“Self-perceived health in older Europeans: Does the choice of survey matter?”

Simone Croezen
Erasmus MC, University Medical Center, Rotterdam, The Netherlands
The Netherlands Institute for Social Research | SCP, The Hague, The Netherlands

Alex Burdorf
Erasmus MC, University Medical Center, Rotterdam, The Netherlands

Frank J van Lenthe
Erasmus MC, University Medical Center, Rotterdam, The Netherlands

Presented in session: “The SWB of older Europeans”
at the 3rd International ESS Conference, 13-15th July 2016, Lausanne, Switzerland

Abstract
Background: Cross-national comparisons of health in European countries provide crucial information to monitor health and disease, to describe health inequalities within and between countries and to inform policy and research priorities. However, variations in estimates might occur when information from cross-national European surveys with different characteristics are used. We compared the prevalence of very good or good self-perceived health across ten European countries according to three European surveys and investigated which survey characteristics of these surveys contributed to differences in prevalence estimates of health.

Methods: We used aggregate data from 2004/2005 of respondents aged 55-64 years from the European Social Survey (ESS), the Survey of Health, Ageing and Retirement in Europe (SHARE) and the European Union Statistics on Income and Living Conditions (EU-SILC). Across the surveys, self-perceived was assessed by the same question with response options ranging from very good to very bad.

Results: Despite a good correlation between the surveys (intraclass correlation coefficient: 0.77), significant differences were found in prevalence estimates of very good or good self-perceived health. The survey response, sample size and survey mode contributed statistically significantly to the differences between the surveys. Multilevel linear regression analyses, adjusted for survey characteristics, showed a higher prevalence of very good or good self-perceived health for SHARE (+6.96, 95% confidence interval: 3.14, 10.8) and a lower prevalence (-3.12, 95% confidence interval: -7.11, 0.86) for ESS, with EU-SILC as the reference survey. Furthermore, the agreement in health estimates between surveys varied across countries, making country-specific comparisons between surveys less reliable.

Conclusion: Three important health surveys in Europe showed substantial differences for presence of very good or good self-perceived health. These differences limit the usefulness for direct comparisons across studies in health policies for Europe.

Keywords
Health surveys, International comparisons, Self-perceived health, Older adults
Acknowledgements
This article uses data from (i) ESS Round 2: European Social Survey Round 2 Data (2004). Data file edition 3.3. Norwegian Social Science Data Services, Norway - Data Archive and distributor of ESS data and (ii) SHARE wave 4 release 1.1.1, as of 28 March 2013 (DOI: 10.6103/SHARE.w4.111) or SHARE wave 1 and 2 release 2.6.0, as of 29 November 2013 (DOI: 10.6103/SHARE.w1.260 and 10.6103/SHARE.w2.260) or SHARELIFE release 1, as of 24 November 2010 (DOI: 10.6103/SHARE.w3.100). The SHARE data collection has been primarily funded by the European Commission through the Fifth Framework Programme (project QLK6-CT-2001-00360 in the thematic programme Quality of Life), through the Sixth Framework Programme (projects SHARE-I3, RII-CT-2006-062193, COMPARE, CIT5- CT-2005-028857 and SHARELIFE, CIT4-CT- 2006-028812) and through the Seventh Framework Programme (SHARE-PREP, No 211909, SHARE-LEAP, No 227822 and SHARE M4, No 261982). Additional funding from the US National Institute on Aging (U01AG09740-13S2, P01 AG005842, P01 AG08291, P30 AG12815, R21AG025169, Y1-AG-4553-01, IAG BSR06-11 and OGHA 04-064) and the German Ministry of Education and Research as well as from various national sources is gratefully acknowledged (see www.shareproject.org for a full list of funding institutions).


Section 1 (Introduction, research question, hypothesis)
Self-perceived health is a widely used indicator for health. It deals with the subjective assessment that a person makes about one’s own health state and serves as an independent predictor for morbidity, health service use and mortality and is often used as an indicator to compute healthy life expectancy (Idler & Benyamini 1997, Stiefel et al. 2010).

Currently, there are three important surveys in Europe, which provide cross-nationally comparable information on self-perceived health of European adults and have their information publicly available: EU Statistics on Income and Living Conditions (EU-SILC), the Survey of Health, Ageing and Retirement in Europe (SHARE) and the European Social Survey (ESS) (Borsch-Supan et al. 2013, ESS 2002, Santourian & Ntakou 2014). EU-SILC, SHARE and ESS have been performed in several European countries, and their information on health status has contributed to several important publications. However, differences between surveys in a multitude of factors, such as study size, sampling strategy and collection methods, can affect the representation of the target population and the measurement of the survey estimate, leading to different estimates of health and thus to different interpretations of the population health status (Aromaa et al. 2003). Potential disagreement in prevalence estimates of health between the surveys could limit the usefulness for direct comparisons across studies in health policies for Europe. Therefore, understanding differences between EU-SILC, SHARE and ESS in terms of their estimation of self-perceived health will provide valuable insights in the comparability of health status information in Europe. Initial comparisons between EU-
SILC, SHARE and ESS showed overall systematic differences in prevalence of self-perceived health, but insight into the determinants of these differences is lacking (Börsch-Supan & Mariuzzo 2005).

