



Published in final edited form as:

*J Aging Health*. 2011 March ; 23(2): 242–266. doi:10.1177/0898264310383421.

## Measuring Health Status: Self-, Interviewer, and Physician Reports of Overall Health

Kimberly V. Smith, PhD<sup>a,\*</sup> and Noreen Goldman, PhD<sup>b</sup>

<sup>a</sup>Mathematica Policy Research, Princeton, NJ

<sup>b</sup>Princeton University, Office of Population Research, Princeton, NJ

### Abstract

**Objectives**—This study examines and compares respondent, interviewer, and physician ratings of overall health.

**Methods**—Data are from the 2006 Social Environment and Biomarkers of Aging Study, a nationally-representative survey of older adults in Taiwan. Ordered probit models are used to examine factors associated with self- and external assessments of health and discordant health ratings.

**Results**—Our results suggest similarities and differences in factors influencing health ratings across evaluators, but a high level of inter-evaluator disagreement in ratings. Discrepancies in ratings between physicians and both respondents and interviewers are associated with the greater weight given to functional limitations and psychological well-being in interviewer and respondent ratings and to the importance of clinical measures or risk factors of illness and mortality in physician assessments.

**Discussion**—Interviewer and physician assessments may be complementary to self-assessed health measures. The importance and implications of these findings for future research are discussed.

### Keywords

Measuring health status; aging; older adult health; self-reported health; Taiwan

## INTRODUCTION

Self-assessed health status (SAH) is one of the most frequently used measures in analyses of health trends, determinants, and inequalities, as well as in assessments of the need for health care resources. The measure is based on a survey question that asks respondents to rate their overall health on a four- or five-point scale that typically runs from excellent to poor. The resulting ordinal variable has been shown to be an independent predictor of a range of health outcomes, including morbidity, use of health services, and mortality (Idler & Benyamini, 1997; Benyamini & Idler, 1999; Idler & Kasl, 1995). Its predictive power, and the ease with which it can be collected, make SAH an important health indicator among older adults. However, many questions remain about the self-evaluation process that underlies SAH ratings, what dimensions of health it captures, and how reporting varies across population sub-groups.

\*Corresponding author. Kimberly V. Smith, Mathematica Policy Research, 600 Alexander Park, Princeton, NJ 08540. ksmith@mathematica-mpr.com. Phone: 609-945-3354.

A large literature, dating back to the 1950s, suggests that SAH ratings are based on a complex aggregation of information on multiple aspects of health—including physical and mental health, physical functioning, health service use, and health behaviors—that is mediated by demographic, socioeconomic, psychosocial, and environmental factors (e.g., Ferraro 1980, Hall, Epstein, & McNeil, 1989; Benyamini, Leventhal, & Leventhal, 1999). Although most of this research focuses on Western developed countries, several studies in Asia suggest that self-assessments of health in non-Western countries involve judgments across these same health domains and take into account similar factors (Goldman, Gleib, & Chang, 2003; Zimmer, Natividad, Lin, & Chayovan, 2000). Recent studies in both the U.S. and Taiwan have found a significant association between biomarkers and SAH, suggesting that biological aspects of health not captured in traditional health indicators are also factored into SAH ratings (Jylhä, Volpato, & Guralnik, 2006; Goldman et al., 2003). However, even after the inclusion of biomarkers in models of SAH, much of the variation in this measure remains unexplained.

Several studies have documented variations in the association between SAH and what are considered more objective indicators of health among older adults. A repeated finding is that among individuals with a similar level of chronic disease or functional limitations, older adults tend to report better health than younger adults (Ferraro, 1980; Idler, 1993). This finding has been attributed to lower self-expectations for health and physical functioning at older ages, the use of age peers as a comparison group, and greater adaptation to illness among older adults (Cockerham, Sharp, & Wilcox 1983; Groot, 2000). Research into factors that influence self-evaluations of health suggests that the referents and criteria used to assess and report overall health vary by age, with older individuals placing greater weight on physical functioning, or their ability to function independently, relative to their peers (Levkoff, Cleary, & Wetle, 1987; Krause & Jay 1994; Benyamini et al., 1999; Benyamini, Leventhal, & Leventhal, 2003). Other studies highlight the importance of non-health factors, such as attitudes, psychosocial well-being, and quality of life, in older adults' conceptualizations of health (e.g., Borawski, Kinney, & Kahana, 1996; Chipperfield, 1993).

In this paper, we gain further insight into self-perceptions of health among older adults by comparing SAH ratings to overall health ratings made by interviewers and physicians in the same survey. To our knowledge, this is the first study to consider interviewer assessments by non-medical personnel as a potential source of information on respondent health. Few studies have examined perceptions of respondent health by any type of survey administrator. One exception is a study conducted almost 30 years ago that compares ratings of physical and mental health by older bereaved persons to ratings by nurse interviewers (Valanis & Yeaworth, 1982). Recent research on perceived age and mortality suggests that even non-medical personnel may provide valuable insights into the health of survey respondents. Christensen, Thinggaard, McGue, Rexbye, et al. (2009), for example, find that untrained strangers' estimates of a person's age—based on facial photographs—are strongly and significantly correlated with survival, physical and cognitive functioning, and leukocyte telomere length. The fact that interviewers in our study have the opportunity to assess not only respondents' appearances, but also their physical, psychological, cognitive and social functioning, suggests that their evaluations are likely to impart considerable information.

Numerous studies have compared morbidity measures based on respondent self-reports to physician data (Krueger, 1957; Kriegsman, Penninx, van Eijk, & Boeke, 1996; Ferraro & Su, 2000), but few have contrasted self and physician ratings of overall health. Strein, Suchman, and Phillips (1958) and Friedsam & Martin (1963) – two early examples of studies comparing self and physician ratings of overall health – find large discrepancies between the two evaluations but come to opposite conclusions about the direction of the relationship. A more recent study using data from the 1982-1984 Hispanic Health and

Nutrition Examination Survey (H-HANES) of middle-aged and older Mexican Americans finds that, on average, physicians judged respondents' health to be better than the respondents did themselves (Markides, Lee, Ray, & Black, 1993).

The objectives of our analysis are to identify the types of information that are associated with respondent, interviewer, and physician health ratings and, in doing so, to determine whether external assessments may provide complementary information on respondent health that could increase the accuracy and predictive power of overall health measurement in population surveys. We further investigate the magnitude and direction of differences in overall health ratings across evaluators and identify the factors driving observed discrepancies. In contrast to some earlier studies, we do not make assumptions about which overall health measure is most accurate. We recognize that responses may vary because respondents, interviewers, and physicians have access to different types of information and bring distinct biases into their assessments. For example, whereas reports by medical personnel are often considered the "gold standard," some studies show that self-reports have greater predictive utility than these more "objective" measures (e.g., Mossey & Shapiro, 1982; Ferraro & Su, 2000). Markides et al. (1993) also find variability in the validity of physician's assessments of respondents' overall health in the H-HANES, suggesting that physician's assessments may not be as "objective" as is often assumed.

