CESifo Working Papers
www.cesifo.org/wp

Do they Know what's at Risk? Health Risk Perception among the Obese

Joachim Winter
Amelie Wuppermann

CESIFO WORKING PAPER NO. 3864
CATEGORY 1: PUBLIC FINANCE
JUNE 2012

CESifo *Working Paper No. 3864*

Do they Know what's at Risk? Health Risk Perception among the Obese

Abstract

The perception of disease risks and risky health behaviors are closely associated. In this paper, we investigate the accuracy of disease risk perceptions among obese individuals. We compare subjective risk perceptions for various diseases elicited in the American Life Panel (ALP) to individual's objective risks of the same diseases. We find that obese individuals significantly underestimate their 5-year risks of diabetes, arthritis or rheumatism, and hypertension, while they systematically overestimate their 5-year risks of a heart attack and a stroke. Obese individuals are thus aware of some but not all obesity-related risks. For given diseases, we document substantial heterogeneities in the accuracy of expectations across individuals.

JEL-Code: I100, I180, D840.

Keywords: obesity, health risk, subjective expectations.

Joachim Winter
Department of Economics
University of Munich
Munich / Germany
winter@lmu.de

Amelie Wuppermann
Department of Law and Economics
University of Mainz
Mainz / Germany
wuppermann@uni-mainz.de

June 2012

We would like to thank Kjell Arne Brekke, John Cawley, Joshua Gottlieb, Michael Hurd, Gábor Kézdi, Darius Lakdawalla, Jürgen Maurer, Pierre-Carl Michaud, Andrew Parker, and James Smith for helpful comments and suggestions. This research has further profited from comments by seminar and conference participants at the Central European University Budapest, the University of Lausanne, the University of Mainz, the Research Institute for the Economics of Aging (MEA) Mannheim, the Helmholtz Center Munich, the iHEAWorld Congress in Toronto, the Annual Congress of the European Economic Association in Oslo, the Conference on "Internet Interviewing and the HRS" in Aspen, and the annual meeting of the Council for Health Economics of the Verein für Socialpolitik in Berlin. Financial support from the National Institute on Aging (NIA), grant no. R01 AG 020717 ("Internet Interviewing and the HRS") is gratefully acknowledged.

1 Introduction

Obesity – that is being too heavy for your height, typically defined as having a Body Mass Index (BMI) of 30 or higher[1] – is a risk factor for various diseases. In particular, it increases the risks of type 2 diabetes, cardiovascular disease, several cancers, arthritis, and psychological problems (Haslam and James, 2005; Dixon, 2010). Furthermore, obesity is responsible for a large share of medical expenditures (Cawley and Meyerhoefer, 2012). Reducing excess weight would not only benefit the affected individuals' health but could also help to improve the financial situation of social security systems. While a reduction in obesity prevalence might increase longevity and thus increase the public annuity burden, it would reduce health care costs. As the reduction in health care costs due to lower obesity would occur earlier than the costs of greater longevity, a reduction in obesity prevalence would likely improve the public financial situation (Goldman et al., 2010).

Whether individuals are obese depends to a large extent on lifestyle choices: One can choose how many calories to consume and whether to engage in high or low calorie expenditure. Obesity-related health risks and thus also perceptions of these risks may play a role in the decision processes as the risks affect the costs of the different options. Indeed, Kan and Tsai (2004) show that obese men in Taiwan react to information on obesity-related risks by decreasing they weight. The authors thus call for public programs that increase the awareness of the harmful effects of obesity. Similarly, Cawley and Ruhm (2012) state that government intervention could be warranted if individuals underestimate the risks associated with unhealthy behavior.

In this paper, we analyze the accuracy of health risk perceptions among the obese in the US. We focus on the question of whether middle-aged individuals with excess body weight accurately assess their individual risks of diabetes, a stroke, a heart attack, chronic lung disease, hypertension, and arthritis or rheumatism. We also investigate how health risk perceptions among the obese compare to those of individuals who have a healthy weight.

Previous studies on health risk perception among the obese indicate that obese individuals are not aware of obesity-related health risks. Gregory et al. (2008) show that a large share of obese adults in the US do not rate their weight as a health risk factor. Similarly, middle-aged obese individuals in the US on average overestimate their chances of living up to the age of 75 (Falba and Busch, 2005). Our contribution to this literature is twofold. First, we focus on the risks of specific diseases rather than general health risks or

[1]The BMI is defined as weight in kg divided by the square of height in meters $\frac{weight(kg)}{height^2(m^2)}$.

life expectancy. Second, we study the *accuracy* of individual health risk perception instead of just over- or underestimation of risks by comparing individuals' subjective risk expectations to their personal objective risks. Overall this allows for a more detailed description of health risk perception among the obese than available up to now.

Our data on subjective disease risk expectations come from the American Life Panel (ALP), a panel study administered by RAND. In a 2010 ALP survey, we asked respondents to assess their chances of developing different diseases within the next 5 years on a $0-100$ scale. The probabilistic design we use in our health expectation questions has been employed to elicit subjective expectations in a variety of domains; see Manski (2004) and Hurd (2009) for reviews of the approach and assessments of the validity of probabilistic expectation questions.

In order to gauge how well individuals in the ALP are informed about their health risks, we compare individuals' subjective risk assessments with their objective risks for the same diseases. We adopt an approach proposed by Khwaja et al. (2009) in their study of health risk perceptions among smokers: We predict objective risks for individuals in our dataset using disease risk models whose parameters are estimated on a different dataset. More specifically, we estimate prediction models for disease onset using data from the Health and Retirement Study (HRS) that comprise individual characteristics, such as age, sex, and BMI, in a baseline year (2002) and subsequently realized disease onsets.[2] Assuming that the relationships between individual characteristics and disease risks are the same in the ALP today as they were in the HRS a few years ago, we use the estimated prediction model to calculate objective risks for individuals in our ALP sample.

In our individual-level comparison of subjective and objective risks, we find interesting differences in accuracy between diseases. In line with findings reported in the earlier literature, our study provides evidence for underestimation of some risks among the obese. In particular, they underestimate their risks of hypertension, arthritis and rheumatism, and diabetes. However, the risks of a heart attack, a stroke or chronic lung disease are significantly overestimated by the obese. Moreover, we document important heterogeneities in accuracy across individuals for given diseases.

The paper is structured as follows. Section 2 describes the datasets and methods that we use in our analysis. Section 3 presents the results. In Section 4, the robustness of the results is explored, and Section 5 discusses the results and concludes.

[2]In addition, we use data from the National Health and Nutrition Examination Survey (NHANES) to calculate objective health risks for younger populations that are not covered by the HRS; see Section 2 for details.

2 Data and methods

The empirical results we report in this paper are based on three datasets. The main analysis is based on data obtained from a survey on subjective health expectations we conducted in the American Life Panel (ALP). Data from the Health and Retirement Study (HRS), which covers individuals aged 50 and older, are used to calculate objective health risks for the ALP respondents in our sample in that age range. As a third dataset, we use the National Health and Nutrition Examination Survey (NHANES) to calculate objective risks for the whole adult age range (albeit in a more limited fashion than is possible for the HRS population).

The ALP is an internet survey of about 3,200 American adults administered by RAND (see *http://rand.org/labor/roybalfd/american_life.html* for a full description).[3] We developed a survey for the ALP that was in the field from July 2010 to May 2011. The total number of respondents who completed the survey was 2,913; these are 90.07% of those ALP members who were invited to participate in this specific study, so unit nonresponse was rather low.

The purpose of our ALP survey is primarily to elicit individuals' subjective expectations of developing certain diseases in the future. We elicit subjective expectations as numerical probabilities, using the question "What do you think is the percent chance that you will develop, or re-develop if you have already been diagnosed with it, the following conditions in the next 5 years and ever in you lifetime?", which is followed by a response grid containing a list of diseases.[4] The question is preceded by a text that introduces the probability response scale and explains that it ranges from 0 to 100, where 0 means that there is absolutely no chance, or 0 percent, and 100 means the event is absolutely sure to happen, or 100 percent. Similar types of questions on survival expectations have proven useful in eliciting subjective probabilities that have predictive power for future outcomes (see Hurd, 2009, for a summary). In addition to the subjective expectations, the survey elicits

[3]ALP members are recruited using offline methods and samples, so participation is not conditional on being an internet user at the time of sampling; those individuals who are willing to participate but are not internet users are provided with internet access by RAND. This procedure alleviates selectivity concerns that are often raised against internet surveys. To correct for remaining selectivity, RAND provides weights that adjust key demographic margins to those of the Current Population Survey (CPS). These weights are used in the descriptive statistical analysis presented in this paper. Couper et al. (2007) provide a discussion of sample selection in internet panels such as the ALP.

[4]Individuals who report that they have already been diagnosed with chronic diseases, such as diabetes or chronic lung disease, are not asked about their chances of developing the specific conditions. For conditions that may relapse, like a stroke or a heart attack, all individuals are asked about the chances independent of their prior history of the disease.

whether the respondent has already been diagnosed with the diseases that we study, as well as risk factors for these diseases, including questions on co-morbidities, family history, and lifestyle. Additional covariates are available from other surveys conducted in the ALP with the same respondents.

The second dataset we use, the HRS, was started in 1992 as a representative panel study of the US non-institutionalized population born between 1931 and 1941 (see Juster and Suzman, 1995). Blacks, Hispanics and residents of Florida were over-sampled. Sampling weights are provided to correct for this over-sampling. Data have been collected every two years since 1992. Additional cohorts have been added to the HRS after the first wave. The last wave whose data were available at the time of writing this paper was conducted in 2008. Our strategy, described in detail below, is to use information on individual characteristics in a baseline wave and on self-reported onset of diseases in future HRS waves to estimate prediction models for the onset of diseases. The estimated parameters from these models are then used to predict objective individual risks for individuals in the ALP. As this procedure requires that the relationships between individual characteristics and disease hazards do not change over time, we use the most recent 5-year period for which HRS data are available. In our analysis the baseline year is thus 2002, so that the 5-year spell is fully observed in the 2004, 2006, and 2008 waves.

In estimating the relationship between individual characteristics and the probability of disease onset in the HRS, we adapt an approach proposed by Khwaja et al. (2009).[5] We estimate duration models separately for the different diseases. This assumes that the different risks are independent of each other and more importantly, that the risks are independent of other competing health risks, such as death. This assumption is particularly crucial for older populations for whom life expectancy might be lower than the 5-year time horizon that we analyze. Therefore, we use only data on relatively young individuals to estimate the prediction models. More specifically, in our main analysis we use HRS data on individuals aged 50 to 62. As a consequence, we can only predict objective risks for individuals in the same age range in the ALP.[6] The results are presented in Section 3.

Since the HRS does not cover individuals younger than 50, we also use data from NHANES to estimate objective risks for the entire adult age range. NHANES is a program of repeated cross-sectional surveys that are designed to assess the health and nutritional status of the US population at all ages. While individuals 60 and older, African Americans

[5]Appendix A gives more details on the estimation strategy.

