Title: "The Demographics of Self-Rated Health: A Systematic Analysis of the Dynamics of Who Reports What When (and Why)"

Authors: Dean Lillard, Dongyue Ying (Ohio State University)

Extended abstract

Many surveys routinely ask respondents to assess their own physical and mental health and to rate it. The surveys collect these data using a five-point Likert scale that usually ranges from very poor to excellent. Although it is natural to question whether people accurately report their own health, social scientists use self-reported health (SRH) data as a proxy for each individual's true underlying health. Empirical evidence documents that an individual's future risk of mortality is strongly associated with current SRH (Idler & Benyamini, 1997; Benyamini & Idler, 1999). While there is evidence that the predictive value of SRH seems to be increasing over calendar time (Schnittker & Bacak, 2014), there is also evidence that the association between SRH and future mortality varies across groups characterized by various socio-economic characteristics (e.g. Dalen et al., 2012; Woo & Zajacova, 2017).

The empirical evidence correlates an individual's SRH at a point in time with his or her future mortality. Most such studies rely on SRH from a subsample of pooled cross-sectional NHIS data whose respondent records have been linked to the corresponding records from the US Mortality Files.

In this paper we look both forward and back to investigate what level of SRH an individual reports at a given time and how that level systematically varies with particular health events, past and future. We model SRH as a function of the past because current health at any time is a reflection of a long series of past investments and health production. We rely on this idea when we model SRH because, as in the literature on mortality, if SRH conveys information, then observed differences in SRH must predict future health conditions and events.

This work builds on earlier evidence showing that early childhood exposure to income inequality is correlated with worse SRH in later life (Lillard et al. 2015; Burkhauser et al. 2016). We describe SRH and it's evolution for the general population, by age, sex, and race/ethnic groups. We also describe how SRH has evolved over the 1980s, 1990s, 2000s, and 2010s.

We use data from the Panel Study of Income Dynamics (PSID). The PSID asked "heads" and "wives" to report SRH annually from 1984-1997 and biennially ever since. In every odd-numbered year since 1999 the PSID also asks respondents about a set of physical, mental, and emotional health conditions. The set includes arthritis, asthma, high blood pressure, cancer, diabetes, heart attack, heart disease, hypertension, lung disease, stroke, emotional/nervous/psychiatric problems, learning disorder, or a permanent loss of memory or mental ability. For each condition, respondents report whether or not a doctor ever diagnosed them with the condition, the age he or she was first diagnosed, the degree to which the condition limits normal daily activities, and the duration of the condition (in years, months, weeks, and days). Starting in 2001 the PSID asks a series of questions about depression. Starting in 2007 the PSID also asks heads and wives to report whether they experienced each of a long list of conditions from birth to age 17 (year-by-year) and the age the condition ended (if it has). The childhood health conditions list includes allergies, asthma, chicken pox, chronic ear problems, depression, difficulty seeing, drug or alcohol problems, epilepsy/seizures, heart trouble, high blood pressure, measles, mumps, other emotional/psychological problems, respiratory disorders, severe headaches/migraines, speech impairment, and stomach problems. Starting in 2005 the PSID asks about other chronic conditions and, in 2011, began to ask respondents to list the specific chronic conditions they suffer.

We code information about all the conditions asked since 1999 because respondents report retrospectively the age they were first diagnosed and the duration of the condition. We add all information about the childhood conditions (from birth to age 17) because respondents report a year-by-year history.

We use these data to construct each respondent's health history. We code indicator variables for each condition, in every year of a respondent's life. The variable takes a value of "0" if the person said he or she had did not experience the condition at that age and a value of "1" if he or she experienced the condition. We merge to each respondent's life history, his or her self-reported health in every year from 1984 to 2015 if he or she reported it.

Our dependent variable is self-reported health. The PSID asks respondents to rate their current health and assign it to one of five Likert-scale categories. The categories range from "Excellent" (value 1) to "Poor" (value 5).

In modeling how current self-reported health varies with health conditions, we depart from the literature in an important way. Much of the literature collapses the five-category SRH data into a binary variable that equals "0" if a person is in "excellent," "very good," or "good" health and "1" if a person is in "good," "fair," or "poor" health. (We ignore studies that run ordinary least squares on SRH data because their results are uninterpretable.) This treatment of the data has two major drawbacks we avoid. First, it throws away information. As importantly, by collapsing the data (or by using OLS) one implicitly imposes unstated assumptions about movements across categories.

