Health and Wealth

Like Deng
Georgia Institute of Technology
Atlanta, Georgia
ldeng31@gatech.edu

Sahil Dhingra
Georgia Institute of Technology
Atlanta, Georgia
sahil.dhingra@gatech.edu

Shan Huang
Georgia Institute of Technology
Atlanta, Georgia
shuang418@gatech.edu

Shuangke Li
Georgia Institute of Technology
Atlanta, Georgia
shuangke94@gatech.edu

ABSTRACT

Health and wealth are two of the key pillars of modern human life. In this project, we intend to answer questions regarding the relationship between health and wealth using advanced analytical tools while incorporating interactive visualizations to effectively represent our findings.

1 OBJECTIVES

The purpose of this project is to reveal the relationship between health and wealth and to construct an interactive visualization that represents our findings effectively. Intuitively, one may assume that those with relatively poor health conditions are more likely to suffer from a poor financial situation such as low income due to job limitations or small savings due to large medical costs. We will test such hypotheses using data obtained from various organizations.

2 CURRENT PRACTICE/ SURVEY

There are numerous and extensive studies on the relationship between health and wealth that measure the cumulative effects of health shocks over life-cycle using structural models, finding that the effects of non-pecuniary are large for poor health [11]. Unlike their computationally burdensome structural model, statistical asset testing shows positive rule using the Medicaid data [13]. Several papers have found large disparities in wealth between the (latent) healthy and the unhealthy individuals in the Health and Retirement Study (HRS) dataset (See e.g.[4], [7], [8], [9], [15]), but in only partial aspects.

Moreover, Murphy et al. [10] and Smith [19] developed a research framework to record the correlation between health and household wealth based on individual's willingness to spend on healthcare. This framework could guide us in capturing the main features we care. However, it merely gives quantitative analysis, which can be compensated by the work of Poterba et al. [16] which uses sophisticated diff-in-diff and quantile methods. Such quantitative methodologies give us a clear insight into the logic behind the data. However, Poterba's work focuses on cross-sectional data while we will be using time series methods.

Ozkan [12] investigated differences in healthcare spending of high and low income individuals. However, individual's wealth is measured only in terms of income. Hajat et.al[6] looked at the relationship between wealth and health conditions such as obesity. As stated in the paper, the shortcomings of this research is incorporating self-reported data, which may make its result less reliable. Pertaining to the motivation of our project, Pollack et al. [14] concluded that studies on health should include considerations on the individuals' wealth.

Viscusi [20] explored approaches to assign appropriate economic values to risks in life when they are related to accidental fatalities. Cutler et al. [3] quantified the effects of health status and risks on labor supply, asset accumulation and welfare. These objectives are relevant to our analysis and they elaborate on the effect of health on various aspects of life. However, it doesn't offer any prescriptive solutions on how individuals with poor health or wealth can avoid adverse circumstances. Capatina [1] shows that socioeconomic status (SES) and health are strongly related. Childhood health could also affect the wealth and health in the adulthood [2]. This could help us come up with initial hypotheses for the data analysis. However, the exact mechanics underlying the link between SES and health are not completely clear.

Current research also helps us gain a more macro perspective on potential clients. [17] serves as a guide for the government to research policy implementations that can improve public health and the financial industry, which will greatly benefit all of us.

3 METHOD AND EVALUATION

Since the HRS data contains a larger number of older individuals, compared to PSID data, we combine these two together to form a large dataset(HRS.csv) with 12434 columns and 42053 rows. In this way the older and younger households are well balanced. In this CSV file, each individual is given a unique ID and each attribute is repeat in 13 waves range from 1992 to 2016. By merging the data of respondent level and family level, the wealth and stock investment of respondent are extracted from the family level. To track the health status, we use the self-reported health of respondent. Note that In the PSID and HRS, individuals are asked to rank their health as excellent, very good, good, fair or poor. We aggregate these answers into a binary measure of health: individuals who report their health to be in the first three categories are classified as healthy or in good health, while individuals who report being in fair or poor health are classified as unhealthy or in bad health.

First we are interested in the survival probability of healthy and unhealthy people, since the observations after death should be omitted in the following analysis. we do survival analysis to measure the lifetimes using the survival status of respondents over interview waves. The estimated survival probability with different health and sex is shown in Figure 1. Intuitively, individuals in bad health would have higher probability to die than healthy individuals.

