al. 2016: A global assessment of the gender gap in self-reported health with survey data from 59 countries. In: BMC Public Health 16. https://doi.org/10.1186/s12889-016-3352-y

Bombak, Andrea E. 2013: Self-Rated Health and Public Health: A Critical Perspective. In: Frontiers Public Health 1. https://doi.org/10.3389/fpubh.2013.00015

Börsch-Supan, Axel 2022a: Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 1. Release version: 8.0.0. SHARE-ERI. Data set. https://doi.org/10.6103/SHARE.w1.800

Börsch-Supan, Axel 2022b: Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 2. Release version: 8.0.0. SHARE-ERIC. Data set. https://doi.org/10.6103/SHARE.w2.800

Börsch-Supan, Axel 2022c: Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 4. Release version: 8.0.0. SHARE-ERIC. Data set. https://doi.org/10.6103/SHARE.w4.800

Börsch-Supan, Axel 2022d: Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 5. Release version: 8.0.0. SHARE-ERIC. Data set. https://doi.org/10.6103/SHARE.w5.800

Börsch-Supan, Axel et al. 2013: Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). In: International Journal of Epidemiology 42,4: 992-1001. https://doi.org/10.1093/ije/dyt088

Cantuaria, Manuella L.; Brandt, Jørgen; Blanes-Vidal, Victoria 2023: Exposure to multiple environmental stressors, emotional and physical well-being, and self-rated health: An analysis of relationships using latent variable structural equation modelling. In: Environment Research 227:115770. https://doi.org/10.1016/j.envres.2023.115770

Capistrant, Beatrix D.; Glymour, Maria M.; Berkman, Lisa F. 2014: Assessing Mobility Difficulties for Cross-National Comparisons: Results from the World Health Organization Study on Global Ageing and Adult Health. In: Journal of the American Geriatrics Society 62,2: 329-335. https://doi.org/10.1111/jgs.12633

Cawthon, Peggy M. et al. 2021: What Cut-Point in Gait Speed Best Discriminates Community-Dwelling Older Adults With Mobility Complaints From Those Without? A Pooled Analysis From the Sarcopenia Definitions and Outcomes Consortium. In: The Journals of Gerontology Series A 76,10: e321-e327. https://doi.org/10.1093/gerona/glab183

Chandra, Subhash; Kumar, Rajesh; Gautam, Sarita 2022: Correlation between Human Development Indices and Epidemiological Transition Ratio among Indian States. In: International Journal of Health Sciences and Research 12,11: 78-85. https://doi.org/10.52403/ijhsr.20221112

Christensen, Kaare et al. 2013: Physical and cognitive functioning of people older than 90 years: A comparison of two Danish cohorts born 10 years apart. In: The Lancet 382,9903: 1507-1513. https://doi.org/10.1016/S0140-6736(13)60777-1

Cislaghi, Beniamino; Cislaghi, Cesare 2019: Self-rated health as a valid indicator for health-equity analyses: evidence from the Italian health interview survey. In: BMC Public Health 19:533. https://doi.org/10.1186/s12889-019-6839-5

Cleary, Paul D. 1997: Subjective and objective measures of health: Which is better when? In: Journal of Health Services Research & Policy 2,1: 3-4. https://doi.org/10.1177/135581969700200102

Costa, Claudia et al. 2019: Evaluation of data availability on population health indicators at the regional level across the European Union. In: Population Health Metrics 17. https://doi.org/10.1186/s12963-019-0188-6

