Chair of Public Finance and Social Policy
of the
Free University Berlin

Master Thesis

The health–wealth nexus over the life cycle

Marcelo Rainho Avila

Supervisor: Prof. Dr. Carsten Schröder

Address: Seestraße 100, 13353, Berlin
Email: m.avila@fu-berlin.de
Submission: 07.02.2024
Matriculation-Nr.: 4679876

Contents

List of Figures

List of Tables

1 Introduction

"The first wealth is health." In 1860, the philosopher Ralph Waldo Emerson posited that good health is a precondition to flourishment and living a fulfilling life. Since then, researchers have shown that people with higher socioeconomic status live healthier and longer lives (see A. Deaton, 2003; Case and A. S. Deaton, 2005; Chetty et al., 2016).

While these results replicate across diverse societies and periods, the exact mechanisms are still open to discussion. To what extent wealth causes health or, conversely, health causes wealth is not trivial to disentangle given the nature of these variables. In a meta-analysis, Kröger et al. find evidence in both directions (Kröger et al., 2015). It is not difficult to think of possible mechanisms. People with higher socioeconomic status might, for example, have access to better health care or live in areas with more opportunities for following healthier life styles. On the other hand, good health might be a prerequisite for achieving financial success, as it enables individuals to participate fully in the workforce and pursue educational and career opportunities. Despite these broad associations, the specific pathways through which health influences wealth accumulation remain complex. This study aims to contribute to this discussion by focusing specifically on the pathway from health to wealth accumulation, quantifying how adverse health outcomes impact individuals' wealth trajectories over time. Further, by analyzing mental and physical health shocks separately, the effects from both domains can be compared to one another.

Figure 1.1 serves to illustrate the difference in wealth trajectory over the life cycle between two groups: those who experience a significant adverse health outcome over the life cycle, and those who do not. Although the starting values, on average, are on the same level earlier in life, one group manages to build up considerably more wealth than the other. By age 60, the group with a more stable health trajectory accumulates 35% (gross) and 40% (net) more wealth than their counterpart. This work aims to explore if, and to what extent, an adverse health outcome, whether in the mental or physical domain, helps explain the observed difference in this figure.

On these accounts, this analysis employs a Difference-in-Differences (DiD) method with multiple time periods and variation in treatment timing to estimate the causal effect of an adverse health outcome on the wealth trajectory of adults in Germany. The unbalanced panel data employed in this work are provided by the German Socio-Economic Panel

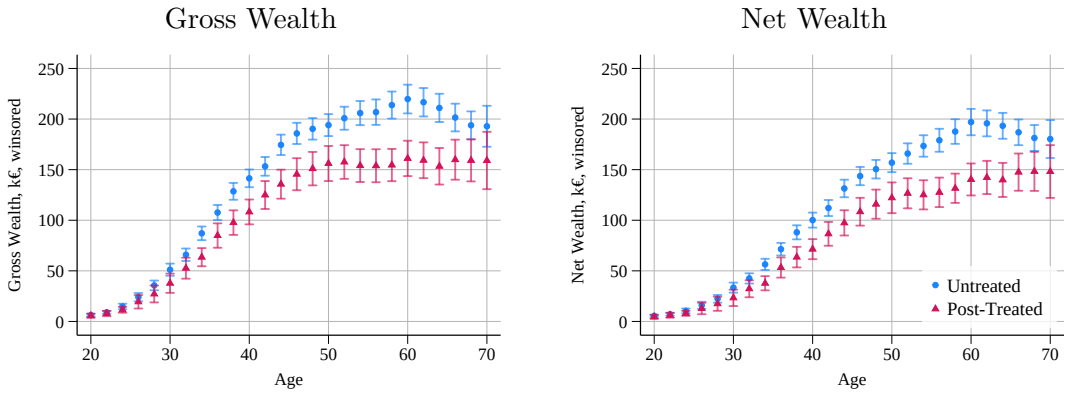


Figure 1.1 Wealth trajectory by treatment status

Notes: Figure shows two-year moving average with whiskers depicting the 99% confidence intervals. Untreated refer to those that did not experience a adverse health outcome up to a given age, while post-treated refer to those treated from the respective certain age onward. For brevity, only the treatment relative to mental health domain is shown. Sample composition changes over life cycle, as the longest time span for a single individual is 19 years. Data source: SOEPv37

(SOEP). The final sample encompasses people aged between 18 and 75 and provides, among other variables, health and wealth information since 2002. It includes over 16 thousand individuals and more than 80 thousand observations.

This work is structured as follows: Chapter 2 presents an overview of the DiD framework proposed by Callaway and Sant'Anna (2021). Chapter 3 describes the data and focal variables used in this study. Chapter 4 focuses on the measurement of health and, as a secondary analysis, evaluates some of the criticisms raised in the literature about uncorrelated health summary scores. Further, the alternative method for computing the summary scores used in this analysis is introduced. Chapter 5 covers the empirical approach and focuses on the results obtained in the main analysis. Also in this chapter, secondary analyses are presented. They explore wealth accumulation links and evaluate effect heterogeneity by education attainment and age groups, as well as a section on robustness checks. Chapter 6 presents a discussion on the obtained results, including a section on threat to internal validity and another on extensions to the current study. Finally, Chapter 7 summarizes the findings in this study.

2 Overview of DiD with Staggered Treatment from Callaway and Sant'Anna (2021)

In recent years, numerous authors have advanced the Difference-in-Differences design literature by addressing the "forbidden comparisons" issue, term coined and explained in Borusyak et al. (2024, p. 14). In short, it arises when comparing later treated to early treated units in a staggered DiD setting in the presence of effect heterogeneity. One of these approaches was proposed by Callaway and Sant'Anna (2021), which is not only robust to this issue but offers other compelling features, as discussed below. For an overview on the recent methods refer to the extensive survey from de Chaisemartin and D'Haultfoeuille (2022).

In the following section I present an overview of the estimator as proposed from Callaway and Sant'Anna (2021). I include only the sections that are relevant to this analysis, while variations and other considerations are left out. For consistency, I borrow the terminology from the authors for this analysis. In particular, I use terms such as "comparison" units instead of commonly used "control" units. I also use the term "group" which can be exchanged for "cohort".

2.1 Building Blocks

In short, the proposed estimator can be divided into two steps. First, compute the differences of valid comparison groups in several simple, balanced 2-periods by 2-groups DiD designs. Then, aggregate these estimations in a meaningful way. Adapting the notation of Eq. 2.8 from (ibid.), the groups in the first step can be defined by their respective treatment period g and the measurement time t. So each group-time Average Treatment Effects on the Treated, hereafter $ATT(g,t)$ is defined as

$$ATT(g,t) = \mathbb{E}\left[Y_t - Y_{g-1} | G_g = 1\right] - \mathbb{E}\left[Y_t - Y_{g-1} | C = 1\right], \tag{2.1}$$

where the conditional terms $G_g = 1$ identifies the group first treated at period g and $C = 1$ determines the comparison units.[1] This is the *group-time average treatment effect*, as defined by the authors (ibid., Section 2.2). Since each one of these groups is by construction,

[1] Note that the authors introduce two approaches that differs in terms of comparison units. One uses never-treated only, while the other uses also not-yet-treated units as comparison. I focus here on the never-treated specification, which is used in the main analysis, due to the risk of anticipation effects biasing the results.

a canonical DiD design, they are valid comparisons and do not suffer from the "negative weights" issue.

In the second step, the Callaway and Sant'Anna (2021) propose different ways of aggregating all $ATT(g,t)$'s. In the main analysis, I follow the event study aggregation. For that, a relative-time variable $e = t - g$ is defined, which identifies time elapsed since treatment adoption. The target parameter is

$$\theta_{\text{es}}(e) = \sum_{g \in \mathcal{G}} \mathbb{1}[g + e \leq \mathcal{T}] P(G = g | G + e \leq \mathcal{T}) ATT(g, g + e), \tag{2.2}$$

where $\mathcal{T}$ is the last period in the analysis. The two new terms in the above summation are, first, the indicator function that restricts the group-time average treatment effects. And second, the term $P(.)$ are the group weights, so that the summation produces an average - weighted by group size - of all $ATT(g,t)$'s included in the aggregation.

Note that this aggregation may suffer from compositional changes, which will have important implications, as explained thoroughly in Callaway and Sant'Anna (ibid., see Equation 3.5). In short, the authors show that there are two approaches to handling it: either restricting the analysis to a balanced sample or, alternatively, relying on the additional assumption that the effects are equal across groups. That is, $ATT(g, g + e)$ does not vary with g for any $e \geq 0$. In the main analysis, I make the equal-effects assumption, but robustness checks with sample restrictions do not change the results considerably.

2.2 Summary Measures

After computing the $ATT(g,t)$'s, there are a few choices of overall aggregation to obtain a single parameter estimating the effect being assigned to treatment. As proposed by Callaway and Sant'Anna (ibid., Section 3.2), an intuitive one would be

$$\bar{\theta}_{\text{simple}} = \frac{1}{k} \sum_{g \in \mathcal{G}} \sum_{t=2}^{\mathcal{T}} \mathbb{1}[t \geq g] ATT(g,t) P(G = g | G \leq \mathcal{T}), \tag{2.3}$$

where $k = \sum_{g \in \mathcal{G}} \sum_{t=2}^{\mathcal{T}} \mathbb{1}[t \geq g] ATT(g,t) P(G = g | G \leq \mathcal{T})$. Another aggregation target proposed is relevant for the event-study structure applied here. We can obtain an overall treatment effect by averaging $\theta_{es}(e)$ over all event times, as in

$$\bar{\theta}_{\text{es}} = \frac{1}{\mathcal{T} - 1} \sum_{e=0}^{\mathcal{T}-2} \theta_{es}(e). \tag{2.4}$$

While the authors propose other summary measures, these are the relevant ones examined in this analysis and presented in Section 5.2.

2.3 Doubly-Robustness

Callaway and Sant'Anna propose three approaches to achieve an estimator capable of conducting inference on the group-time average treatment effects (Callaway and Sant'Anna, 2021, Sec. 4). In short, the Outcome Regression, a generalization of the approach in Heckman et al. (1997), requires the evolution of the response variable of the comparison group to be modeled correctly to ensure conditional parallel trends. The second approach, building on inverse probability weighting (IPW) proposed by Abadie (2005), requires correct estimation of the conditional probability of being in group g.

The doubly-robust method, extending the simpler DiD estimator from Sant'Anna and Zhao (2020) to allow for multiple *group-times*, is particularly appealing because it can be seen as a combination of both approaches, modeling the response variable evolution as well as the propensity score, but it only requires one to be correctly specified. Thus the term "doubly robust".

For a more detailed view on the proposed estimator, see Appendix A.1. A closer look at Equation (A.2) shows that the treated group are not re-weighted (or all receive unit weights). Further, the subscripts in $\widehat{ATT}_{dr}(g,t)$, in Equation (A.1), hints that this procedure is conducted for each one of the 2-by-2 group-time ATT, which entails important implications: the same comparison units, those that are never treated in this case, are used for all group-time estimations but receive different weights in each *step*.

The implication is that it is not trivial to extract and store the weights from the IPW procedure.[2] This means that while this procedure doubly robust and automatically "balances" treatment and comparison groups, the researcher cannot access the same weights to double check if the estimator is doing a good job at that. This is a considerable drawback, because it reduces the ability of the research to better explore some of the identifying assumptions of causality.

2.4 Key Identifying Assumption

As in conventional DiD settings, the key identifying assumption is of common trends. That is, the response variable of the treated units would follow the same trend of that from comparison units also in the absence of treatment. If the assumption holds, we have a credible counterfactual with which we can estimate causal treatment effects. Since this assumption is not testable, we refer to pre-treatment trends to assess its plausibility. If, prior to treatment, the eventually treated and comparison units follow a similar trend, it is reasonable to infer that the trend would have remained parallel after treatment time also

[2]The main author of the `csdid` package for Stata (Rios-Avila, Sant'Anna et al., 2023) confirms this in the Statalist forum (Rios-Avila, 2023). One could extract the weights from each treatment time cohort, and use them for further checks. In a setting with more than a few cohorts, however, this quickly becomes infeasible.

in the *counterfactual world* where the in the treated units were not treated.

It is also important to note that while pre-treatment balance can boost the plausibility of the parallel trends assumption, balanced groups are not actually a requirement. This strategy relies on the assumption that solely the trends are parallel, that is, the evolution of the response variable are the same across groups, regardless of starting values.

The pre-treatment periods can also be useful to assess the presence of reverse-causality. In this setting, reverse causality would imply that wealth outcomes (or lack thereof) in early periods actually causing the steeper health degradation, which ultimately causes treatment assignment eventually. Given the nature of these two variables at hand, how they might, a priori, interact and slowly evolve over the life time, this is a serious reason of concern.

3 Data

In this section, I first describe the data source and highlight the focal variables. Subsequently, I outline the working sample and present descriptive statistics, focusing on the group differences across treatment status.

3.1 The Socio-Economic Panel

This analysis is based on the Socio-Economic Panel (SOEP), a survey of individuals and households in Germany conducted yearly since 1984 (Goebel et al., 2019, Liebig et al., 2022). The two main components for the current study are the modules screening health indicators and the one wealth at the individual level.

Health variables In 2002, the health module was revised and since then it has been conducted every second year (Grabka, 2022). The items in this module target whether the respondent's daily activities are impaired due to physical or mental health related issues, as well as a subjective assessment on their current health. An overview on the questionnaire is presente in Table A.1. This resource also provides detailed information on the resulting variables and conducted transformations before using them in the factor model. In addition, the distribution of each variable is also presented in Figure A.2.

From these variables, I compute alternative Physical and Mental Component Summary scores, PCS and MCS, respectively. An in depth explanation on these constructs is presented in Chapter 4. These summary scores are the key variables in this work. They are used to assess the individual's health trajectory and to determine their treatment status.

Wealth variables The wealth module is also available since 2002, but conducted less frequently, namely in 2002, 2005, 2012, 2017 and 2019. The focal variables are overall gross wealth and overall net wealth as covered in the *personal assets and liabilities* section of the individual questionnaire. It covers real state assets and debts, savings and investments, building loan contracts, financial assets, tangible assets, and consumer debt. In 2017, two categories covering automotive and educational loans were added. These items, however, are not included in this analysis in order to maintain a consistent scope over the whole period. Furthermore, the wealth measures were deflated to 2020 values by the CPI index provided by Germany's Federal Statistical Office (n.d.).

Since the wealth and health modules are conducted in different years, I apply a linear interpolation to the wealth variables in order have wealth measures matching the frequency

of the health module. To illustrate, if a participant has valid wealth information in, say, years 2002, 2007 and 2012, I assume that any variation in wealth behaves linearly within the sub-ranges. The years with valid information can be considered fixed *knots* in this exercise. With that, I impute values for 2004 and 2006, as well as 2008 and 2010, such that it varies linearly between 2002–2007 and, likely with a different change rate, between 2007–2012. I refrain from doing extrapolations, with the exception of 2020, in which year the values are taken from 2019. The assumption is that wealth measures do not vary too drastically within one year. That implies that people with only one completed wealth module are not included in this analysis. In conclusion, the resulting data comprise 10 waves from 2002 to 2020 with a two-year frequency.

Further variables of interest In secondary analyses, I use variables related to the labor market, such as annual labor earnings, current employment status, as well as full-time experience and unemployment experience since entering the labor market. These variables are used to evaluate possible wealth accumulation channels.

Furthermore, variables on subjective well-being are also used in secondary analyses. These include satisfaction with health, work, income, and life overall.

In addtition, variables covering concrete health diagnoses that can be clearly mapped onto physical or mental health are also employed. These assess whether the respondent has ever been diagnosed with back pain, heart conditions, depression, or sleep disorder.

Note that these items are only asked from 2009 onward, so the assumption is made that, prior to 2009, people were not diagnosed with any of these health conditions. This is arguably a strong assumption but is somewhat remedied by the question formulation, which asks if the respondent has "ever" been diagnosed with X. In a more detailed analysis, however, those that stated yes should have been dropped from the analysis, since the timing would be unclear. This was not done here, however, with the implicit assumption that the diagnosis happened on that very year.

In addition, personal characteristics such as age, gender, marital status, federal state of residence, and years of education are used as covariates in the main analyses. These variables are employed in the DiD_{DR} framework to compute the inverse probabilities of treatment. The workings of this procedure are explored in more detail in Chapter 2. Regarding age, a restricted cubic spline with five equally spaced knots is employed to better capture nonlinearities in the health and wealth trajectories over the life cycle. To conclude, the covariates used in the main estimation specification are gender, age spline, federal state residence, legal disability, marital status, and years of education.

3.2 Final Working Sample

From the health module, the data consist of 240,770 observations with valid PCS and MCS scores. Out of these, 167,567 can be merged with valid observations from the wealth modules. After restricting to those with valid information on the variables of interest (described above), there remain 163,938 observations.

Finally, I restrict the sample to adults aged between 18 and 75. The age restriction is done so that the models can still capture the effect of an adverse event happening around the retirement age, and how it evolves into the early retirement years. Those older than 75 years, however, it is likely to display a distinct (dis)accumulation pattern, which would likely be better captured by separate evaluations. Finally, after the merging process, it is required by the DiD framework that each individual has valid data in at least two periods. Those with only one valid observation dropped from the analysis.

With that, the final sample consists of 141,337 observations. A few considerations are necessary, though. First, the always-treated units are dropped from the estimation. Those account for 7,866 observations in the physical and 8,395 in the mental domain. Further, each of the 2 by 2 DiD calculations is performed with balanced observations for that specific calculation. More details are presented in Chapter 2, but as a brief illustration, in order to compute a single ATT, say, in 2010 for the group treated in 2004, only those individuals present in 2004 and in 2010 can be used. The final number of unique and total observations ultimately used in the estimations are presented in the table of results for each specification.