## METHODS

The data used for this analysis are from the second (2006) wave of the 2000-2006 Social Environment and Biomarkers of Aging Study (SEBAS). SEBAS is based on a national subsample of respondents aged 54 and older who were interviewed for the Survey of Health and Living Status of the Near Elderly and Elderly in Taiwan, also referred to as the Taiwan Longitudinal Survey of Aging (TLSA). The TLSA began in 1989 with a nationally-representative sample of adults aged 60 and older and, in 1996, was expanded to include an additional sample of adults aged 50 to 66. The SEBAS 2000 wave is based on a randomly-selected subsample of participants in the 1999 wave of TLSA. Older adults (71 and older) and residents of urban areas were oversampled. The 2006 SEBAS sample includes surviving respondents who completed the 2000 SEBAS survey and a sub-sample of younger respondents (aged 53 to 60) who were interviewed for the first time for the 2003 wave of TLSA. Details about TLSA and SEBAS can be found elsewhere (Taiwan Provincial Institute of Family Planning 1989 and 1997; Chang, Gleib, Goldman, & Weinstein, 2007).

The 2006 SEBAS includes a home interview, during which a physical performance assessment is administered, and a hospital visit during which respondents undergo a medical examination. A total of 66 interviewers with extensive training administered the home interview, which had an average length of 76 minutes. The training consisted of an explanation of the study, theory of the home interview, and protocols for obtaining consent, administering the questionnaire, and conducting the physical performance assessment. No particular guidance was provided on how to rate the respondent's overall health. During the hospital visit, survey staff collected information on family disease history and health-related behaviors and took blood pressure and anthropometric measurements. Medical personnel then conducted a physical exam, administered an abdominal ultrasound, and drew blood samples. A total of 54 physicians conducted physical exams and provided assessments of respondents' overall health.

For the 2006 SEBAS, 1,284 respondents (87% response rate) provided home interviews and 1,036 (81% of those interviewed) participated in the medical examination. Participation in the exam was lower among the youngest (aged 53-59) and oldest (80+) respondents, the less educated, those with one or more activity of daily living (ADL) limitations, and those who

had a medical exam within the preceding 3 months. Participation was not related to self-reported health status.

The current analysis is based on respondents who completed both the home interview and medical examination. We control for variables associated with selection into the medical examination in our models to minimize bias in our results. Out of the 1,036 respondents who completed the home interview and examination, 14 (1.4%) were proxy respondents. Proxy respondents were not asked the self-assessed health question and, therefore, are excluded from this analysis. Observations with missing values on health ratings or explanatory variables were also dropped, leaving a total sample size of 838. Conditional on completion of the home interview and medical exam, the probability of having one or more missing values on an outcome or explanatory variable was not significantly related to any demographic, socioeconomic, or health-related variable, with the exception of mobility. Respondents with a higher degree of mobility limitation were more likely to be excluded from this analysis.

### Measures and Descriptive Statistics

**Outcome measures**—The SAH measure comes from the following question asked of respondents (in Mandarin) during the home interview: “Regarding your current state of health, do you feel it is excellent [5], good [4], average [3], not so good [2], or poor [1]?” This question was asked towards the beginning of the home interview, following questions pertaining to the respondent’s background but before other health questions and the physical performance assessment.

The interviewer-assessed health (IAH) measure uses the response from a very similar question asked of interviewers: “Regarding the respondent’s current state of health, do you (INTERVIEWER) feel it is excellent [5], good [4], average [3], not so good [2], or poor [1]?” Interviewers were asked this question at the end of the home interview and after the performance assessments had been completed. Similarly, physicians were asked to rate the respondent’s (patient’s) health at the end of the medical examination, using the same scale. Their responses are used as the physician-assessed health (PAH) measure. It should be noted that the physicians conducting the SEBAS medical exam were not the respondents’ regular physicians; respondents met the SEBAS physician for the first time on the day of their hospital visit. Therefore, physicians’ assessments were based only on the physical exam and the medical history form completed prior to the exam. The medical history form focuses primarily on chronic illnesses, long-term medication use, and health behaviors. Information on physical functioning and psychological well-being was not collected through the medical history or exam forms. However, this information may have been attained through observation and interaction with the respondent during the exam. At the time of their rating, physicians did not have knowledge of any laboratory results or of responses to the home interview.

Table 1 gives the frequency distributions of the three summary health measures. Approximately 11% of respondents consider themselves in excellent health whereas 27% of interviewers and 4% of physicians rate the respondent’s health as such. While very few respondents (0.6% – 3.2%) are rated in poor health by any source, the proportion rated in not-so-good health is higher for self-assessments (21%) than for interviewer (7%) and physician assessments (11%). The frequency distributions also suggest that respondents and physicians favor the middle response category (“Average”), which accounts for over 40% of ratings in both cases. In contrast, only 21% of interviewers’ ratings fall in the “Average” category; interviewers appear to favor the “Good” response category, which comprises 44% of interviewers’ ratings.

**Explanatory variables**—We examine a range of factors that may affect individuals' self-evaluations of their overall health. Indicators of physical health and functioning and psychological well-being, as well as health behaviors and demographic, socioeconomic, and social factors, are based on respondents' reports. We also examine more objective measures of health, including biomarkers associated with chronic disease and health conditions, physical functioning measures derived from the in-home performance tests administered by the interviewer, and clinical measures based on the medical exam conducted by physicians. We initially considered a more extensive set of explanatory variables within each of these categories, but, in light of the modest sample size on which this analysis is based, we excluded variables that were not significant at the 10% level in any model. Measures that were excluded comprise: self-reported ulcer, having one or more activity of daily living (ADL) limitations, frequency of physical exercise, weak grip strength, and abnormalities of the rectum, limb and breast detected during the medical exam. In addition, three biomarkers collected during the medical exam were excluded: ratio of total to high-density lipoprotein (HDL) cholesterol, glycosylated hemoglobin (HbA1c), and systolic blood pressure. The lack of significance of some of these measures may result from relatively low prevalence (e.g., only 5% of respondents have any ADL limitation) and, hence, lack of statistical power. Table 2 provides summary statistics for the covariates included in the final analyses, which are described below.

**Self-reported measures of physical health**—We examine the following eight self-reported indicators of chronic and/or current illness: hypertension and/or use of anti-hypertensive medication, diabetes, heart disease, cancer, respiratory tract illness, liver or gall bladder disease, kidney disease, and gout. Approximately 40% of respondents report having at least one chronic condition. The most prevalent are high blood pressure (33%), diabetes (17%) and heart disease (15%).