Deeg, Dorly J. H.; Kriegsman, Didi M. W. 2003: Concepts of self-rated health: Specifying the gender difference in mortality risk. In: The Gerontologist 43,3: 376-386. https://doi.org/10.1093/geront/43.3.376

di Lego, Vanessa 2021: Health expectancy indicators: what do they measure? In: Cadernos Saúde Coletiva 29. https://doi.org/10.1590/1414-462x202199010376

di Lego, Vanessa; Di Giulio, Paola; Luy, Marc 2020a: Gender Differences in Healthy and Unhealthy Life Expectancy. In: *Jagger, Carol et al.* (Eds.): International Handbook of Health Expectancies. Cham: Springer: 151-172. http://dx.doi.org/10.1007/978-3-030-37668-0_11

di Lego, Vanessa; Lazarevič, Patrick; Luy, Marc 2020b: The Male-Female Health-Mortality Paradox. In: Encyclopedia of Gerontology and Population Aging. Cham: Springer: 1-8. http://dx.doi.org/10.1007/978-3-319-69892-2_798-2

di Lego, Vanessa; Sánchez-Romero, Miguel; Prskawetz, Alexia 2022: The impact of COVID-19 vaccines on the Case Fatality Rate: The importance of monitoring breakthrough infections. In: International Journal of Infectious Diseases 119: 178-183. https://doi.org/10.1016/j.ijid.2022.03.059

di Lego, Vanessa; Sauerberg, Markus 2023: The Sensitivity of the Healthy Life Years Indicator: Approaches for Dealing with Age-Specific Prevalence Data. In: Comparative Population Studies 48. https://doi.org/10.12765/CPoS-2023-06

Eikemo, Terje Andreas et al. 2008: Welfare state regimes and differences in self-perceived health in Europe: A multilevel analysis. In: Social Science & Medicine 66,11: 2281-2295. https://doi.org/10.1016/j.socscimed.2008.01.022

Fors, Stefan; Thorslund, Mats; Parker, Marti G. 2006: Do actions speak louder than words? Self-assessed and performance-based measures of physical and visual function among old people. In: European Journal of Ageing 3: 15-21. https://doi.org/10.1007/s10433-006-0021-5

Galenkamp, Henrike et al. 2020: Self-Rated Health: When and How to Use It in Studies Among Older People? In: *Jagger, Carol et al.* (Eds.): International Handbook of Health Expectancies Cham: Springer: 173-181. https://doi.org/10.1007/978-3-030-37668-0_12

Galenkamp, Henrike et al. 2012: Estimates of prospective change in self-rated health in older people were biased owing to potential recalibration response shift. In: Journal of Clinical Epidemiology 65,9: 978-988. https://doi.org/10.1016/j.jclinepi.2012.03.010

Gallagher, Jane E. et al. 2016: Factors associated with self-reported health: implications for screening level community-based health and environmental studies. In: BMC Public Health 16:640. https://doi.org/10.1186/s12889-016-3321-5

Gardner, Raquel C.; Langa, Kenneth M.; Yaffe, Kristine 2017: Subjective and objective cognitive function among older adults with a history of traumatic brain injury: A population-based cohort study. In: PLOS Medicine 14,3: e1002246. https://doi.org/10.1371/journal.pmed.1002246

Glass, Thomas A. 1998: Conjugating the "Tenses" of Function: Discordance Among Hypothetical, Experimental, and Enacted Function in Older Adults. In: Gerontologist 38,1: 101-112. https://doi.org/10.1093/geront/38.1.101

Grodstein, Francine et al. 2001: Type 2 Diabetes and Cognitive Function in Community-Dwelling Elderly Women. In: Diabetes Care 24,6: 1060-1065. https://doi.org/10.2337/diacare.24.6.1060

Guralnik, Jack M. et al. 1989: Physical Performance Measures in Aging Research. In: Journal of Gerontology 44,5: M141–M146. https://doi.org/10.1093/geronj/44.5.M141

Guyatt, Gordon; Walter, Stephen; Norman, Geoff 1987: Measuring change over time: Assessing the usefulness of evaluative instruments. In: Journal of Chronic Diseases 40,2: 171-178. https://doi.org/10.1016/0021-9681(87)90069-5

Hardy, Susan E. et al. 2011: Ability to Walk 1/4 Mile Predicts Subsequent Disability, Mortality, and Health Care Costs. In: Journal of General Internal Medicine 26: 130-135. https://doi.org/10.1007/s11606-010-1543-2