3.2.1 Descriptive Statistics

Table 3.1 presents statistics comparing the eventually-treated and never-treated groups in the physical and mental domains. In both cases, the data refer to the information from the first survey year of each respondent.

When comparing the physical and mental domains, one can observe that the groups exhibit greater similarity in the mental health domain. For instance, the difference in starting values of net wealth in the mental domain is only around 13 thousand, whereas in the physical domain, it differs by about 27 thousand. Similarly, the number of months in unemployment differs by around 2.5 in the physical domain, whereas in the mental domain, it only differs by around 1.5.

Interestingly, the MCS value in the mental health domain, which is the variable used in the treatment assignment rule, is one of the few that does not show a statistically significant difference between the eventually-treated and never-treated groups. This indicates that when first entering the SOEP, the eventually-treated individuals are not significantly dissimilar from the never-treated group to start with. This might be one reason why, as later shown, the models in the mental health domain are more supportive of the parallel trends assumption. It is worth emphasizing that the data correspond to that apt to be

Table 3.1 Descriptive statistics by treatment group for both health domains

	Physical Domain			Mental Domain		
	Eventually Treated (PCS)			Eventually Treated (MCS)		
	no N: 13,786 (80.6%)	yes N: 3,316 (19.4%)	Test	no N: 12,933 (78.5%)	yes N: 3,545 (21.5%)	Test
Age	43.88 (14.39)	43.24 (12.97)	0.018	44.59 (14.42)	43.46 (13.47)	<0.001
Years of Education	12.68 (2.79)	12.13 (2.50)	<0.001	12.44 (2.73)	12.28 (2.59)	0.001
Physical Health (oblique)	55.29 (5.84)	54.66 (4.51)	<0.001	53.35 (8.13)	51.57 (8.77)	<0.001
Mental Health (oblique)	52.51 (7.96)	50.37 (8.72)	<0.001	54.63 (6.24)	54.79 (3.72)	0.142
Gross Wealth (log)	3.46 (2.19)	3.14 (2.17)	<0.001	3.42 (2.19)	3.30 (2.17)	0.004
Net Wealth (neglog)	3.05 (2.46)	2.63 (2.54)	<0.001	2.99 (2.49)	2.87 (2.44)	0.018
Gross Wealth (k€, winsored)	144.57 (233.24)	114.35 (195.83)	<0.001	139.76 (228.88)	125.85 (206.47)	0.001
Net Wealth (k€, winsored)	117.19 (204.92)	90.63 (172.95)	<0.001	113.98 (202.34)	100.13 (179.60)	<0.001
Unemployment exp. (months)	7.46 (21.96)	10.01 (24.42)	<0.001	8.15 (23.64)	9.70 (25.66)	<0.001
Full-Time exp. (months)	182.89 (157.53)	186.90 (149.14)	0.183	189.78 (159.47)	186.62 (152.77)	0.291
Gender						
Male	6,573 (47.7%)	1,518 (45.8%)	0.049	6,178 (47.8%)	1,623 (45.8%)	0.036
Female	7,213 (52.3%)	1,798 (54.2%)		6,755 (52.2%)	1,922 (54.2%)	
Marital Status						
Married	8,683 (63.0%)	2,099 (63.3%)	0.007	8,380 (64.8%)	2,264 (63.9%)	0.245
Single	3,332 (24.2%)	747 (22.5%)		2,890 (22.3%)	840 (23.7%)	
Widowed	418 (3.0%)	90 (2.7%)		417 (3.2%)	95 (2.7%)	
Divorced	1,047 (7.6%)	309 (9.3%)		994 (7.7%)	273 (7.7%)	
Separated	306 (2.2%)	71 (2.1%)		252 (1.9%)	73 (2.1%)	
Education Attainment						
Less than High School	1,582 (11.5%)	414 (12.5%)	<0.001	1,609 (12.4%)	434 (12.2%)	0.02
High School	8,114 (58.9%)	2,147 (64.7%)		7,885 (61.0%)	2,246 (63.4%)	
More than High School	4,082 (29.6%)	755 (22.8%)		3,432 (26.6%)	863 (24.4%)	
Employment Status						
Not Employed	3,445 (25.0%)	699 (21.1%)	<0.001	3,403 (26.3%)	854 (24.1%)	0.007
Employed	10,341 (75.0%)	2,617 (78.9%)		9,530 (73.7%)	2,691 (75.9%)	

Notes: Number unique observations and proportion (in parenthesis) by treatment group are indicated at the table's header. Those refer to unique individuals present in the main model, after dropping always-treated and dealing with missing values of covariates. The statistics refer to the first survey year of each individual. Continuous variables are summarized by their mean and, in parenthesis, standard deviation. Categorical variables are summarized by count and, in parenthesis, group percentage. The fourth column of each domain shows the p-values of pooled t-tests for continuous variables and Pearson's χ^2 tests for categorical variables.

using the estimation procedure. This means that the always-treated units are discarded. If those would be considered, the differences would be much larger, specially at the respective first survey year.

With a large number of observations as in this case, even small deviations from the mean across subgroups would render them to be statistically different to a high confidence level. To assess how the groups differ over the entire distribution, a univariate kernel density estimation, is presented in Figure 3.1. Restricting the data to the first survey year of each participant, we observe that the groups, divided by never- and eventually-treated, are not too dissimilar, as the test statistics in Table 3.1 would suggest. Focusing on the selected variables Age, Gross, and Net Wealth, we see well-overlapping support in both health domains, while in the mental case, the distributions among treated and untreated are more similar.

In this context, I also conduct a logistic regression of *eventually-treated* on the same covariates as in the main analysis. This aims to estimate the probability of being (eventually) treated, conditioned on gender, age spline, federal state residence, legal disability, marital status, and years of education. The goal is to better asses if there is enough overlap in the key variables used in the estimations.

A Common overlap is one of the assumptions of the doubly-robust specification of Callaway and Sant'Anna (2021, see Assumption 6, "Overlap condition"). The results, as shown in Figures 3.1 A.a and 3.1 B.a, corroborate this assumption. Over the whole probability range (0–1) both groups either display considerable density or, at the extremes, is very thinly populated. There are no areas where the treatment probability of one group is densely estimated and the other is not. In the mental domain, again, the densities are more similar to one another when compared to the results from the physical domain.

Note that the resulting estimated probabilities, or their inverse, are not the same set of weights used in the DID_{DR} framework. In this exercise, only a single probability of eventual treatment is estimated. In the DiD_{DR}, in contrast, the estimation targets the probability of being first treated at time g. With that, several values for each individual are computed and used in each of respective 2x2 blocks. This exercise, however, serves to enhance the credibility that the overlap condition would also be met within the estimation of each block.

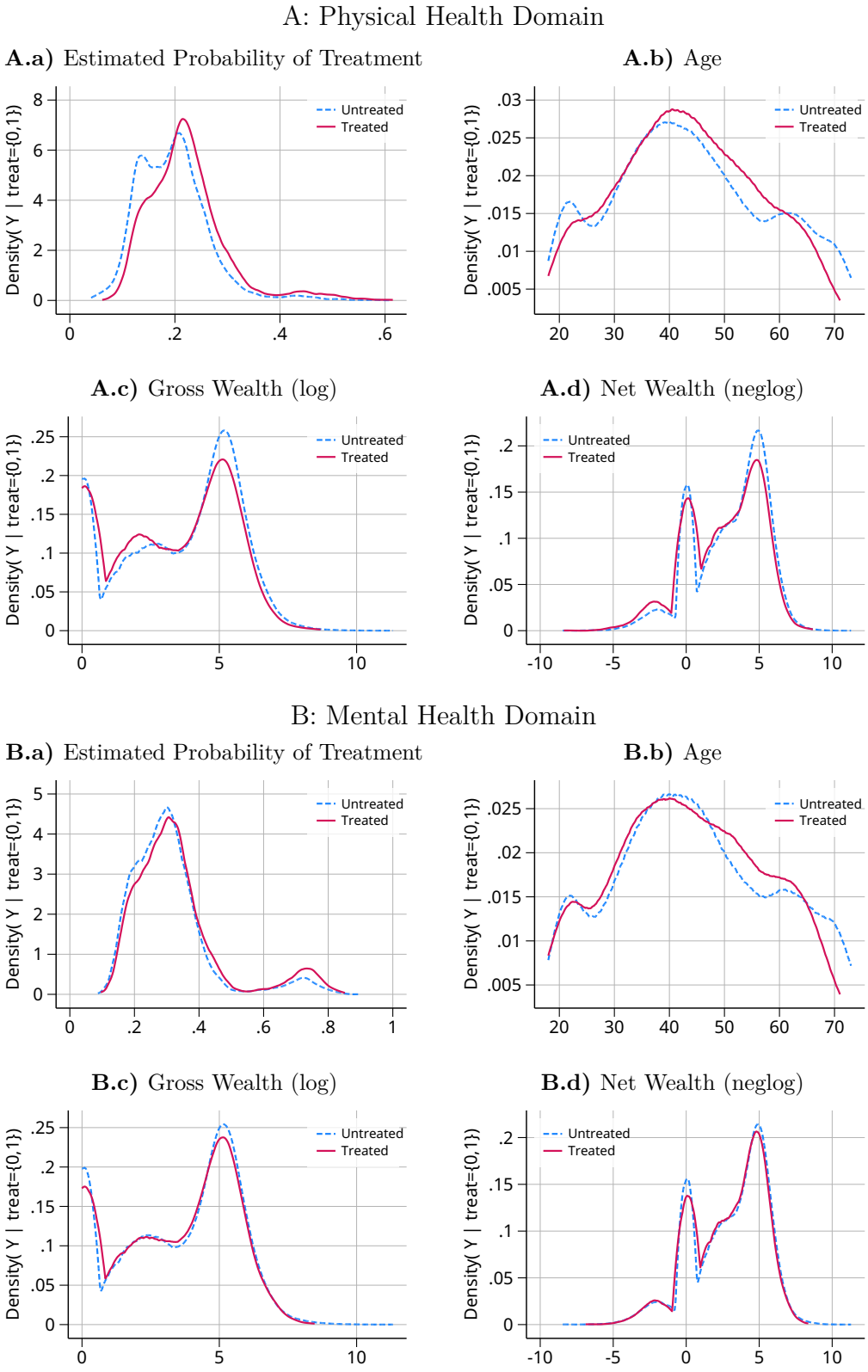


Figure 3.1 Univariate kernel density estimation by eventual treatment groups

Notes: The probability of being eventually-treated (panels A.a and B.a) was estimated via a logistic regression on the same covariates from the main DiD_{DR} models, with eventually-treated as the response variable. The distribution of all variables shown above indicate a good overlap across treatment status, with the mental domain displaying a slightly better overlap. Data restricted to the first survey year of each participant.

4 Measuring Health

Since 2002, as part of the health module in the individual questionnaire, SOEP includes questions modeled on the 12-item Short Form Health Survey version 2 (hereafter SF-12). Based on the SF-12 methodology, SOEP provides the Physical (PCS) and the Mental Component Summary (MCS) scores, as described by Andersen et al. (2007, p. 176).

Briefly, the SF-12 methodology, which is a shorter version of SF-36, consists of formulating twelve questions, from which eight sub-scales over different aspects on health-related quality of life are computed. Four sub-scales refer to the physical and four to the mental health domain. Each domain is summarized by a single score, the Physical (PCS) and the Mental Component Summary (MCS). In this analysis, I follow the same concept and use the same input variables, but recreate the summary scores with two key distinctions from the methodology proposed by Ware et al. (2002) and described by Andersen et al. (2007) specifically to the SOEP case.

First, instead of calculating the mean of each sub-scale and then applying a factor analysis, or Principal Component Analysis (PCA) to extract the scores, I input all twelve variables in the factor model.[1] This allows for a finer grained exploration of some properties of the model. For example, one can check how each item behave and whether items from the same concept are clustered together.

Second, I apply an oblique rotation after estimating the common factor model instead of a orthogonal one. The main reason being that the orthogonal rotation implies uncorrelated factors. This, in turn, entails mental and physical health to be uncorrelated. Without delving too deep into the details, the main idea of applying a rotation, after conducting the PCA, and, in this case, after reducing the data to two dimensions, is to *simplify* its structure by rotating the axes so that, at best, each axis can be mapped onto a single health domain. This procedure becomes clear by visually comparing the figures before rotation, as presented in Figure 4.1 a to those after rotation, shown in Figures 4.1 b and 4.1 c. Ideally, the variables should be loaded strongly onto one factor and weakly onto the other, so that a clear separation emerges and each axis can be clearly mapped onto a health domain. In the orthogonal rotation, however, the axes must remain at a 90° angle. In contrast, an oblique rotation allows for correlated factors. Visually, one can picture the axes closing or opening like a pair of scissors to better accommodate the loadings.

In addition, regarding the construction of sub-scales constructed from two input variables,

[1] The terminology in this field is a "minefield", as Nick Cox (2005) describes in a Statalist post. In this analysis, I call the the SF-12 method a *PCA followed by a varimax rotation* and the alternative method a *Common Factor Analysis followed by an oblique rotation*, following Fabrigar and Wegener, 2012.

instead of list-wise deletion, I impute the missing item with the same value from the corresponding item within the same sub-scale. Since the measures within sub-scales are very highly correlated, doing so poses no credible risk of bias, while retaining most of the available data. For an overview on the scales, input variables and questionnaire formulation, see Table A.1.

The implication of employing a PCA followed by an orthogonal rotation is that the generated scores are uncorrelated. In our setting, this means one would expect mental and physical health to be uncorrelated or, at least, that the PCA can extract a portion of the variation of those measures that are uncorrelated with one another. It has been argued, however, that this is not the case. Widaman (1993) affirms that PCA "should not be used ... to obtain parameters reflecting latent constructs or factors". Other authors also suggests that PCA should be seen solely as dimensionality reduction method. Furthermore, Fabrigar and Wegener (2012, p. 31) state that PCA do not correspond to "meaningful latent constructs" but rather "represent efficient methods of capturing information in the measured variables". Also regarding health, and specifically the SF-12 or SF36 methodology, several authors have raised criticism and proposed alternatives (see Wilson et al. (2000), Tucker et al. (2013), Hann and Reeves (2008) and Hagell et al. (2017)). While most of these works focus on the longer SF-36 version, I show that the results replicate in the SF-12 case. On that account, I estimate alternative summary scores after an oblique rotation and compare against the the SF-12 method.

4.1 Evaluation of PCS and MCS scores

Crucial to this endeavor is that the input variables of these models capture the information that they are targeting. This can be confirmed by looking at the factor loadings depicted in Figures 4.1 b and 4.1 c. We can see two well-defined clusters, which are separated by the physical and mental health domain. That is, those in the sub-scale of General Health (GH), Physical Function (PF), Bodily Pain (BP) and Role Physical (RP) are strongly loaded onto factor one, and weakly onto factor two. The opposite is true for Mental Health (MH), Vitality (VT), Role Emotional (RE) and Social Function (SF). With that, we can confidently characterize factor one as the physical domain and factor two as the mental one. Furthermore, in the alternative method (panel c), we see that each individual variable belonging to the same domain are very close to one another. This indicates that those questions on the same sub-scale are capturing similar concepts. We can also see that the alternative method, being more flexible, allows for a *simpler* structure. That is, the variables are strongly related to one of the two factors and more weakly related to the other. Finally, vitality is located differently in the alternative method. It is not strongly related to any of the factors, while in the SF-12 method, it is relatively strongly loaded on factor two. Maybe this reflects that the underlying item is not capturing very well what

is being targeted. The wording of the item asks if the respondent feels "energetic". It can be unclear if one should, a priori, expect it to be in the physical or the mental sense. In conclusion, both methods are able to capture a similar pattern and well discriminate the physical from the mental domain.

Turning our attention to Table 4.1, we can see the values of the factor loadings in the first two columns as well as the *Uniqueness* of each variable in the third column. The factor loadings are the same as depicted in Figure 4.1. The uniqueness tells us the proportion of the variance that is unique to that variable in the factor model. It is the opposite of *Communality*, another commonly reported statistic, where Uniqueness = (1 − Communality), with that being the portion of the variance that is shared in the factor model.

Finally, in the last two columns, we can see the score coefficients obtained with both methods. These are the coefficients that, multiplied with the health items, generate the physical and mental health summary scales, as in

$$pcs_i = \boldsymbol{\lambda}_1 \boldsymbol{h}_i \qquad \text{and} \qquad mcs_i = \boldsymbol{\lambda}_2 \boldsymbol{h}_i \tag{4.1}$$

where the $\boldsymbol{\lambda}_{\{1;2\}}$ are vectors of score coefficients relative to factors 1 and 2, while $\boldsymbol{h}_i$ represent the vector of health items from individual i. The same applies to the SF-12 method, but taking into consideration the intermediate step of computing the subscales from domains consisting of two items and then applying the procedure as above, thus with eight items in each vector instead of twelve. The coefficients respective of each health item in the alternative method and relative to the health sub-scales in the SF-12 method are presented in in Table 4.1.

At this point it is worth taking a look at each score coefficient and compare the two methods. Here we can see the culprit of the *agreement problem*, pointed out in Tucker et al. (2013). Namely, the presence of negative score coefficients imply that a high value in the input variable will be strongly and negatively correlated with the respective summary score relative to that factor. To illustrate, take the coefficients respective of Physical Function (Table 4.1b). They amount to 0.4 in the first and nearly −0.2 in the second factor. Interpreting the first factor as the physical domain and the second as the mental, the above implies that, ceteris paribus, a one unit increase in the physical function of an individual implies a 0.4 units increase in their PCS and, concurrently, 0.2 units reduction in their MCS score. The ratio of both effects is one half, but in opposing directions. The same and in roughly the same magnitude, but in reverse, applies to the coefficients of the Mental Health sub-scale. Further, we can see that all score coefficients are positively correlated to one factor and moderately to strongly correlated to the other factor.