As a measure of physical functioning, we include the number of mobility limitations, a count of how many of the following nine activities the respondent reported having any difficulty performing: standing for 15 minutes, standing for two hours, squatting, raising both hands over the head, grasping/turning objects with fingers, running 20-30 meters, walking 200-300 meters, and walking up three flights of stairs. The average number of mobility limitations is 1.7. The number of days spent in a hospital (mean=1.47) is used as an indicator of medical service use. Lastly, we include a four-point ordinal variable that measures the level of pain/discomfort that the respondent experienced in the preceding month (none=0, mild=1, moderate=2, or severe=3). Approximately 47% of respondents report having bodily pain during the preceding month and the mean level of pain is 0.68.

We include two indicators of health-related behaviors: daily smoking in the past six months and frequency of engagement in any of the following relaxation techniques: Qigong, Tai Chi, meditation, yoga, and activities similar to Chi Kung (never, < 1 time/week, 1-2 times/week, 3-5 times/week, and 6+ times/week).

**Measures of psychological well-being**—We also include three indicators of psychological well-being. The first is a measure of depressive symptoms based on a 10-item, shortened form of the Center for Epidemiological Studies Depression (CES-D) scale, which results in an index ranging from 0 to 30 (Goldman et al., 2003). The second is a perceived stress index that is based on responses to seven questions that ask what level of stress/anxiety the respondent currently feels about his/her health, financial situation, relations with other family members, as well as about his/her family members' financial situation, job, or marital situation. Possible responses are none=0, some stress/anxiety=1, and a lot of stress/anxiety=2. The index is calculated by summing across all items, if there are at least four valid items, and dividing by the number of items included in the index (Goldman, Gleib,

Seplaki, Liu, & Weinstein, 2005). The potential range for this index is 0 to 2 and the mean index score for our sample is 0.28. Our final measure in this health domain is a sleep index, which is based on a subset of questions used in the Pittsburgh Sleep Quality Index (PSQI) and constructed based on standard practice (Buysse, Reynolds, Monk, Berman, & Kupfer, 1989). Respondents were asked a series of seven questions related to sleep quality, latency (how long it takes to fall asleep), duration (number of hours of sleep per night), efficiency (hours of sleep/hours in bed), and daytime dysfunction (trouble staying awake during the day). Responses are scored to form a sleep quality index ranging from 0 (high quality) to 15 (low quality).

**Biomarkers**—We explored the inclusion of four biomarkers widely used in clinical practice: body mass index (BMI), systolic blood pressure, ratio of total to high-density lipoprotein (HDL) cholesterol, and glycosylated hemoglobin (HbA1c). These measures are associated with chronic disease, physical functioning, and/or mortality and have been used as indicators of health risk factors in numerous studies (e.g., Jylhä et al., 2006). Biomarkers can serve as more objective or complementary indicators of physical health and may also signal the severity of self-reported chronic conditions. In our models, BMI was the only biomarker significant at the 10% level and, therefore, is the only biomarker included in the final models presented here. BMI was calculated as weight in kilograms divided by height in meters squared based on measures collected during the physical exam. We include indicators for whether the respondent was underweight (BMI < 18.5) or obese (BMI ≥ 30) at the time of the survey.

**Physical performance and exam measures**—Physical performance measures are based on four in-home performance tests: peak expiratory flow, measured walking, and chair stands. These tests are widely used to capture objective assessments of physical functioning and have been associated with morbidity and increased risk of death (Guralnik, Seeman, Tinetti, Nevitt, & Berkman, 1994; Reuben, Siu, & Kimpau, 1992). The information derived from performance tests is designed to be complementary to self-reported and medical exam measures and has been shown to be particularly useful in detecting disability (Guralnik, Ferrucci, Pierper, Leveille, Markides, et al., 2000; Fried, Herdman, Kuhn, Rubin, & Turano, 1991). Peak flow is obtained from a meter that measures the maximum airflow during an expiration delivered with maximum force (Quanjer, Lebowitz, Gregg, Miller, & Pedersen, 1997). The maximum value from three trials is used as the indicator of peak flow. For the walking test, interviewers measured the time it took respondents to walk three meters (a shorter distance was used for nine respondents that had space limitations). Walking speed is calculated as seconds per meter for the better of two trials. Lastly, interviewers recorded the time it took participants to sit down and stand up from a chair five times in succession. To normalize the performance measures for sex and height, performance is measured with standardized residuals and associated quartiles (Van Frago, Gahbauer, Van Ness, Concato, & Gill, 2008; Quanjer et al., 1997). For each performance test, we include an indicator for whether the respondent's performance was in the lowest quartile as a measure of disability. Respondents who could not attempt or complete a performance test for reasons related to physical health are included in the lowest-performing quartile (Guralnik et al., 2000). The fraction of respondents unable to attempt or complete a test exceeded 2% only for the chair stand test (5% did not complete this test due to physical limitations). For the chair stand test, we include a variable indicating if the respondent was unable to complete the test, an indicator of more severe disability.

Lastly, we include a dichotomous indicator for detection of a heart abnormality during the medical examination, which occurred in 7% of respondents.

**Socio-demographic variables**—Other covariates include sex, age, perceived social position, and urban/rural residence. Perceived social position is derived from respondents' ranking of their current situation (in terms of money, education, and occupation) relative to all other people in Taiwan, based on a 10-rung ladder where higher values indicate better standing. This measure, known as the MacArthur Scale of Subjective Social Status, may be a better indicator of social position among older Taiwanese than conventional measures of socioeconomic status (SES) (Goldman, Cornman, & Chang, 2006). The mean social standing level was 4.4. Lastly, we examine an indicator of social networks, measured as the number (0-8) of the following social activities in which the respondent participates: neighborhood, religious, farmers', political, social service, and village/lineage associations; elderly clubs; and continuing education centers.

### Statistical Methods

The first step of our analysis is to assess levels of inter-evaluator agreement in health ratings based on Cohen's kappa statistic, a commonly-used measure of the magnitude of agreement between raters. We calculate both unweighted and weighted kappa statistics. The weighted kappa statistic takes into account the distance between ratings (on the 5-point rating scale) and assigns less weight to greater distances<sup>1</sup>. We use the cut-offs proposed by Landis and Koch (1977) to interpret the level of agreement associated with a given kappa statistic: 0.00-0.20=Slight; 0.21-0.40=Fair; 0.41-0.60=Moderate; 0.61-0.80=Substantial; 0.81-1.0=Almost perfect.

We next use ordered probit models to examine correlates associated with the three (ordinal) summary health measures: SAH, IAH, and PAH. At an early stage of the analysis, we estimated two models for each outcome, one that included only self-reported data and a second model that included measures from the performance tests and medical exam. We found that the magnitude and significance of the coefficients on the self-reported measures changed only slightly between the first and second model specification, suggesting that the performance and clinical measures are largely complementary to the battery of self-reported measures. We, therefore, present results for only the second model. As noted previously, in the interest of model parsimony, the final specification includes only variables that were significant at the 10% level in one or more models. Excluding insignificant variables had little effect on coefficient estimates or model fit.