[6]For comparison, we also estimate all models for the sample of individuals aged at least 50. The results are shown in Appendix B.

and Hispanics are over-sampled, sampling weights are provided to ensure representativeness for the US population. Since 1999, NHANES data is released in biannual waves. As NHANES is not a panel dataset, we cannot use it to estimate risk prediction models to infer individual-specific risk in the ALP. However, the data allows to estimate age-specific risks for the overall population. In a first step, we estimate age-specific disease incidence by pooling information from the 2005–06, 2007–08 and 2009–10 NHANES surveys. In a second step, we use these age-specific incidence estimates in a life table to predict 5-year risks.[7] This procedure provides estimates of the average 5-year risks of different diseases at different ages. We compare these age-specific objective risks to average subjective and objective risks of individuals in the ALP.[8]

Table 1 displays the fraction of individuals who report a prior diagnosis of the conditions that we consider in this paper. This table is based on all ALP respondents aged 20 to 80 (2,871 of the 2,913 respondents). Estimates for the US population between 20 and 80 based on the NHANES data are reported for comparison; these are based on 16,657 individuals. The fractions of individuals with prior diagnoses of the different conditions are almost identical in the two datasets. The weights used for the ALP (which are based only on socio-demographic variables) seem to be rather successful at insuring representativeness of the data.[9]

Table 1: Prior diagnosis of diseases in ALP and NHANES

	ALP		NHANES	
	Mean	Std. Dev.	Mean	Std. Dev.
Diabetes	0.10	0.30	0.08	0.27
Hypertension	0.28	0.45	0.29	0.46
Arthritis/RA	0.22	0.42	0.24	0.43
Stroke	0.02	0.15	0.03	0.16
Heart Attack	0.03	0.16	0.03	0.18
Chronic lung disease	0.07	0.25	0.07	0.25

Notes: Individuals aged 20-80 in both samples. Sampling weights and NHANES sampling design taken into account in the estimation of means and standard deviations.

[7] See appendix C for details.

[8] Making use of additional NHANES waves from 1999 onwards allows to also calculate age-specific average objective risks for different BMI categories, such as individuals with normal weight, and overweight or obese individuals. Results indicate overestimation of all disease risks but arthritis; they are available upon request.

[9] As the information on disease prevalence in both datasets is self-reported, both ALP and NHANES estimates might suffer from reporting biases or undiagnosed conditions.

Table 2: Descriptive statistics

	HRS 2002		ALP 2010	
	Mean	Std. Dev.	Mean	Std. Dev.
Demographics				
Age	57.99	2.46	55.52	3.70
Male	0.47	0.50	0.49	0.50
Married	0.71	0.45	0.68	0.47
White	0.85	0.35	0.83	0.37
Education				
Less than High School	0.15	0.36	0.06	0.23
High School or equiv.	0.35	0.48	0.32	0.46
Some College	0.24	0.43	0.30	0.46
BA or equiv.	0.13	0.34	0.21	0.41
More than BA	0.12	0.33	0.12	0.33
Self-rated health				
Excellent	0.17	0.37	0.08	0.27
Very good	0.33	0.47	0.41	0.49
Good	0.29	0.45	0.33	0.47
Fair	0.15	0.36	0.15	0.36
Poor	0.06	0.25	0.04	0.19
BMI				
Normal weight (BMI$<$ 25)	0.27	0.44	0.25	0.43
Overweight ($25 \leq$ BMI < 30)	0.39	0.49	0.36	0.48
Obese 1 ($30 \leq$ BMI < 35)	0.19	0.40	0.22	0.41
Obese 2 ($35 \leq$ BMI < 40)	0.08	0.27	0.07	0.26
Obese 3 ($40 \leq$ BMI)	0.06	0.24	0.08	0.27
Smoking Status				
Current	0.20	0.40	0.21	0.40
Former	0.41	0.49	0.32	0.47
Never	0.38	0.49	0.48	0.50
Disease History				
Diabetes	0.14	0.34	0.15	0.35
Hypertension	0.42	0.49	0.42	0.49
Arthritis/RA	0.46	0.50	0.32	0.47
Stroke	0.04	0.20	0.03	0.17
Heart Attack	0.02	0.12	0.03	0.18
Chronic lung disease	0.07	0.25	0.08	0.27
N	4776		953	

Notes: Sampling weights used in estimation of means and standard deviations. Results based on individuals aged 50-62 in both samples.

In our main analysis, we restrict the sample to individuals aged 50 to 62. As discussed above, the lower limit derives from the availability of HRS data that we need to predict objective risks; the upper limit reflects our concern that treating disease risks as independent is likely to involve larger distortions for older individuals. There are 953 individuals in this age group in our ALP survey. In table 2, weighted means and standard deviations of relevant health variables, risk factors, and demographics are displayed for these 953 individuals. For means of comparison, descriptive statistics are also displayed for individuals aged 50 to 62 in the HRS 2002 sample used to estimate objective risks.

While the HRS and ALP samples seem to be similar with respect to demographics, there are differences in educational attainment, self-assessed health, BMI categories, smoking status, and health history. The ALP sample is better educated, has lower self-assessed health, a higher fraction of obese individuals, and a higher fraction of individuals who never smoked. Furthermore, a lower fraction of individuals in the ALP report having been diagnosed with arthritis or rheumatism. The differences in education, BMI and smoking status could possibly reflect trends in educational attainment and changes in prevalence of smoking and obesity over time. While the prevalence of smoking in the US decreased, the prevalence of obesity and the average educational attainment increased.[10]

For the results of our analysis, the differences in individual characteristics between the samples do not matter as long as the relationships between the individual characteristics and the probabilities of developing different diseases in the future are constant over time. As mentioned above, we use the latest possible HRS wave (2002) in order to increase the plausibility of this assumption. While disease histories are not included as such in the prediction models, they might be important for our results as they determine which individuals are included in the estimation of objective risks: For each condition, only individuals who report no prior diagnosis of the condition are used to estimate the prediction model. We thus only use data from relatively healthy individuals. We also only use the results of the prediction models to predict objective risks for individuals without a prior diagnosis. The same selection rule thus applies in both models.

The subjective expectation questions in the ALP play a central role for our study. From other studies on subjective expectations, we know that the answers to such questions are often rounded to focal values, such as 5%, 10%, or 25%. Furthermore, a large fraction of responses typically occurs at 0%, 50% and 100% (see Manski and Molinari, 2010; Bruine de

[10]See National Center for Health Statistics (2010), page 24ff. for trends in smoking and obesity in the US and US Census data available at *www.census.gov/hhes/socdemo/education/data/cps/historical/index.html.* for the trend in educational attainment.

Table 3: Subjective 5-year risks in the ALP

		Fraction of Responses								
	N	NR	0	1 – 4	50	96 – 99	100	M10	M5	Other
Diabetes	836	0.03	0.34	0.10	0.06	0.00	0.00	0.22	0.23	0.01
Hypertension	580	0.02	0.32	0.09	0.07	0.00	0.00	0.27	0.22	0.01
Arthritis/RA	627	0.03	0.27	0.07	0.04	0.00	0.06	0.29	0.18	0.05
Stroke	930	0.03	0.29	0.10	0.08	0.00	0.00	0.23	0.26	0.01
Heart Attack	929	0.03	0.28	0.09	0.08	0.00	0.00	0.26	0.26	0.01
Chronic Lung Disease	895	0.04	0.42	0.10	0.05	0.00	0.00	0.18	0.20	0.01

Notes: Sampling weights used to estimate means. Results based on individuals aged 50-62 without prior diagnosis of specific condition. NR = nonresponse, $M10$ = answer is multiple of 10 other than 50, $M5$ = answer is multiple of 5 that is not also a multiple of 10. *Other* indicates any other answers, i.e. not integers between 0 and 4, 96 and 99, 50, 100, or multiples of 5 or 10.

Bruin and Carman, 2012; Bruine de Bruin et al., 2011). Figure 1 and table 3 characterize the distributions of the responses to the expectation questions in the ALP.[11] Figure 1 displays histograms of the responses for individuals in the ALP who are between 50 and 62 years old and report no prior diagnosis of the specific disease. Table 3 delivers more details on the different fractions of responses in the same sample. Next to nonresponse, NR, table 3 contains information on the fraction of responses that are 0, 50, 100, multiples of 10 other than 50, multiples of 5 that are not also multiples of 10, integers between 1 and 4 or 96 and 99, and other values.

Nonresponse does not seem to be a big issue in our data, as only between 2 and 4 % of responses are missing. The histograms in figure 1 show that the distributions of responses to the expectation questions are positively skewed with large spikes at 0. In addition, there seem to be spikes at 50 and at multiples of 10. Table 3 shows that indeed between 27% of individuals for arthritis, rheumatism or rheumatoid arthritis and 42% for chronic lung disease report a zero chance of developing the different conditions in the next 5 years.

[11] Subjective expectations on developing arthritis or rheumatism in the future are elicited by two separate questions in the ALP. The first question asks individuals about their chances of developing arthritis or rheumatism except rheumatoid arthritis, $Pr(arthr)$, and the second question asks about the chances of developing rheumatoid arthritis, $Pr(RA)$. In the HRS, however, there is only information on arthritis and rheumatism including rheumatoid arthritis. We therefore combine the answers to the two subjective expectations question in the ALP. As we are interested in whether overweight and obese individuals underestimate their risks of developing different diseases, we aggregate the two probabilities in a way that results in the largest possible subjective probability of developing arthritis. Specifically, we set $Pr(arthritis) = min\{100; Pr(arthr) + Pr(RA)\}$.