Looking at the coefficients obtained from the alternative method, we see that fewer are negative. Further, they are considerable smaller in magnitude. So, while the oblique rotation does not completely fix the agreement problem, it ameliorates it substantially. The

implications of this issue can be analyzed in Figure 4.2, where the computed PCS and MCS are plotted following the SF-12 and the alternative method.

The SF-12 method results in scores with, in some sense, nicer statistical properties: The distribution—specially looking at each dimension separately—is considerably less skewed. In contrast, in the alternative method, the data is more heavily concentrated in the upper-right corner. Those are the people with good physical and mental health. Note that in both methods, the data are shifted and scaled to achieve a mean of 50 and standard deviation of 10. Other moments, however, differ considerably over the two methods.

One effect of the agreement problem can be observed here as well. Figure 4.2 a, referring to the SF-12 method, shows that there are no points in the lower-left and upper-right corners. Although the scores range from 0 to 80 in each dimension, we do not observe people with score of over 60 in both dimensions concurrently. Likewise in the lower-left corner, there are no observations with very low scores in both dimensions. Further, the healthier an individual seems to get in one dimension, the sicker they appear in the other dimension, as indicated by the data points in the upper-left (high MCS and low PCS) and lower-right (low MCS and high PCS) corners. This seems to be, as Tucker et al. (2013) argue, a "mathematical artifact" and might not reflect the true relationship between physical and mental health.

In the alternative method, as shown in Figure 4.2 b, the data points are more compact in a square shape, although still slightly tilted, indicating that the artifact is still present, although restricted to the very end of the distributions. For a better grasp on the sub-scales and how they behave in relation to the summary scores scales, see Figure A.1, where I present the development each variable over age.

It is worth noting that the SF-12 method is a common approach in the literature and that the scores have been validated in several settings (see Gill et al., 2007; Vilagut et al., 2013; Christensen et al., 2013). The issues discussed might be confined to the tails of the distribution and, on the whole, the scores still captures valid information. Having said that, since stark variation in health scores is central to this analysis and the extremes of the distributions are key areas of interest, I opt to use the alternative as the main method for the remainder of this work.

Table 4.1 Factor loadings and score coefficients from alternative and sf-12 method

a) Alternative method

	Rotated factor loadings			Score coefficients	
	Factor 1	Factor 2	Uniqueness	Factor 1	Factor 2
General Health	**0.647**	0.195	0.462	0.132	0.020
Physical Function 1	**0.761**	0.007	0.417	0.206	−0.069
Physical Function 2	**0.800**	0.005	0.358	0.245	−0.086
Bodily Pain	**0.645**	0.198	0.462	0.108	0.014
Role Physical 1	**0.723**	0.273	0.275	0.225	0.027
Role Physical 2	**0.755**	0.252	0.244	0.285	0.002
Mental Health 1	0.099	**0.608**	0.582	−0.014	0.141
Mental Health 2	0.020	**0.501**	0.742	−0.015	0.133
Vitality	0.284	**0.340**	0.741	0.023	0.084
Role Emotional 1	0.132	**0.823**	0.235	−0.067	0.377
Role Emotional 2	0.120	**0.808**	0.271	−0.058	0.314
Social Function	0.251	**0.643**	0.419	0.004	0.139

Notes (a): Factor loadings from a Common Factor model followed by an oblique rotation with promax(1.6). Promax value chosen with the intent to allow for reasonably correlated factors (around .5) while maintaining any negative score coefficients lower than 0.1. Score coefficients obtained via regression method. Bold digits mark loadings bigger than 0.3.

b) SF-12 method

	Rotated factor loadings			Score coefficients	
	Factor 1	Factor 2	Uniqueness	Factor 1	Factor 2
General Health	**0.768**	0.295	0.323	0.314	−0.080
Physical Function	**0.845**	0.144	0.265	0.408	−0.199
Bodily Pain	**0.786**	0.254	0.317	0.338	−0.111
Role Physical	**0.789**	0.377	0.236	0.298	−0.036
Mental Health	0.090	**0.829**	0.305	−0.234	0.482
Vitality	0.282	**0.571**	0.594	−0.042	0.257
Role Emotional	0.297	**0.782**	0.301	−0.106	0.382
Social Function	0.345	**0.737**	0.338	−0.065	0.338

Notes: Factor loadings from a Principal Components model followed by a varimax rotation with Kaiser normalization. Some score coefficients are relatively high, such as those from Physical Function and Mental Health. This implies, for example, that a positive value in the Physical Function strongly affects Factor 1 (physical), but also impacts Factor 2 about half as strongly, but in the negative direction. Analogously, but in reversed directions, the same happens in the Mental Health case. Score coefficients obtained via regression method. Bold digits mark loadings bigger than 0.3.

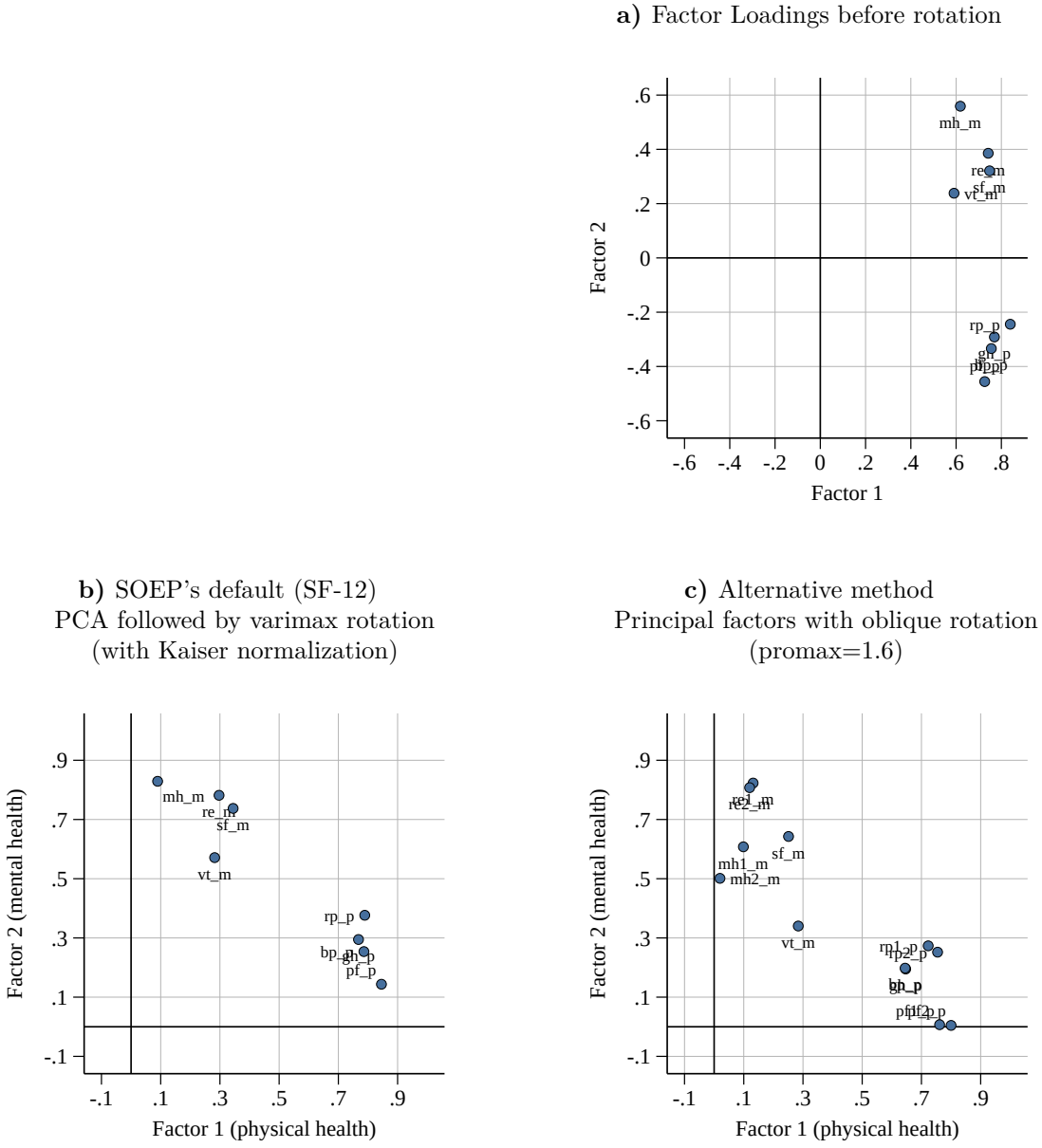
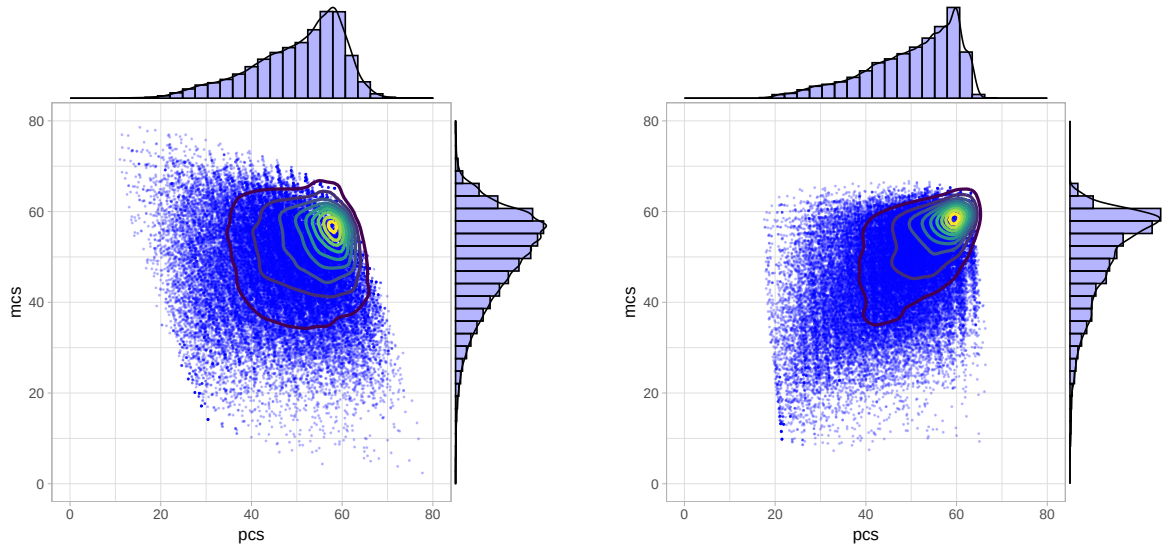


Figure 4.1 Factor loadings comparison between SF-12 and alternative methods

Notes: The graphs show in both versions that the variables are meaningfully clustered in their respective group and the clusters are well separated. However, allowing for an oblique rotation results in a clearer factor separation. The loadings of each variable are closer to one of the axes, indicating that a variable strongly affects one but not the other factor. Further, by using all variables instead of the mean of grouped variables in the alternative model we can confirm that variables from the same category are indeed closely clustered. One drawback with the alternative model is that `vt` (Vitality) is only weakly loaded from both factors. Variables with suffix m belong to the mental domain, whereas those with p to the physical one. For an overview on the variables see Table A.1.



a) SF-12 methodology

b) Alternative methodology

Figure 4.2 Bivariate density estimation of mental and physical health scores

Note: This figure plots the health scores of each individual and shows a bi-variate density on the canvas and a histogram of each domain on the opposite axis. Each point depicts the pcs (x-axis) and mcs (y-axis) for each individual. In panel a, the SF-12 methodology is shown and, in b, an alternative procedure with oblique rotation. In both methods scores are shifted and scaled to have a mean of 50 and a standard deviation of 10.

5 Empirical Methodology and Results

In this section, revisiting the theoretical groundwork explored in Chapter 2, I present the empirical methodology applied to the current setting.

5.1 Event Study Framework

In the main analysis, the aim is to estimate the dynamics of wealth accumulation which can be attributed to a significant health degradation in the physical and mental domain. That is, how does experiencing a negative health outcome affects wealth accumulation over subsequent years in a relatively long term. With this in mind, the parameter of interest is[1]

$$\theta_{\text{es}}(e) = \sum_{g \in \mathcal{G}} \mathbb{1}\big[g + e \leq \mathcal{T}\big] P(G = g | G + e \leq \mathcal{T}) ATT(g, g + e), \qquad \text{(2.2 revisited)}$$

where $e = t - g$, $t \in \{2002, 2004, ..., 2020\}$, $g \in \mathcal{G}$, where $\mathcal{G} = \{2004, 2006, ..., 2020\}$ and $\mathcal{T} = 2020$. Note that the first item of $\mathcal{G}$ is 2004, because always-treated units are dropped from the analysis. Additionally, people assigned to treatment in their first survey year are also not considered in the analysis.

The relative time window is restricted to $e = \{-10, -8, ..., 10, 12\}$ years, as estimates outside this range become very imprecise. The time of treatment, i.e., when an individual first experiences a negative health outcome happens between $e = -2$ and $e = 0$, but is only becomes observable in the data when $e = 0$. The base or reference period, is set to $e = -2$, which is the last period before treatment.

5.1.1 Treatment Assignment Rule

Everything in what follows is conducted twice: once in relation to the physical domain and once in relation to the mental domain. For brevity, however, I subsequently describe the rule using a general term *health outcome*.

An individual i is assigned to group $g = t$ if they experience a relatively severe adverse health outcome compared to people of the same age and gender, as measured by the respective component summary score (PCS for the physical and MCS for the mental domain). Specifically, if the individual's summary score is half a standard deviation worse than the median score of people of the same age and gender, they become a candidate for treatment

[1] This might be an unfamiliar presentation on the event study framework, for an analogous version based on a regression-based design, refer to Appendix A.2.

assignment.[2] If this individual experiences a negative health outcome at least once more (though not necessarily consecutively), they are indeed assigned to treatment in the period when they first experience the adverse health outcome. This decision has two primary reasons: first, to prevent treatment assignment due to a small fluctuation around the threshold, and second, by employing a less strict assignment rule, approximately half of the sample would eventually be assigned to treatment. Thus, by restricting to those experiencing it twice, the analysis focuses on individuals with a more severe health deterioration. Further, the age–gender adjustment, instead of using the simple difference was done to account for gender–age specific reporting heterogeneity, as indicated by Ziebarth (2010).

The threshold of one-half standard deviation aligns with the findings of Vilagut et al. (2013, p. 568), who identify 45.6 as the optimal cutoff point for MCS_{sf12} to evaluate 30-day depressive disorders. Given that our health components have a mean of 50 and a standard deviation of 10, half a standard deviation is approximately equal to the identified cutoff point. On the other hand, the comparison (or control) group consists of people with more stable health trajectories, experiencing a health outcome worse than the threshold at most once.

Note that, for clarity, I also use the term "shock" throughout this work to denote the adverse event. However, it must be emphasized that not all treatment assignments occur due to a drastic score decrease from one period to the next for a given person. A continuous but slow degradation of their health also characterizes treatment. Further, also regarding terminology, I use conventional terms in the causal inference literature borrowed from the medical domain and since this analysis deals with health outcomes, this can lead to confusion. To clarify, "treatment assignment" here simply means that the individual experiences a bad health outcome, without taking into account any medical treatments, in the usual common sense of the world.

5.1.2 Wealth Measures

For this analysis, I use four measures of wealth: gross and net wealth, as well as these measured in levels and in logs. The variables in levels are winsored at the 1st and 99th percentile. In the case of net wealth, the *neglog* transformation was applied, where

$$W^{\text{neglog}} = \text{sign}(W) \cdot \log\left(1 + |W|\right). \tag{5.1}$$

The neglog transformation possesses several interesting features, beyond its primary goal of reducing the skewness of the input data. Notably, it retains the value of zero due to the sign function. Moreover, for large positive values of W, it behaves as $\log(W)$, and for large negative values, it behaves like $-\log(W)$. For values around 0, it approximates W linearly. The same transformation is applied to gross wealth, but as it is non-negative, the

[2] Alternative threshold and rules are tested in robustness checks.

transformation simplifies to $\log(1+W)$. Additionally, winsoring the variables at the 1st and 99th percentiles has been shown to enhance the stability of the models.

On the parallel trends assumption Crucial to effect identification is the assumption of Parallel Trends (PT). Dealing with health accumulation over time, it's essential to critically consider what this assumption entails. In principle, one could question whether trends are more likely to be parallel when dealing with the wealth variable in absolute or relative terms. This boils down to whether an additive or a multiplicative wealth accumulation process better models the wealth trajectory over the life cycle. Hence, I consider wealth in levels and in logs, but a priori, find it plausible that for different starting values of wealth, the multiplicative process is likely to capture the accumulation path of different groups better. Concretely, it seems more credible that richer and poorer people can accumulate a similar percentage of their current wealth, rather than expecting their wealth to increase by values in absolute term.

On wealth measure and transformation choice Both net and gross measures offer compelling reasons for being the key variable of interest. Net wealth might better capture effective changes in the wealth trajectory. For instance, if an individual borrows money to buy a house, this would increase gross wealth by the value of the house but leave net wealth intact until they start paying off the loan. Moreover, experiencing negative net wealth might be an intriguing aspect in this analysis, given its potential linkages with stress and anxiety.

On the other hand, the data reveal a considerable number of people with large negative net wealth. Consider an individual who takes a substantial loan to open a business or another who incurs student loans. If the value of their assets when opening the business or by the end of their studies is considerably smaller than the value of their acquired debt, they would be located at the very low end of the net wealth distribution. However, one would hardly argue that they are the poorest people in the population. On this account, I focus on gross wealth as the main measure to capture the socio-economic status of the individual.