Instead we fit ordered probit models that estimate the probability a person categorizes his/her health in each of the five possible categories. The ordered probit method also estimates where in the distribution of the probability index that each category starts and ends.

Our data vary across persons (i) who belong to one of 15 birth cohorts (c) and who report their health in successive calendar years (t). Our self-reported health data, h_{it}, represent the continuously distributed underlying state of true health, h_{it}^*, in five categories as:

$$h_{ict} = \begin{cases} 1 & if & & h_{ict}^* & \leq 0 \\ 2 & if & 0 < & h_{ict}^* & \leq \mu_1 \\ 3 & if & \mu_1 < & h_{ict}^* & \leq \mu_2 \\ 4 & if & \mu_2 < & h_{ict}^* & \leq \mu_3 \\ 5 & if & \mu_3 < & h_{ict}^* & \end{cases} \quad (1)$$

While we do not observe the underlying true health, we assume that it can, in principle, varies with the set of current, past, and future health conditions/events that a person reports. We denote these as H_{ict}, $\sum_{k=1}^{t-1} H_{ick}$, and $\sum_{j=1}^{T} H_{icj}$ respectively. We include an index of childhood health conditions intended to capture an individual's overall health during childhood, $\hat{I}_{t-k}$. We also include, in the extended model, a vector of individual characteristics, X_{it}, that may include both time invariant (e.g. genetic make-up) and elements that change over time (e.g. health behaviors) and permanent family income, $\widehat{FAMinc}_{it-2}$. Formally we suppose that true current health is given by:

$$h_{ict}^* = \gamma_1 H_{ict} + \gamma_2 \sum_{k=1}^{t-1} H_{ick} + \gamma_3 \sum_{k=1}^{t-1} (t-k) * H_{ick} + \gamma_4 \sum_{j=1}^{T} H_{icj} + \gamma_5 \sum_{j=1}^{T} (t+j) * H_{icj} + \gamma_6 \widehat{H}_{ic18} + \beta X_{ict} + \gamma_7 \widehat{FAMinc}_{ict-2} + \epsilon_{ict} \quad (2)$$

where ϵ_{ict} is a normally distributed error term with mean zero that captures stochastic, individual-specific shocks to health in each period. In fact, all of our health covariates are vectors that include as many elements as

conditions we observe. The vector of future health conditions in principle varies up to the last year of a person's life. For both the past and future health event history, we add the interaction term that is a polynomial of the number of years that have elapsed since (remain) until the person was diagnosed with (will get) the particular condition.

Note that our sample is right-censored and subject to a mortality-specific selection bias because we do not observe health conditions for PSID respondents who died between 1984 (when they first reported SRH) and 1999 (2007) when they reported on the health conditions (health conditions in childhood). We also systematically fail to observe the evolution of health conditions (i.e. future health) of PSID respondents who answered the surveys starting in 1999 but who subsequently suffered more severe and debilitating health events. For those respondents our health histories are incomplete. (Note – we are able to impute some of those histories using the PSID "Death" file that lists the date and up to six causes of death for PSID respondents who attrited because they died.)

To use categorical health data researchers follow a standard latent variable approach because they do not observe data on the continuously distributed true health of individuals. To use the latent variable approach we must assume that the underlying true health is a linear function of determinants.

We order the data so that the first category corresponds to "poor" health and the last category to "excellent" health. We follow the standard approach and assume that the error term in (2) is normally distributed so that we can model the probability that one observes an individual in the j-th category as an ordered probit model.

$$P\left(h_{ict}^{*} = h_{ict}(j)\right) = Pr(\mu_{j-1} < (expression\ in\ (2) \leq \mu_j) \qquad (3)$$

We estimate ordered-probit models that restrict various combinations of the coefficients in (2) to zero. We first describe SRH controlling only for basic demographic covariates (age, sex, race/ethnicity). We then show how current SRH varies with the current set of conditions a person suffers. In successive models we add the history of each person's health, the time elapsed since experiencing the conditions/events, the future health events/conditions he/she will experience and the time until the diagnosis or onset of those conditions/events. We show how the estimates vary when one conditions on an index of experienced childhood health conditions.