Hoeymans, Nancy et al. 1996: Measuring functional status: Cross-sectional and longitudinal associations between performance and self-report (Zutphen Elderly Study 1990-1993). In: Journal of Clinical Epidemiology 49,10: 1103-1110. https://doi.org/10.1016/0895-4356(96)00210-7

Huang, Zhiyong; Maurer, Jürgen 2019: Validity of Self-Rated Memory Among Middle-Aged and Older Chinese Adults: Results From the China Health and Retirement Longitudinal Study (CHARLS). In: Assessment 26,8: 1582-1593. https://doi.org/10.1177/1073191117741188

Huppert, Felicia A.; Whittington, Joyce E. 1993: Longitudinal changes in mental state and personality measures. In: *Cox, Brian D.; Huppert, Felicia A.; Whichelow, Margaret J.* (Eds): The Health and Lifestyle Survey: Seven years on. Aldershot: England: Dartmouth: 133-172.

Idler, Ellen L. 2003: Discussion: Gender Differences in Self-Rated Health, in Mortality, and in the Relationship Between the Two. In: The Gerontologist 43,3: 372-375. https://doi.org/10.1093/geront/43.3.372

Jürges, Hendrik 2007: True health vs response styles: exploring cross-country differences in self-reported health. In: Health Economics 16,2: 163-178. https://doi.org/10.1002/HEC.1134

Kislaya, Irina et al. 2021: Combining self-reported and objectively measured survey data to improve hypertension prevalence estimates: Portuguese experience. In: Archives of Public Health 79:45. https://doi.org/10.1186/s13690-021-00562-y

Kislaya, Irina et al. 2019: Do self-reported data accurately measure health inequalities in risk factors for cardiovascular disease? In: International Journal of Public Health 64: 721-729. https://doi.org/10.1007/s00038-019-01232-1

Kramer, Michael S.; Feinstein, Alvan R. 1981: Clinical biostatistics: LIV. The biostatistics of concordance. In: Clinical Pharmacology & Therapeutics 29,1: 111-123. https://doi.org/10.1038/clpt.1981.18

Kwon, E.; Hu, Peifeng 2018: Harmonization of Cross-National Studies of Aging to the Health and Retirement Study User Guide: Physical and Anthropometric Measurement.

Lan, Tzuo-Yun et al. 2002: Performance Tests and Disability: Developing an Objective Index of Mobility-Related Limitation in Older Populations. In: The Journals of Gerontology 57,5: M294-M301. https://doi.org/10.1093/gerona/57.5.M294

Layes, Audrey; Asada, Yukiko; Kepart, George 2012: Whiners and deniers – What does self-rated health measure? In: Social Science & Medicine 75,1: 1-9. https://doi.org/10.1016/j.socscimed.2011.10.030

Lazarevič, Patrick 2023: Biases in Assertions of Self-Rated Health. In: Comparative Population Studies 48. https://doi.org/10.12765/CPoS-2023-04

Lazarevič ,Patrick; Brandt, Martina 2020: Diverging ideas of health? Comparing the basis of health ratings across gender, age, and country. In: Social Science & Medicine 267:112913. https://doi.org/10.1016/j.socscimed.2020.112913

Lazarevič, Patrick et al. 2018: Non-Health Influences on Generic Health Ratings: Comparing the Susceptibility of Self-Rated Health (SRH) and the Minimum European Health Module (MEHM) to Biases Due to Optimism, Hypochondriasis, and Social Desirability. In: PUMA (Ed.): PUMA Survey V.2: Ergebnisberichte der einzelnen PUMA- Befragungsmodule. Plattform für Umfragen, Methoden und empirische Analysen (PUMA): 5-8 [https:// www. puma-plattf orm.at/ f ileadmin/ user_upload/ p_puma/ Modulberichte_PUMA_Survey_V .2_01.pdf, 14.02.2023].