Figure 1: Distribution of subjective 5-year risks

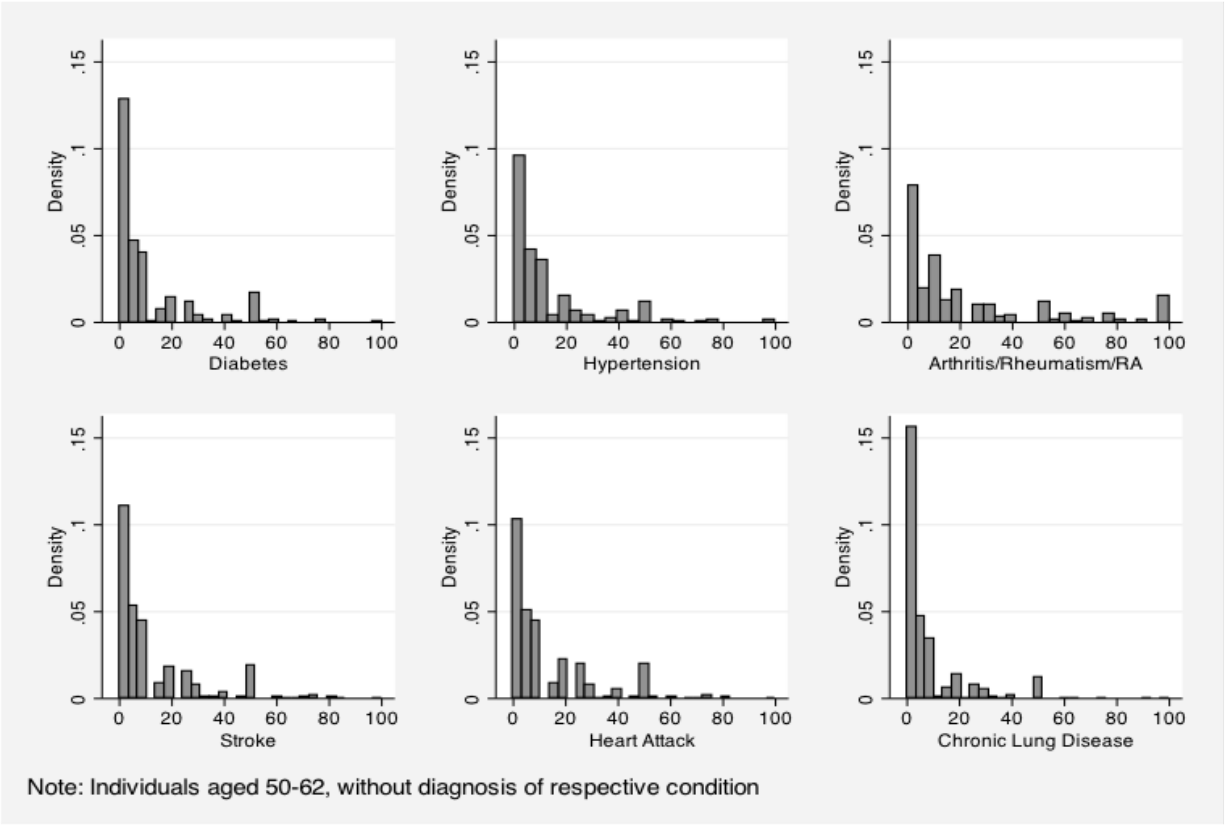
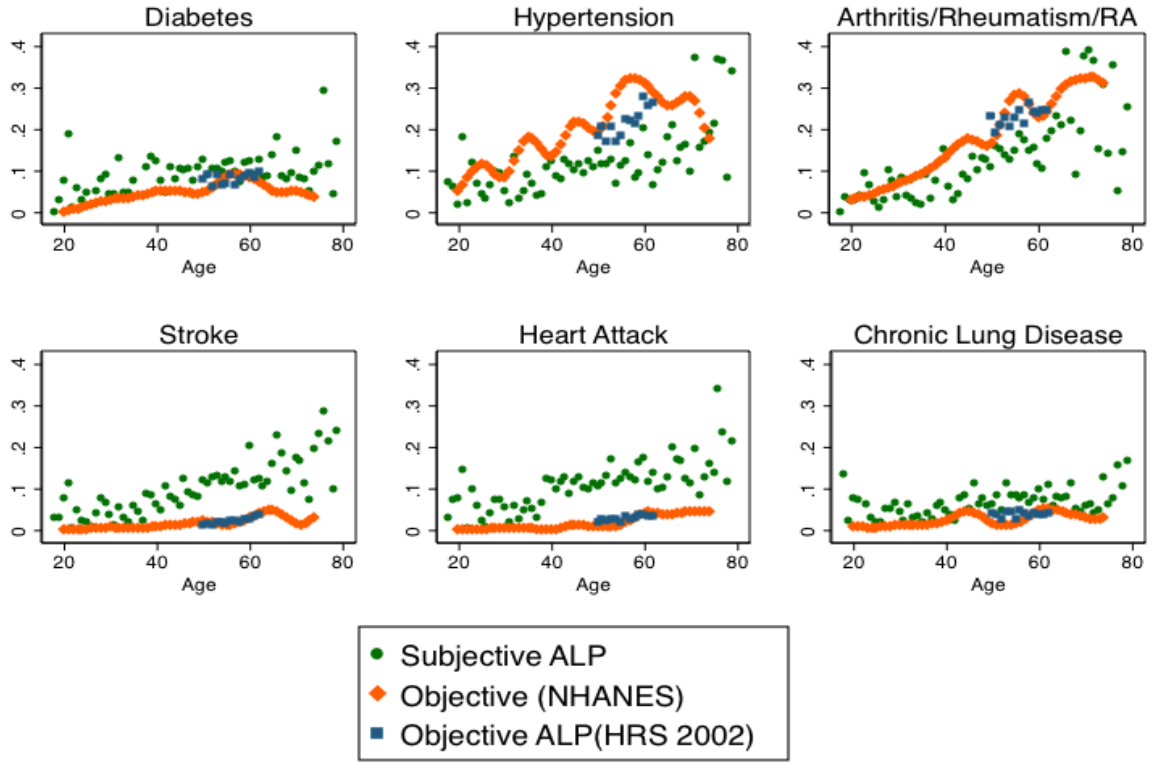


Figure 2: Objective and subjective 5-year risk by Age



Furthermore, large fractions of individuals report values that are multiples of 5 or 10. This suggests that individuals might be rounding their responses to these values. In the main analysis whose results are presented in the next section, we nevertheless follow the tradition in the literature on subjective expectations and take the responses at face value. In a robustness check reported in Section 4, we follow an idea of Manski and Molinari (2010) and construct rounding intervals for each individual that depend on her answers to all the 5-year expectation questions, which makes the analysis less sensitive to implicit assumptions about rounding.

3 Results

Figure 2 contains age-specific subjective 5-year risks for individuals in the ALP, age-specific objective 5-year risks based on NHANES data, and age-specific objective 5-year risks for individuals in the ALP who are between 50 and 62 years old. All the graphs included

in figure 2 display age on the horizontal axis and the different risks on the vertical axis. Concerning the average subjective risks, there is an increase with age for all diseases, in particular for hypertension, arthritis, stroke, and heart attack. Given that age is a known risk factor for these diseases, the increase might reflect some general knowledge of the nature of these risks.[12] Similarly, the increases with age in the objective risks based on the NHANES data support age as a risk factor.

Comparing the subjective risks in the ALP and objective age-specific risks based on NHANES we see that individuals seem to overestimate the risks of most diseases except for hypertension and arthritis. Furthermore, the overestimation of risks increases with age, in particular for the risks of stroke and heart attack. Individuals thus seem to overestimate the steepness of the age gradient. Additionally, figure 2 indicates that our different estimates for the objective risks are fairly similar. While the objective risks based on NHANES are population estimates, those based on HRS are weighted risks for individuals in the ALP. It is reassuring that these measures are similar despite the different estimation procedures.

The results displayed in table 4 are based on the individual-specific subjective and objective risk measures for individuals in the ALP and shed light on the relationship between the accuracy of subjective risk assessments and obesity. In each panel of the table, results for the risk of one specific disease are shown. The first two sets of results in each panel investigate how the subjective risks of the different diseases vary between BMI categories. Column (1) shows the estimated coefficients of regressions of subjective risk on a constant and dummies for being overweight, that is having a BMI between 25 and 30, and three obesity categories, 30 $\leq$BMI $<$ 35, 35 $\leq$ BMI $<$ 40, and 40 $\leq$ BMI. In column (2) additional control variables are included in these estimations. In particular, all variables that are used in the HRS risk prediction models (age, sex, race, marital status, educational achievement, smoking status, and self-rated health) are included as additional controls. How the accuracy of the subjective risk measures varies between different BMI categories is investigated in column (3). It displays the results of OLS regressions with the difference in individual subjective and objective risk as dependent variables and a constant and dummies for the BMI categories as explanatory variables.[13]

The results displayed in column (1) of table 4 indicate that for all diseases but arthritis and chronic lung disease, individuals with excess weight rate their risks significantly higher

[12]For general risk factors of the different diseases see for example the webpage of the Centers of Disease Control and Prevention (CDC) at http://www.cdc.gov/az/a.html.

[13]See Appendix D for more details on the different equations that are estimated.

Table 4: Subjective 5-year risks and obesity

	Subjective risk (S) (1)		S and additional controls (2)		Difference ($S - \widetilde{O}$) (3)	
Diabetes (N=817)						
25 ≤BMI<30	3.158***	(1.105)	3.076***	(1.127)	-0.013	(0.011)
30 ≤BMI<35	7.013***	(1.639)	5.684***	(1.633)	-0.028*	(0.016)
35 ≤BMI<40	4.830**	(2.082)	3.069	(2.025)	-0.136***	(0.020)
40 ≤ BMI	14.180***	(3.985)	11.568***	(3.819)	-0.008	(0.037)
Constant	6.729***	(0.753)	0.192	(8.563)	0.049***	(0.008)
R^2	0.047		0.096		0.048	
Hypertension (N= 567)						
25 ≤BMI<30	3.443**	(1.421)	3.275**	(1.458)	-0.057***	(0.014)
30 ≤BMI<35	5.114**	(2.271)	4.550*	(2.417)	-0.083***	(0.024)
35 ≤BMI<40	12.887***	(4.284)	11.561***	(4.281)	-0.061	(0.044)
40 ≤BMI	10.413**	(4.945)	9.285*	(5.087)	-0.076	(0.046)
Constant	8.180***	(0.872)	2.237	(11.779)	-0.045***	(0.009)
R^2	0.041		0.077		0.036	
Arthritis/RA (N = 615)						
25 ≤BMI<30	-1.037	(2.768)	-0.687	(2.789)	0.019	(0.027)
30 ≤BMI<35	-1.545	(3.429)	-2.743	(3.447)	-0.028	(0.034)
35 ≤BMI<40	2.506	(5.710)	-0.624	(5.639)	-0.064	(0.057)
40 ≤BMI	4.832	(6.317)	0.173	(6.357)	-0.069	(0.061)
Constant	22.265***	(2.214)	-3.675	(18.901)	0.002	(0.022)
R^2	0.003		0.069		0.010	
Stroke (N = 911)						
25 ≤BMI<30	1.917	(1.202)	1.559	(1.188)	0.018	(0.012)
30 ≤BMI<35	4.125***	(1.585)	2.369	(1.580)	0.033**	(0.016)
35 ≤BMI<40	7.594***	(2.693)	5.236**	(2.631)	0.076***	(0.027)
40 ≤BMI	4.455*	(2.466)	0.594	(2.787)	0.020	(0.025)
Constant	9.580***	(0.871)	-7.728	(8.926)	0.082***	(0.009)
R^2	0.018		0.073		0.015	
Heart Attack (N = 911)						
25 ≤BMI<30	3.114**	(1.251)	2.843**	(1.228)	0.028**	(0.012)
30 ≤BMI<35	4.042**	(1.573)	2.252	(1.597)	0.022	(0.015)
35 ≤BMI<40	8.195***	(2.535)	5.661**	(2.486)	0.065***	(0.025)
40 ≤BMI	5.908**	(2.422)	2.224	(2.651)	0.042*	(0.025)
Constant	9.833***	(0.924)	-14.897*	(8.352)	0.082***	(0.009)
R^2	0.021		0.102		0.013	
Lung Disease (N = 874)						
25 ≤BMI<30	0.178	(1.110)	-0.035	(1.052)	0.002	(0.010)
30 ≤BMI<35	2.215	(1.368)	1.358	(1.396)	0.014	(0.013)
35 ≤BMI<40	0.410	(1.618)	-0.798	(1.576)	-0.015	(0.015)
40 ≤BMI	-0.061	(1.602)	-1.648	(1.829)	-0.019	(0.017)
Constant	7.277***	(0.830)	-12.211	(7.486)	0.045***	(0.008)
R^2	0.004		0.106		0.005	

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: Robust standard errors in parentheses. N indicates the number of observations used in all estimations displayed in the respective panel. It reflects the number of individuals aged 50-62 in the ALP without prior diagnosis of the respective condition. Columns (1) and (2) use subjective 5-year risk as dependent variable. In (1) BMI categories as controls, in (2) BMI categories plus age, sex, race, marital status, education and self assessed health as controls. In (3) difference between subjective and objective risk as dependent variable, BMI categories only controls.

than individuals whose weight is normal, on average.[14] The same is true when additional variables that predict objective risks are included as controls, as can be seen in column (2). The increase in subjective risks of diabetes, hypertension, stroke and heart attack by BMI category is in line with excess weight being a risk factor for these diseases. Similarly, the absence of significant differences in subjective risks of chronic lung disease between the different BMI categories aligns with the fact that weight is not a major risk factor for this disease. Excess weight, however, is known to increase the risks of arthritis/rheumatism and rheumatoid arthritis. Finding no BMI gradient in subjective risks of these diseases is thus surprising.[15] Whether the missing gradient reflects a lack of awareness for weight as a risk factor among individuals with excess weight, however, cannot be inferred from these results. In order to investigate this issue further, we compare the subjective measures with objective ones.