Furthermore, due to the crossing over 0, the interpretation of coefficients from the neglog-transformed models is not trivial and cannot be simply taken as percentage or proportional effects. While the neglog transformation might be interesting for modeling and technical reasons, interpretation and economic implications based on those results are harder to convey. With this reasoning, I consider the model with log-transformed gross wealth to be the main model in this analysis and focus on those results. However, for completeness, all results are presented.

5.2 Results

In this section, I first present the results of the estimated event-study coefficients $\widehat{\theta}_{es}(e)$ for each of the eight main models. These models are from the physical and mental domains, encompassing specifications in levels and logarithms, and employing gross and net wealth as the response variable. Subsequently, I present the results of a secondary analysis aiming to explore possible mechanisms through which wealth accumulation is affected due to physical or mental health degradation.

In all the following figures, the panels depict the resulting coefficient estimates for two models simultaneously, facilitating the comparison of the same model across different health dimensions. Thus, the single lines do not depict a treated vs. non-treated development but directly represent the ATT coefficients for the model in the physical and mental domains.

It's important to note that, for easier interpretation, a transformation was applied to the coefficients in the (neg)log models so that they already represent the effect in percentage terms. More details on the transformation and a caveat on the interpretation of the neglog transformation are discussed in Appendix A.3.

5.2.1 Effects of Health Deterioration on Wealth Accumulation

In Figure 5.1 and the corresponding Table 5.1, I present the main results of this analysis. In broad terms, the results are similar across different wealth dimensions, net and gross, and transformations, whether in levels or in (negative) logarithms. The *SimpleATT*, as defined in Equation (2.3), is negative and statistically significant at least to the $\alpha = 0.05$ level in all but the physical domain with neglog-transformed net wealth (model P2).

The average pre-treatment trend (*Pre average*) is not statistically significantly different from 0 in all specification. In contrast, the average post-treatment (*Post average*) is statistically significant in all but P2. Most single pre-treatment effects are not statistically different from 0, apart from the last period before the reference period ($e = 4$). This suggests that the negative health outcome is not experienced completely unexpectedly. There is an indication of an anticipation one periods before crossing the threshold defined in the treatment assignment rule.

The χ^2 Pretrend test, which is a joint significance test with the null hypothesis (H_0) that all pre-treatment $ATT(g,t)'s$ are equal to a constant k is clearly rejected ($p \leq 0.001$) in level specifications from the physical domain (models P3 and P4) and on the verge of rejection ($p = 0.069$) in the neglog specification (P2). The pre-treatment average and aforementioned tests corroborate the parallel trends assumptions of the log-transformed gross wealth in both physical and mental health domains. Therefore, I give more weight on those specifications (P1 and M1) in further considerations.

Table 5.1 Main results: table of coefficients

	Physical Health				Mental Health			
	(neg)log		level		(neg)log		level	
	gross (%) (P1)	net (%)[1] (P2)	gross (P3)	net (P4)	gross (%) (M1)	net (%)[1] (M2)	gross (M3)	net (M4)
SimpleATT	−6.51**	−5.65	−6.96**	−6.55***	−8.34***	−12.93***	−6.79**	−6.50**
	(2.14)	(2.98)	(2.18)	(1.96)	(2.00)	(2.50)	(2.33)	(2.09)
Pre average	3.06	5.14	4.86	3.92	2.13	8.44	−1.70	0.31
	(3.81)	(5.47)	(3.85)	(3.44)	(3.51)	(5.41)	(4.17)	(3.52)
Post average	−8.44**	−5.86	−8.47**	−7.85**	−10.30***	−15.64***	−9.09**	−8.40***
	(2.56)	(3.72)	(3.00)	(2.65)	(2.41)	(2.88)	(2.83)	(2.55)
$\hat{\theta}_{es}(-10)$	1.33	3.30	5.79	3.85	4.68	10.78	−8.17	0.11
	(8.37)	(11.81)	(8.47)	(7.08)	(8.27)	(11.57)	(10.20)	(8.63)
$\hat{\theta}_{es}(-8)$	7.76	8.01	6.48	5.69	−1.48	7.71	−1.14	−0.16
	(6.28)	(8.17)	(5.73)	(4.91)	(5.52)	(8.18)	(6.18)	(5.33)
$\hat{\theta}_{es}(-6)$	0.41	4.32	2.51	1.97	1.74	8.93	−0.23	−1.36
	(3.36)	(5.31)	(3.59)	(3.05)	(3.50)	(4.79)	(3.51)	(3.36)
$\hat{\theta}_{es}(-4)$	2.90	4.99*	4.67***	4.18***	3.69*	6.39**	2.75*	2.65*
	(1.54)	(2.42)	(1.22)	(1.11)	(1.68)	(2.36)	(1.37)	(1.20)
$\hat{\theta}_{es}(0)$	−0.75	−3.26	−2.21*	−2.61***	−1.08	−3.33*	−1.83	−1.87*
	(1.14)	(1.79)	(0.91)	(0.78)	(1.20)	(1.46)	(0.98)	(0.87)
$\hat{\theta}_{es}(2)$	−1.78	−3.97	−4.61*	−5.08**	−3.40	−6.15*	−3.04	−3.20
	(1.90)	(2.98)	(1.83)	(1.63)	(1.85)	(2.36)	(1.89)	(1.64)
$\hat{\theta}_{es}(4)$	−7.10**	−7.21	−7.32**	−6.80**	−10.81***	−14.55***	−6.76*	−6.82**
	(2.64)	(3.65)	(2.55)	(2.29)	(2.38)	(3.03)	(2.64)	(2.47)
$\hat{\theta}_{es}(6)$	−9.86**	−9.76*	−9.07**	−8.04**	−11.39***	−18.61***	−8.04*	−8.26**
	(3.12)	(4.13)	(3.18)	(2.82)	(2.63)	(3.50)	(3.14)	(2.91)
$\hat{\theta}_{es}(8)$	−12.53***	−5.99	−10.45**	−8.28*	−15.35***	−21.85***	−9.99*	−9.76**
	(3.52)	(5.28)	(3.78)	(3.41)	(3.26)	(3.79)	(4.00)	(3.54)
$\hat{\theta}_{es}(10)$	−14.17**	−6.25	−12.95**	−11.15*	−13.16**	−19.27***	−12.73*	−11.12*
	(4.23)	(6.46)	(4.94)	(4.64)	(3.80)	(4.79)	(5.26)	(4.69)
$\hat{\theta}_{es}(12)$	−11.96*	−4.42	−12.66*	−13.03*	−15.80**	−23.62***	−21.26***	−17.74**
	(5.50)	(7.95)	(6.31)	(5.60)	(4.65)	(5.63)	(6.36)	(6.03)
N	90,207	90,207	90,207	90,207	84,925	84,925	84,925	84,925
Unique N	17,581	17,581	17,581	17,581	16,881	16,881	16,881	16,881
Pretrend χ^2 (df)	28.5 (22)	32.5 (22)	47.0 (22)	59.5 (22)	17.5 (22)	24.8 (22)	17.2 (22)	15.2 (22)
Pretrend p-value	0.161	0.069	0.001	0.000	0.735	0.308	0.753	0.852

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table depicts the event study coefficients $\theta_{\text{es}}(e)$ as in Equation (2.2). *SimpleATT* depicts the average of all ATT(g,t)'s weighted by group size (see Equation (2.3)). *Pre average* and *Post average* are the average pre- and post-treatment effects with equal weights for each e see Equation (2.4)). One can evaluate the pre-trends by looking at the pretrend χ^2 test, the coefficients of *Pre average* or at each ATT for $e < 0$. Standard error shown in parenthesis are estimated via multiplicative Wild Bootstrap with 999 replications and clustered at the individual level. Covariates used for doubly robust procedure was age spline, federal state, legal disability, marital status, gender, and years of education. Using never treated as control group to avoid bias due to anticipation effects.

[1]Coefficients (and standard errors) of (neg)log specifications are transformed to represent the effect in percentage terms, but such interpretation of the neglog transformation might be biased (see Appendix A.3)

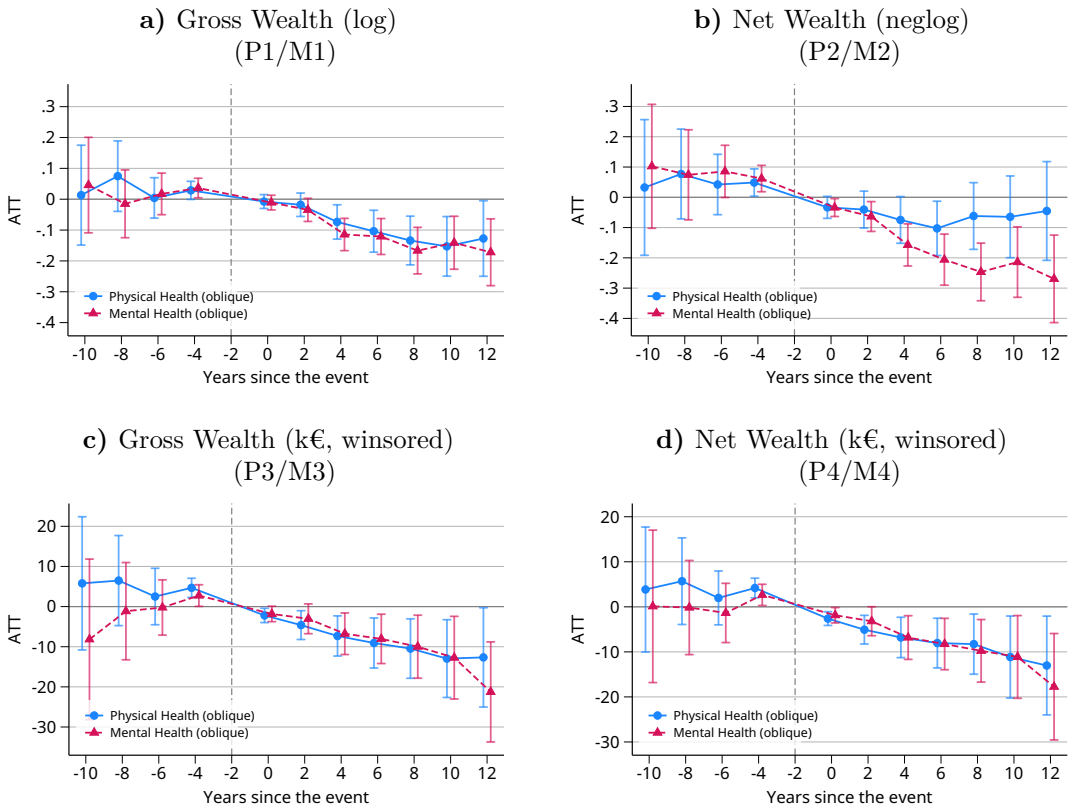


Figure 5.1 Main results

Notes: This figure illustrates the dynamic Average Treatment on the Treated (ATT) of physical health (in blue) and mental health (in red) deterioration on different measures and transformations of wealth. The time window is restricted to ten periods before and twelve after treatment due to a drastic decrease in statistical precision outside this range. The time of treatment is denoted as $e = 0$, and the reference period is set to $e = -2$, depicted with a vertical dashed line. Whiskers depict the 95% confidence interval based on Wild Bootstrap standard errors clustered at the individual level. This panel is a visual representation of the eight models as displayed in Table 5.1. Panel a depicts models P1 and M1, panel b M2 and P2, panel c P3 and M3, while panel d depicts models P4 and M4.

Physical domain The average treatment effect for the treated sub-population (ATT) from experiencing an adverse physical health outcome averages at about -6.5% or -8.5% gross wealth build-up, depending on the aggregation choice. Ten to twelve years post-treatment, the effect reaches around -12% to -14%.

Note that the *post average* tends to display stronger effects than the *SimpleATT* because each post-treatment period is equally weighted, whereas the simple average is weighted by each group size. Due to attrition, the group sizes shortly after and before treatment are bigger than those that remain longer in SOEP. Since the effects shortly after treatment are smaller in magnitude, the observed difference arises.

In absolute terms (models P3 and P4), experiencing a negative health outcome is associated with accumulating €7,000 (*SimpleATT*) to €8,500 (*post average*) less wealth than the control units. This difference reaches up to €12,600 (gross) and €13,000 (net) within twelve years from the event. However, due to the strong rejection of the pre-trend tests, the

validity of a causal interpretation is less credible in these two specifications. Note also that the effect is measured compared to the untreated group. Therefore, it is not necessarily the case that they experiences a decrease in their wealth; rather, they fail to follow the same accumulation path relative to the control group. These patterns can be observed in Figures B.3 and B.4.

Mental domain Looking at the mental health domain, the effects tend to be stronger. Those that experience an negative mental health outcome accumulate, on average, −8.3 to −10.3% gross wealth, depending on the aggregation choice. We also see an indication that the effects might arise faster. Four years after the event ($e = 4$), we see already an effect of −10%, and reaching −15% within the twelve years window.

In absolute terms (models M3 and M4), the impact reaches −21.3 (gross) and −17.7 (net) thousand Euros twelve years after the event. Note, however, that there is a large, unexpected increase (in absolute terms) in the last period of the event window, and the confidence interval is quite large at that stage.

In general terms, the impact of experiencing a negative mental health outcome seems to be stronger than that in the physical domain. Further, the models also seem to behave better in the mental domain when evaluating the pre-treatment trends and the precision of estimates after the event.

One caveat, though, is that while we cannot reject the pre-treatment tests and most of the pre-treatment ATT's confidence intervals (CI) include 0, the pre-treatment CI's are still quite large. This implies that, in a less optimistic view, one cannot exclude the possibility of diverging pretrend paths. To illustrate, a straight line crossing from pre-treatment to post-treatment (and diverging from 0) would imply different wealth accumulation paths. In that being true, this challenges the parallel trends assumption and, thus, the estimated coefficients would not portray a causal effect of experiencing a bad health outcome.

5.2.2 Exploring Wealth Accumulation Channels

To understand potential reasons for the observed differences in wealth accumulation, I explore the effect the same adverse health outcomes on labor market statuses. I focus on four key variables: full-time employment experience, unemployment experience (both measured in months since entering the labor market), current individual annual labor earnings, as well as current employment status. In the labor earnings specification, unemployed people are kept in the analysis with a labor earnings of 0.

Full-time experience As shown in Figure 5.2 a, one can observe a slightly different trajectory after treatment when comparing the physical and mental health models. Physical health shock suggests a stronger effect, resulting in over 7 months less full-time employment experience twelve years after the event. In contrast, the mental health impact is nearly

5 months. However, the pre-treatment trend in the physical health model, although with confidence intervals that include 0, hits at a linear pre-trend.

Unemployment experience In Figure 5.2 b, a distinct pattern emerges. Similarly to the full-time experience case, the physical health model also shows a pre-trend divergence, but this time more clearly. In contrast, the mental health model shows a distinct break-point at $e = -2$. The effects in both domains are of similar magnitude, reaching around 4.5 months of additional unemployment experience when compared to the respective untreated groups.

Labor earnings Figure 5.2 c shows a similar pre-treatment trend in the physical and mental domains, supporting a common pre-trend only up to period $e = -6$. Before that, although with confidence intervals still covering 0, one could argue for a divergence in pre-treatment trend which is detrimental to the parallel trends assumption.

After the event, the physical health shock has a more pronounced impact compared to the mental health case. The post-average effect aggregates to -2.6 thousand Euros in annual labor income. Twelve years after the event, it reaches -4 thousand Euros. In the mental dimension, the *post average* aggregates to -1.6, and reaches nearly -2.7 thousand Euros twelve years after the event.

Employment status In Figure 5.2 d, we observe that the pre-treatment trend is quite similar in both models, with the centers of estimates relatively close to 0. Post treatment, on the other hand, the shock in the physical domain is stronger. Given that the response variable is binary, the estimates can be interpreted as in a linear probability model. In the physical domain, the average post-treatment effect is a nearly 5% reduction in the employment rate for the affected group, reaching 7.5% twelve years after the event.

In the mental health domain, the effect is similarly strong to that in the physical case in the first two periods after the event, reaching a -3.5% reduction in the employment rate at $e = 4$, and stabilizes in that range over the rest of the event window. In this case, the simple average and post average are roughly the same, equaling around a 2.8% reduction in the employment rate for the affected group.

5.2.3 Effect Heterogeneity

So far, the analysis covered the entire sample. However one might anticipate effect heterogeneity based on key characteristics of the population. To explore this, I replicate the the main analysis after dividing the sample into two subgroups, one based on age and the other on educational attainment.