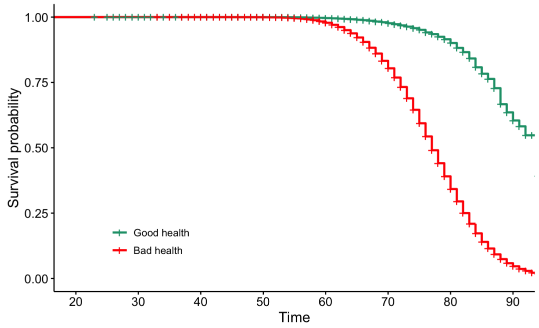


Figure 1: Survival probability.

Conditional on survival, it is interesting to test the causal effect of health on financial variables such as medical expenditure, labor income, stock investment, and wealth, we apply the unpaired two-samples t-test to compare the means of these financial variables of healthy and unhealthy respondents. The Null hypothesis H_0 is set as

$$H_0 : \quad m_{0,i} \neq m_{1,i} \tag{1}$$

where $m_{0,i}$ ($m_{1,i}$) is the first moment of tested financial variable of healthy (unhealthy) respondents among age group i, where there are 4 divided age groups: 20∼40 age group, 40∼50 age group, 50∼60 age group, and 65+ age group. Except that the medical expenditures of healthy and unhealthy respondents with age between 20 and 40 show no significant difference, all the other t-test results yield p-value smaller than 0.001, which implies rejecting the Null hypothesis H_0 but accepting that there exists distinguished finance difference between healthy and unhealthy respondents. Moreover, it turns out that bad health comes along with lower higher medical expenditure, lower labor income, lower stock investment, and lower accumulative wealth. The group difference of moment is shown in Figure 2.

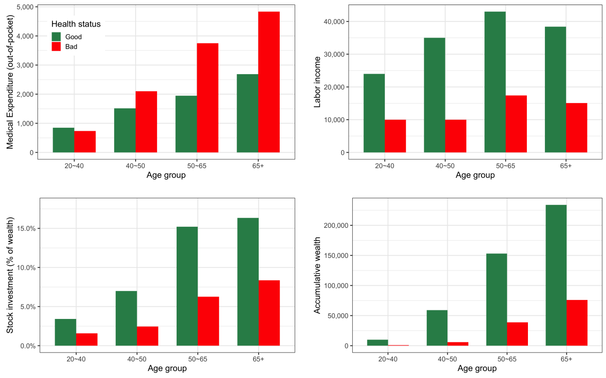


Figure 2: Group difference.

We have found evidence from the data that health status significantly affects individual's wealth allocations. The analysis above also indicates that health affects wealth allocation mainly through two different channels. In the first channel, bad health decreases the net in-pocket cash flows. Note that bad health requires higher out-of-pocket medical costs and in the mean time decreases the labor income (see e.g. decreases the efficient working hours), yielding lower net in-pocket cash flows. the shortage of net in-pocket cash flows impedes the accumulation of wealth (both financial wealth and non-financial wealth like house equity). The first channel seems trivial and is easy to understand. Inventively, we find a second channel regrading of growth rate, which is nontrivial and meaningful to current 401K retirement plan. In the second channel, due to the high demand of liquid money for medical coverage in the presence of low in-pocket cash flows, unhealthy investors would weigh money market account (such as risk-less bonds with low returns) more than risky stocks with high returns, inducing lower stock investments as shown in the third panel of Figure 2. By calculation, the wealth growth return is the weighted sum of stock return and bond (or cash) return where the weight equals to the corresponding proportional allocation. Thus, individuals staying in a long-time bad health have lower wealth growth return than healthy individuals. The lower wealth growth return slows the accumulation of wealth over life cycle. The lower stock investment in young and unhealthy group indicates higher risk aversions for poor investors in bad health, contrary to the belief of sophisticated wisdom that young and poor investors are more aggressive.