Although the exam instrument on which the physician ratings are based differed from the home interview administered by interviewers, we include the same variables in the SAH, IAH, and PAH models for two reasons. First, physicians may have gleaned additional information through observation and discussion with the respondent, such as slow walking speed and level of pain/discomfort, that was not explicitly addressed by the exam instrument. Similarly, in the case of BMI, which is based on anthropometric measures taken during the medical exam, the respondent and interviewer would likely be aware of whether the respondent was underweight or obese. Second, this analytic strategy allows us to determine the relative importance of various factors underlying the three health evaluations by comparing standardized coefficients across models.<sup>2</sup>

Lastly, we identify the factors that account for discordant ratings between evaluators. We estimate ordered probit models in which the outcome is the simple difference between two ratings (ranging from 4 to -4). The resulting parameter estimates indicate whether covariates

<sup>1</sup>The weights are given by  $[1 - (|i - j| / (k-1))]$ , where  $i$  and  $j$  indicate the rows and columns of the ratings by the two raters and  $k$  is the number of possible ratings. These weights result in perfect agreement being assigned a weight of one, a one-point rating difference a weight of 0.75, a two-point difference a weight of 0.50, a three-point difference a weight of 0.25, and a four-point difference a weight of 0.0.

are associated with relatively better or worse ratings by one evaluator versus another. For these models, as well as the previous set of ordered probit models, we test the joint significance of each category of variables (e.g., chronic conditions) using Wald tests.

To account for the sampling design, all descriptive statistics are based on weighted data. In the regression models, we control for age and urban/rural residence and calculate robust standard errors clustered at the township level. Statistical analyses are performed using STATA, version 11.

## RESULTS

### Level of Agreement

The descriptive statistics in Table 1 indicate that, on average, interviewers and physicians perceive a respondent's health to be better than the respondent, and interviewers rate respondent's health better than physicians. The standard deviations shown in Table 1 also suggest that there is less variation around the mean health rating for physicians than respondents and interviewers. However, these descriptive statistics do not provide an indication of the extent of agreement between raters. To assess interrater agreement in health ratings, we calculate both unweighted and weighted kappa statistics.

Table 3 indicates that the weights have a considerable impact on the kappa statistic and corresponding agreement rate. However, based on the Landis and Koch (1977) cut-points and categorization scheme for the kappa statistic, the substantive results are generally the same regardless of whether weights are used. There is only "slight" (0.00-0.20) agreement across all inter-evaluator pairs, except in the case of the weighted kappa statistic for respondent versus interviewer ratings, in which case the agreement level improves marginally to "fair".

### Determinants of Health Ratings

The parameter estimates and standardized coefficients from the ordered probit models of the three summary health measures are found in Table 4. Because the outcome variable ranges from excellent (5) to poor (1), a positive coefficient indicates that a higher value of the variable is associated with a better health rating whereas a negative coefficient indicates that it is associated with a worse health rating.

As expected, our results suggest that self-reported chronic diseases are important factors in health ratings across evaluators – i.e., the chronic disease coefficients are jointly significant in all models and associated with worse health ratings. However, the significance and relative importance (as suggested by standardized coefficients) of specific illnesses vary across evaluators. Only diabetes is statistically significant at the 5% level or higher for all three ratings. The standardized coefficients imply that diabetes is one of the most important factors in health ratings across evaluators.

Mobility limitations are significant predictors of (worse) respondent and interviewer ratings only. Similarly, psychological well-being measures are jointly significant ( $p < 0.01$ ) and associated with worse ratings in respondent and interviewer—but not physician—models.

---

<sup>2</sup>As a robustness test, we estimated models that included only variables reflecting information that the evaluator is known to have had at the time of the assessment through either the survey instrument (in the case of the interviewer and physician) or through self-knowledge (in the case of the respondent). This involved excluding between one (IAH model) and seven (PAH model) covariates. The coefficient estimates for the variables included in both the full and restricted model were similar across models. As an additional test, we estimated equations that included only the common set of variables in the SAH, IAH, and PAH restricted models and found that this also had little impact on our substantive results. The results are available from the authors upon request.



The standardized coefficients suggest that mobility limitations and depressive symptoms are among the most important factors in SAH and IAH ratings. Perceived stress is a significant factor in respondent ratings only. Among the other health indicators, level of pain/discomfort—associated with worse health ratings—is a significant and relatively important factor in both respondent and physician ratings.

Despite being a powerful and well-known risk factor for illness and mortality, smoking is significantly associated with worse health ratings only in the physician model. Smoking is insignificant in respondent and interviewer models even if we remove other measures of pulmonary health, such as self-reported respiratory illness and peak airflow, from the models. The indicator of obesity has a significant and negative effect on health ratings only in the physician model.

All three of the physical performance indicators are significant in the interviewer model, and the standardized coefficients suggest that these measures are among the most heavily weighted in interviewers' overall health assessments. Low peak flow, slow walking speed, and inability to participate in the chair stand test are significant predictors of worse interviewer ratings. Although some aspects of physical performance may be captured in the self-reported mobility and respiratory illness measures, it is interesting that none of the performance measures are significant in the respondent model. Low walking speed is the only significant physical performance measure in physicians' ratings, perhaps because this aspect of physical performance is relatively easy to detect through observation.

The medical exam outcome—detection of a heart abnormality—is significantly associated only with worse PAH ratings, despite the expectation that respondents would know about some heart abnormalities. This variable may be picking up under-reporting or level of severity of heart disease.

Lastly, socio-demographic factors, as a group, are significantly related to health assessments by all three evaluators. However, the significance of the individual measures varies across evaluators. For example, the coefficient on perceived social position is significantly associated with (better) health ratings only for respondents, the coefficient on female is significantly associated with (worse) ratings only for interviewers, and the coefficient on age is significantly associated with (worse) ratings only for physicians. These coefficients may be capturing unmeasured health dimensions, such as more debilitating health conditions among women, older individuals or those of lower SES. However, they could also reflect reporting biases, such as more optimistic self-perceptions of health among those with higher social standing (Shmueli, 2003). Unfortunately, our data do not allow us to distinguish among these explanations.

### **Determinants of Inter-evaluator Differences in Health Ratings**

The above results suggest substantial differences in the factors that influence evaluations of overall health by respondents, interviewers, and physicians. We next examine which factors are significant predictors of inter-evaluator disagreement. We estimate ordered probit models with the same covariates as in Table 4 to determine the effect of measured respondent characteristics on simple differences in ratings. The estimated coefficients and standardized coefficients are presented in Table 5. In the first two columns, a positive coefficient indicates that a higher value of the variable is associated with better health ratings by respondents than interviewers/physicians; in the third column, a positive coefficient indicates that a higher value of the variable is associated with interviewers providing better ratings than physicians.