Note that in these specifications, the analysis focuses on comparing treated and control units within the same subgroup rather than comparing outcomes across different groups. To clarify, for the higher education group, the figures illustrate the difference in the evolution of

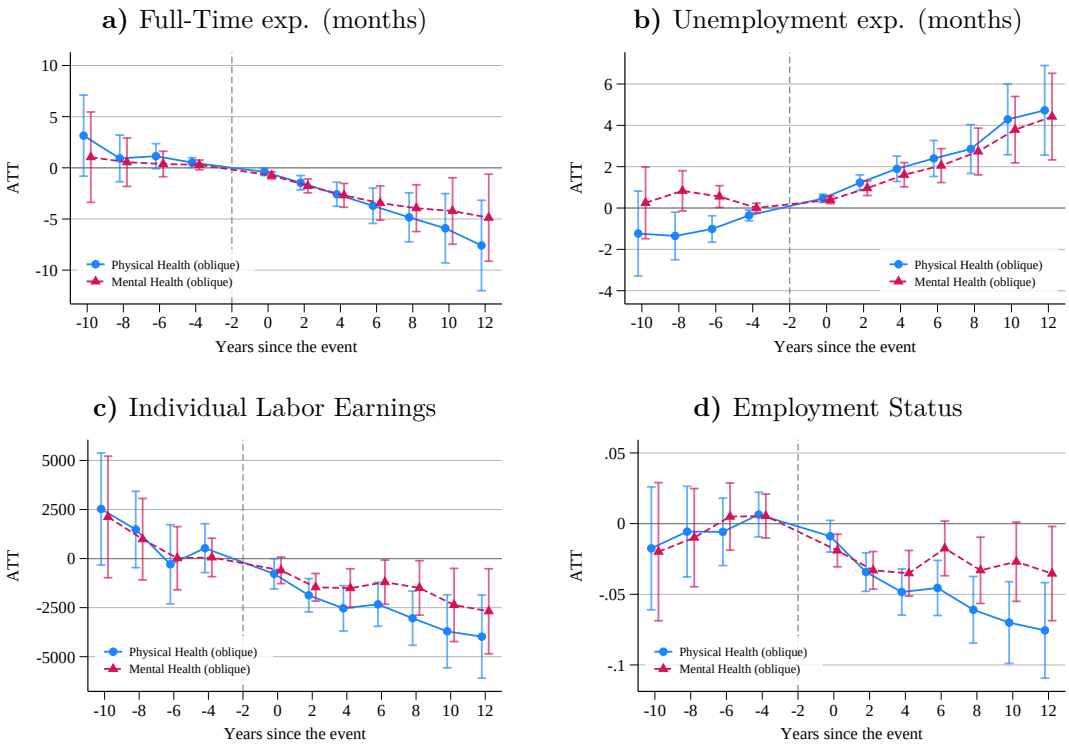


Figure 5.2 Labor market outcomes comparison

Notes: The same DiD_{DR} framework was applied to other response variables, using the same covariates and treatment rule as specified in the main analysis. Whiskers depict the 95% confidence interval

the response variable between treated and untreated units, all within the higher education group. If the trends of each subgroup are significantly different, it provides evidence of effect heterogeneity with respect to the grouping variable. For easier comparison between different models and specifications, sub-group figures from the physical domain are presented in shades of blue, and those from the mental domain are in variations of red.

Heterogeneity by educational attainment The sample is divided into two groups: those who attained at most a high-school degree and those who with a higher education certificate. On average, those in the first group completed 11 years of schooling, while those in the latter completed nearly 16 years.

The results of the effect on wealth accumulation, measured by the logarithm of gross wealth, as depicted in Figure 5.3 a, are similar to those obtained in the main analysis (compare with Figure 5.1 a). Further, we observe no substantial evidence of differences between the two groups, suggesting no effect heterogeneity by educational attainment.

When focusing on the unemployment experience, a group distinction becomes clearer, as depicted in Figure 5.3 b. The general pattern in the physical domain is similar to that depicted in the main analysis, illustrating a distinct pre-treatment trend that extends linearly into the post-treatment period. However, there is a variation in magnitude between both subgroups, with the less educated experiencing an effect twice as large as the one faced

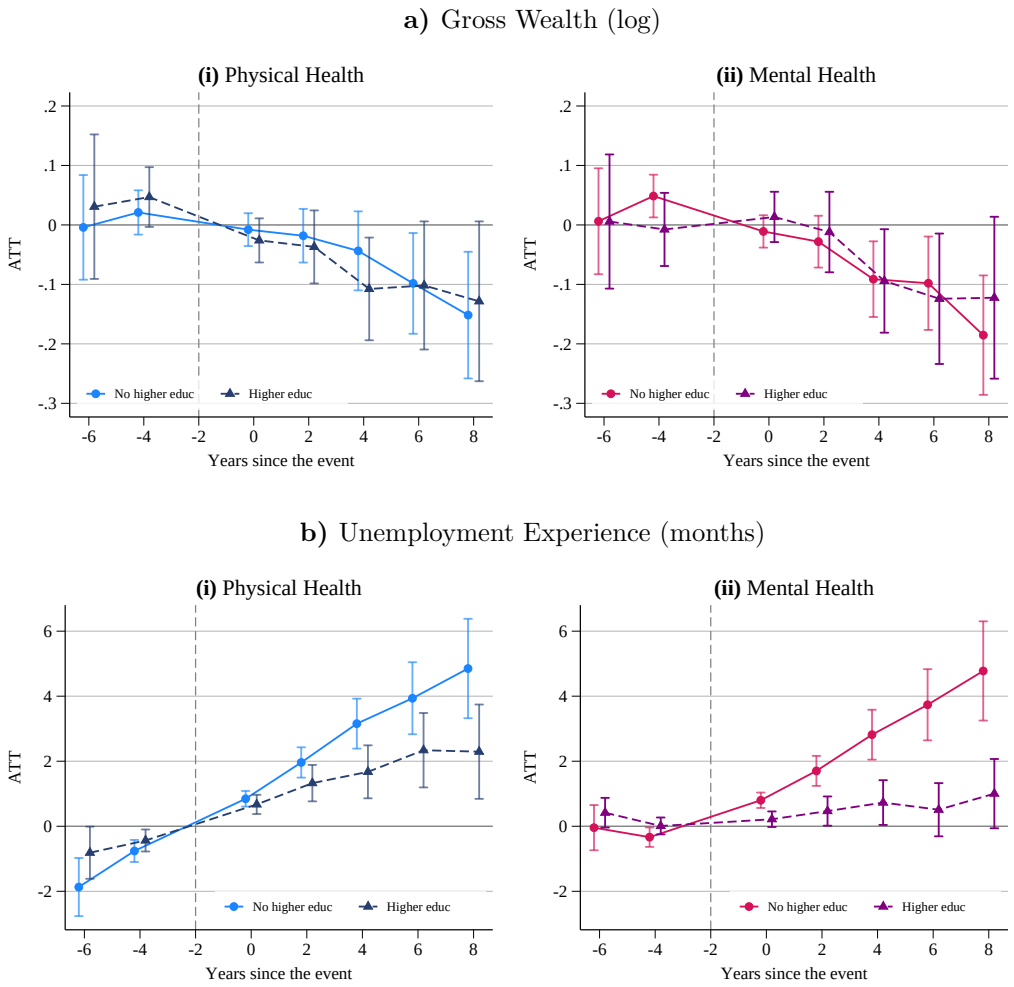


Figure 5.3 Effect heterogeneity by educational attainment

Note: Whiskers depict the 95% confidence interval. Reduced event window due to decrease in precision after dividing the sample into subgrous.

by the highly educated.

Turning our attention to the mental health domain, the general pattern also mirrors that of the main analysis, where the assumption of a common trend is more plausible. In this case, the effect is primarily driven by the less educated group, manifesting an impact of around five months of unemployment experience eight years after the event. In contrast, for the highly educated, the effect amounts to approximately 1.5 months of unemployment in the same time frame.

Heterogeneity by age group One can anticipate differences in wealth accumulation and (dis)savings patterns over the life cycle. To explore heterogeneity by age, three groups are defined: younger (aged 18 to 39), prime-age (40 to 55), and older adults (56 to 75). These cutoff points were chosen to achieve groups roughly equal in sample size, but also aiming on capturing three distinct stages in the wealth accumulation trajectory. The first group begins, on average, with minimal wealth but can rapidly build it up in this time window.

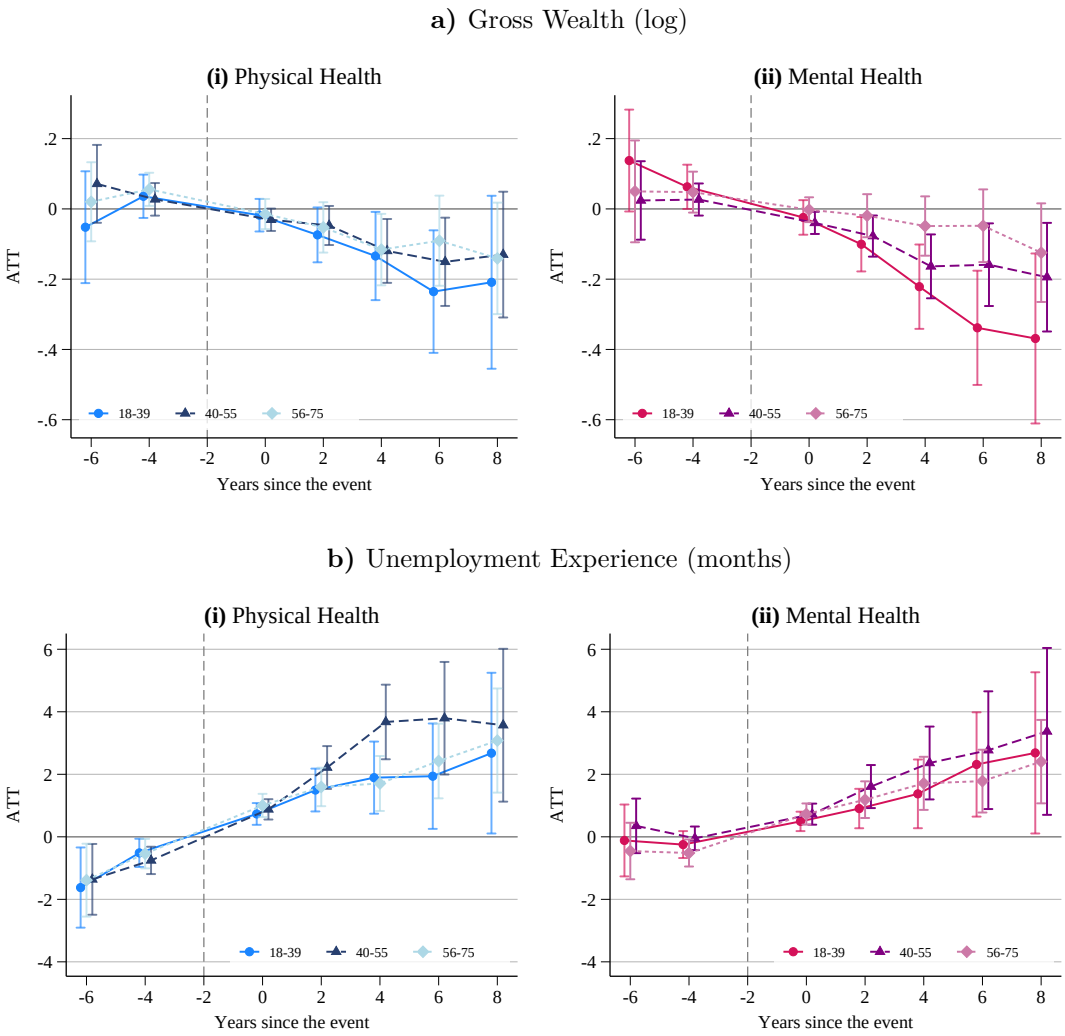


Figure 5.4 Effect heterogeneity by age groups

Note: Whiskers depict the 95% confidence interval. Reduced event window due to decrease in precision after dividing the sample into subgrous.

The second group, on average, starts with an already consolidated wealth value but can still experience further accumulation, stabilization, or disaccumulation. Meanwhile, the third group is likely to start experiencing dissavings to a lesser or higher extent.

When dividing the sample into three groups, one should exercise caution in interpretation due to the statistical imprecision that arises. With that in mind, examining the wealth panels in Figure 5.4 a, the general pattern appears similar to that of the main analysis. However, the parallel trends assumption, as judged by the pre-treatment trends, may seem plausible for some groups but less so for the others. Even refraining from interpreting a causal effect, one can clearly observe distinct subgroup accumulation paths, which could be the subject of deeper examination.

When looking at the unemployment experience, as depicted in Figure 5.4 b, a result similar to that in the main analysis becomes apparent. There is a more pronounced hint —

stronger in the physical domain and less so in the mental domain — that the prime-age group is more strongly affected by the adverse health outcome. Similar to the main analysis, however, only in the mental health domain does the pre-treatment trend corroborate the parallel trends assumption.

5.2.4 Validation and Robustness Checks

Treatment assignment rule To validate the treatment assignment rule, as described in detail in Section 5.1.1, I reapply the procedure to variables that are intrinsically related to the measures of physical and mental health. With that, strong correlations with the adverse health outcome are to be expected. Clearly, the aim here is not to evaluate the impact on such measures, since they are basically different facets of the same concept, but to validate the assignment rule and better understand the model. For that, I examine a few measures of subjective health and well-being, as well as specific health diagnoses.

All four variables depicted in Figure 5.5 are measured on an 11-point Likert scale, ranging from very unsatisfied to very satisfied with the current item. A few insights emerge. First, the effect is long-lasting in both domains when looking at life and health satisfaction, meaning that the effect does not bounce back to 0 after the adverse event. Satisfaction with income or work, on the other hand, shows a slow trend back towards 0.

Another insight, when comparing the two health domains and leaving aside the difference in magnitude, is that the mental health shock has a stronger impact on life satisfaction than on health satisfaction. This suggests that the items measured by the MCS weigh more heavily on overall life satisfaction, whereas physical health is more saliently measured by the PCS. The magnitude of the mental health shock is similar concerning life and health satisfaction. In contrast, the physical health shock shows a considerably stronger impact on health than on life satisfaction.

Focusing on the pre-trends, especially in the mental health domain in panels a, b, and d, a pattern emerges where the Average Treatment Effects on the Treated (ATT's) converge towards 0 . This suggests the existence of anticipation a few periods before treatment (as defined by the chosen treatment rule). An appealing extension would be to set the treatment time one or two periods before experiencing the more drastic adverse health outcome to account for this anticipation. Nevertheless, the visible kink around $e = -2$ does indicate that there was a considerable divergence in the paths between treated and control units occurring due to (or for other reasons, but concurrently) the health shock.

The variables depicted in Figure 5.6 target four specific health diagnoses. Two are from the physical domain, Back Pain and Cardiopathy, and two are from the mental domain, Depression and Sleep Disorder. The response variables are binary indicators of ever being diagnosed with the respective condition, so the coefficients can be interpreted as in the linear probability model. Illustratively, a coefficient of 0.05 means a 5% increase in the diagnosis rate of a given condition.

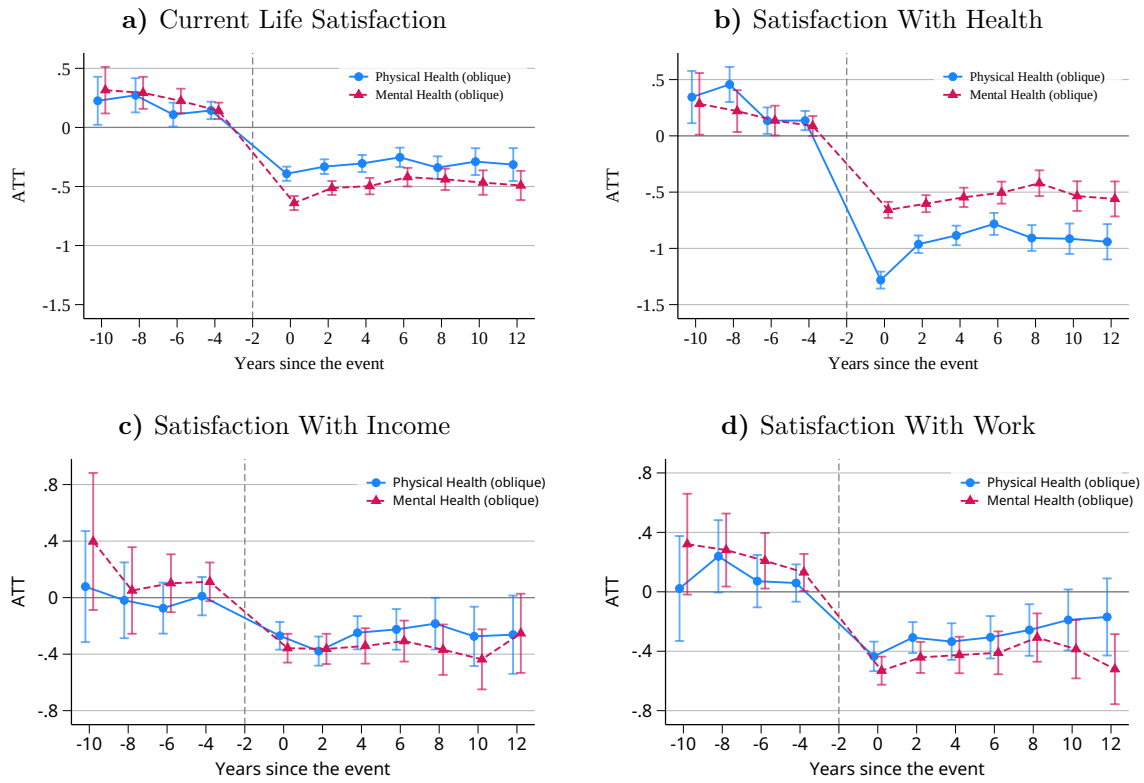


Figure 5.5 Validation on subjective health and well-being measures

Notes: Variables measured in a 11-point Likert scale. Whiskers depict the 95% confidence interval.

The four panels show a few patterns that are useful for assessing the inner workings of the models. The Figures 5.6 a and 5.6 b, for example, confirm that the health shock in the physical domain is more strongly associated with back pain and cardiopathy than the shock in the mental domain. On a more technical note, we observe a pre-trend that is parallel to the 0 line but significantly different from 0. This indicates an anticipation of one period but not longer, since we are measuring the ATT's against the reference period when $e = -2$. If that is true, adjusting for anticipation by assigning the treatment time one period prior to the current rule could better capture the treatment effect. In that case, the post-treatment effect would be higher than currently shown.

In Figures 5.6 c and 5.6 d, with variables related to mental health, we observe that the shock from both domains seems to capture a similar increase in the diagnostic rate after the event. While this could raise concerns about using health scores obtained from obliquely rotated factors, the results are very similar when using scores obtained from orthogonal rotation following a Principal Components Analysis (see Figure C.2 on page 55).