The different patterns of stock investment over life-cycle sheds lights on the optimal strategic asset allocation (i.e., target dated fund) in 401K retirement plan according to the employee's health status. A target date fund (TDF) — also known as a life-cycle, dynamic-risk or age-based fund — is a collective investment scheme that mimics the optimal strategic asset allocation of single investor over life cycle. We compare the stock allocations (as percent of wealth) over life cycle found in our dataset with the Vanguard Target Retirement Fund glide path[1] in Figure 3. We find that the Vanguard TDF shifts from riskier assets in the early years to less risky assets in the later ones, contrary to our empirical findings. Vanguard TDF argues that the risk tolerance as investors declines with age. When we are young, time is on our side and we can handle more risk. While nearing retirement, things reverse and we become risk-averse. This gives their decreasing patterns of stock allocation over life cycle. However, we argue that this is not true for young and poor investors. The effect of wealth shortage on stock investment in early age (most young individuals starts with zero wealth) is indispensable. As we have found in our dataset, indeed, young and poor investors are not risk-tolerant as expected. The stock allocation shows a hump shape over life cycle, rather a simple decreasing pattern as manually constructed in Vanguard TDF. The health status also has significant effect on life-cycle stock allocations. Our analysis provide more considerable risk factors for TDF.

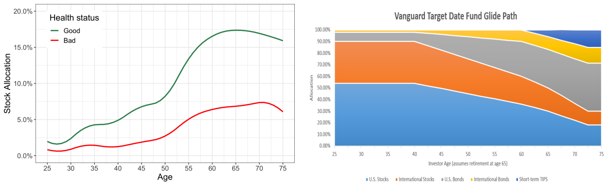


Figure 3: Group difference.

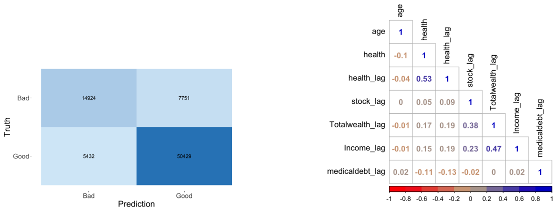
As there exists significant relation between health and wealth as tested above, inventively we can predict one's health status (good or bad) simply using his (her) financial reports in last time period, instead of referring to those complicated health-related reports such as medical records (See e.g. [18]). Using the four financial variables tested above and age as independent variables, we build a logistic regression model on health status mainly for prediction purpose:

$$\begin{aligned} &\log\left(\frac{P(health_t = Good)}{P(health_t = Bad)}\right) = \beta_0 + \beta_1 \cdot Age_t + health_{t-1} \\ &+\beta_2 \cdot MedicalCost_{t-1} + \beta_3 \cdot Income_{t-1} \\ &+\beta_4 \cdot StockInvestment_{t-1} + \beta_5 \cdot Wealth_{t-1}, \end{aligned} \tag{2}$$

where $\{\beta_i\}_{i=0}^5$ are parameters to be estimated. The result shows that all betas are significant nonzero (that is, all independent variables are statistically important). Our logistic regression model achieves accuracy of 0.8321 and gains AUC score of 0.7635. We show the confusion matrix and correlation matrix in Figure 4. Notice that respondents' health status will change over years so that in 13 waves totally we have great samples of data for both healthy and unhealthy people. The number of observations of bad and good health are 2:5 (i.e., $(14924 + 7751)/(5432 + 50429) \approx 0.4$ by the confusion matrix in Figure 4). Therefore, there is no issue of extreme imbalance in our dataset. By the correaltion matrix, we also find that the health status in last wave has

[1] From https://paulmerriman.com/the-ultimate-target-date-fund-portfolio/

high positive correlation with the health to be predicted in current wave. This is intuitive since we know that health status has long-term duration such that people in bad health in the past two (or more) years will be more likely to stay in bad health than those individuals with healthy historical records. We also find that health status (good health comes with higher score) are positively correlated with wealth and income significantly, consistent with our findings in Figure 4.



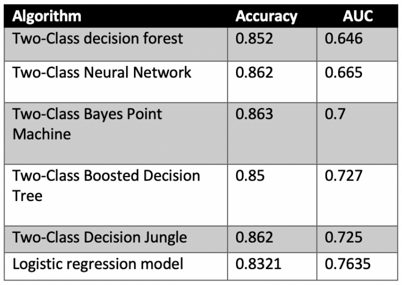
	Bad	Good
Bad	14924	7751
Good	5432	50429

	age	health	health_lag	stock_lag	Totalwealth_lag	Income_lag	medicaldebt_lag
age	1						
health	-0.1	1					
health_lag	-0.04	0.53	1				
stock_lag	0	0.05	0.09	1			
Totalwealth_lag	-0.01	0.17	0.19	0.38	1		
Income_lag	-0.01	0.15	0.19	0.23	0.47	1	
medicaldebt_lag	0.02	-0.11	-0.13	-0.02	0	0.02	1

Figure 4: Confusion and correlation matrix.