The Wald tests suggest that differences in the importance of socio-demographic factors and chronic diseases in health ratings across evaluators are significant determinants of ratings' disagreement for all evaluator pairs. In addition, disagreements between physicians and both respondents and interviewers are significantly associated with the greater weight placed on functional/physical limitations and psychological well-being in respondent/interviewer assessments, and the detection of a heart abnormality in physician assessments. This may reflect the limited information available to physicians on physical functioning and psychological well-being at the time of their health assessment. Another explanation may be that physicians do not perceive these health domains as being as important as clinical factors. The difference in health ratings between respondents and interviewers is significantly associated with the greater weight given to the performance assessment measures in interviewer ratings, in particular the inability to perform chair stands, as well as the importance placed on pain/discomfort in respondent ratings.

## DISCUSSION

Although a great deal of research has focused on understanding SAH ratings, a large portion of the variance in SAH remains unexplained. We use a unique dataset that contains extensive self-reported physical and mental health indicators, as well as "objective" health measures, to try to shed light on what factors underlie SAH and how they compare to those incorporated in overall assessments by interviewer and physicians. In doing so, we provide suggestive evidence that external assessments of respondent health may provide information on respondent health that is complementary to SAH ratings.

More specifically, our results suggest as many differences as similarities in the factors incorporated into health ratings by respondents, interviewers, and physicians. We hypothesize that these discrepancies are due in part to inter-evaluator variation in what factors are considered most important to overall health status, as well as to differences in the type and level of information available to the different evaluators. However, our findings indicate that external evaluators may take into account some aspects of health that are not significant or important in respondent's ratings, suggesting that interviewer and physician assessments may be complementary to self-assessed health measures. For example, our results suggest that interviewers place relatively more weight on various aspects of physical functioning than respondents. Our results also imply that, not surprisingly, differences in ratings between respondents and physicians are driven by differences in the weight given to clinical measures or risk factors of illness and mortality, most notably indicators of cardiovascular conditions, BMI, and smoking status. These results suggest that, by taking into account several important measures of health that are given little weight in respondent's self-assessments, external evaluators impart information on respondent health that is not captured in the SAH measure. Thus, these findings also underscore some of the deficiencies of the SAH measure.

The emphasis given to various aspects of health by the different types of evaluators was undoubtedly influenced by the information available to each of them at the time of their response to the overall health rating question. Respondents participated in the physical performance assessments and the medical exam subsequent to providing their health ratings; interviewers never had access to information from the medical exams; and physicians never had access to the household interview information or the physical performance assessments. With regard to information on psychological well-being, respondents presumably had extensive knowledge about their own mental health, interviewers had responses to a modest number of questions in the home interview and physicians had only the insights they inferred during the medical exam. However, these differences in data availability across evaluators do not provide a complete explanation for the differential weighting of factors in

health ratings across evaluators. Aspects of physical functioning (such as low walking speed) or physical health (such as being underweight or obese) are likely to be gleaned without a physical performance assessment or clinical measurement, and information on some health factors (such as smoking status) was provided to both external evaluators.

Our results show that differences in the factors incorporated into health ratings across evaluators result in extensive inter-evaluator disagreement in ratings. In our sample, interviewers and physicians tended to perceive a respondent's health to be better than the respondent, and interviewers rated respondent's health better than physicians.

We also find suggestive evidence of differences in reporting styles, with respondents and physicians favoring the middle rating ("Average") and interviewers favoring "Good". In addition, we find that perceived social position, sex, and age are significant predictors of respondent, interviewer, and physician ratings, respectively, even after controlling for extensive self-reported and "objective" health measures. This may indicate that these socio-demographic characteristics are capturing unmeasured aspects of health, but may also reflect reporting biases in evaluator ratings. For instance, the finding that higher perceived social position and advanced age are associated with better health ratings by respondents than physicians is consistent with research suggesting the existence of reporting heterogeneity by age and SES, whereby older and higher SES individuals report better subjective health than objective measures of health status indicate (Groot, 2000; Humphris & Van Doorslaer, 2000; Shmueli, 2003).

At the time of this analysis, health outcome or survival data for respondents subsequent to the 2006 survey were not available. Therefore, we are not able to assess the validity of each of the summary health measures by estimating its explanatory power in predicting future health outcomes. However, these preliminary findings lead us to expect that including IAH and PAH ratings in models of health and survival will increase the predictive power of the models. In addition, interviewer and physician ratings could be used to improve summary health indices. Given the ease and low cost of collecting interviewer assessments (i.e., adding a single question at the end of the interview), and the increasing use of clinical assessments and medical personnel in population surveys in several developed and developing countries, these are important areas for future research.

## Acknowledgments

This work was supported by the Demography and Epidemiology Unit of the Behavioral and Social Research Program of the National Institute of Aging (Grant R01AG16790) and the Eunice Kennedy Shriver National Institute of Child Health and Human Development (Grant R24HDO47879). We are grateful to Dana Gleib, James Trussell, and Germán Rodríguez for helpful comments and suggestions.

## References

- Benyamini Y, Idler EL. Community studies reporting association between self-rated health and mortality: Additional studies, 1995 to 1998. *Research on Aging*. 1999; 1999:392–401.
- Benyamini Y, Leventhal H, Leventhal EA. Self-assessments of health: What do people know that predicts their mortality? *Research on Aging*. 1999; 27:477–500.
- Benyamini Y, Leventhal EA, Leventhal H. Elderly people's ratings of the importance of health-related factors to their self-assessments of health. *Social Science & Medicine*. 2003; 56:1661–68. [PubMed: 12639583]
- Borawski EA, Kinney JM, Kahana E. The meaning of older adults' health appraisals: Congruence with health status and determinant of mortality. *Journal of Gerontology: Social Sciences*. 1996; 51B(3):S157–S170.

- Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh sleep quality index: A new instrument for psychiatric practice and research. *Psychiatry Research*. 1989; 28:193–213. [PubMed: 2748771]
- Chang, M.; Glei, D.; Goldman, N.; Weinstein, M. The Taiwan biomarker project. In: Weinstein, M.; Vaupel, JW.; Wachter, KW., editors. *Biosocial surveys. Committee on advances in collecting and utilizing biological indicators and genetic information in social science surveys*, Committee on Population, Division of Behavioral and Social Sciences and Education. Washington, D.C.: The National Academies Press; 2007. p. 60-77.
- Chipperfield JG. Incongruence between health perceptions and health problems: Implications for survival among seniors. *Journal of Aging and Health*. 1993; 5(4):475–496.
- Christensen K, Thinggaard M, McGue M, Rexbye H, Hjelmberg JvB, Aviv A, et al. Perceived age as clinically useful biomarker of ageing: Cohort study. *BMJ*. 2009; 339(Dec 11\_2):b5262. [PubMed: 20008378]
- Cockerham WC, Sharp K, Wilcox JA. Aging and perceived health status. *Journal of Gerontology*. 1983; 38:349–355. [PubMed: 6841932]
- Ferraro K. Self-ratings of health among the old and old-old. *Journal of Health and Social Behavior*. 1980; 21:377–383. [PubMed: 7204931]
- Ferraro KF, Su Y. Physician-evaluated and self-reported morbidity for predicting disability. *American Journal of Public Health*. 2000; 90:103–108. [PubMed: 10630145]
- Fried LP, Herdman SJ, Kuhn KE, Rubin G, Turano K. Preclinical disability: Hypotheses about the bottom of the iceberg. *Journal of Aging and Health*. 1991; 3:285–300.
- Friedsam HJ, Martin HW. A comparison of self and physicians' health ratings in an older population. *Journal of Health and Human Behavior*. 1963; 4(3):179–183. [PubMed: 14066576]
- Goldman N, Cornman JC, Chang M. Measuring subjective social status: A case study of older Taiwanese. *Journal of Cross-Cultural Gerontology*. 2006; 21:71–89. [PubMed: 17106645]
- Goldman N, Glei D, Chang M-C. The role of clinical risk factors in understanding self-rated health. *Annals of Epidemiology*. 2003; 14(1):49–57. [PubMed: 14664780]
- Goldman N, Glei D, Seplaki C, Liu I-Wen, Weinstein M. Perceived stress and physiological dysregulation in older adults. *Stress*. 2005; 8(2):95–105. [PubMed: 16019601]
- Goldman N, Lin I-F, Weinstein M, Lin Y-H. Evaluating the quality of self-reports of hypertension and diabetes. *Journal of Clinical Epidemiology*. 2003; 56(2):148–154. [PubMed: 12654409]
- Groot W. Adaption and scale of reference bias in self-assessments of quality of life. *Journal of Health Economics*. 2000; 19(3):403–420. [PubMed: 10977197]
- Guralnik JM, Seeman TE, Tinetti ME, Nevitt MC, Berkman LF. Validation and use of performance measures of functioning in a non-disabled older population: MacArthur Studies of Successful Aging. *Aging*. 1994; 6:410–419. [PubMed: 7748914]
- Guralnik JM, Ferrucci L, Pierper CF, Leveille SG, Markides KS, Ostir GV, Studenski S, Berkman LF, Wallace RB. Lower extremity function and subsequent disability: Consistency across studies, predictive models, and value of gait speed alone compared with the short physical performance battery. *Journal of Gerontology*. 2000; 4:M221–M231.
- Hall JA, Epstein AM, McNeil BJ. Multidimensionality of health status in an elderly population: Construct validity of a measurement battery. *Medical Care*. 1989; 27(3):S168–S177. [PubMed: 2921884]
- Humphris K, van Doorslaer E. Income-related health inequality in Canada. *Social Science & Medicine*. 2000; 50:663–671. [PubMed: 10658847]
- Idler EL. Age differences in self-assessments of health: Age changes, cohort differences, or survivorship? *Journal of Gerontology*. 1993; 48(6):S289–S300. [PubMed: 8228003]
- Idler EL, Benyamini Y. Self-rated health and mortality: A review of twenty-seven community studies. *Journal of Health and Social Behavior*. 1997; 38:21–37. [PubMed: 9097506]
- Idler EL, Kasl SV. Self-ratings of health: Do they predict change in functional ability? *Journals of Gerontology B: Psychological Sciences and Social Sciences*. 1995; 50:S344–S353.
- Jylhä M, Volpato S, Guralnik JM. Self-rated health showed a graded association with frequently used biomarkers in a large population sample. *Journal of Clinical Epidemiology*. 2006; 59(5):465–471. [PubMed: 16632134]

- Krause NM, Jay GM. What do global self-rated health items measure? *Medical Care*. 1994; 32(9): 930–942. [PubMed: 8090045]
- Kriegsman DMW, Penninx BWJH, Van Eijk JTM, Boeke AJP, Deeg DJH. Self-reports and general practitioner information on the presence of chronic diseases in community dwelling elderly : A study on the accuracy of patients' self-reports and on determinants of inaccuracy. *Journal of Clinical Epidemiology*. 1996; 49(12):1407–1417. [PubMed: 8970491]
- Krueger DE. Measurement of prevalence of chronic disease by household interviews and clinical evaluations. *American Journal of Public Health*. 1957; 47:953–60. [PubMed: 13444505]
- Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977; 33:159–174. [PubMed: 843571]
- Levkoff SE, Cleary PD, Wetle T. Differences in the appraisal of health between aged and middle-aged adults. *Journal of Gerontology*. 1987; 42:114–120. [PubMed: 3794190]
- Markides KS, Lee DJ, Ray LA, Black SA. Physicians' ratings of health in middle and old age: A cautionary note. *Journals of Gerontology: Social Sciences*. 1993; 48(1):S24–S27.
- Mossey JM, Shapiro E. Self-rated health: A predictor of mortality among the elderly. *American Journal of Public Health*. 1982; 72:800–808. [PubMed: 7091475]
- Quanjer PH, Lebowitz MD, Gregg I, Miller MR, Pedersen OF. Peak expiratory flow: Conclusions and recommendations of a Working Party of the European Respiratory Society. *The European Respiratory Journal*. 1997; 24:2S–8S. [PubMed: 9098701]
- Reuben B, Siu AL, Kimpau S. The predictive validity of self-report and performance-based measures of function and health. *Journal of Gerontology*. 1992; 47:M106–M110. [PubMed: 1624692]
- Shmueli A. Socio-economic and demographic variation in health and its measures: The issue of reporting heterogeneity. *Social Science & Medicine*. 2003; 57:125–134. [PubMed: 12753821]
- Suchman EA, Phillips BS, Streib Gordon F. An analysis of the validity of health questionnaires. *Social Forces*. 1958; 36(3):223–232.
- Taiwan Provincial Institute of Family Planning; Population Studies Center and Institute of Gerontology, University of Michigan. 1989 Survey of Health and Living Status of the Elderly in Taiwan: Questionnaire and survey design. Ann Arbor: Population Studies Center, University of Michigan; 1989.
- Taiwan Provincial Institute of Family Planning; Population Studies Center and Institute of Gerontology, University of Michigan. 1996 Survey of Health and Living Status of the Elderly and Middle Aged in Taiwan (Comparative study of the elderly in four Asian countries, Taiwan aging studies series: Res Rep Nos 7-1 & 7-2). Ann Arbor: Population Studies Center, University of Michigan; 1997.
- Valanis BG, Yeaworth R. Ratings of physical and mental health in the older bereaved. *Research in Nursing Health*. 1982; 5(3):137–146. [PubMed: 6923442]
- Van Frago, Gahbauer EA, Van Ness PH, Concato J, Gill TM. Peak expiratory flow as a predictor of subsequent disability and death in community-living older persons. *Journal of the American Geriatrics Society*. 2008; 56(6):1014–1020. [PubMed: 18422951]
- Zimmer Z, Natividad J, Hui-Sheng L, Chayovan N. A cross-sectional analysis of the determinants of self-assessed health. *Journal of Health and Social Behavior*. 2000; 41(4):465–481. [PubMed: 11198569]