We aim to investigate the agreement between the surveys by describing systematic differences in self-perceived health across countries, and investigating the role of survey characteristics such as survey response, sample size, sampling strategy and survey mode as possible determinants of observed differences between the surveys.

Section 2 (Data and methods)

Surveys
EU-SILC was launched in 2003 and provides annually collected data on income distribution and social inclusion in the European Union based on nationally representative probability samples of the population residing in private households. Only household members aged 16 years and older were interviewed (Eurostat 2007). SHARE is a cross-national panel database on health, socio-economic status and social and family networks, started in 2004-05. Based on probability samples in all participating countries, SHARE represents the noninstitutionalized population aged 50 years and older (born in 1954 or earlier) (Borsch-Supan et al. 2013). ESS is a cross-sectional social survey aimed at investigating Europe's changing institutions and the attitudes, beliefs and behaviour patterns of the European population. The first wave was fielded in 2002–03. ESS used random probability samples based on full coverage of the eligible residential populations aged 15 years and older (ESS 2002).

To avoid hampering the comparability between the surveys because of potential selection bias due to loss to follow-up in SHARE, we focussed our comparisons on information from the baseline measurement in SHARE (2004–05). This corresponded to the cross-sectional information from ESS wave 2 (2004–05), and EU-SILC 2005. Across the surveys, 10 countries were commonly included: Austria (AT), Belgium (BE), Germany (DE), Denmark (DK), Greece (GR), Spain (ES), France (FR), Italy (IT), Netherlands (NL) and Sweden (SE).

All three surveys used probability sampling as their primary sampling strategy. For most countries, the response was between 60 and 69%. The largest sample sizes were obtained in EU-SILC, with at least 5000 respondents for each country, whereas the country-specific sample sizes in SHARE and ESS were between 1000 and 4000 respondents. SHARE and ESS predominantly used computer-assisted personal interview (CAPI) to collect the information from their respondents. In EU-SILC, survey modes varied between and even within countries.

Self-perceived health
EU-SILC, SHARE and ESS included the same measure of self-perceived health, based on the question: “How is your health in general?”. Also, similar response options were used across the surveys: (i) very good; (ii) good; (iii) fair; (iv) bad; (v) very bad. This assessment of self-perceived health was proposed by the WHO (De Bruin et al. 1996)
and adopted by the Minimum European Health Module (The EHEMU/EHLEIS team 2010). For our current analysis and in line with common practice in presenting the prevalence rates in research and policy documents, response options for self-perceived health have been dichotomized across all surveys, with the response categories ‘very good’ and ‘good’ into one ‘very good or good’ category.

**Statistical analysis**

In each survey information was retrieved on self-perceived health among respondents aged 55–64 years for 10 countries. We focussed on the age group 55–64 years for the comparability of self-perceived health between the surveys. For EU-SILC, population-weighted aggregate data or self-perceived health were only publicly available by 10-year age groups through the Eurostat website. We calculated comparable group-level information based on individual-level data, which were available after registration on their project websites for SHARE and for ESS. Like in EU-SILC, weights were applied in SHARE and ESS to represent the national populations based distributions of age and gender.

First, weighted prevalence estimates of very good or good self-perceived health by gender and educational level were described for each survey. All three surveys followed the 1997 International Standard Classification of Education ISCED-97 and educational level was presented by three groups: (i) pre-primary, primary and lower secondary education, first and second stage of basic education (ISCED 0-2); (ii) upper and post-secondary education (ISCED 3-4); (iii) first and second stage of tertiary education (ISCED 5-6).

Second, we established whether systematic differences in self-perceived health between surveys were present across countries by using Bland and Altman plots. The Bland and Altman technique enabled to quantify the agreement in self-perceived health between surveys for each country separately, by calculating the mean difference between two surveys, and 95% limits of agreement as the mean difference ±1.96 SD (Bland & Altman 1986). The mean difference reflects the systematic difference between the surveys. The smaller the range between the two limits of agreement, the better the agreement is.

Third, the intraclass correlation coefficient (ICC) was calculated as an indication of the correlation of the prevalence estimates of self-perceived health between the surveys within a country. The total variance is reflected by variations between countries and by variations within countries. The ICC is the proportion of total variance that is attributed to differences between countries. As the surveys are nested within the countries, the variance within countries is reflected by variance between the surveys. Thus, a large ICC implies substantial differences between countries and therefore relatively small differences between surveys within countries.