Column (3) explicitly takes the individuals' objective risk measure into account. The results indicate that the reference group, i.e. those individuals whose weight is normal, overestimate the risks of diabetes and chronic lung disease by approximately 5 percentage points, and the risks of a stroke and a heart attack by around 8 percentage points. The risk of arthritis is on average correctly assessed among individuals with normal weight, while the risk of hypertension is underestimated by approximately 5 percentage points.

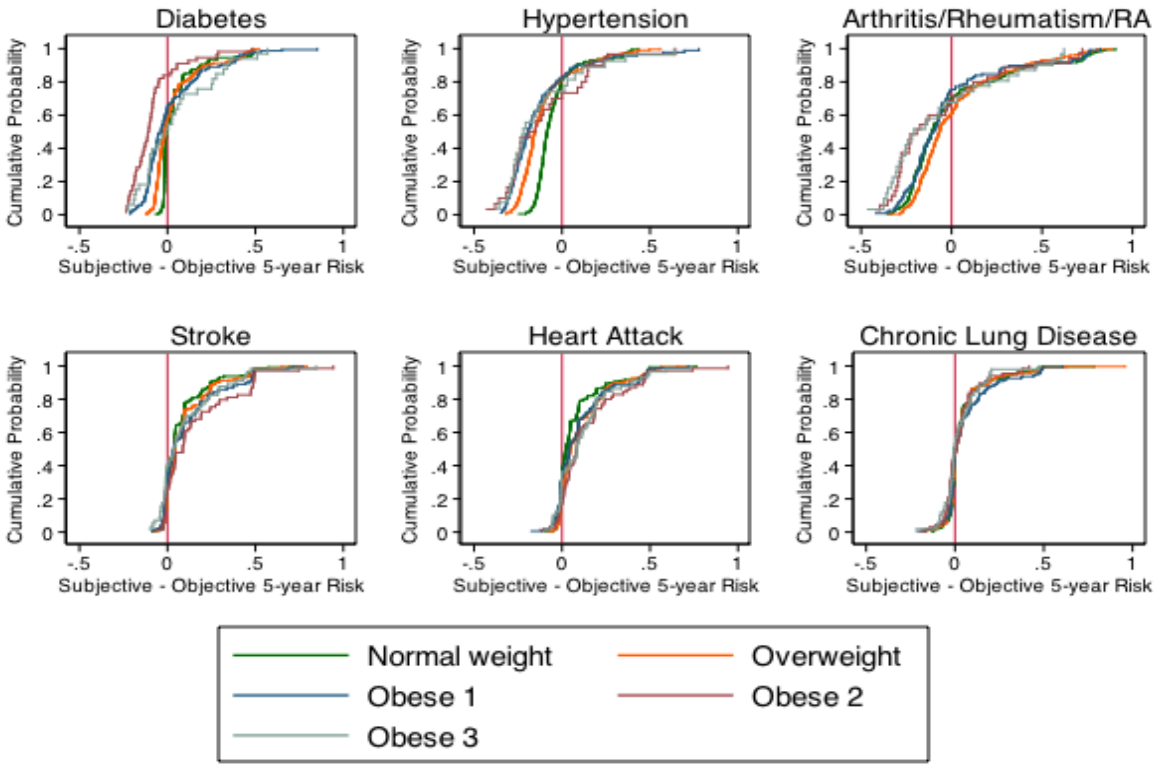
The coefficient estimates for the different BMI categories in column (3) can be interpreted as the difference in the average deviations of subjective from objective risk between the respective BMI group and normal-weight individuals. The results thus indicate that obese individuals overestimate the risks of a stroke and a heart attack even more than normal weight individuals. Furthermore, obese individuals overestimate the risk of chronic lung disease as much as normal weight individuals but underestimate the risk of diabetes, at least in one of the obesity categories. Similar to the normal weight, the average assessment of arthritis risk is accurate among the obese, while the risk of hypertension is underestimated.

Taken together, the results in table 4 indicate that the risks of a heart attack, a stroke and chronic lung disease are generally overestimated, with obese individuals overestimating the BMI gradient in the risks of heart attack and stroke. The risk of diabetes is on average

[14]Technically, the group of individuals with a BMI < 25 also includes individuals who are underweight, defined as having a BMI lower than 18.5. As these are only 2% of individuals in our data, we refer to the overall category as "normal" weight.

[15]This result does not depend on the aggregation of the subjective risk of arthritis and rheumatism and the risk of rheumatoid arthritis. Neither of the two separate risks significantly differs between the BMI categories.

Figure 3: Differences between subjective and objective risks



overestimated by the normal weight but underestimated among the obese, while the risk of hypertension is generally underestimated. Furthermore, on average the risk of arthritis is correctly assessed.

Averages, however, might hide important details as within a BMI group, over- and underestimation could cancel out between individuals. To investigate possible heterogeneities within BMI groups, figure 3 displays the cumulative distribution functions of the difference between subjective and objective risks for different diseases, stratified by BMI category. In addition, tables 5 and 6 report the results of quantile regressions. The different quantiles analyzed are the 10th, 25th, 50th, 75th and 90th quantile. Results for all quantiles within one disease were estimated jointly. Table 5 reports the results for the 10th and 25th quantile and table 6 the results for the remaining quantiles. The dependent variable in all estimations is the difference between subjective and objective 5-year risk.

Figure 3 and tables 5 and 6 indicate that the subjective assessment of the risks of a

heart attack, a stroke and chronic lung disease are more or less accurate for almost 80% of individuals, while the other 20% overestimate their risk by a large amount. The average overestimation thus hides that a large fraction of individuals seems to be able to assess their risks fairly accurately. For example, table 6 shows that the 10th percentile of the distribution of the difference between subjective and objective risk of a stroke among the normal weight is -1.3 percentage points while the median lies at 3.8 percentage points. The subjective risk of 40% of normal weight individuals is thus only off by between -1.3 and $+3.8$ percentage points. The 90th quantile, however, lies at 25 percentage points, indicating that 10% of normal weight individuals overestimate the risk of a stroke by at least this amount.

In the case of diabetes, a large fraction of normal weight individuals seems to be able to assess their risk rather well. Only 10% of the normal weight individuals underestimate their risks by more than 3 percentage points. Among the obese, however, 25% of individuals underestimate their risk by more than 10 percentage points. 25% of individuals in the category of BMI between 35 and 40 even underestimate their diabetes risk by more than 18 percentage points. In this latter BMI category, 75% of individuals underestimate their risks significantly.

Figure 3 and the two quantile regression tables confirm the average underestimation of hypertension risk in all weight categories. Only a small fraction of individuals overestimates their risk of hypertension. 50% of normal weight individuals underestimate their hypertension risk by at least 8 percentage points. Among the obese, the underestimation at the median is significantly larger than among the normal weight.

In contrast, the average results for arthritis hide that 50% of individuals of all BMI categories underestimate their risk by at least 10 percentage points with half of these individuals underestimating the arthritis risk by at least 20 percentage points. In the average result this is outweighed by a large overestimation of arthritis risk of at least 42 percentage points among 10% of individuals.

All in all, the results of the quantile regressions provide three key insights. First, a large fraction of individuals – also among the obese – is able to assess their risks of a heart attack, a stroke and chronic lung disease rather well. But some individuals in all BMI groups significantly overestimate these risks. Second, a large fraction of normal weight individuals does a good job at assessing their risk of diabetes, while a large fraction of obese individuals significantly underestimates this risk. Third, the risks of hypertension and arthritis are significantly underestimated by a large fraction of individuals in all BMI

Table 5: 10th and 25th quantile

	(1) Diabetes	(2) Hypertension	(3) Arthritis	(4) Heart Attack	(5) Stroke	(6) Lung Disease
10th Quantile						
25 ≤BMI<30	-0.0491***	-0.0975***	0.0327**	0.00488	0.00226	-0.00737
	(0.003)	(0.009)	(0.013)	(0.004)	(0.003)	(0.008)
30 ≤BMI<35	-0.109***	-0.155***	-0.0358**	-0.0200***	-0.00521	-0.00348
	(0.009)	(0.013)	(0.018)	(0.008)	(0.004)	(0.013)
35 ≤BMI<40	-0.199***	-0.171***	-0.0936***	0.00153	0.00143	-0.0294*
	(0.012)	(0.041)	(0.033)	(0.010)	(0.005)	(0.016)
40 ≤BMI	-0.165***	-0.165***	-0.123***	-0.0342*	-0.0270	-0.0509***
	(0.029)	(0.029)	(0.027)	(0.018)	(0.017)	(0.019)
Constant	-0.0206***	-0.146***	-0.236***	-0.0166***	-0.0128***	-0.0289***
	(0.001)	(0.006)	(0.009)	(0.003)	(0.002)	(0.006)
25th Quantile						
25 ≤BMI<30	-0.0348***	-0.0816***	0.0490***	0.00251	0.00261	-0.00113
	(0.002)	(0.009)	(0.018)	(0.004)	(0.003)	(0.003)
30 ≤BMI<35	-0.0834***	-0.144***	0.00689	-0.00430	-0.00475	-0.00558*
	(0.006)	(0.013)	(0.018)	(0.003)	(0.003)	(0.003)
35 ≤BMI<40	-0.161***	-0.155***	-0.0867***	0.0168	0.00530	-0.0197***
	(0.019)	(0.023)	(0.022)	(0.012)	(0.010)	(0.006)
40 ≤BMI	-0.0875***	-0.150***	-0.106***	-0.00264	-0.00966	-0.0160
	(0.027)	(0.020)	(0.032)	(0.008)	(0.007)	(0.010)
Constant	-0.0169***	-0.120***	-0.199***	-0.00383***	-0.00428***	-0.0101***
	(0.001)	(0.005)	(0.013)	(0.001)	(0.001)	(0.002)
N	817	567	615	911	911	874