Lazarevič, Patrick; Quesnel-Vallée, Amélie 2023: Rating Health and Rating Change: How Canadians Rate Their Health and Its Changes. In: Journal of Aging and Health 35,7-8: 535-542. https://doi.org/10.1177/08982643221119654

Lee, Jinkook; Phillips, Drystan; Wilkens, Jenny 2021: Gateway to Global Aging Data: Resources for Cross-National Comparisons of Family, Social Environment, and Healthy Aging. In: The Journals of Gerontology 76: S5-S16. https://doi.org/10.1093/geronb/gbab050

Long, J. Scott; Ervin, Laurie H. 2000: Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model. In: The American Statistican 54,3: 217-224. https://doi.org/10.2307/2685594

Luy, Marc; Di Giulio, Paola 2006: The impact of health behaviors and life quality on gender differences in mortality. In: *Geppert, Jochen; Kühl, Jutta* (Eds.): Gender und Lebenserwartung. Bielefeld: Kleine: 113-147.

Luy, Marc; Di Giulio, Paola; Minagawa, Yuka 2023: The impact of interpersonal reporting heterogeneity on cross-country differences in Healthy Life Years in Europe. In: European Journal of Public Health 33,6: 1060-1064. https://doi.org/10.1093/eurpub/ckad142

Maddox, George L. 1962: Some Correlates of Differences in Self-assessment of Health Status Among the Elderly. In: Journal of Gerontology 17,2: 180-185. https://doi.org/10.1093/geronj/17.2.180

Mathers, Colin D.; Robine, Jean-Marie; Wilkins, Russell 1994: Health expectancy indicators: recommendations for terminology. In: *Mathers, Colin D.; McCallum, John; Robine, Jean-Marie* (Eds.): Advances in health expectancies: Proceedings of the 7th Meeting of the International Network on Health Expectancy (REVES). Canberra: Australian Institute of Health and Welfare: 34-41.

Middleton, Addie; Fritz, Stacy L.; Lusardi, Michelle 2015: Walking Speed: The Functional Vital Sign. In: Journal of Aging and Physical Activity 23,2: 314-322. https://doi.org/10.1123/japa.2013-0236

Molina, Teresa 2016: Reporting Heterogeneity and Health Disparities Across Gender and Education Levels: Evidence From Four Countries. In: Demography 53,2: 295-323. https://doi.org/10.1007/s13524-016-0456-z

Newman, Anne B. et al. 2006: Association of Long-Distance Corridor Walk Performance With Mortality, Cardiovascular Disease, Mobility Limitation, and Disability. In: JAMA 295,17: 2018-2026. https://doi.org/10.1001/jama.295.17.2018

OECD 2019: Measuring What Matters: the Patient-Reported Indicator Surveys.

Oksuzyan, Anna et al. 2019: Is the story about sensitive women and stoical men true? Gender differences in health after adjustment for reporting behavior. In: Social Science and Medicine 228: 41-50. https://doi.org/10.1016/j.socscimed.2019.03.002

Peel, Nancye M.; Kuys, Suzanne S.; Klein, Kerenaftali 2013: Gait Speed as a Measure in Geriatric Assessment in Clinical Settings: A Systematic Review. In: The Journals of Gerontology: Series A 68,1: 39-46. https://doi.org/10.1093/gerona/gls174

Pfarr, Christian; Schmid, Andreas; Schneider, Udo 2012: Reporting heterogeneity in self-assessed health among elderly Europeans. In: Health Economics Review 2. https://doi.org/10.1186/2191-1991-2-21

Picavet, Emmanuel 2017: Perceived individual freedom and collectively provided care. In: *Guillemin, Francis et al.* (Eds.): Perceived Health and Adaptation in Chronic Disease. New York, NY: Routledge: 265-275.