Finally, we investigated the role of survey characteristics as possible determinants of disagreement between surveys using multi-level modelling with countries as the higher level, based on the assumption that the surveys were clustered within each country. The prevalence of very good or good self-perceived health (in percentages) was our continuous dependent variable, and regressed on a variable including a category for each survey and several other variables, each representing a specific characteristic of the survey, i.e. response (continuous variable), sample size (continuous variable), sampling strategy (simple random sampling; multi-stage sampling; single or multi-stage sampling using telephone directories; other) and survey mode [computer-assisted personal interview (CAPI), computer-assisted telephone interview (CATI), paper and
pencil interview (PAPI) and self-administered]. Each possible determinant was included in separate models, plus in a fully adjusted model. With three surveys and 10 countries, the sample size included 30 observations (prevalence estimates) of very good or good self-perceived health. All analyses were performed using Stata/SE 13.1.

Section 3 (Results)

Across all surveys, very good or good self-perceived health was least prevalent in Austria, Germany, Spain and Italy, with the lowest prevalence found among lower educated German men in EU-SILC (34%) and among lower educated Austrian men in SHARE (35%). The highest prevalence of very good or good health was reported for higher educated Greek men in SHARE (93%) and for higher educated Austrian women in EU-SILC (87%).

Figure 1 presents Bland and Altman plots showing the agreement in prevalence estimates of very good or good self-perceived health between EU-SILC, SHARE and ESS, separately for men and women. When comparing the surveys, SHARE’s prevalence estimates of very good or good self-perceived health are systematically higher than the estimates obtained in EU-SILC or ESS. The mean differences were 9.8% in men and 6.7% in women between SHARE and EU-SILC and were 6.3% in men and 10.0% in women between SHARE and ESS. Between EU-SILC and ESS, the agreement was better, indicated by a lower prevalence estimate by EU-SILC in men (mean difference = -3.5%) and a higher estimate by EU-SILC in women (mean difference 3.2%). When looking at the individual countries, it seems that the agreement between surveys is better for those countries with a relatively high prevalence of very good or good self-perceived health, such as Denmark, Sweden, the Netherlands, Belgium and Greece.

According to the ICC, the overall level of agreement between the surveys within countries was fairly high (ICC = 0.77). Table 1 presents the results of the multilevel analysis, which showed a statistically significantly higher prevalence of very good or good self-perceived health in SHARE as compared with EU-SILC (+8.22% points). There was no statistical significant difference between ESS and EU-SILC (+0.16% points). After adjustment for the survey characteristics, we still found a statistical significant difference between SHARE and EU-SILC (+6.96% points, 95% CI: 3.14 to 10.8) and an increased but still not significant difference between ESS and EU-SILC (-3.12% points, 95% CI, -7.11 to 0.86). In the fully adjusted model, the prevalence of self-perceived health was 0.13% points higher when the survey response was increased by 1% point. On the other hand, sample size was negatively associated with the prevalence of self-perceived health; an increase of 1000 respondents resulted in a 0.79% points decrease in the prevalence of very good or good health. When compared with the CAPI survey mode, CATI and PAPI resulted in higher prevalence estimates of self-perceived health (+6.71% points and +5.36% points, respectively). Sampling strategy did not contribute significantly to the self-perceived health prevalence.
Section 4 (Conclusion/Discussion)

This study showed that the prevalence of very good or good self-perceived health assessed by different surveys varies substantially across the surveys. When taking into account survey characteristics such as survey response, sample size, sampling strategy and survey mode, we found that prevalence estimates of very good or good self-perceived health ranged almost 10% points between the surveys on the country level.

Survey response, sample size and survey mode were all associated with the prevalence of very good or good self-perceived health. A low response could indicate a selective population taking part in the survey, and if health plays an important role in this selection process, surveys or countries with lower response may have higher prevalence estimates of very good or good health. Furthermore, a lower response may enhance the probability that respondents differ from non-respondents in their characteristics, which may weaken the external validity of the survey results. Furthermore, it could be that the response bias in our estimates may be differential between surveys, countries or age groups, as factors behind non-response may differ. The response variable in our study was based on the response of the total study population, whereas the analyses only focussed on those aged 55–64 years. If the non-response was age specific, we might be able to explain more variation between the surveys. But, we question whether age-specific non-response would have different effects on the variation across the surveys.