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: Bootstrapped standard errors (400 reps) in parentheses

Table 6: Median, 75th and 90th quantile

	(1) Diabetes	(2) Hypertension	(3) Arthritis	(4) Heart Attack	(5) Stroke	(6) Lung Disease
Median						
25 ≤BMI<30	-0.00912	-0.0688***	0.0440**	0.0231**	0.00451	0.00434
	(0.007)	(0.010)	(0.018)	(0.009)	(0.006)	(0.006)
30 ≤BMI<35	-0.0307**	-0.110***	0.0111	0.0184	-0.00294	0.00376
	(0.015)	(0.018)	(0.024)	(0.013)	(0.012)	(0.009)
35 ≤BMI<40	-0.101***	-0.0649	-0.0824	0.0598***	0.0491**	0.00657
	(0.013)	(0.057)	(0.075)	(0.018)	(0.025)	(0.015)
40 ≤BMI	0.00101	-0.136***	-0.0976	0.0615***	-0.00391	-0.00669
	(0.044)	(0.044)	(0.078)	(0.023)	(0.022)	(0.012)
Constant	-0.00479	-0.0840***	-0.109***	0.0240***	0.0382***	0.00114
	(0.005)	(0.006)	(0.013)	(0.006)	(0.005)	(0.004)
75th Quantile						
25 ≤BMI<30	0.0125	-0.0516**	0.0541	0.0885***	0.0409	0.0118
	(0.024)	(0.024)	(0.072)	(0.019)	(0.036)	(0.020)
30 ≤BMI<35	0.0546	-0.0583	-0.0514	0.0726**	0.0870***	0.0428
	(0.038)	(0.040)	(0.075)	(0.030)	(0.025)	(0.029)
35 ≤BMI<40	-0.113***	0.106	0.0337	0.127**	0.150*	0.0115
	(0.033)	(0.089)	(0.125)	(0.055)	(0.085)	(0.023)
40 ≤BMI	0.126	0.0318	0.0731	0.0929***	0.0866	0.0283
	(0.096)	(0.095)	(0.158)	(0.033)	(0.055)	(0.028)
Constant	0.0428**	-0.0183*	0.0482	0.0961***	0.0962***	0.0457***
	(0.019)	(0.011)	(0.063)	(0.009)	(0.010)	(0.014)
90th Quantile						
25 ≤BMI<30	0.0230	0.0510	-0.0201	0.0443	0.0308	0.0131
	(0.074)	(0.061)	(0.143)	(0.083)	(0.061)	(0.051)
30 ≤BMI<35	0.119	-0.00799	0.0452	0.138	0.215***	0.0613
	(0.082)	(0.101)	(0.224)	(0.102)	(0.061)	(0.078)
35 ≤BMI<40	-0.135	0.0581	0.0420	0.162*	0.245***	-0.00888
	(0.094)	(0.116)	(0.239)	(0.085)	(0.041)	(0.070)
40 ≤BMI	0.149	0.154	0.0404	0.100	0.144*	-0.0124
	(0.092)	(0.182)	(0.199)	(0.105)	(0.074)	(0.057)
Constant	0.190***	0.0891*	0.422***	0.298***	0.245***	0.179***
	(0.039)	(0.048)	(0.131)	(0.066)	(0.026)	(0.043)
N	817	567	615	911	911	874

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: Bootstrapped standard errors (400 reps) in parentheses

groups, particularly so among the obese.

4 Robustness analysis

In this section, we present several robustness analyses. First, we explore how sensitive our results are to assumptions used in the HRS risk prediction model. Second, we investigate whether our results are sensitive to weighting. Third, we report estimates using the relative difference between subjective and objective risk instead of the absolute difference as dependent variable. Fourth, the sensitivity of our results to the inclusion of additional control variables is explored. Fifth, we group BMI in three instead of five categories, combining the three obesity categories into one. Finally, we investigate how rounding might affect our results.

The results of our robustness analysis are displayed in tables 7 and 8. Each set of results is based on a regression of the difference between subjective and objective risk on a constant and dummies for the different excess weight categories. They are thus comparable to the results displayed in column (3) of table 4.

Column (1) of table 7 shows that our results are not sensitive to the use of duration models as risk prediction models in the HRS. The results in column (1) use objective risks based on probit models instead of the ones based on duration models. In the probit prediction models, we use an indicator variable that takes on the value one if an individual is diagnosed with the specific condition sometime before the 2008 wave of the HRS and is zero otherwise as dependent variable instead of the time to disease onset. The results confirm conclusions drawn in the main analysis. Obese individuals underestimate the risk of diabetes and hypertension. Moreover, the results provide direct evidence that obese individuals also underestimate the risk of arthritis or rheumatism.

The sensitivity of our results to changes in the weighting procedure is investigated in columns (2) and (3) of table 7. In (2) we restrict the sample in the HRS to those individuals who report using the internet regularly in order to make the sample more comparable to the internet-based ALP data. In (3) we use ALP sampling weights in estimating the results. Both set of results are very similar to the results in column (3) of table 4. Furthermore, they are in line with our conclusion from the quantile regression analysis that obese individuals not only underestimate the risk of hypertension and diabetes but also the risk of arthritis or rheumatism.

Column (4) of table 7 reports the estimated coefficients of a regression model with the

Table 7: Sensitivity to choice of sample and functional form

	(1) HRS Probit		(2) HRS Internet		(3) ALP Weights		(4) Relative Risk	
Diabetes (N=817)								
25 ≤BMI<30	-0.026**	(0.011)	-0.025**	(0.011)	-0.011	(0.015)	-2.401***	(0.464)
30 ≤BMI<35	-0.056***	(0.017)	-0.044***	(0.017)	-0.025	(0.021)	-2.798***	(0.463)
35 ≤BMI<40	-0.193***	(0.020)	-0.147***	(0.021)	-0.159***	(0.021)	-3.500***	(0.451)
40 ≤BMI	-0.042	(0.036)	-0.009	(0.038)	-0.077*	(0.042)	-2.877***	(0.488)
Constant	0.044***	(0.008)	0.056***	(0.008)	0.050***	(0.011)	3.045***	(0.444)
R^2	0.093		0.055		0.067		0.078	
Hypertension (N= 567)								
25 ≤BMI<30	-0.072***	(0.015)	-0.107***	(0.015)	-0.065***	(0.021)	-0.136	(0.092)
30 ≤BMI<35	-0.119***	(0.025)	-0.115***	(0.024)	-0.092***	(0.032)	-0.116	(0.119)
35 ≤BMI<40	-0.094**	(0.045)	-0.019	(0.044)	-0.059	(0.052)	0.013	(0.152)
40 ≤BMI	-0.116**	(0.046)	-0.058	(0.045)	-0.103**	(0.052)	-0.106	(0.157)
Constant	-0.078***	(0.009)	-0.021**	(0.009)	-0.038**	(0.016)	-0.332***	(0.075)
R^2	0.067		0.086		0.045		0.005	
Arthritis/RA (N = 615)								
25 ≤BMI<30	0.012	(0.027)	-0.012	(0.027)	0.001	(0.035)	0.095	(0.144)
30 ≤BMI<35	-0.030	(0.034)	-0.080**	(0.034)	-0.055	(0.038)	-0.163	(0.157)
35 ≤BMI<40	-0.137**	(0.057)	-0.149***	(0.056)	-0.131**	(0.052)	-0.246	(0.204)
40 ≤BMI	-0.117*	(0.062)	-0.095	(0.065)	-0.104	(0.087)	-0.277	(0.194)
Constant	-0.038*	(0.022)	0.012	(0.022)	0.010	(0.028)	0.052	(0.112)
R^2	0.022		0.023		0.023		0.008	
Stroke (N = 911)								
25 ≤BMI<30	0.017	(0.012)	0.015	(0.012)	0.009	(0.018)	2.985	(1.834)
30 ≤BMI<35	0.029*	(0.016)	0.037**	(0.016)	0.036	(0.025)	-1.812	(1.788)
35 ≤BMI<40	0.079***	(0.027)	0.080***	(0.027)	0.076**	(0.037)	17.965**	(7.210)
40 ≤BMI	0.013	(0.026)	0.018	(0.026)	0.010	(0.042)	-2.068	(3.562)
Constant	0.077***	(0.009)	0.086***	(0.009)	0.088***	(0.014)	10.181***	(1.291)
R^2	0.015		0.017		0.014		0.034	
Heart Attack (N = 911)								
25 ≤BMI<30	0.026**	(0.012)	0.026**	(0.012)	0.035**	(0.017)	-1.055	(2.943)
30 ≤BMI<35	0.016	(0.015)	0.024	(0.015)	0.026	(0.023)	-8.550***	(2.610)
35 ≤BMI<40	0.059**	(0.025)	0.059**	(0.025)	0.040	(0.031)	-7.185**	(2.790)
40 ≤BMI	0.037	(0.025)	0.045*	(0.025)	0.033	(0.037)	2.193	(5.939)
Constant	0.077***	(0.009)	0.081***	(0.009)	0.080***	(0.013)	15.635***	(2.288)
R^2	0.011		0.011		0.008		0.012	
Lung Disease (N = 874)								
25 ≤BMI<30	0.003	(0.010)	0.006	(0.011)	0.006	(0.017)	0.084	(0.816)
30 ≤BMI<35	0.013	(0.013)	0.020	(0.013)	0.002	(0.017)	0.819	(1.224)
35 ≤BMI<40	-0.026*	(0.015)	-0.023	(0.016)	-0.029*	(0.018)	-2.052**	(0.799)
40 ≤BMI	-0.030*	(0.017)	-0.040**	(0.016)	-0.041**	(0.018)	-0.853	(1.331)
Constant	0.037***	(0.008)	0.046***	(0.008)	0.046***	(0.013)	3.579***	(0.646)
R^2	0.009		0.013		0.010		0.005	

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: Robust standard errors in parentheses. Dependent variable in columns (1), (2) and (3) is difference between subjective and objective risk. Column (1) uses objective risk based on probit model in the HRS, (2) uses only HRS participants who state that they use the internet to estimate the weibull duration model, (3) uses the main duration model and HRS sample but the ALP sampling weights. Dependent variable in column (4) is the difference between subjective and objective risk relative to the objective risk.