**Table 1**

Distribution of summary health measures (%), SEBAS 2006

	Self-assessed health	Interviewer-assessed health	Physician-assessed health
Excellent [5]	10.8	26.5	4.2
Good [4]	22.1	44.1	40.3
Average [3]	43.0	21.0	44.2
Not so good [2]	20.9	7.4	10.7
Poor [1]	3.2	1.1	0.6
Mean rating (s.d)	3.2 (0.98)	3.9 (0.92)	3.4 (0.75)

Notes: N=848. Percentages are based on weighted data. Standard deviations are shown in parentheses.

**Table 2**

Summary statistics, SEBAS 2006

	<b>Mean</b>	<b>Std. Dev.</b>
<b>Socio-demographic variables</b>		
Female	0.47	0.50
Age	65.02	9.26
Perceived social position	4.36	1.76
Urban	0.50	0.50
No. of social activities	0.87	1.11
<b>Chronic disease</b>		
High blood pressure or use of antihypertensive agents	0.33	0.47
Diabetes	0.17	0.38
Heart disease	0.15	0.36
Cancer	0.01	0.12
Respiratory illness	0.06	0.24
Liver	0.08	0.28
Kidney disease	0.05	0.21
Gout	0.08	0.27
<b>Functional limitations</b>		
No. of mobility limitations	1.65	2.21
<b>Other Health Indicators</b>		
No. of hospital days	1.47	5.94
Level of pain/discomfort	0.68	0.86
<b>Health Behaviors</b>		
Smoke daily in past 6 months	0.18	0.38
Frequency of relaxation activities	0.88	1.55
<b>Psychological well-being</b>		
CES-Depression scale	4.50	5.29
Stress index	0.28	0.36
Sleep index	3.96	3.23
<b>Biomarkers</b>		
BMI <= 18.5 (underweight)	0.03	0.16
BMI >= 30 (obese)	0.06	0.25
<b>Performance assessment measures</b>		
Low peak flow	0.27	0.44
Low walking speed	0.25	0.43
Unable to perform chair stand	0.05	0.21
<b>Results of physical exam</b>		
Heart abnormality detected	0.07	0.26

Notes: N=848. Descriptive statistics are based on weighted data.

**Table 3**

## Interrater Agreement - Kappa statistics

	Kappa	Agreement	Expected Agreement
<b>Panel A: Unweighted Kappa statistic</b>			
SAH vs. IAH	0.128 <sup>**</sup>	33.0%	23.2%
SAH vs. PAH	0.095 <sup>**</sup>	37.3%	30.7%
PAH vs. IAH	0.085 <sup>**</sup>	34.8%	28.8%
<b>Panel B: Weighted Kappa statistic</b>			
SAH vs. IAH	0.271 <sup>**</sup>	77.9%	69.7%
SAH vs. PAH	0.176 <sup>**</sup>	80.3%	76.1%
PAH vs. IAH	0.154 <sup>**</sup>	78.7%	74.9%

Notes: The weighted kappa is based on the following weights:  $1 - (|i-j|/(k-1))$ , where  $i$  and  $j$  indicate the rows and columns of the ratings by the two raters and  $k$  is the number of possible ratings. The following interpretation of Kappa statistics is commonly used (Landis and Koch, 1977): 0.00-0.20=slight; 0.21-0.40=fair; 0.41-0.60=moderate; 0.61-0.80=substantial; and 0.81-1.0=almost perfect

<sup>\*\*</sup>  
p < 0.01



**Table 4**  
 Estimated coefficients from ordered probit models of three summary health measures, SEBAS 2006

	Coefficients			Standardized coefficients		
	SAH	IAH	PAH	SAH	IAH	PAH
<b>Socio-demographic variables</b>						
Female	-0.003	-0.247**	0.085	-0.002	-0.123**	0.042
Age	0.010 <sup>+</sup>	0.001	-0.015**	0.094 <sup>+</sup>	0.005	-0.151**
Perceived social position	0.068*	0.015	-0.005	0.122*	0.027	-0.009
Urban	0.176*	0.254 <sup>+</sup>	0.117	0.086*	0.124 <sup>+</sup>	0.057
No. of social activities	0.051 <sup>+</sup>	0.092**	0.039	0.056 <sup>+</sup>	0.102**	0.043
	<i>Joint significance [<math>\chi^2</math>]</i>					
	13.79*	18.43**	28.21**			
<b>Chronic disease</b>						
High blood pressure/use of antihypertensive agents	-0.188*	-0.192 <sup>+</sup>	-0.317**	-0.090*	-0.092 <sup>+</sup>	-0.151**
Diabetes	-0.331**	-0.543**	-0.654**	-0.124**	-0.204**	-0.246**
Heart disease	-0.312*	0.018	-0.246*	-0.116*	0.007	-0.092*
Cancer	-1.050*	-0.899 <sup>+</sup>	-0.559	-0.119*	-0.102 <sup>+</sup>	-0.063
Respiratory illness	-0.436**	-0.211	-0.117	-0.107**	-0.052	-0.029
Liver	-0.079	-0.173	-0.255*	-0.022	-0.048	-0.071*
Kidney disease	-0.011	-0.304 <sup>+</sup>	-0.453**	-0.002	-0.064 <sup>+</sup>	-0.095**
Gout	-0.208 <sup>+</sup>	-0.316*	0.090	-0.057 <sup>+</sup>	-0.087*	0.025
	<i>Joint significance [<math>\chi^2</math>]</i>					
	41.90**	56.27**	127.77**			
<b>Functional limitations</b>						
No. of mobility limitations	-0.098**	-0.149**	-0.011	-0.217**	-0.331**	-0.025
<b>Other Health Indicators</b>						
No. of hospital days	-0.015	-0.014 <sup>+</sup>	-0.009	-0.084	-0.077 <sup>+</sup>	-0.051
Level of pain/discomfort	-0.188**	-0.062	-0.125*	-0.161**	-0.053	-0.108*
	<i>Joint significance [<math>\chi^2</math>]</i>					
	14.12**	3.46	6.92*			
<b>Health Behaviors</b>						