For sample size, we found opposite associations: a higher sample size was associated with lower prevalence estimates of very good or good self-perceived health. Although response and sample size helped explain the observed differences in the prevalence estimates of self-perceived health between the surveys, the impact of non-response bias on our results could be minimal as we based our analyses on the age stratum of 55–64 years, thereby minimizing the influence of age-distribution on the observed differences in prevalence estimates. Furthermore, subgroup analyses were performed within this age group by gender and educational level, for which similar results were obtained. For survey mode, our results showed that for surveys using CATI, more positive scores were obtained than when using CAPI or when questionnaires were self-administered. Our findings are in line with previous research on health-related quality of life, where telephone administration yielded more positive scores than the self-administered mode, but are contradicting with a study comparing four surveys, which found that telephone interviewing led to a higher prevalence of poor self-perceived health than face-to-face interviews (Salomon et al. 2009; Hanmer et al. 2007; Hays et al 2009). As CATI was used only in a few countries in EU-SILC in our study, more research is needed to replicate our findings and to give insight in the possible impact of CATI on the health estimates.

A major strength of this research is that the same question on health was used in comparable populations of the three surveys: non-institutionalized persons, aged 55–64 years. The comparisons between surveys that were made in this study were based on aggregate data in older adults aged 55–64 years. We deliberately chose to use this age group, as for this age group the numbers of respondents were highest and probably gave most reliable estimates across the surveys. Furthermore, all three surveys used weights to take into account the original national population. Also, all surveys had an overlapping data collection period, 2005, in which the assessments of self-perceived health took place.
A different question order may have hampered the comparability between the surveys (Bowling & Windsor 2008; Lumsdaine & Exterkate 2013). Especially for a subjective indicator such as self-perceived health, respondents’ answers may depend on the questions preceding the self-perceived health question. Respondents may consider a broader view and reflect more thoroughly on the concept of health when more nearby questions are posed (Wikman 2007). Based on the survey questionnaires, we found that the order was quite consistent across EU-SILC and ESS; the first question of the health-related module. However, the place of the health-related module within the total questionnaire differed between the surveys, as they have different scopes. Only in SHARE, the order within the health-related module was randomly assigned to the respondents, as an experiment. Based on earlier research on this topic, which found that health was negatively influenced by the battery of health questions preceding the assessment (Lumsdaine & Exterkate 2013), we hypothesized that SHARE would underestimate the level of very good or good self-perceived health when compared with the other surveys, but the opposite is true. Therefore, we do not feel that the order determines the differences found in our study.

This article provides valuable insights in the comparability and the causes of incomparability of health information across European surveys. The goal of this study was not so much to provide absolute estimates of self-perceived health for policy purposes, but to examine the influence of several survey characteristics on the observed differences in estimates of health. As differences in health estimates between three major European population-based surveys seem substantial, they could thus lead to different interpretations of the population health status. When using SHARE information, population health might be estimated more favourably than, when using EU-SILC or ESS. However, we do not know whether SHARE overestimates, or ESS underestimates self-perceived health when compared with the other surveys. Further steps need to be taken to improve understanding of observed disagreement between European surveys and to evaluate whether disagreement in occurrence of other health indicators could be explained by survey characteristics. If the disagreement across health indicators is differential, this might influence the identification of research and policy priorities. To conclude, our findings offer an important reminder that disagreement in health prevalence estimates between surveys may limit the usefulness for direct comparisons across studies in health policies for Europe.

References


Figure 1 Bland and Altman plots presenting agreement in prevalence estimates of very good or good self-perceived health with mean differences and limits of agreement (1.96 SD) between EU-SILC, SHARE and ESS in men (first row) and women (second row)
Table 1 Multilevel analysis for very good or good self-perceived health (random intercept for country), N=30

<table>
<thead>
<tr>
<th></th>
<th>Crude model</th>
<th>Fully adjusted model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>95% CI</td>
</tr>
<tr>
<td>Survey</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EU-SILC</td>
<td>1</td>
<td>4.83 to 11.61</td>
</tr>
<tr>
<td>SHARE</td>
<td>8.22</td>
<td>-3.23 to 3.55</td>
</tr>
<tr>
<td>ESS</td>
<td>0.16</td>
<td>-3.23 to 3.55</td>
</tr>
<tr>
<td>Response</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample size (*1000)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Survey mode</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAPI</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>CATI</td>
<td>6.71</td>
<td>1.03 to 12.39</td>
</tr>
<tr>
<td>PAPI</td>
<td>5.36</td>
<td></td>
</tr>
<tr>
<td>Self-administered</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>59.49</td>
<td></td>
</tr>
<tr>
<td>Variance (country)</td>
<td>49.07</td>
<td></td>
</tr>
<tr>
<td>Variance (residual)</td>
<td>14.99</td>
<td></td>
</tr>
</tbody>
</table>

Sampling strategy did not statistically significantly contribute to the fully-adjusted model, and was therefore dropped.