Purser, Jama L. et al. 2005: Mild Cognitive Impairment and 10-Year Trajectories of Disability in the Iowa Established Populations for Epidemiologic Studies of the Elderly Cohort. In: Journal of the American Geriatrics Society 53,11: 1966-1972. https://doi.org/10.1111/j.1532-5415.2005.53566.x

Ravaghi, Hamid et al. 2020: Models and methods for determining the optimal number of beds in hospitals and regions: a systematic scoping review. In: BMC Health Services Research 20,186. https://doi.org/10.1186/s12913-020-5023-z

Reuben, David B.; Siu, Albert L.; Kimpau, Sokkun 1992: The Predictive Validity of Self-Report and Performance-based Measures of Function and Health. In: Journal of Gerontology 47,4: M106-M110. https://doi.org/10.1093/geronj/47.4.M106

Robine Jean-Marie; Jagger, Carol 2003: Creating a coherent set of indicators to monitor health across Europe: the Euro-REVES 2 project. In: European Journal of Public Health 13: 6-14. https://doi.org/10.1093/EURPUB/13.SUPPL_1.6

Salthouse, Timothy A. 2016: Aging Cognition Unconfounded by Prior Test Experience. In: The Journals of Gerontology 71,1: 49-58. https://doi.org/10.1093/geronb/gbu063

Santos, João Vasco et al. 2021: Health expectancies in the European Union: same concept, different methods, different results. In: Journal of Epidemiology & Community Health 75: 764-771. https://doi.org/10.1136/jech-2020-213791

Schwartz, Carolyn E.; Sprangers, Mirjam A. G. 1999: Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. In: Social Science & Medicine 48,11: 1531-1548. https://doi.org/10.1016/S0277-9536(99)00047-7

Shumway, Robert H.; Stoffer, David S. 2011: Time Series Analysis and Its Applications. With R Examples. Cham: Springer. https://doi.org/10.1007/978-3-319-52452-8

Simpson Crystal F. et al. 2004: Agreement Between Self-Report of Disease Diagnoses and Medical Record Validation in Disabled Older Women: Factors That Modify Agreement. In: Journal of the American Geriatrics Society 52,1: 123-127. https://doi.org/10.1111/j.1532-5415.2004.52021.x

Sohel, Nazmul et al. 2016: Factors influencing discrepancies in self-reported memory and performance on memory recall in the Canadian Community Health Survey – Healthy Aging, 2008-09. In: Age and Ageing 45,2: 280-286. https://doi.org/10.1093/ageing/afv163

Spitzer, Sonja; Shaikh, Mujaheed 2022: Health misperception and healthcare utilisation among older Europeans. In: The Journal of the Economics of Ageing 22:100383. https://doi.org/10.1016/j.jeoa.2022.100383

Spitzer, Sonja; Shaikh, Mujaheed; Weber, Daniela 2022: Older Europeans' health perception and their adaptive behaviour during the COVID-19 pandemic. In: European Journal of Public Health 32,2: 322-327. https://doi.org/10.1093/eurpub/ckab221

Spitzer, Sonja; Weber, Daniela 2019: Reporting biases in self-assessed physical and cognitive health status of older Europeans. In: PLoS One 14: e0223526. https://doi.org/10.1371/journal.pone.0223526

Sprangers, Mirjam A. G.; Schwartz, Carolyn E. 1999: Integrating response shift into health-related quality of life research: a theoretical model. In: Social Science & Medicine 48,11: 1507-1515. https://doi.org/10.1016/S0277-9536(99)00045-3

Spuling, Svenja M.; Wolff, Julia K.; Wurm, Susanne 2017: Response shift in self-rated health after serious health events in old age. In: Social Science & Medicine 192: 85-93. https://doi.org/10.1016/j.socscimed.2017.09.026

Steel, Nicholas et al. 2003: Physical and cognitive function. In: *Marmot, Michael et al.* (Eds.): Health, Wealth and Lifestyles of the Older Population of England: The 2002 English Longitudinal Study of Ageing. London: The Institute for Fiscal Studies: 249-300.