We can extend our regression model to general Machine Learning framework with more input variables. We selected 20 attributes from HRS file and calculated their feature importance. According to the feature importance result, we picked the top 5 features used to train our model. We trained five models by using five different algorithms. The accuracy of each model is shown in Figure 5.

In an alternate approach for the HRS dataset, we tested predicting the same health status binary variable by using current as well as lagging values of the variables as predictors in the model, including health status of the respondent in the previous survey, which is the original value ranging from 1 to 5. Other variables include age, gender, years of education, household income, non housing financial wealth and total wealth.



Algorithm	Accuracy	AUC
Two-Class decision forest	0.852	0.646
Two-Class Neural Network	0.862	0.665
Two-Class Bayes Point Machine	0.863	0.7
Two-Class Boosted Decision Tree	0.85	0.727
Two-Class Decision Jungle	0.862	0.725
Logistic regression model	0.8321	0.7635

Figure 5: Models' accuracy comparison

By training a decision tree classifier, we are able to achieve an AUC score of 0.79 on the test dataset. An XGBoost model also achieves a similar AUC score. It is interesting to note that the lagging health status (reported on the previous wave) is the most significant predictor in the model and excluding this variable causes the AUC score of the model to drop to 0.68. Other significant variables include household income, years of education and age, while in the second model non housing financial wealth and lagging income are selected instead of age. This tells us that at a high level, the health status of individuals does not change significantly over subsequent waves and they're less likely to transition from good health to bad or vice versa, however transient analysis would reveal more interesting insights and trends.

To track the composition of different kinds of groups, we have tested a clustering approach using KMeans clustering algorithm. In this approach, we initialized clusters using the first survey observations, and tracked cluster statistics over the course of subsequent waves. The wealthy group, represents a small chunk of the survey population, and shows higher income, number of years in education and better health than the rest of the clusters.

4 INNOVATION

- We test the casual effect of health status on individual's financial status via unpaired two-samples t-test.
- We find that the health status not only affects the net in-pocket cash flows, but also affects individual's stock allocation. The latter one is nontrivial is ignored by many conventional wisdom.
- We find a hump-shape stock allocation over life cycle, contrary to the manually constructed allocations in Vanguard target dated fund that is popular in the market currently. Our analysis provides more considerable risk factors for TDF construction in 401K retirement plan.
- We developed multiple models such as Logistic regression, XGBoost, and Neutral Network to predict health status.

5 VISUALIZATION

We developed four visualizations from our analysis to convey our findings in an intuitive form of representation.

- In the clustering visualization, we initialize 5 clusters for the first survey wave and draw bivariate scatterplots of relevant health and wealth attributes over the following 12 HRS surveys, where each cluster is represented by a bubble and the number of people at the time of the survey represented by its size. Another important aspect of this scatterplot visualization is the ability to do a temporal bivariate analysis between any two of the seven attributes. To keep a track of each cluster's trends, the cluster also leaves behind a trail, which is of a smaller size to maintain differentiability. The survey wave can be selected using the slider at the top and variables for x and y axes can be selected using the first and second dropdowns at the bottom respectively. It should be noted that a cluster at any point consists of only the surviving participants from the respective initialized cluster.

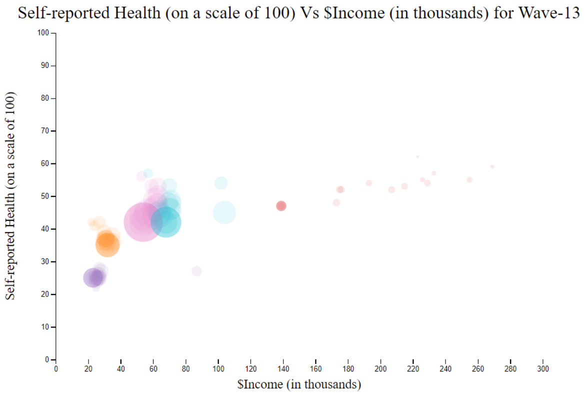


Figure 6: Health vs Income Plot for 5 Clusters with Trails for Previous Surveys

- We plot the average medical cost, labor income, stock investment, and accumulative wealth grouped by age and health status. The different financial variables can be selected using the drop down at the top. The corresponding static plots can be found in figure 2.
- The figure 7 shows how gender is related to health and how health condition changes over time amongst females and males – each grid is composed of 100 rectangles and each rectangle represents 1% of population in a specific gender group. The number of light blue rectangles indicates percentage of unhealthy people, while the number of dark blue rectangles indicates percentage of healthy population.