Table 8: Sensitivity to further controls, categorization of obesity, and rounding

	(1) Optimism		(2) Education		(3) Categorization		(4) Rounding	
Diabetes (N=817)								
25 ≤BMI<30	-0.012	(0.011)	-0.014	(0.011)	-0.014	(0.011)	-1.942	(1.470)
30 ≤BMI<35	-0.027	(0.017)	-0.031*	(0.017)			-1.500	(1.990)
35 ≤BMI<40	-0.138***	(0.020)	-0.136***	(0.019)			-15.916***	(2.214)
40 ≤BMI	-0.008	(0.037)	-0.009	(0.037)			-2.434	(3.763)
30 ≤BMI					-0.052***	(0.014)		
Constant	0.121***	(0.037)	0.046***	(0.017)	0.050***	(0.008)	10.703***	(1.082)
R^2	0.057		0.052		0.020		0.043	
Hypertension (N= 567)								
25 ≤BMI<30	-0.057***	(0.014)	-0.055***	(0.014)	-0.057***	(0.014)	-6.086***	(1.826)
30 ≤BMI<35	-0.082***	(0.024)	-0.076***	(0.025)			-6.501**	(2.743)
35 ≤BMI<40	-0.061	(0.044)	-0.053	(0.043)			-7.057	(4.577)
40 ≤BMI	-0.075	(0.046)	-0.070	(0.047)			-7.061	(4.870)
30 ≤BMI					-0.077***	(0.020)		
Constant	-0.016	(0.051)	-0.076***	(0.024)	-0.045***	(0.009)	1.341	(1.274)
R^2	0.037		0.052		0.035		0.023	
Arthritis/RA (N = 615)								
25 ≤BMI<30	0.018	(0.027)	0.023	(0.027)	0.020	(0.027)	0.606	(2.765)
30 ≤BMI<35	-0.029	(0.034)	-0.026	(0.034)			-2.625	(3.420)
35 ≤BMI<40	-0.068	(0.059)	-0.065	(0.056)			-6.802	(5.910)
40 ≤BMI	-0.073	(0.060)	-0.055	(0.060)			-6.413	(6.043)
30 ≤BMI					-0.041	(0.031)		
Constant	0.018	(0.075)	-0.064*	(0.038)	0.001	(0.022)	6.938***	(2.207)
R^2	0.012		0.034		0.008		0.009	
Stroke (N = 911)								
25 ≤BMI<30	0.019	(0.012)	0.018	(0.012)	0.019	(0.012)	0.916	(1.471)
30 ≤BMI<35	0.034**	(0.016)	0.026*	(0.016)			4.307**	(1.873)
35 ≤BMI<40	0.071***	(0.026)	0.073***	(0.027)			5.320*	(2.864)
40 ≤BMI	0.016	(0.025)	0.013	(0.025)			1.792	(2.840)
30 ≤BMI					0.041***	(0.014)		
Constant	0.166***	(0.039)	0.107***	(0.019)	0.082***	(0.009)	14.156***	(1.102)
R^2	0.035		0.028		0.010		0.009	
Heart Attack (N = 911)								
25 ≤BMI<30	0.028**	(0.012)	0.027**	(0.012)	0.028**	(0.012)	1.765	(1.516)
30 ≤BMI<35	0.022	(0.015)	0.016	(0.016)			2.926	(1.877)
35 ≤BMI<40	0.058**	(0.025)	0.062**	(0.025)			3.967	(2.619)
40 ≤BMI	0.037	(0.025)	0.038	(0.025)			4.827*	(2.770)
30 ≤BMI					0.035***	(0.014)		
Constant	0.124***	(0.035)	0.106***	(0.019)	0.082***	(0.009)	14.280***	(1.164)
R^2	0.031		0.020		0.008		0.006	
Lung Disease (N = 874)								
25 ≤BMI<30	0.002	(0.010)	0.001	(0.010)	0.002	(0.010)	-0.840	(1.433)
30 ≤BMI<35	0.014	(0.013)	0.013	(0.013)			2.204	(1.767)
35 ≤BMI<40	-0.015	(0.015)	-0.016	(0.015)			-3.863**	(1.850)
40 ≤BMI	-0.020	(0.017)	-0.019	(0.017)			-1.730	(2.434)
30 ≤BMI					0.002	(0.011)		
Constant	0.059*	(0.031)	0.037**	(0.016)	0.045***	(0.008)	10.524***	(1.090)
R^2	0.006		0.008		0.000		0.009	

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: Robust standard errors in parentheses. Columns (1), (2) and (3) use difference between subjective and objective risk as dependent variable. Column (1) includes a measures of optimism and pessimism as controls, column (2) includes education categories, column (3) only uses objective risk based on only one category of obesity (BMI≥ 30) and also uses only this category in estimation, and column (4) uses upper bound of rounding interval instead of face-value of subjective risk in dependent variable.

relative difference of subjective and objective risk, $\frac{S_i^d - \widetilde{O_i^d}}{\widetilde{O_i^d}}$, instead of the absolute difference as dependent variable. The results are similar to the results of the main analysis. Individuals who are normal- or underweight overestimate the risks of most diseases significantly. As the percent rather than percentage point metric puts an emphasis on deviations from smaller objective values, it comes as no surprise that there is overestimation between 350 and almost 1600% for diseases that show relatively low risks among normal- and underweight individuals. As in the main results, the risk of hypertension is underestimated and the risk of arthritis is correctly assessed among the normal- and underweight. Since the relative risk measure puts less weight on deviations from higher objective risks, the evidence for underestimation of risks among the obese is weaker than in the main results. Nevertheless, obese individuals underestimate their risk of hypertension to the same extent as normal weight individuals.

The first two sets of results in table 8 include additional control variables. In column (1) measures of optimism and pessimism based on Scheier et al. (1994) are included as additional controls to investigate whether systematic differences in these measures between BMI categories drive our results. The coefficient estimates are the same as in column (3) of table 4, indicating that differences in optimism or pessimism between respondents do not drive our results. Similarly, the results in column (2) indicate that including information on educational attainment as additional control does not change the estimated coefficients of the BMI categories. Systematic differences in educational attainment between BMI categories thus do not seem to be able to explain our results.

The results presented in column (3) of table 8 investigate the robustness of our main results with respect to the definition of obesity. Instead of three obesity categories, one category for individuals with a BMI of 30 or larger is included as explanatory variable in the model. Compared to the results in our main specification, the only qualitative difference occurs for the risk of diabetes. The underestimation of this risk among the obese disappears when collapsing the three categories to one.

The final robustness check investigates whether our results are sensitive to rounding. We follow Manski and Molinari (2010) and construct intervals for each individual's subjective expectation responses that depend on the answers that individuals give to all 5-year expectation questions. For example, if individuals answer all of these questions with 0 or 100 we infer that a response of 0 means they think the risk is not higher than 50 percent, i.e. the stated 0 corresponds to an interval of [0, 50]. Similarly, if all of an individual's answers are multiples of 10, we assume that she is rounding her responses to the nearest

10, and thus a stated risk of 10 could mean any number between 5 and 15. As we are particularly interested in underestimation of risks, we use only the upper bound (UB_i) of these rounding intervals in our robustness check. The results in column (4) use the upper bound of the rounding interval for the subjective risk instead of the subjective risk measure itself, i.e. the dependent variable is $UB_i^d - \widetilde{O_i^d}$ instead of $S_i - \widetilde{O_i^d}$. Not surprisingly, the results indicate larger overestimation of risks among the normal weight compared to the main analysis. Furthermore, the results of this specification indicate that normal weight individuals do not underestimate the risk of hypertension while obese individuals still underestimate the risks of hypertension and diabetes. Underestimation of these risks among the obese thus seems to be robust to taking rounding into account.

A final concern we should address is related to our result that respondents tend to overestimate the incidence probability of diseases with small (predicted) risk while they underestimate the probability of developing diseases with high risk. This feature of our data could be purely mechanical if respondents tend to give similar probability answers (say, 5 percent) for several diseases, perhaps due to rounding. The robustness analysis with respect to rounding that we just presented does not fully alleviate this concern, but we conducted a number of additional checks. These indicate that our results do not change when estimated only for individuals who assign different risks to at least two of the studied diseases (diabetes, hypertension, arthritis, rheumatoid arthritis, stroke, heart attack, and lung disease). Similarly, restricting the estimation sample to individuals with a range of at least 5 percentage points between the subjective risks of the studied diseases does not affect the results. Full results for these robustness checks are available on request. Taken together, they lead us to conclude that our main finding of systematic over- and underreporting is not a purely mechanical effect of response behavior. But even if it were, as long as respondents use the probabilities they provide in our survey when they make health-related decisions, which is the premise of the entire literature on the elicitation of subjective probabilities, our substantive conclusions would not change.

Overall, the robustness analysis confirms our main findings that obese individuals underestimate the 5-year risk of hypertension and diabetes. The underestimation of diabetes risks, however, is confined to individuals in one of the three obesity groups, namely individuals with a BMI between 35 and 40. This section additionally supports the conclusions drawn from the quantile regression results that obese individuals also underestimate the risk of arthritis and rheumatism.

5 Discussion and conclusions

In this paper, we add to the literature on risk perception among the obese by comparing subjective probabilities of developing different diseases within the following 5 years to objective probabilities of the same events. While the earlier literature shows that the obese are often not aware of their increased health risks in a general sense, our results indicate that the obese are aware of some health risks but not of others. In particular, they even overestimate their risks of getting a heart attack or a stroke within the next 5 years. Other risks, however, such as the risk of diabetes, hypertension or arthritis are underestimated among the obese.

In general, the obese thus overestimate relatively small risks such as the risk of a heart attack and a stroke and underestimate large risks, in particular the risk of hypertension. Our results indicate a similar pattern of under- and overestimation among individuals with a normal weight. These results are in line with findings in the psychological literature on risk perception (see e.g. Lichtenstein et al., 1978) and could be explained by Tversky's and Kahneman's "availability heuristic", according to which "people assess [...] the probability of an event by the ease with which instances or occurrences can be brought to mind"(Tversky and Kahneman, 1974) .

Differences in availability between the diseases could result from different levels of salience.[16] Heart attack and stroke are leading killers and might thus be more salient than hypertension, which is often symptom-free and has less dramatic immediate consequences. For example, direct treatment costs amounted to USD 5,112 in 2008 for an average patient with stroke and to USD 4,114 for an average patient with heart disease, whereas the direct costs of treating hypertension were only USD 858 per patient on average.[17]

While the general patterns of over- and underestimation are similar among normal weight and obese individuals, there are important level differences. The obese overestimate small risks and underestimate large risks to a larger extent than the normal weight. These differences are robust to controlling for other covariates, such as educational achievement. Similar heterogeneities have also been documented in the literature on health risk perception among smokers (see Cawley and Ruhm, 2012, for a summary). Furthermore,

[16]We are not aware of an individual-level dataset that contains information on both subjective risk perceptions and awareness of population health risks. However, at the aggregate level, internet search volumes tracked by Google Trends (*http://trends.google.com*) show that relative to their population prevalence, diseases such as hypertension are much less often searched for than heart attack or stroke (details are available in request). The underestimation of disease risks might thus reflect a general lack of salience.

[17]These data on disease costs come from the Medical Expenditure Panel Survey (MEPS), *http://www.meps.ahrq.gov/mepsweb/*.

we document heterogeneities in accuracy of health risk perceptions within BMI groups. As in other contexts in economics, understanding the sources of heterogeneities in subjective expectations and behavior constitutes an important topic for future research (see Hurd, 2009).