In conclusion, we gather that the assignment rule is able to capture variations in subjective well-being measurements as well as in specific health diagnoses in the physical and mental domain.

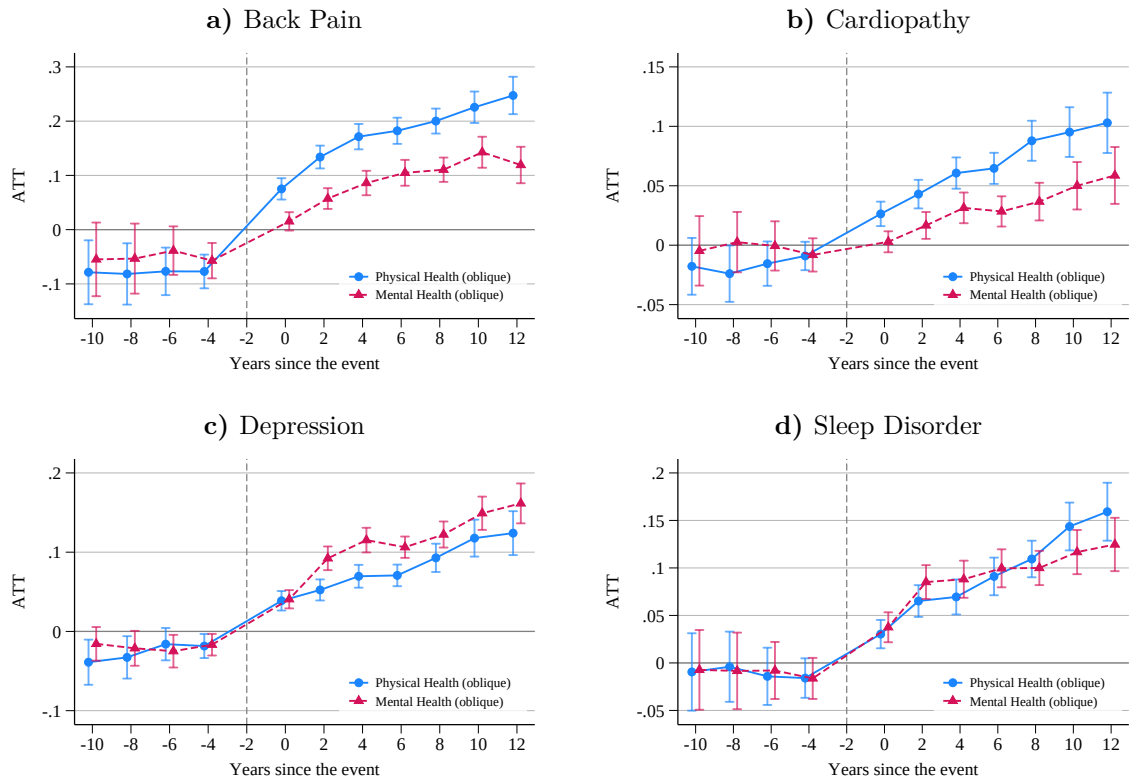


Figure 5.6 Validation with selected measures of health diagnoses

Notes: Response variables are binary indicators of being ever diagnosed with the respective condition. Panels have different scale on the y-axis. Whiskers depict the 95% confidence interval

5.2.5 Robustness Checks

In the following, I present a series of alternative specifications and model variations conducted to assess the robustness of the results obtained.

Model specification In order to assess model stability, I compare the main results with four other specifications. These alternative results are presented in Figure B.1 and Table B.1. Note that all robustness models are to be compared against the primary specification from the main analysis; that is, P1 and M1 are to be compared against $P1_{(a)}$ to $P1_{(d)}$ and $M1_{(a)}$ to $M1_{(d)}$, respectively.

The specifications vary in terms of the comparison units, by keeping not-yet-treated units in the control group (models tagged with subscript a). Furthermore, a model without covariates was run, which implies a simpler *DiD* without the Doubly-Robustness property, since no covariates are used to predict treatment probabilities (tagged with subscript b). Additionally, to account for attrition, the inverse probability of remaining in SOEP is used as weights in one specification (tagged with subscript c). An evidence of panel attrition differentially affecting the participation length over wealth quintiles is presented in Figure B.5. Finally, a model with a *balanced* panel is also estimated, where the sample is restricted to people that remains in SOEP from 2002 to 2020 (tagged with subscript d). In this spec-

ification, there are 2,371 unique individuals and 22,687 total observations in the physical health domain. In the mental domain, the figures are 2,110 and 20,010, respectively. Note that some individuals are automatically dropped in the estimation procedure due to, e.g., restricting the event window so that the number of total observations is not exactly $N \times T$.

The results are stable across specifications, and similar to those obtained in the main analysis, with the balanced specification ($P1_d$ and $M1_d$) being the closest ones. The coefficients in the other models are lower by about two to three percentage points, but the general pre and post-treatment remain similar.

Treatment assignment rule One possible source of vulnerability in the main specification is the discretionary choice of the treatment rule. To assess if this poses a risk to the interpretation of the main analysis, I replicate it with a different treatment assignment rule. In this case, individuals are assigned to treatment after only one experience of an adverse health outcome, while the threshold is set to one standard deviation lower than the median health score values of people of the same age and gender. The higher threshold is chosen so that not too many people are assigned to treatment at a point or another.

The results, as presented in detail in Table B.2 and Figure B.2, and focusing in the main specification with log of gross wealth as response variable, show a reduction of about half in the physical health domain. The *Simple ATT* of $P1_r$ reduces to -3.44% (from 6.5% in P1) and the post average aggregates to -4.8%, albeit with a relatively large standard error of 2.49 deeming it statistically insignificant.

In the mental domain, the results are more in line with those obtained in the main analysis. The *simple ATT* of $M1_r$ is estimated at -7.36% (compared to -8.34% from M1), and the post-average is -9.66% (from -10.30%) and still significant at any conventional level. In the longer term, the center of the estimated coefficients are slightly reduced in the longer term. In the $M1_r$, the coefficients six to twelve years after the event of stabilizes at around -12% (compared to -13 to -15% from M1).

In conclusion, the results are considerable stable in the mental health domain over different specifications. In the physical domain, we gather evidence of some model dependency, meaning that the results are more susceptible to variation depending on the exact modeling specifications. In the secondary models P2–P4 (those with net wealth or wealth measured in levels), the results are more stable. However, due to pre-trend issues and the expectation that the parallel trends hold in relative but not in absolute terms, I give more weight to the primary models, which capture the evolution of gross wealth in relative terms.

6 Discussion

The interactions between health and wealth are complex. Both variables evolve gradually over the life cycle and their interdependence is not trivial to disentangle. In this study, I focus on the proposition that health causes wealth. There are several plausible mechanisms, through which this premise would prove to be true. A health shock can have, for example, a direct impact on one's wealth through expenses to cover the medical costs. In addition, such a shock, can impair one's labor capacity to a lesser or greater extent. The ability to pursue career and educational objectives can also be impaired by a unstable health trajectory. These examples illustrate how a health shock can lead to a different wealth accumulation trajectory over the long run.

Conversely, it is also easy to think of mechanisms where wealth causes health. While Meer et al. (2003), via an instrumental variable approach, find evidence of a negligible effect in the short run, over a longer time span, it is credible that wealthier people could be able to achieve a better health status through access to better health care, or make preferable lifestyle choices due to easier access to health-promoting activities, to name a few. In addition, there are a set of factors, which might be unobserved by the researcher, that show a causal pathway to both health and wealth. A behavioral or genetic trait might be the cause of, for example, being better equipped to attain higher education and, thus, accumulate more wealth while, at the same time, avoid unhealthy behaviors such as alcohol abuse or cigarettes consumption. On these accounts, I present in the following section key threats to the internal validity of this work.

6.1 Threats to Internal Validity

The framework used in this study accounts for the issue of unobserved factors, by differencing out the outcome of the same individual over different points in time. Assuming these characteristics remain constant, these hidden factors cancel themselves out. With that, the obtained coefficients are valid causal estimates of the effect even in the presence of heterogeneity that, in a cross-sectional setting, would only grant a correlational interpretation.

A stronger threat to internal validity in this setting, emerges when the researcher subscribes only to the *social causation* hypothesis, which states that health is caused by, but does not cause wealth. In this setting, they could argue that those that are eventually treated are different from the non-treated in such key aspects, that regardless of experiencing an adverse health outcome, their wealth trajectory would not follow a parallel line to the path of those with more stable health. Differently to the hidden factors case, the

applied framework does not mechanically account for reverse-causation, as implied by the social causation hypothesis. If this premise truly represents the underlying mechanism, the results obtained would remain the same. Only our interpretation and their implications would be wrong.

To address this concern, in the availability of several pre-treatment periods, by evaluating their patterns, the researcher can adjust their beliefs on the veracity of such hypotheses.

In this concrete case, the pre-treatment trends show, as covered in depth in Section 5.2, mixed results. In some specifications, the pre-trends being close to 0 indicates no difference in the outcome changes across treated and untreated groups previous to treatment event. Furthermore, the kink observed around the treatment event, corroborates to the hypothesis that the adverse health outcome does impact the wealth accumulation. With that said, the confidence intervals over the pre-treatment periods are large in some specifications. With that, one could argue that the pre-treatment tests do not have enough power to rule out the hypothesis of a diverting pre-treatment trends. In some cases, such as the physical health shock on unemployment experience (see Figure 5.2 b), this is indeed the most credible interpretation. With this in mind, the work of Rambachan and Roth (2023), where they propose a flexible specification of the pre-trend tests, could prove to find an interesting application in this setting.

While this discussion may appear to be a solely philosophical consideration, the distinction has clear real life implications. If a society, for example, finds it to be unfair that people experience high socioeconomic disparities at later stages of life and policy makers wish to remedy that, one interpretation would grant more room for policy instruments to be used than the other. Precisely, the causal interpretation implies that a policy aiming to minimize severe health variation for the working-age population would have the additional benefit of reducing wealth inequality at retirement age, as an example. If the correlation interpretation is taken, the policy would have to act directly on reducing wealth inequality at retirement in order to achieve its goals.

To conclude, the specifications in the mental health domain shows better support for a causal identification, specially in the primary models with log of gross wealth as the response variable. One could, therefore, hypothesize that there is a higher degree of exogeneity in the mental than in the physical dimension. It appears plausible, a priori, that in the physical domain, the health deterioration happens gradually and, therefore, can be anticipated in terms of reduced working hours or earlier retirement. And this would lead to a lower wealth build-up over time. Similarly, it also seems credible that a portion of mental health deterioration happens unexpectedly and, therefore, cannot be fully anticipated. If that is true, this would explain the difference in pre-trend pattern when comparing both health domains and why the mental health specifications portray a more credible causal interpretation.

6.2 Extensions and Outlook

In this section, I list a few conceivable extensions to this study in terms of its content as well as concerning its technical nature. Contentwise, one could enlarge this study's scope by including the partner's and household data. This could prove to be a significant gain by capturing wealth patterns within a household. The current setting focuses solely on the individual wealth trajectory.

Another improvement could be achieved by including family history data, in order to capture characteristics inheritable from parents. This could include socioeconomic variables such as the father's and the mother's educational attainment. In addition, received inheritance could be a proxy of parent's wealth. In the health dimension, the parent's age of decease could be a proxy for their health, which would function as genetic proxy. These variables would capture an inter-generational dimension that focusing solely on the individual trajectory neglects. Preliminary checks have shown, however, that while these items are asked by the SOEP, the data availability is rather sparse. The options would be focusing on the subset with valid data or imputing the missing values in order to use most of the otherwise available information.

The current study can also be extended in a technical way. A matching estimator, for example, that assures balanced groups prior to treatment, based on a rich set of covariates provided by the SOEP, would ensure that the treated and control units are indeed very similar to one another. If the results are similar, this procedure would increase the credibility of causal identification. In the current setting, however, a considerable threat to validity is the possibility of differential trends due to different levels of wealth between groups in the pre-treatment period. Daw and Hatfield (2018) show the risks introducing *regression to the mean* bias by matching on variables that are correlated with the response variable, including, clearly, the response variable itself.

In addition, Rambachan and Roth (2023) propose an extension to DiD designs where the parallel trends assumption might be violated. By not restricting the trends between groups to be perfectly parallel, the authors present novel tools for inference over the bounds of the causal effect, relative to the extent of the difference in trends, justified by the observed pre-treatment trend.

To conclude, in the current setting, the health summary scales, which are continuous variables, are dichotomized into a binary treatment variable, as is required by the applied DiD framework. Recent addition to the literature, however, generalizes this approach to a framework with continuous variables (see Callaway, Goodman-Bacon et al., 2024). The possible improvement is two-fold: First, it removes the discretionary decision of the treatment rule threshold and, second, it is able to distinguish effect intensities.

7 Conclusion

In this study, I assess the impact of an adverse health outcome on wealth accumulation over a substantial time span. By dissecting the analysis into mental and physical health dimensions, I explore variations in effects across these domains. The investigation encompasses ten bi-yearly waves of the German Socio-Economic Panel, spanning from 2002 to 2020, and addresses the wealth accumulation trajectories of individuals aged between 18 and 75 years old. Concentrating on primary specifications that measure gross wealth in relative terms, the key findings can be summarized as follows:

An adverse health outcome in the physical domain reduces the gross wealth accumulation by 6.5%, and reaches 12% twelve years after the event. In the mental health domain, the effect averages at 8.3% less gross wealth than what is accumulated by the control group. Twelve years after the event, it accounts for 15% reduced wealth accumulation.

In exploring possible effect channels, I show that the increase in unemployment after the adverse health event, measured by unemployment experience, employment status or lack of full-time employment experience, plausibly demonstrate the effect mechanisms.

Furthermore, I find no clear evidence on effect heterogeneity by education attainment on wealth accumulation directly, although, in terms of unemployment experience—specially in the mental domain—the less educated are capturing the majority of the effect.

When analyzing by age groups, there are hints of effect heterogeneity on wealth accumulation, with the younger population (18–39) experiencing a stronger impact. The difference among groups is bigger in the mental domain. The effect on unemployment experience, in contrast, is stronger for those aged between 40 and 55.

In a secondary analysis, I confirm that the items from SOEP's health module, which are based on the SF-12v2 methodology, captures the intended concepts. The items, as one would expect, are well clustered in the factor space and a clear separation between physical and mental health dimensions emerges. However, I could also reproduce some of the raised criticisms on the *agreement problem* between sub-scales and their respective summary scores, when using the SF-12 with SOEP data. On this account, an alternative exploratory factor analysis model followed by an oblique rotation is adopted to construct the summary scores used in the main analysis.

Conceivable extensions to this work include integrating partner's information to better account for within-household wealth structures. In addition, family-related data such as parent's health and proxies for socioeconomic status such as education attainment could prove to be key explanatory factors to the individual's wealth accumulation trajectory.

References

Abadie, A. (2005). 'Semiparametric Difference-in-Differences Estimators'. In: *The Review of Economic Studies* 72.1, pp. 1–19.

Andersen, H. H. et al. (2007). 'Computation of Standard Values for Physical and Mental Health Scale Scores Using the SOEP Version of SF-12v2'. In: *Schmollers Jahrbuch*, pp. 171–182.

Borusyak, K., X. Jaravel and J. Spiess (Jan. 2024). *Revisiting Event Study Designs: Robust and Efficient Estimation*. Version 5. DOI: 10.48550/arXiv.2108.12419. preprint.

Callaway, B., A. Goodman-Bacon and P. H. C. Sant'Anna (Feb. 2024). *Event-Studies with a Continuous Treatment*. DOI: 10.3386/w32118. preprint.

Callaway, B. and P. H. C. Sant'Anna (2021). 'Difference-in-Differences with Multiple Time Periods'. In: *Journal of Econometrics* 225.2, pp. 200–230. DOI: 10.1016/j.jeconom.2020.12.001.

Case, A. and A. S. Deaton (2005). 'Broken Down by Work and Sex: How Our Health Declines'. In: *NBER Chapters*, pp. 185–212.

Chetty, R. et al. (2016). 'The Association Between Income and Life Expectancy in the United States, 2001-2014'. In: *JAMA* 315.16, pp. 1750–1766. DOI: 10.1001/jama.2016.4226.

Christensen, L. N. et al. (2013). 'Validation of the 12 Item Short Form Health Survey in a Sample from Region Central Jutland'. In: *Social Indicators Research* 114.2, pp. 513–521.

Daw, J. R. and L. A. Hatfield (2018). 'Matching and Regression to the Mean in Difference-in-Differences Analysis'. In: *Health Services Research* 53.6, pp. 4138–4156. DOI: 10.1111/1475-6773.12993.

Deaton, A. (2003). 'Health, Inequality, and Economic Development'. In: *Journal of Economic Literature*, p. 46.

De Chaisemartin, C. and X. D'Haultfoeuille (May 2022). *Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey*. DOI: 10.3386/w29734. preprint.

Fabrigar, L. R. and D. T. Wegener (2012). *Exploratory Factor Analysis*. Understanding Statistics. Oxford ; New York: Oxford University Press. 159 pp.

Federal Statistical Office (n.d.). *Consumer Price Index: Germany, Years (Series 61111-0001)*. Genesis-Online.

Gill, S. C. et al. (2007). 'Validity of the Mental Health Component Scale of the 12-Item Short-Form Health Survey (MCS-12) as Measure of Common Mental Disorders in the General Population'. In: *Psychiatry Research* 152.1, pp. 63–71. DOI: 10.1016/j.psychres.2006.11.005.

Goebel, J. et al. (2019). 'The German Socio-Economic Panel (SOEP)'. In: *Jahrbücher für Nationalökonomie und Statistik* 239.2, pp. 345–360. DOI: 10.1515/jbnst-2018-0022.

Grabka, M. M. (2022). *SOEP-Core V37 - HEALTH*. Research Report 1181. SOEP Survey Papers.