Strawbridge, William J.; Wallhagen, Margaret I. 1999: Self-Rated Health and Mortality Over Three Decades: Results from a Time-Dependent Covariate Analysis. In: Research on Aging 21,3: 402-416. https://doi.org/10.1177/0164027599213003

Studenski, Stephanie et al. 2011: Gait Speed and Survival in Older Adults. In: JAMA: The Journal of the American Medical Association 305,1: 50-58. https://doi.org/10.1001/JAMA.2010.1923

Suchman, Edward A.; Streib, Gordon F.; Phillips, Bernard S. 1958: An Analysis of the Validity of Health Questionnaires. In: Social Forces 36,3: 223-232. https://doi.org/10.2307/2573809

Tolonen, Hanna et al. 2021: Cross-national comparisons of health indicators require standardized definitions and common data sources. In: Archives of Public Health 79,208: 1-14. https://doi.org/10.1186/S13690-021-00734-W

United Nations Development Programme 2022: Human Development Report 2021-2022. Human Development Index (HDI). New York.

Verbrugge, Lois M.; Jette, Alan M. 1994: The disablement process. In: Social Science and Medicine 38,1: 1-14. https://doi.org/10.1016/0277-9536(94)90294-1

Warner, Lisa M. et al. 2012: Health-specific optimism mediates between objective and perceived physical functioning in older adults. In: Journal of Behavioral Medicine 35: 400-406. https://doi.org/10.1007/s10865-011-9368-y

White, Daniel K. et al. 2013: Trajectories of Gait Speed Predict Mortality in Well-Functioning Older Adults: The Health, Aging and Body Composition Study. In: The Journal of Gerontology 68,4: 456-464. https://doi.org/10.1093/gerona/gls197

White, Halbert 1980: A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. In: Econometrica 48,4: 817-838. https://doi.org/10.2307/1912934

Wong, Rebeca; Peláez, Martha; Palloni, Alberto 2005: Self-reported general health in older adults in Latin America and the Caribbean: Usefulness of the indicator. In: Revista Panamericana de Salud Publica/Pan American Journal of Public Health 17,5-6: 323-332. https://doi.org/10.1590/s1020-49892005000500004

World Bank 2022: World Bank Annual Report 2022: Helping Countries Adapt to a Changing World. Washington, D.C.: World Bank.

World Health Organization 2018: 2018 Global reference list of 100 core health indicators (plus health-related SDGs). World Health Organization.

World Health Organization 2005: Constitution of the World Health Organization: World Health Organization, Basic documents. Geneva.

Wuorela, Maarit et al. 2020: Self-rated health and objective health status as predictors of all-cause mortality among older people: a prospective study with a 5-, 10-, and 27-year follow-up. In: BMC Geriatrics 20,120. https://doi.org/10.1186/s12877-020-01516-9

Young, Codi et al. 2021: Harmonization of Cross-National Studies of Aging to the Health and Retirement Study User Guide Physical Functional Health. CESR-Schaeffer Working Paper No. 001.

Date of submission: 08.09.2023 *Date of acceptance: 26.06.2024*

Dr. Vanessa di Lego (✉). Department of Demography, Faculty of Economics, Universidade Federal de Minas Gerais. Belo Horizonte, Brazil. E-mail: vdilego@cedeplar.ufmg.br
URL: https://vdilego.github.io

Dr. Sonja Spitzer. Department of Demography, University of Vienna, Wittgenstein Centre (IIASA, OeAW, University of Vienna). Vienna. Austria. E-mail: sonja.spitzer@univie.ac.at
URL: https://ufind.univie.ac.at/en/person.html?id=112190

Dr. Patrick Lazarevič. Statistik Austria. E-mail: Patrick.Lazarevic@statistik.gv.at
URL: https://www.researchgate.net/profile/Patrick-Lazarevic

Comparative Population Studies

www.comparativepopulationstudies.de

ISSN: 1869-8980 (Print) – 1869-8999 (Internet)

Published by
Federal Institute for Population Research (BiB)
65180 Wiesbaden / Germany

Managing Publisher
Dr. Nikola Sander