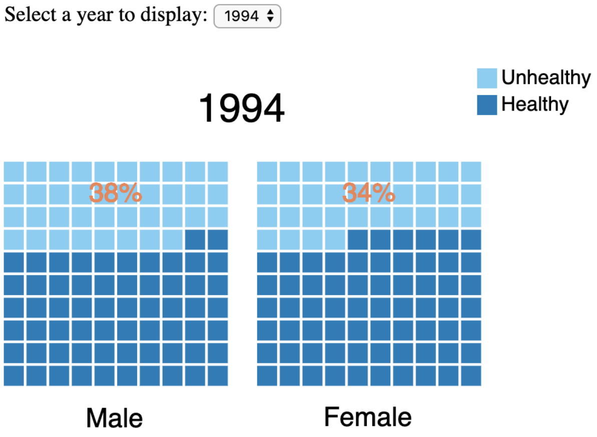


Figure 7: Gender and Health

- The figure 8 shows U.S. choropleth map of overall wealth-health condition. The data used in choropleth is from PSID. We extracted data from PSID on each individual's age, state, health, stock, medical debt, house equity and financial wealth. By combining each individual's stock, medical debt, house equity and financial wealth, we can get overall wealth. Then we used Pandas library in python to perform a series of data integration and cleaning tasks to get the final file for the plot. The color index used in choropleth is the ratio of average overall wealth and average health in each state, where darker color indicates better wealth-health condition of the state. Another implementation of this plot, by age (20s - 90s), is also included in our submission.

6 CONCLUSION AND DISCUSSION

In this project, we carried out several statistical tests and experiences on the following questions: 1. what is the survival probability of respondents in our dataset? 2. Does significant statistical causal effect of health on financial variables exist? 3. Do simple financial variables such medical expenditure, labor income, stock investment, and wealth have significant predictive power on health status? 4. Can the prediction model be extended to more complicated ML models with more variables available? 5. How does the composition of different

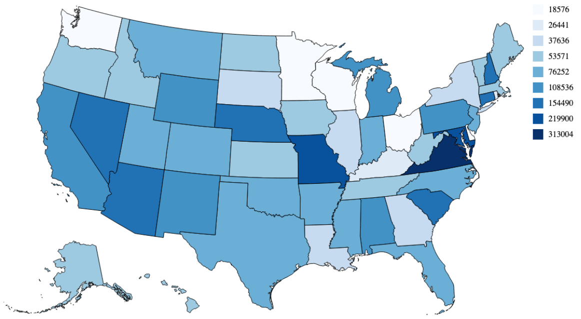


Figure 8: Wealth-Health Ratio By State

kinds of health groups change over time? 6. How is gender related to health and how does health condition change over time? 7. What is the overall wealth-health condition in 52 states? Our findings can be summarized below:

- From figure 8, we can conclude that Virginia has highest overall wealth-health ratio. In general, Northeast region has highest overall wealth-health ratio, then West region follow by South, Pacific Northwest and Midwest. This observation aligns close to the U.S. wealth data by state. Traditionally, Northeastern states have been known to have a good health care system with a higher percentage of wealthy population. However, we didn't factor in the cost of living for each state which can significantly affect the wealth-health index as well.
- From figure 6, it can be observed that clusters with average income close to or higher than the national average, report good health, while clusters with lower incomes report bad or poor health. Over time, health status for all clusters deteriorates, except the one that started with poor health, however clusters with good income still report better health than clusters with lower incomes and a trend in health can be observed against income.
- From figure 7, we found that except 2014, the ratio of unhealthy population in males is higher than that in females. Also, the ratio of unhealthy population in both male and female tends to increase over time.
- In Section 3, we have found evidence from the data that health status significantly affects individual's wealth allocations. The simple logistic regression in (2) shows high predictive power of financial variables on health status. Healthy individuals exhibit higher income and wealth, whereas unhealthy individuals require higher out-of-pocket medical costs. In Figure 3, we also find that health status has significant effect on life-cycle stock allocations. Our analysis provides more considerable risk factors (see e.g. health status) for target date fund in one's 401K retirement plan.