Hagell, P., A. Westergren and K. Årestedt (2017). 'Beware of the Origin of Numbers: Standard Scoring of the SF-12 and SF-36 Summary Measures Distorts Measurement and Score Interpretations'. In: *Research in Nursing & Health* 40.4, pp. 378–386. DOI: 10.1002/nur.21806.

Hann, M. and D. Reeves (2008). 'The SF-36 Scales Are Not Accurately Summarised by Independent Physical and Mental Component Scores'. In: *Quality of Life Research* 17.3, pp. 413–423. DOI: 10.1007/s11136-008-9310-0.

Heckman, J. J., H. Ichimura and P. E. Todd (1997). 'Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme'. In: *The Review of Economic Studies* 64.4, pp. 605–654. DOI: 10.2307/2971733.

Kröger, H., E. Pakpahan and R. Hoffmann (2015). 'What Causes Health Inequality? A Systematic Review on the Relative Importance of Social Causation and Health Selection'. In: *European Journal of Public Health* 25.6, pp. 951–960. DOI: 10.1093/eurpub/ckv111.

Liebig, S. et al. (2022). *Socio-Economic Panel, Data from 1984-2020, (SOEP-Core, V37, EU Edition)*. In collab. with Kantar Deutschland GmbH. Version v37. SOEP Socio-Economic Panel Study. DOI: 10.5684/SOEP.CORE.V37EU.

Meer, J., D. L. Miller and H. S. Rosen (2003). 'Exploring the Health–Wealth Nexus'. In: *Journal of Health Economics* 22.5, pp. 713–730. DOI: 10.1016/S0167-6296(03)00059-6.

Nick Cox (11th Sept. 2005). *St: RE: -Factor Pcf- vs -Pca- (Was Factor Score Postestimation)*. URL: https://www.stata.com/statalist/archive/2005-09/msg00297.html (visited on 12/12/2023).

Rambachan, A. and J. Roth (2023). 'A More Credible Approach to Parallel Trends'. In: *Review of Economic Studies* 90.5, pp. 2555–2591. DOI: 10.1093/restud/rdad018.

Rios-Avila, F. (29th Apr. 2023). *Does CSDID Store Weight (Dripw)? - Statalist*. URL: https://www.statalist.org/forums/forum/general-stata-discussion/general/1711680#post1711687 (visited on 23/01/2024).

Sant'Anna, P. H. C. and J. Zhao (2020). 'Doubly Robust Difference-in-Differences Estimators'. In: *Journal of Econometrics* 219.1, pp. 101–122. DOI: 10.1016/j.jeconom.2020.06.003.

Tucker, G., R. Adams and D. Wilson (2013). 'Observed Agreement Problems between Sub-Scales and Summary Components of the SF-36 Version 2 - An Alternative Scoring Method Can Correct the Problem'. In: *PLOS ONE* 8.4, e61191. DOI: 10.1371/journal.pone.0061191.

Vilagut, G. et al. (2013). 'The Mental Component of the Short-Form 12 Health Survey (SF-12) as a Measure of Depressive Disorders in the General Population: Results with Three Alternative Scoring Methods'. In: *Value in Health* 16.4, pp. 564–573. DOI: 10.1016/j.jval.2013.01.006.

Ware, J. E. et al. (2002). *How to Score Version 2 of the SF-12 Health Survey (with a Supplement Documenting Version 1)*. Lincoln, RI: Qualitymetric Incorporated.

Widaman, K. F. (1993). 'Common Factor Analysis versus Principal Component Analysis: Differential Bias in Representing Model Parameters?' In: *Multivariate Behavioral Research* 28.3, pp. 263–311. DOI: 10.1207/s15327906mbr2803_1.

Wilson, D., J. Parsons and G. Tucker (2000). 'The SF-36 Summary Scales: Problems and Solutions'. In: *Sozial- und Präventivmedizin* 45.6, pp. 239–246. DOI: 10.1007/BF01591686.

Ziebarth, N. (2010). 'Measurement of Health, Health Inequality, and Reporting Heterogeneity'. In: *Social Science & Medicine* 71.1, pp. 116–124. DOI: 10.1016/j.socscimed.2010.03.016.

Key Software

Correia, S. (21st Aug. 2023). *REGHDFE: Stata Module to Perform Linear or Instrumental-Variable Regression Absorbing Any Number of High-Dimensional Fixed Effects*. Version 6.12.3.

Jann, B. (2007). 'Making Regression Tables Simplified'. In: *The Stata Journal: Promoting communications on statistics and Stata* 7.2, pp. 227–244.

Rios-Avila, F., P. H. C. Sant'Anna and B. Callaway (25th Feb. 2023). *CSDID: Stata Module for the Estimation of Difference-in-Difference Models with Multiple Time Periods*. Version 1.73.

Rios-Avila, F., P. H. C. Sant'Anna and B. Callaway (4th Jan. 2024). *CSDID2: New Version of CSDID. All in Mata*. Version 1.21.

StataCorp. (2023). *Stata Statistical Software*. Version Release 18. College Station, TX: StataCorp LLC.

Appendix A

Additional Resources

A.1 Doubly Robust Estimator

In this section, for completeness, I present the $\widehat{ATT}_{dr}(g,t)$ as described in Callaway and Sant'Anna (2021, Section 4). Only the equations for the never-treated design is presented here, as this was the one used in the main analysis. Further, since in the main analysis I do not adjust for anticipation, the version presented here is also simplified in this respect.

$$\widehat{ATT}_{dr}(g,t) = \mathbb{E}_n\left[\left(\widehat{w}_g^{\text{treat}} - \widehat{w}_g^{\text{comp}}\right)\left(Y_t - Y_{g-1} - \widehat{m}_{g,t}\left(X;\hat{\beta}_{g,t}\right)\right)\right], \tag{A.1}$$

where

$$\widehat{w}_g^{\text{treat}} = \frac{G_g}{\mathbb{E}_n\left[G_g\right]}, \qquad \widehat{w}_g^{\text{comp}} = \frac{\frac{\hat{p}_g(X;\widehat{\pi}_g)C}{1-\hat{p}_g(X;\widehat{\pi}_g)}}{\mathbb{E}_n\left[\frac{\hat{p}_g(X;\widehat{\pi}_g)C}{1-\hat{p}_g(X;\widehat{\pi}_g)}\right]} \tag{A.2}$$

A.2 TWFE Alternative Design

In the following, I present the alternative model featuring a TWFE regression design.

$$Y_{i,t} = \sum_{e\in\mathcal{E}} \theta_e^{\text{FE}} \mathbb{1}(\text{health degradation})_{i,t} + \eta X_{i,t} + \delta_i + \gamma_t + u_{i,t}, \tag{A.3}$$

where the relatvei event times $e \in \mathcal{E} = \{-10, -8, -6, -4, 0, 2, ..., 12\}$. Note that relative event time $e = -2$ is dropped from the estimation, in order to be able to identify the ATT in a full set of individual and year fixed effects. The indicator function $\mathbb{1}(\text{health degradation})_{i,t}$ tracks the relative time since/until experiencing a significant health degradation. The vector X incorporates (time varying) covariates used in the model, and δ_i and γ_t represent the individual and year fixed effects, respectively.

A.3 Transformation of Log-Linear Specifications

As explained in `estout`'s manual (Jann, 2007), a function, $f(b)$, and its first derivative, $\partial f(b)$, can be used to transform the coefficients to be presented. Concretely, the following functions were used for the log specifications:

$$f(b) = 100 \cdot (e^b - 1), \qquad \partial f(b) = 100 \cdot (e^b). \tag{A.4}$$

According to the package's manual, the first derivative is used to transform the standard errors as in:

$$\text{se} \cdot \partial f(b) \tag{A.5}$$

With regards to the log of *net* wealth, however, the original coefficients cannot be interpreted as percentage changes due to crossing over from negative to positive and vice-versa. To illustrate, if an individual's net worth is valued at negative 10.000€, and in the next period it is valued at positive 20.000€. What would the percentage change in their net wealth? Only if one assumes very few people cross the €0 threshold could one interpret the coefficients approximately as a percentage change. As a matter of fact, even the interpretation of gross wealth is not, at first, trivial due to the possibility of owning 0 wealth. This issue is, however, circumvented with the unproblematic assumption that everyone has something valued for at least 1€. On these accounts, the transformation was applied to both gross and net wealth specifications, but the reader should be aware of this caveat when interpreting the results of the neglog transformation.

A.4 Health Scales and Sub-Scales Over the Life Cycle

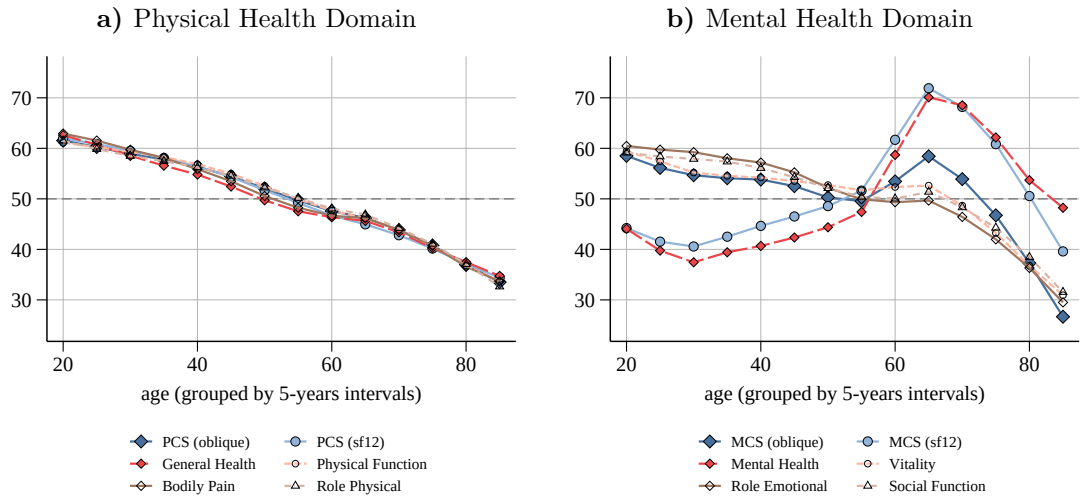


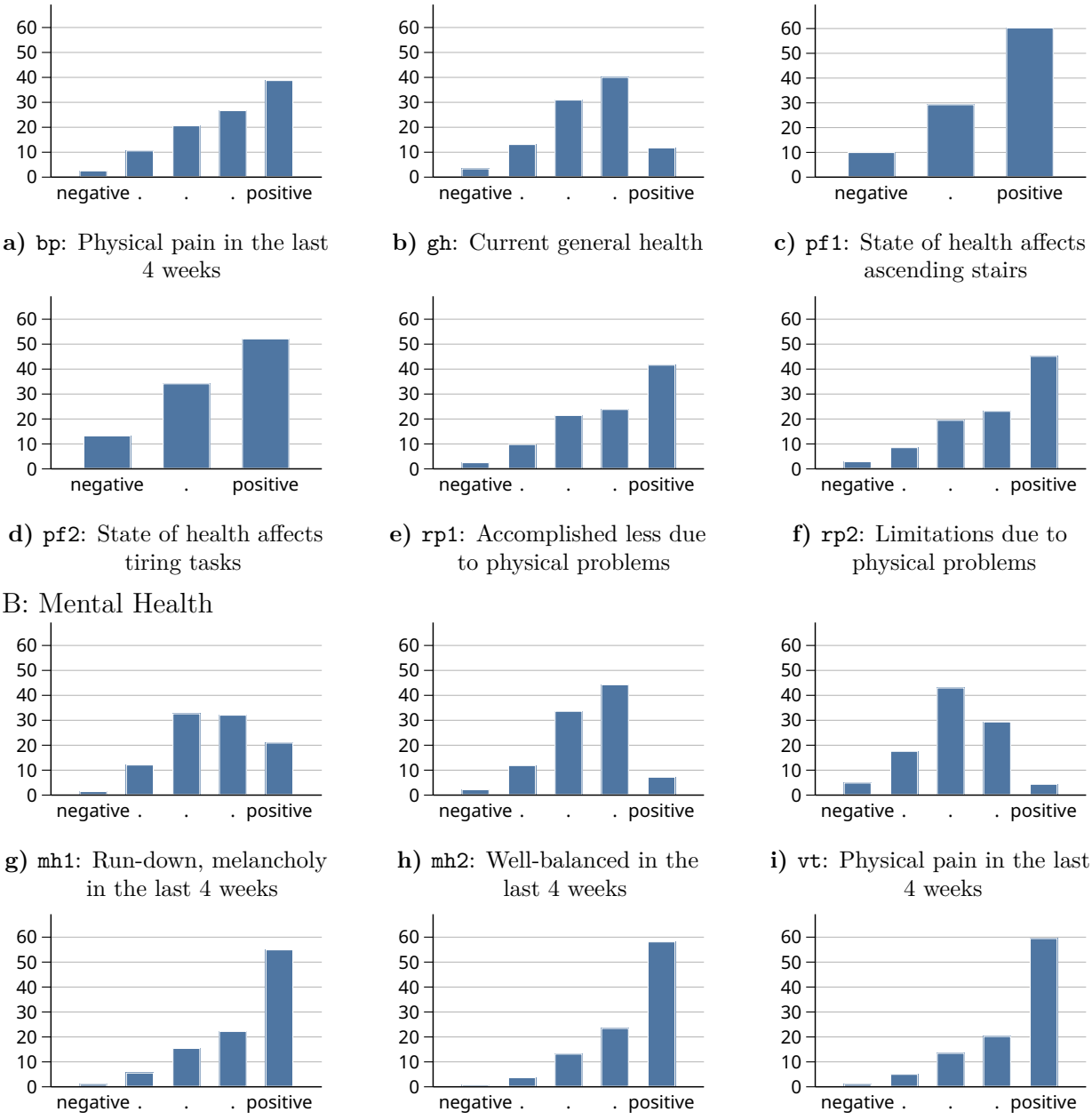
Figure A.1 Comparison of average health summary scores and sub-scales by age group

Notes: This figure shows the sub-scales superimposed with the summary scores constructed following the SF-12 method as well as the alternative method with an oblique rotation. Since age is a key variable that captures most of the variation in health outcomes, this visualization helps us better grasp how the summary scores behave relative to each sub-scale. Note that *Mental Health* is the denomination of one of the sub-scales, while MCS is the summary score in the mental domain.

Panel a shows the PCS (from both methodologies) and the sub-scales from the physical health dimension. It becomes apparent that in this domain, the summary scores are coherent with all sub-scales in both methodologies.

In the Mental domain, depicted in panel b, one sub-scale (Mental Health) shows a different pattern to the other three. MCS_{sf21} follows this single variable very closely, in detriment of the other three. It is worth noting that the interpretation of how the mental health unfolds over the life cycle differs considerably over the two methods. MCS_{sf21} would suggest a rather low starting point, then a slow increase over the time, a big increase around retirement age and, finally, a rapid decrease subsequently. In the oblique case, in contrast, we see a relatively high starting point, then a slow decrease over time, then only a moderate increase around retirement and, finally, also a rapid decrease later on.

A: Physical Health



a) `bp`: Physical pain in the last 4 weeks

b) `gh`: Current general health

c) `pf1`: State of health affects ascending stairs

d) `pf2`: State of health affects tiring tasks

e) `rp1`: Accomplished less due to physical problems

f) `rp2`: Limitations due to physical problems

B: Mental Health

g) `mh1`: Run-down, melancholy in the last 4 weeks

h) `mh2`: Well-balanced in the last 4 weeks

i) `vt`: Physical pain in the last 4 weeks

j) `re1`: Accomplished less due to emotional problems

k) `re2`: Less careful due to emotional problems

l) `sf`: Limited socially due to health

Figure A.2 Histogram of physical and mental health variables used in the factor model

Notes: Graphs depict proportion of categories so that they sum up to 100. Labels recoded to order from 'negative' to 'positive' in terms of health outcome, independently of specific wording. For a detailed description on question wording, framing and recoding, see Table A.1.

Table A.1 Overview of module health module from individual questionnaire

Category	Variable	Question
Physical Health Domain		
Physical Function 1	`pf1` (`ple0004`)	$^{(b)}$When you have to climb several flights of stairs on foot, does your health limit you greatly, somewhat, or not at all?
Physical Function 2	`pf2` (`ple0005`)	$^{(b)}$And what about other demanding everyday activities, such as when you have to lift something heavy or do something requiring physical mobility: Does your health limit you greatly, somewhat, or not at all?
General Health	`gh` (`ple0008`)	$^{(c,rev)}$How would you describe your current health?
Bodily Pain	`bp` (`ple0030`)	$^{(a)}$have severe physical pain?
Role Physical 1	`rp1` (`ple0031`)	$^{(a)}$feel that due to physical health problems you achieved less than you wanted to at work or in everyday activities?
Role Physical 2	`rp2` (`ple0032`)	$^{(a)}$feel that due to physical health problems you were limited in some way at work or in everyday activities?
Mental Health Domain		
Stress	`st` (`ple0026`)	$^{(a,e)}$feel rushed or pressed for time?
Mental Health 1	`mh1` (`ple0027`)	$^{(a)}$feel down and gloomy?
Mental Health 2	`mh2` (`ple0028`)	$^{(a,rev)}$feel calm and relaxed?
Vitality	`vt` (`ple0029`)	$^{(a,rev)}$feel energetic?
Role Emotional 1	`re1` (`ple0033`)	$^{(a)}$feel that due to mental health or emotional problems you achieved less than you wanted to at work or in everyday activities?
Role Emotional 2	`re2` (`ple0034`)	$^{(a)}$feel that due to mental health or emotional problems you carried out your work or everyday tasks less thoroughly than usual?
Social Function	`sf` (`ple0035`)	$^{(a)}$feel that due to physical or mental health problems you were limited socially, that is, in contact with friends, acquaintances, or relatives?