	Coefficients				Standardized coefficients			
	SAH	IAH	PAH		SAH	IAH	PAH	
Smoke daily in past 6 months	0.140	-0.050	-0.228*		0.0529	-0.019	-0.086*	
Frequency of relaxation activities	0.041 <sup>+</sup>	0.098**	-0.006		0.064 <sup>+</sup>	0.153**	-0.009	
	<i>Joint significance [<math>\chi^2</math>]</i>							
	5.58 <sup>+</sup>	16.33**	3.93					
<b>Psychological well-being</b>								
CES-Depression scale	-0.042**	-0.035**	-0.006		-0.220**	-0.185**	-0.034	
Stress index	-0.246**	-0.028	0.117		-0.087**	-0.010	0.041	
Sleep index	-0.056**	-0.046*	-0.025*		-0.182**	-0.150*	-0.081*	
	<i>Joint significance [<math>\chi^2</math>]</i>							
	77.91**	27.51**	7.35 <sup>+</sup>					
<b>Biomarkers</b>								
BMI ≤ 18.5 (underweight)	0.049	-0.186	-0.491 <sup>+</sup>		0.008	-0.030	-0.080 <sup>+</sup>	
BMI ≥ 30 (obese)	-0.107	-0.178	-0.345*		-0.026	-0.044	-0.085*	
	<i>Joint significance [<math>\chi^2</math>]</i>							
	1.48	2.87	7.23*					
<b>Performance assessment measures</b>								
Low peak flow	-0.111	-0.256**	-0.055		-0.049	-0.114**	-0.024	
Low walking speed	-0.132	-0.245**	-0.227*		-0.058	-0.107**	-0.099*	
Unable to perform chair stand	0.116	-0.556*	0.021		0.025	-0.121*	0.005	
	<i>Joint significance [<math>\chi^2</math>]</i>							
	2.83	33.57**	10.92*					
<b>Results of physical exam</b>								
Heart	-0.095	-0.143	-0.900**		-0.025	-0.038	-0.241**	
Pseudo R-sq	0.163	0.219	0.137		0.163	0.219	0.137	
N	848	848	848		848	848	848	

Notes: Dependent variable is a 5-point summary health measure that ranges from 5=excellent to 1=poor. Robust standard errors, clustered at the township level, are calculated.  $\chi^2$  is based on Wald tests of joint significance of the preceding set of variables.

<sup>+</sup> p<0.10,

\* p<0.05,

\*\* p<0.01

**Table 5**

Estimated coefficients from ordered probit models of ratings disagreement, SEBAS 2006

	Coefficients				Standardized coefficients			
	SAH-IAH	SAH-PAH	IAH-PAH	SAH-IAH	SAH-PAH	IAH-PAH	SAH-PAH	IAH-PAH
<b>Socio-demographic variables</b>								
Female	0.167*	-0.056	-0.216**	0.083*	-0.028	-0.108*		
Age	0.006	0.017**	0.012*	0.062	0.167**	0.113*		
Perceived social position	0.052**	0.059**	0.011	0.093**	0.105**	0.020		
Urban	-0.020	0.076	0.096	-0.010	0.037	0.047		
No. of social activities	-0.011	0.022	0.034	-0.012	0.024	0.037		
	<i>Joint significance [<math>\chi^2</math>]</i>							
	12.97*	23.03**	24.41**					
<b>Chronic disease</b>								
High blood pressure or use of antihypertensive agents	-0.035	0.041	0.081	-0.017	0.020	0.039		
Diabetes	0.137	0.160*	0.031	0.052	0.060*	0.012		
Heart disease	-0.278*	-0.092	0.174	-0.104*	-0.034	0.065		
Cancer	-0.123	-0.389	-0.310	-0.014	-0.044	-0.035		
Respiratory illness	-0.223+	-0.247	-0.060	-0.055+	-0.061	-0.015		
Liver	0.086	0.110	0.038	0.024	0.031	0.011		
Kidney disease	0.239	0.284	0.075	0.050	0.060	0.016		
Gout	0.057	-0.225+	-0.290+	0.016	-0.062+	-0.080+		
	<i>Joint significance [<math>\chi^2</math>]</i>							
	19.62*	51.52**	16.42*					
<b>Functional limitations</b>								
No. of mobility limitations	0.036	-0.071**	-0.107**	0.081	-0.157**	-0.238**		
<b>Other Health Indicators</b>								
No. of hospital days	-0.001	-0.005	-0.004	-0.003	-0.025	-0.022		
Level of pain/discomfort	-0.117**	-0.071	0.036	-0.100**	-0.061	0.031		
	<i>Joint significance [<math>\chi^2</math>]</i>							
	8.80*	2.10	1.12					
<b>Health Behaviors</b>								
Smoke daily in past 6 months	0.170+	0.263*	0.119	0.064+	0.099*	0.045		

	Coefficients				Standardized coefficients			
	SAH-JAH	SAH-PAH	IAH-PAH	IAH-JAH	SAH-JAH	SAH-PAH	IAH-PAH	IAH-JAH
Frequency of relaxation activities	-0.032	0.034	0.067*	-0.050	0.054	0.104*		
	<i>Joint significance [<math>\chi^2</math>]</i> 5.39 <sup>+</sup> 7.78* 5.19 <sup>+</sup>							
<b>Psychological well-being</b>								
CES-Depression scale	-0.003	-0.025*	-0.024*	-0.016	-0.135*	-0.124*		
Stress index	-0.177 <sup>+</sup>	-0.259**	-0.088	-0.062 <sup>+</sup>	-0.091**	-0.031		
Sleep index	-0.014	-0.028 <sup>+</sup>	-0.016	-0.044	-0.091 <sup>+</sup>	-0.053		
	<i>Joint significance [<math>\chi^2</math>]</i> 6.97 <sup>+</sup> 54.72** 13.98**							
<b>Biomarkers</b>								
BMI <= 18.5 (underweight)	0.180	0.340	0.198	0.029	0.055	0.032		
BMI >= 30 (obese)	0.004	0.124	0.115	0.001	0.031	0.028		
	<i>Joint significance [<math>\chi^2</math>]</i> 0.48 2.72 1.55							
<b>Performance assessment measures</b>								
Low peak flow	0.095	-0.056	-0.153**	0.042	-0.025	-0.068**		
Low walking speed	0.080	0.050	-0.021	0.035	0.022	-0.009		
Unable to perform chair stand	0.618**	0.128	-0.469*	0.134**	0.028	-0.102*		
	<i>Joint significance [<math>\chi^2</math>]</i> 15.63 1.56 12.64**							
<b>Results of physical exam</b>								
Heart	0.009	0.487**	0.517**	0.002	0.131**	0.139**		
Pseudo R-sq	0.031	0.061	0.067	0.031	0.061	0.067		
N	848	848	848	848	848	848		

Notes: Dependent variable is a 5-point summary health measure that ranges from 5=excellent to 1=poor. Robust standard errors, clustered at the township level, are calculated.  $\chi^2$  is based on Wald tests of joint significance of the preceding set of variables.

<sup>+</sup> p<0.10,

\* p<0.05,

\*\* p<0.01