Over successive models we explore how self-reported health in a given year varies with:

1. Basic covariates only (age, sex, race/ethnicity)
2. Contemporaneous health conditions listed above
3. The same health conditions experienced in the past
4. The length of time since the person was first diagnosed with each of the health conditions
5. The same health conditions not yet experienced but that will be diagnosed in a future year
6. The length of time that will pass until a doctor diagnoses the respondent as suffering the health condition.
7. An index of childhood health conditions
8. Additional demographic and and family characteristics (e.g. "permanent" family income)
9. Membership in cohorts born in different five-year periods that date back to the 1920s
10. Membership in demographic groups defined by race, sex, and years of completed schooling.

All models control for age, race/ethnicity, and sex of the individual.

Estimates of γ_1 - γ_6 reveal how current self-reported health systematically reflects not only particular underlying health conditions but also how that association evolved from the time a person first got diagnosed with each

condition. The coefficients on future health conditions/events also help reveal variation in current SRH that (likely) reflects behaviors or health that may or may not be observed currently but that will lead to future health shocks/events. For example, the onset of diabetes in the future will be systematically related to observed (current and past) body mass index and also the probability a person experiences a heart attack in the future.

Our preliminary results show that health conditions, such as hypertension, have a "long arm" from the past. People who developed hypertension early in life systematically report worse current health than those who developed it more recently. The association is large. For other events, e.g. stroke, SRH falls by more but (incompletely) recovers as time passes.

As noted above, we are in the process of accounting for selection bias because some of our conditions (e.g. cancer) cause a differential attrition that systematically selects the sample. Even in our preliminary results we observe this effect – current SRH drops when people report being diagnosed with cancer but the association is short-lived. This result is plausibly due to the early mortality of people diagnosed with severe cancers that kill quickly. We are working to account for this type of selection bias.

Our preliminary results also expose the heterogeneity in SRH that is likely due to unobserved (or only partially observed) behaviors/conditions that will result in future degradations to health. Current SRH is systematically lower for people who (will soon) be diagnosed with diabetes, lung disease and experience a stroke; for diseases such as cancer and learning disorder, the association is weaker and less statistically significant.

We use the coefficients we estimate to then describe the evolution of health over time (from 1984 to 2015) for the whole population (with PSID weights) and by groups defined by sex, race/ethnicity, and years of completed schooling. We will estimate models with and without individual fixed effects.

References

Benyamini Y, Idler EL. (1999). "Community studies reporting association between self-rated health and mortality additional studies, 1995 to 1998." *Research on Aging*, Vol. 21: 392–401.

Burkhauser, RV, Hahn, MH, Lillard, DR, Wilkins, R. (2016.) "Does Income Inequality in Early Childhood Predict Self-Reported Health In Adulthood? A Cross-National Comparison of the United States and Great Britain." Research in Labor Economics, Vol. 43 (Inequality: Causes and Consequences): 407-476.

Dalen, J D, Huijt, T, Krokstad, S, & Eikemo, TA. (2012). "Are there educational differences in the association between self-rated health and mortality in Norway? The HUNT Study." *Scandinavian Journal of Public Health*, Vol. 40: 641–647.

Lillard, DR, Burkhauser, RV, Hahn, MH, Wilkins, R. (2015). "Does Early-Life Income Inequality Predict Self-Reported Health In Later Life? Evidence From the US." Social Science and Medicine, 128(3): 347-355 http://dx.doi.org/10.1016/j.socscimed.2014.12.026.

Idler EL, Benyamini Y. (1997). "Self-rated health and mortality: A review of twenty-seven community studies." *Journal of Health and Social Behavior*, Vol. 38: 21–37.

Schnittker, J, Bacak, V. (2014). "The Increasing predictive validity of self-rated health." *PLoS ONE*, 9.

Woo, H, Zajacova, A. (2017). "Predictive Strength of Self-Rated Health for Mortality Risk Among Older Adults in the United States: Does It Differ by Race and Ethnicity?" *Research on Aging*, Vol. 39(7): 879–905.