7 CONTRIBUTION

- Shan Huang: Concatenate cross-wave observations for PSID data and merge PSID with HRS data; Survival analysis grouped by health status; Unpaired two-samples t-test to test the casual effect of health status on individual's financial status; Analysis on how health status affects financial status via two channels; Empirically find hump-shaped stock allocations over life-cycle compared to Vanguard target dated fund; Health status prediction by logistic regression based on age, lagged health and lagged financial variables. Integrate data needed for visualizations.
- Shuangke Li: Finishing the data transformation and computing the feature importance of 20 selected features. Calculating the correlation values between health condition and selected features. Based on feature importance, selecting top five important features use to build models. Building the GridView to illustrate the relationship between health and gender and display how does health condition change over time.
- Sahil Dhingra: Cleaning and preparing the HRS dataset for model development. Testing model performance using Decision trees and XGBoost algorithms. EDA to study static distributions of health status vs total wealth. Clustering analysis using Kmeans to track cluster statistics over time. Cluster visualization scatterplot with slider and dropdowns for survey wave and attribute selection.
- Like Deng: Update and finalize report. Revise schedule and assignment. Cleaning and integrate dataset for choropleth visualization. Implement choropleth visualization with slide bar to select age group.

8 BIBLIOGRAPHY

[1] E. Capatina. Life-cycle effects of health risk. *Journal of Monetary Economics*, 2015.

[2] Paxson C. Case, A. Causes and consequences of early-life health. *Demography, 47(1), S65-S85.*, 2010.

[3] D. M. Cutler, A. Lleras-Muney, and T. Vogl. Socioeconomic status and health: Dimensions and mechanisms. *National Bureau of Economic Research*, 2008.

[4] C. Dobkin, A. Finkelstein, R. Kluender, and M. J. Notowidigdo. The economic consequences of hospital admissions. *American Economic Review, 108(2), 308-52*, 2018.

[5] Timothy Gubler and Lamar. Pierce. Healthy, wealthy, and wise: Retirement planning predicts employee health improvements. *Psychological Science*, 2014.

[6] A. Hajat, J.S. Kaufman, K.M. Rose, A. Siddiqi, and J.C. Thomas. Do the wealthy have a health advantage? cardiovascular disease risk factors and wealth. *Social Science Medicine*, 2010.

[7] T. Halliday. Heterogeneity, state dependence, state dependence and health. *Econometrics Journal*, 2008.

[8] T. Halliday. The evolution of latent health over the life course. *Maternal and Child Health Journal*, 2012.

[9] P. Krusell and A. Smith. Income and wealth heterogeneity in the macroeconomy. *Journal of Political Economy*, 1998.

[10] Kevin M. Murphy and Robert H. Topel. The value of health and longevity. *University of Chicago and National Bureau of Economic Research*, 2005.

[11] M. D. Nardi, S. Pashchenko, and P. Porapakkarm. The lifetime costs of bad health. *NATIONAL BUREAU OF ECONOMIC RESEARCH*, 2017.

[12] Serdar Ozkan. A macroeconomic analysis of health careover the life cycle. *University of Toronto*, 2017.

[13] S. Pashchenko and P. Porapakkarm. Work incentives of medicaid beneficiaries and the role of asset testing? *International Economic Review*, 2017.

[14] Craig Evan Pollack, Sekai Chideya, Catherine Cubbin, Brie Williams, Mercedes Dekker, and Paula Braveman. Should health studies measure wealth? a systematic review. *American Journal of Preventive Medicine*, 2007.

[15] J. Poterba and S. Venti. The asset cost on poor health. *The Journal of the Economics of Ageing*, 2017.

[16] Venti F. Steven Poterba, M. .J. and W.F. David. The asset cost of poor health. *National Bureau of Economic Research*, 2010.

[17] Jason Q. Purnell and Anjum Hajat. Asset funders network - the health and wealth connection - opportunities for investment across the life course. 2017.

[18] B. Seligman, S. Tuljapurkar, and D. Rehkopf. Machine learning approaches to the social determinants of health in the health and retirement study. *SSM - Population Health*, 2018.

[19] J.P. Smith. Health bodies and thick wallets. *Journal of Economic Perspectives 13.2 (1999): 145-166*, 1999.

[20] W. Viscusi. The value of risks to life and health. *American Economic Association*, 1993.

[21] S. Wu. The effects of health events on the economic status of married couples. *Journal of Human Resources 38.1 (2003): 219-230*, 2003.