Concerning the design of policies to decrease obesity prevalence, our results suggest that a general increase in public information on health risks related to obesity is unlikely to be effective. Already today the risks of severe diseases such as a heart attack and a stroke are overestimated among the obese. However, increasing the information on the risks of hypertension, arthritis and diabetes among the obese might be beneficial. While to our knowledge no causal evidence for a relationship between risk perception and calorie intake or expenditures exists, Kan and Tsai (2004) show that an increase in information on obesity-related risks is related to a decrease in BMI among obese men in Taiwan. Furthermore, there are a couple of studies that find relationships between health risk perceptions and other behavior. De Paula et al. (2011) provide evidence that individuals change their sexual behavior following a change in beliefs about their HIV infection status. Moreover, there is correlational evidence that individuals who rate their health risks as higher are more likely to take up preventive measures, such as vaccinations (Brewer et al., 2007; Carman and Kooreman, 2011), hygienic precautionary measures (Bruine de Bruin et al., 2011), and screenings (Carman and Kooreman, 2011). A relation between expectations and behavior has also been found in areas other than health, such as financial decision making (e.g. Hurd et al., 2011).

In general, when it comes to behavioral changes, one might ask whether individuals react to the expected *level* of future health risks that is associated with being obese (as we implicitly assume). Alternatively, one might argue that individuals' behavior is affected more by the expected *change* of future health risks that would be induced by a behavioral change such as gaining or losing weight. Distinguishing between information on risk levels and (potential) risk changes is potentially important when it comes to communicating future health risks to high-risk groups such as the obese in order to affect their behavior. Our data do not directly inform the question of which version works better because we only asked for expected risk levels in our questionnaire, so we leave this issue to future research.

References

Anderson, R. (2000). Method for constructing complete annual US life tables. *Vital and Health Statistics*, 2(129):1–10.

Brewer, N., Chapman, G., Gibbons, F., Gerrard, M., McCaul, K., and Weinstein, N. (2007). Meta-analysis of the relationship between risk perception and health behavior: The example of vaccination. *Health Psychology*, 26(2):136–145.

Bruine de Bruin, W. and Carman, K. (2012). Measuring risk perceptions: What does the excessive use of 50% mean? *Medical Decision Making*, 32:232–236.

Bruine de Bruin, W., Parker, A., and Maurer, J. (2011). Assessing small non-zero perceptions of chance: The case of H1N1 (swine) flu risks. *Journal of Risk and Uncertainty*, 42:145–159.

Carman, K. and Kooreman, P. (2011). Flu shots, mammograms, and the perception of probabilities. IZA Discussion Paper 5739.

Cawley, J. and Meyerhoefer, C. (2012). The Medical Care Costs of Obesity: An Instrumental Variables Approach. *Journal of Health Economics*, 31(1):219–230.

Cawley, J. and Ruhm, C. J. (2012). The Economics of Risky Health Behaviors. In McGuire, T., Pauly, M., and Barros, P., editors, *Handbook of Health Economics*, volume 2, pages 95–199. Elsevier.

Couper, M. P., Kapteyn, A., Schonlau, M., and Winter, J. (2007). Noncoverage and nonresponse in an internet survey. *Social Science Research*, 36(1):131–148.

De Paula, A., Shapira, G., and Todd, P. (2011). How beliefs about HIV status affect risky behaviors: Evidence from Malawi, Sixth Version. PIER Working Paper 11-005.

Dixon, J. (2010). The effect of obesity on health outcomes. *Molecular and Cellular Endocrinology*, 316(2):104–108.

Falba, T. and Busch, S. (2005). Survival expectations of the obese: Is excess mortality reflected in perceptions? *Obesity Research*, 13(4):754–761.

Goldman, D., Michaud, P., Lakdawalla, D., Zheng, Y., Gailey, A., and Vaynman, I. (2010). The Fiscal Consequences of Trends in Population Health. *National Tax Journal*, 63(2):307–330.

Gregory, C., Blanck, H., Gillespie, C., Maynard, L., and Serdula, M. (2008). Perceived health risk of excess body weight among overweight and obese men and women: differences by sex. *Preventive Medicine*, 47(1):46–52.

Haslam, D. and James, W. (2005). Obesity. *The Lancet*, 366(9492):1197–1209.

Hurd, M. (2009). Subjective probabilities in household surveys. *Annual Review of Economics*, 1:543–562.

Hurd, M., van Rooij, M., and Winter, J. (2011). Stock market expectations of dutch households. *Journal of Applied Econometrics*, 26:416–436.

Juster, F. T. and Suzman, R. (1995). An overview of the Health and Retirement Study. *Journal of Human Resources*, 30:S7–S56.

Kan, K. and Tsai, W. (2004). Obesity and risk knowledge. *Journal of Health Economics*, 23(5):907–934.

Khwaja, A., Silverman, D., Sloan, F., and Wang, Y. (2009). Are mature smokers misinformed? *Journal of Health Economics*, 28(2):385–397.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., and Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory*, 4(6):551–578.

Manski, C. (2004). Measuring expectations. *Econometrica*, 72(5):1329–1376.

Manski, C. and Molinari, F. (2010). Rounding probabilistic expectations in surveys. *Journal of Business and Economic Statistics*, 28(2):219–231.

National Center for Health Statistics (2010). *Health, United States, 2009: With Special Feature on Medial Technology*. Hyattsville, MD.

Scheier, M., Carver, C., and Bridges, M. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation of the life orientation test. *Journal of Personality and Social Psychology*, 67(6):1063–1078.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Appendices

A Calculation of objective risk using HRS data

Objective risks for individuals in the ALP are based on prediction models for disease onset estimated using data from the HRS. Similar to Khwaja et al. (2009), we use D duration models to model the relationship between the individual characteristics in a baseline year and the time to onset of each disease d, $d \in \{1, ..., D\}$. We assume that the survivor functions follow Weibull distributions and allow for Gamma distributed unobserved heterogeneity. The survivor function for disease d is

$$R(t_i^d; \mathbf{X_i}, \boldsymbol{\theta}^d, \mu^d | \eta_i^d) = \left(exp(-\lambda_i^d (t_i^d)^{\mu^d}) \right)^{\eta_i^d} \tag{1}$$

where λ_i^d is parameterized as $exp(-\mu^d \mathbf{X}_i' \boldsymbol{\theta}^d)$ so that the survival distribution is a Weibull in accelerated failure time (AFT) metric. η_i^d stands for the unobserved heterogeneity that is $\Gamma(\frac{1}{\sigma^d}, \sigma^d)$ distributed. t_i^d represents the time until individual i is diagnosed with disease d. If individual i does not report a diagnosis of the disease within the time that she is observed in the HRS, i is treated as a right-censored observation. In this case, t_i^d contains the time to censoring, i.e. the time in which i is observed.[18] μ^d is the shape parameter of the Weibull distribution, $\mathbf{X}_i$ represents the vector of individual characteristics and includes information on smoking status, BMI categories, self-rated health, age, sex, educational degree, marital status, and race. $\boldsymbol{\theta}^d$ is the corresponding vector of coefficients.

We estimate μ^d, $\boldsymbol{\theta^d}$, and σ^d by maximum likelihood. Based on these parameter estimates, we calculate each individual i's probability of not getting disease d in the next t years as

$$\widehat{O_i^d}(t; \mathbf{X}_i, \hat{\boldsymbol{\theta}}^d, \hat{\mu}^d) = E_\eta \left(\widehat{O_i^d}(t; \mathbf{X}_i, \hat{\boldsymbol{\theta}}^d, \hat{\mu}^d | \hat{\eta}^d) \right) = \left[1 + \hat{\sigma}^d exp(-\hat{\mu}^d \mathbf{X}_i' \hat{\boldsymbol{\theta}}^d) t^{\hat{\mu}^d} \right]^{-\frac{1}{\hat{\sigma}^d}} \tag{2}$$

Naturally, the probability of getting the disease within the next t years then is

$$\widetilde{O_i^d} = 1 - \widehat{O_i^d}(t; \mathbf{X}_i, \hat{\boldsymbol{\theta}}^d, \hat{\mu}^d). \tag{3}$$

As the time horizon of the subjective risk is 5 years, we set $t = 5$.

[18] The time to disease onset, t_i^d, is measured in years. A report of a new diagnosis in wave A is coded as the time in years between wave 1 and wave A minus 1. As there are two years between consecutive waves, t_i^d is coded as 5, for example, in the case of a new report of a disease d by individual i in wave 4. For right-censored observations t_i^d is measured in years between 2002 and the year of the observation i's last interview.

Table 9: Duration model for disease onsets in HRS

	(1) Diabetes		(2) Hypertension		(3) Arthritis		(4) Stroke		(5) Heart Attack		(6) Lung Disease	
Age	-0.019	(0.019)	-0.036**	(0.016)	-0.017	(0.019)	-0.093**	(0.039)	-0.046	(0.036)	-0.015	(0.027)
Male	0.010	(0.106)	0.162*	(0.085)	0.451***	(0.100)	-0.517**	(0.203)	-0.941***	(0.184)	0.284*	(0.155)
SRH – very good	-0.032	(0.152)	-0.039	(0.110)	-0.232	(0.157)	-0.112	(0.308)	0.117	(0.326)	-0.160	(0.225)
SRH – good	-0.027	(0.160)	-0.168	(0.116)	-0.287*	(0.150)	-0.387	(0.296)	-0.452	(0.320)	-0.646***	(0.244)
SRH – fair	-0.428**	(0.207)	-0.151	(0.164)	-0.671***	(0.185)	-0.574*	(0.340)	-1.125***	(0.322)	-0.943***	(0.239)
SRH – poor	-0.543**	(0.217)	-0.198	(0.201)	-1.068***	(0.293)	-1.415***	(0.477)	-1.950***	(0.378)	-1.465***	(0.338)
White	0.112	(0.128)	0.140	(0.113)	-0.010	(0.122)	0.544**	(0.263)	-0.218	(0.248)	-0.253	(0.190)
Married	-0.102	(0.112)	-0.028	(0.094)	-0.209	(0.142)	0.157	(0.202)	-0.005	(0.232)	0.019	(0.156)
Less than HS	-0.100	(0.145)	-0.144	(0.117)	0.366**	(0.154)	0.084	(0.336)	-0.031	(0.277)	0.137	(0.189)
Some College	0.116	(0.128)	-0.019	(0.103)	0.198	(0.132)	-0.022	(0.257)	-0.301	(0.231)	0.320*	(0.167)
BA or eq.	0.416**	(0.169)	0.097	(0.131)	0.237	(0.154)	-0.037	(0.291)	0.706*	(0.391)	0.572**	(0.265)
More than BA	0.079	(0.166)	0.201	(0.144)	0.149	(0.159)	0.075	(0.362)	0.307	(0.344)	0.514**	(0.259)
Current Smoker	-0.087	(0.130)	0.020	(0.105)	-0.140	(0.133)	-0.817***	(0.295)	-0.444*	(0.247)	-1.485***	(0.187)
Former Smoker	0.103	(0.112)	0.053	(0.088)	-0.287**	(0.111)	-0.048	(0.202)	0.050	(0.209)	-0.584***	(0.170)
Overweight	-0.891***	(0.154)	-0.493***	(0.099)	0.132	(0.141)	0.073	(0.275)	-0.085	(0.248)	-0.027	(0.174)
Obese 1	-1.316***	(0.164)	-0.617***	(0.120)	-0.062	(0.159)	-0.260	(0.283)	-0.439	(0.286)	-0.108	(0.196)
Obese 2	-1.861***	(0.198)	-0.748***	(0.173)	-0.355	(0.220)	0.242	(0.380)	-0.468	(0.369)	-0.200	(0.259)
Obese 3	-1.647***	(0.213)	-0.743***	(0.184)	-0.262	(0.231)	-0.545	(0.393)	-0.118	(0.370)	0.040	(0.290)
Constant	5.037***	(1.211)	4.675***	(0.941)	3.154***	(1.175)	9.421***	(2.077)	7.836***	(2.372)	5.412***	(1.603)
$\ln(\mu)$	0.485***	(0.089)	0.417***	(0.071)	0.554***	(0.166)	0.513**	(0.256)	0.424***	(0.154)	0.476***	(0.109)
$\ln(\sigma)$	1.484***	(0.538)	0.650	(0.540)	1.531**	(0.612)	3.165***	(1.138)	2.866***	(0.595)	2.039***	(0.573)
N	4071		2656		2443		4580		4720		4454	

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: Robust standard errors in parentheses. Weibull AFT parameterization with Gamma distributed heterogeneity. μ shape parameter of the Weibull, σ parameter of Gamma distribution. Estimates based on individuals aged 50–62 in the 2002 HRS with follow-up until 2008. Each estimation only includes individuals who do not report having been diagnosed with specific disease before 2002. Results are weighted using sampling weights provided with the data.

Table 9 reports coefficients and standard errors after estimation of the model specified in equation (1) for the different diseases. As the models are specified in accelerated failure time metric, a positive coefficient indicates that the time to disease onset increases with the associated variable, i.e. the risk decreases with the variable. Conversely, a negative coefficient indicates that the risk increases with an increase in the associated variable. The risks of most diseases significantly increase with age. The disease risks are higher for worse assessments of self-rated health, and they increase with smoking and BMI categories. For most of the diseases the BMI categories are not significant predictors of risk. The coefficients however are significant and of about the same magnitude when the sample is not restricted to individuals aged 50 to 62 (results upon request).

B Results for individuals aged 50 and older in the ALP

While we only use individuals aged 50 to 62 in our main analysis in order to avoid modeling competing risks, it is possible to use the HRS to estimate objective risks for all middle aged to old individuals in the HRS. Table 10 thus displays the same results as table 4 but for individuals aged 50 and older in the ALP. The objective risks in these results are based on risk prediction models estimated on the entire HRS sample. Overall, the results in the two tables are very similar and our results are not driven by restricting the sample to only middle aged individuals.

C Calculation of objective risks using NHANES

In order to calculate objective 5-year risks of developing the analyzed diseases based on NHANES we proceed in two main steps. First, we estimate age-specific disease incidence in NHANES. Second, we use the age-specific incidence in a life table to predict 5-year risks.

We follow the Centers for Disease Control and Prevention (CDC) in calculating disease incidence using the NHANES data.[19] NHANES not only includes information on whether individuals have been diagnosed with different conditions in the past but also on the age at first diagnosis. In addition, there is information on individuals' current age. We take the difference between age at the time of the survey and age at diagnosis to calculate how long individuals have had the specific disease. If the difference is 0, individuals have been

[19]See link Methods and Limitations at http://www.cdc.gov/diabetes/statistics/incidence/fig2.htm.

Table 10: Subjective 5-year risks and obesity – 50+ sample

	Subjective risk (S) (1)		S and additional controls (2)		Difference ($S - \widetilde{O}$) (3)	
Diabetes (N=1269)						
25 ≤BMI<30	3.028***	(0.837)	2.842***	(0.857)	-0.011	(0.008)
30 ≤BMI<35	8.331***	(1.382)	7.393***	(1.396)	-0.016	(0.014)
35 ≤BMI<40	5.403***	(1.634)	4.036**	(1.628)	-0.092***	(0.016)
40 ≤BMI	11.423***	(3.410)	9.665***	(3.346)	-0.026	(0.033)
Constant	6.696***	(0.572)	3.699	(3.944)	0.038***	(0.006)
R^2	0.048		0.074		0.022	
Hypertension (N= 812)						
25 ≤BMI<30	3.649**	(1.440)	3.467**	(1.445)	-0.028**	(0.014)
30 ≤BMI<35	4.918**	(2.007)	4.725**	(2.078)	-0.067***	(0.020)
35 ≤BMI<40	11.880***	(3.989)	10.982***	(3.982)	-0.008	(0.040)
40 ≤BMI	6.324	(4.150)	6.096	(4.182)	-0.070*	(0.041)
Constant	10.147***	(0.967)	-13.532**	(6.864)	-0.075***	(0.010)
R^2	0.023		0.070		0.017	
Arthritis/RA (N = 894)						
25 ≤BMI<30	0.799	(2.302)	0.419	(2.327)	-0.010	(0.023)
30 ≤BMI<35	0.755	(2.998)	-0.374	(3.018)	-0.040	(0.029)
35 ≤BMI<40	3.839	(5.066)	0.784	(5.162)	-0.072	(0.051)
40 ≤BMI	1.911	(5.610)	-1.613	(5.694)	-0.082	(0.054)
Constant	23.395***	(1.789)	-7.326	(9.375)	0.016	(0.018)
R^2	0.001		0.053		0.006	
Stroke (N = 1415)						
25 ≤BMI<30	2.138**	(1.018)	1.851*	(1.005)	0.020**	(0.010)
30 ≤BMI<35	4.958***	(1.359)	3.773***	(1.325)	0.047***	(0.013)
35 ≤BMI<40	5.259**	(2.047)	3.435*	(2.049)	0.056***	(0.020)
40 ≤BMI	4.969**	(2.335)	2.858	(2.499)	0.045*	(0.023)
Constant	11.100***	(0.729)	-15.358***	(4.840)	0.082***	(0.007)
R^2	0.013		0.080		0.013	
Heart Attack (N = 1401)						
25 ≤BMI<30	2.964***	(1.048)	2.619**	(1.030)	0.020**	(0.010)
30 ≤BMI<35	4.576***	(1.350)	3.218**	(1.356)	0.026**	(0.013)
35 ≤BMI<40	4.934**	(1.934)	2.576	(1.974)	0.024	(0.019)
40 ≤BMI	5.981**	(2.454)	3.738	(2.572)	0.042*	(0.024)
Constant	11.114***	(0.767)	-12.713***	(4.570)	0.090***	(0.007)
R^2	0.013		0.092		0.005	
Lung Disease (N = 1362)						
25 ≤BMI<30	0.288	(0.856)	0.150	(0.825)	0.004	(0.008)
30 ≤BMI<35	2.298**	(1.072)	1.789*	(1.081)	0.015	(0.010)
35 ≤BMI<40	-1.070	(1.153)	-1.970*	(1.165)	-0.029**	(0.011)
40 ≤BMI	-0.610	(1.362)	-1.490	(1.454)	-0.018	(0.014)
Constant	7.485***	(0.645)	-2.411	(3.468)	0.044***	(0.006)
R^2	0.006		0.090		0.008	

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: Robust standard errors in parentheses. N indicates the number of observations used in all estimations displayed in the respective panel. It reflects the number of individuals aged 50-62 in the ALP without prior diagnosis of the respective condition. Columns (1) and (2) use subjective 5-year risk as dependent variable. In (1) BMI categories as controls, in (2) BMI categories plus age, sex, race, marital status, education and self assessed health as controls. In (3) difference between subjective and objective risk as dependent variable, BMI categories only controls.

diagnosed with the disease within the last year. Furthermore, following CDC we take half of the individuals with a difference of 1 as having been diagnosed with the condition within the last year.

In order to ensure large enough sample sizes for each age, we pool NHANES waves 2005–2006, 2007–2008, and 2009–2010 (except for hypertension for which information on the age at first diagnosis is not available in 2005/06).[20] Since NHANES over-samples some parts of the population and has a sophisticated sample design we further use sampling weights and take the sample design into account in all estimations involving these data, so as to ensure that the results are representative for the US population.

Disease incidence at age a is calculated as the weighted number of individuals who have been diagnosed with a disease within the last year relative to the weighted number of individuals that were at risk a year before the survey, that is all individuals who report never having been diagnosed with the disease at time of the survey and those individuals who got a diagnosis within the last year.

We further smooth the incidence curve by age using interpolation methods. In particular, we use Beer's ordinary minimized fifth difference formula that is also used in smoothing incidence in life tables (see e.g. Anderson, 2000).

The common range at which information on age at diagnosis is available for all diseases and all survey years is 20–79. We thus calculate incidence for this age range. We then use the age-specific incidence to infer 5-year risks at different ages using life table methods. In particular, we set the life table radix l_{20}, i.e. the number of individuals at risk at age 20, to 100,000. We then use the incidence at age 20 to calculate how many individuals will still be disease free at age 21, namely $l_{20}(1 - Incidence(20))$. In general, we get the number of disease-free individuals at age a as $l_a = l_{a-1}(1 - Incidence(a-1))$. Given l_a for all ages 20–79, we can calculate the probability of being diagnosed with a disease within the next 5 years at age a for ages 20 to 74 as $1 - \frac{l_{a+5}}{l_a}$.

D Obesity and subjective risk

The results in table 4 are based on estimating different equations. Column (1) explores how the subjective 5-year risk (S_i^d) of individual i for disease d varies between BMI categories by estimating

[20] Results based on pooling further waves are very similar and available upon request.

$$S_i^d = \beta_0^d + \beta_1^d OV_i + \beta_2^d OB1_i + \beta_3^d OB2_i + \beta_4^d OB3_i + \epsilon_i^d \tag{4}$$

where OV is a dummy for having a BMI between 25 and 30, and $OB1$, $OB2$, and $OB3$ stand for the three categories of obesity, $30 \leq$ BMI < 35, $35 \leq$ BMI < 40, and BMI≥ 40, respectively.

The results in column (2) are based on a specification that adds a vector of additional control variables to equation (4)

$$S_i^d = \delta_0^d + \delta_1^d OV_i + \delta_2^d OB1_i + \delta_3^d OB2_i + \delta_4^d OB3_i + \mathbf{X}_i'\boldsymbol{\delta} + \eta_i^d \tag{5}$$

where $\mathbf{X}_i$ consists of the additional predictors of the HRS risk prediction models, i.e. smoking status, self-rated health, age, sex, educational degree, marital status, and race. These are additional influences on the subjective risk that potentially vary systematically between different BMI categories.

In the estimation of the last column we follow Khwaja et al. (2009) and regress the difference between subjective and objective risk on dummies for the different obesity categories:

$$S_i^d - \widetilde{O_i^d} = \gamma_0^d + \gamma_1^d OV_i + \gamma_2^d OB1_i + \gamma_3^d OB2_i + \gamma_4^d OB3_i + u_i^d \tag{6}$$

If γ_0^d is significantly positive, individuals who are normal- or underweight significantly overestimate their risk of developing disease d. Significantly positive estimates for $\gamma_1^d, \gamma_2^d, \gamma_3^d$ or γ_4^d indicate that in the respective group the difference between subjective and objective risk is on average larger than among the normal- or underweight individuals. Individuals in group k on average underestimate their risk of disease d if $\gamma_0^d + \gamma_k^d$ is significantly smaller than 0.