Notes: This table shows an overview of the questions used in the factor models to construct the physical and mental health scores. Names of input variables from `pl` dataset in parenthesis.
a: Categorical variable with 5 levels (1: *Always* to 5: *Never*) and time framed as *previous four weeks*.
b: Categorical variable with 3 levels (1: *Greatly* to 3: *Not at all*) and no explicit time frame.
c: Categorical variable with 5 levels (1: *Very good* to 5: *Bad*) and time framed as *currently*.
rev: Categories order reversed so lower values are 'negative' and higher values are 'positive' health outcomes.
e: `st` is not included SF-12 methodology and, for better comparability, also not in the alternative factor model, but displayed here for completeness.

Appendix B

Robustness Checks

B.1 Model specification variation

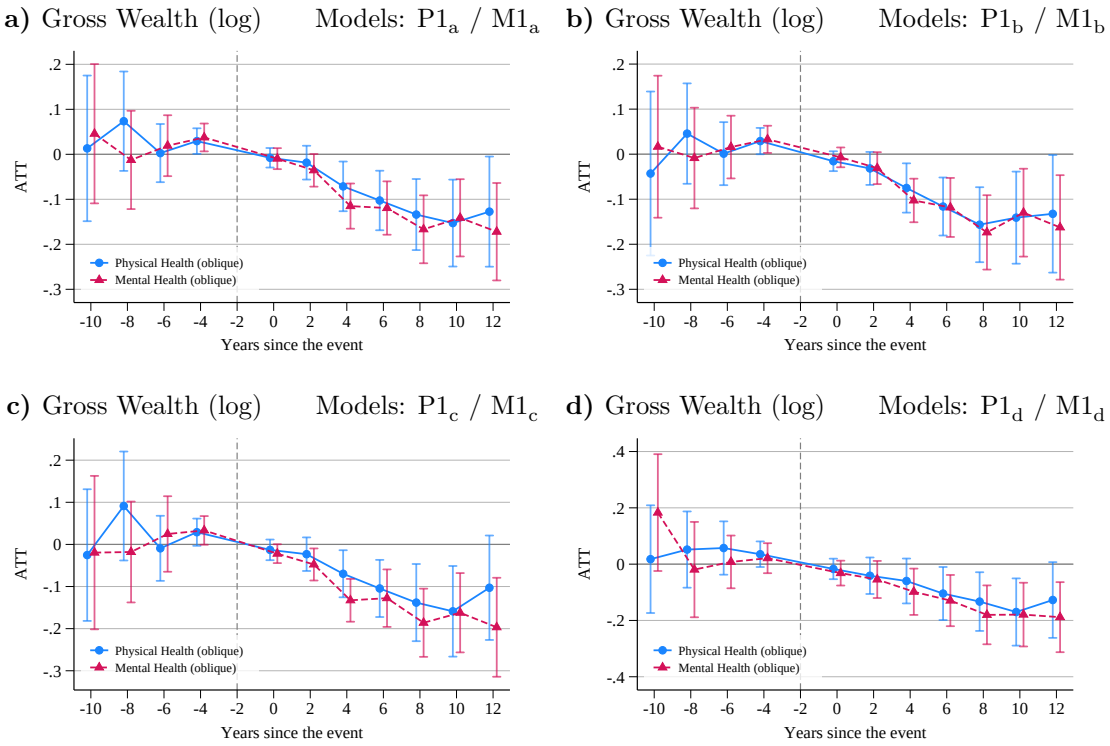


Figure B.1 Varying model specification

Notes: Replications with log($Gross\ Wealth$) as response variable with different specifications to check for model dependence.
Models $P1_a$ and $M1_a$ include not-yet-treated (as well as never-treated) as comparison units.
Models $P1_b$ and $M1_b$ do not include any covariates.
Models $P1_c$ and $M1_c$ include covariates and are weighted by the inverse probability of remaining in the SOEP, to account for sample attrition might have differential impact over the wealth distribution.
Models $P1_d$ and $M1_d$ are balanced, estimated with people present from 2002 to 2020. The coefficients are stable across models and similar to those obtained from the main specifications P1 and M1. The estimated coefficients of the mental health domain are consistently higher than those of the physical domain. Coefficients (and standard errors) already transformed to represent the effect in percentage terms. The same results are also presented in Table B.1.

Table B.1 Varying model specification

	Physical Health Gross Wealth (log)				Mental Health Gross Wealth (log)			
	($P1_a$) (%)	($P1_b$) (%)	($P1_c$) (%)	($P1_d$) (%)	($M1_a$) (%)	($M1_b$) (%)	($M1_c$) (%)	($M1_d$) (%)
SimpleATT	−6.47**	−7.30**	−6.55**	−8.44*	−8.32***	−7.90***	−9.59***	−11.17**
	(2.13)	(2.21)	(2.24)	(3.64)	(1.98)	(2.13)	(2.04)	(3.67)
Pre average	3.02	0.83	2.16	4.13	2.27	1.43	0.51	4.94
	(3.76)	(4.32)	(3.62)	(4.97)	(3.43)	(3.72)	(4.22)	(5.87)
Post average	−8.40**	−9.11**	−8.35**	−8.90*	−10.28***	−9.84***	−11.75***	−11.58**
	(2.64)	(2.67)	(2.92)	(3.52)	(2.39)	(2.67)	(2.43)	(3.60)
$\hat{\theta}_{es}(-10)$	1.33	−4.21	−2.50	1.79	4.68	1.67	−1.93	20.10
	(8.37)	(8.90)	(7.78)	(9.94)	(8.27)	(8.18)	(9.11)	(12.72)
$\hat{\theta}_{es}(-8)$	7.63	4.67	9.55	5.31	−1.24	−0.86	−1.79	−1.92
	(6.07)	(5.95)	(7.23)	(7.28)	(5.50)	(5.65)	(6.00)	(8.48)
$\hat{\theta}_{es}(-6)$	0.27	0.11	−0.93	5.89	1.93	1.59	2.51	0.80
	(3.31)	(3.57)	(3.90)	(5.12)	(3.52)	(3.61)	(4.69)	(4.82)
$\hat{\theta}_{es}(-4)$	2.98*	2.95	2.94	3.58	3.82*	3.35*	3.38	2.14
	(1.49)	(1.53)	(1.70)	(2.40)	(1.64)	(1.59)	(1.80)	(2.78)
$\hat{\theta}_{es}(0)$	−0.77	−1.55	−1.29	−1.69	−0.96	−0.71	−2.18	−3.13
	(1.10)	(1.11)	(1.23)	(1.83)	(1.18)	(1.11)	(1.12)	(2.18)
$\hat{\theta}_{es}(2)$	−1.84	−3.12	−2.30	−4.03	−3.50	−3.06	−4.65*	−5.30
	(1.88)	(1.81)	(1.99)	(3.18)	(1.79)	(1.75)	(1.85)	(3.18)
$\hat{\theta}_{es}(4)$	−6.88*	−7.23**	−6.73*	−5.80	−10.87***	−9.77***	−12.43***	−9.34*
	(2.62)	(2.59)	(2.66)	(3.83)	(2.28)	(2.23)	(2.27)	(3.81)
$\hat{\theta}_{es}(6)$	−9.75**	−10.97***	−9.93**	−9.91*	−11.26***	−11.16***	−11.98***	−12.14**
	(3.05)	(2.93)	(3.12)	(4.33)	(2.68)	(2.97)	(3.06)	(4.08)
$\hat{\theta}_{es}(8)$	−12.53***	−14.49***	−12.90**	−12.45*	−15.35***	−15.94***	−16.97***	−16.48***
	(3.52)	(3.63)	(4.07)	(4.66)	(3.26)	(3.54)	(3.43)	(4.46)
$\hat{\theta}_{es}(10)$	−14.17**	−13.15**	−14.69**	−15.64**	−13.16**	−12.18**	−14.97***	−16.41**
	(4.23)	(4.53)	(4.68)	(5.14)	(3.80)	(4.37)	(4.08)	(4.83)
$\hat{\theta}_{es}(12)$	−11.96*	−12.40*	−9.78	−11.96	−15.80**	−15.01**	−17.86**	−17.17**
	(5.50)	(5.83)	(5.71)	(6.04)	(4.65)	(5.03)	(4.92)	(5.26)
N	90,207	90,207	90,207	22,687	84,925	84,925	84,925	20,010
Unique N	17,581	17,581	17,581	2,371	16,881	16,881	16,881	2,110
Pretrend χ^2 (df)	29 (22)	25.8 (22)	26.6 (22)	32.6 (22)	17.8 (22)	14.3 (22)	17.7 (22)	31.5 (22)
Pretrend p-value	0.146	0.259	0.225	0.068	0.720	0.889	0.726	0.086
Covariates	✓		✓	✓	✓		✓	✓
Inv.Pr(stay)			✓				✓	
Balanced				✓				✓
Not Yet	✓				✓			

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Replications with log($Gross\ Wealth$) as response variable with different specifications to check for model dependence. Wild Bootstrap standard error in parenthesis. Models $P1_a$ and $M1_a$ include not-yet-treated (as well as never-treated) as comparison units.

Models $P1_b$ and $M1_b$ do not include any covariates.

Models $P1_c$ and $M1_c$ include covariates and are weighted by the inverse probability of remaining in the SOEP, to account for sample attrition might have differential impact over the wealth distribution.

Models $P1_d$ and $M1_d$ are balanced, estimated with people present from 2002 to 2020. The coefficients are stable across models and similar to those obtained from the main specifications P1 and M1. The covariates included are gender, age spline, federal state residence, legal disability, marital status, and years of education. Coefficients (and standard errors) already transformed to represent the effect in percentage terms. For a visual representation of the same results see Figure B.1.

Table B.2 Table of results with different treatment rule

	Physical Health				Mental Health			
	(neg)log		level		(neg)log		level	
	gross (%) ($P1_r$)	net (%)[1] ($P2_r$)	gross ($P3_r$)	net ($P4_r$)	gross (%) ($M1_r$)	net (%)[1] ($M2_r$)	gross ($M3_r$)	net ($M4_r$)
SimpleATT	−3.44*	−2.77	−4.07*	−4.47**	−7.36***	−8.40***	−4.52*	−3.97*
	(1.70)	(2.39)	(1.75)	(1.56)	(1.62)	(2.13)	(1.80)	(1.62)
Pre average	1.60	4.74	3.47	2.86	3.00	4.75	1.53	2.16
	(2.24)	(3.41)	(2.12)	(1.92)	(2.42)	(3.12)	(2.73)	(2.26)
Post average	−4.80	−3.25	−6.47*	−6.75**	−9.66***	−10.74***	−6.37*	−5.39*
	(2.49)	(3.63)	(2.70)	(2.44)	(2.17)	(2.91)	(2.83)	(2.31)
$\hat{\theta}_{es}(-10)$	6.04	9.38	5.62	5.11	3.22	5.53	1.15	3.89
	(4.67)	(7.28)	(4.00)	(3.72)	(5.05)	(6.71)	(5.40)	(4.76)
$\hat{\theta}_{es}(-8)$	−1.76	1.73	2.80	1.86	4.50	7.53	1.98	3.48
	(3.27)	(4.85)	(3.20)	(2.86)	(3.47)	(4.81)	(3.54)	(3.12)
$\hat{\theta}_{es}(-6)$	1.22	4.12	4.64*	3.57*	2.41	1.66	1.46	0.19
	(2.09)	(3.04)	(1.97)	(1.71)	(2.10)	(2.89)	(2.22)	(2.03)
$\hat{\theta}_{es}(-4)$	1.03	3.86*	0.81	0.88	1.90	4.37**	1.54	1.09
	(1.12)	(1.63)	(0.95)	(0.86)	(1.13)	(1.50)	(0.97)	(0.86)
$\hat{\theta}_{es}(0)$	−0.69	−2.45	−0.77	−1.19	−2.41**	−2.76*	−1.79*	−1.36*
	(0.77)	(1.34)	(0.74)	(0.73)	(0.80)	(1.16)	(0.71)	(0.69)
$\hat{\theta}_{es}(2)$	−2.56	−3.19	−3.42*	−3.46*	−4.14**	−5.10*	−2.53	−2.49
	(1.58)	(2.50)	(1.51)	(1.39)	(1.53)	(2.28)	(1.51)	(1.42)
$\hat{\theta}_{es}(4)$	−3.94	−1.41	−4.77*	−6.06**	−10.91***	−12.36***	−5.25*	−5.30**
	(2.26)	(3.73)	(2.35)	(2.17)	(2.07)	(2.96)	(2.37)	(2.02)
$\hat{\theta}_{es}(6)$	−6.21*	−3.57	−4.72	−5.79*	−11.42***	−14.55***	−7.17*	−6.24*
	(2.79)	(4.39)	(2.93)	(2.51)	(2.46)	(3.35)	(3.01)	(2.78)
$\hat{\theta}_{es}(8)$	−7.76*	−2.10	−6.49	−5.63	−13.42***	−14.26**	−7.64*	−6.91*
	(3.55)	(5.35)	(3.87)	(3.50)	(3.08)	(4.23)	(3.89)	(3.39)
$\hat{\theta}_{es}(10)$	−5.71	−3.02	−10.44*	−10.35*	−12.58**	−11.91*	−7.71	−6.59
	(4.18)	(6.35)	(5.29)	(4.45)	(3.95)	(4.91)	(5.20)	(4.63)
$\hat{\theta}_{es}(12)$	−6.51	−6.88	−14.66*	−14.77*	−12.13*	−13.50*	−12.48	−8.83
	(6.02)	(7.54)	(6.82)	(5.94)	(4.74)	(6.13)	(6.52)	(5.44)
Obs	90,207	90,207	90,207	90,207	84,925	84,925	84,925	84,925
Pretrend χ^2 (df)	24.8 (26)	18.1 (26)	36.2 (26)	38.9 (26)	23.4 (26)	29.9 (26)	36.6 (26)	31.6 (26)
Pretrend p-value	0.529	0.872	0.088	0.050	0.608	0.272	0.081	0.205

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Results of replicating the primary model with a different assignment rule. Instead of having to experience a negative health outcome at least twice, people experiencing it only once is already assigned to treatment. Further, the threshold is 1 Std. Dev. instead of 1/2 used in the main analysis. In general, the results are similar, but consistently smaller than in the main analysis. The difference is around half to two thirds of estimations in the main results. The only model which is not strongly affected by these parameter changes is the one with gross wealth in logarithms as the response variable, $M1_r$ in the robustness checks and M1 in the main analysis. The covariates included are gender, age spline, federal state residence, legal disability, marital status, and years of education. [1]Coefficients (and standard errors) of log specifications are transformed to represent the effect in percentage terms, but such interpretation of the neglog transformation might be biased (see Appendix A.3) Wild Bootstrap standard error in parenthesis. The main results are presented in Table 5.1 and the corresponding notes on the procedure apply to this table equally.

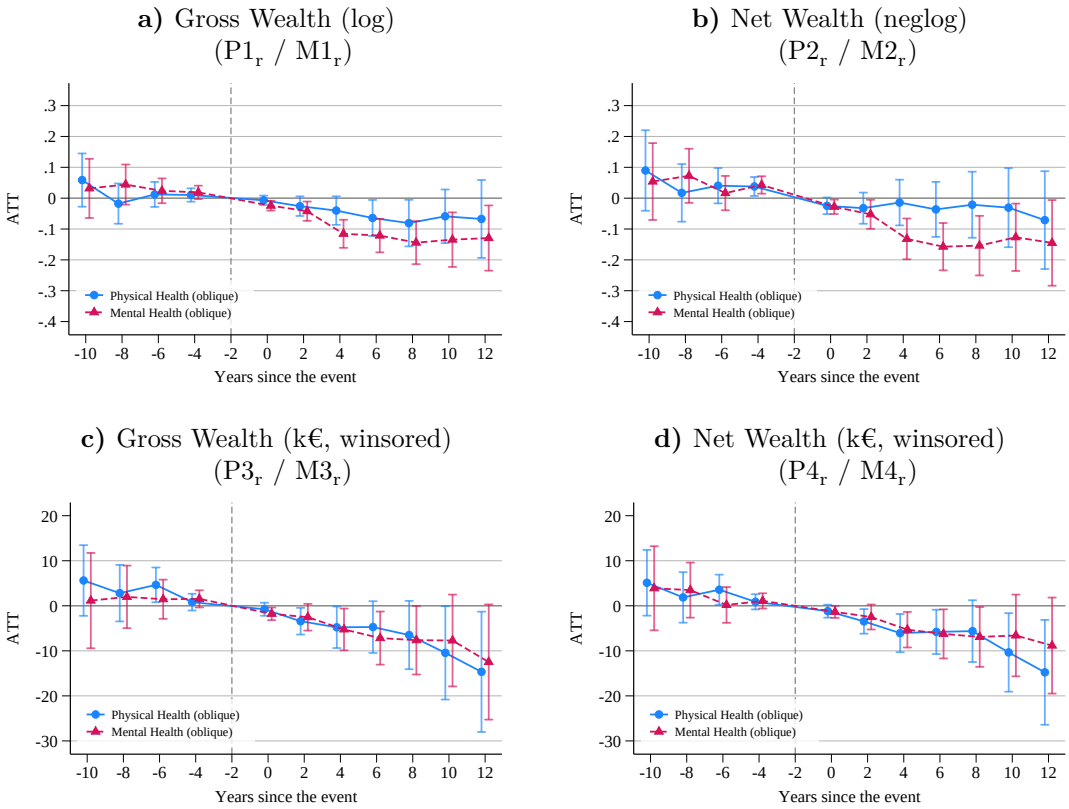


Figure B.2 Event-study results with different treatment assignment rule

Notes: The results corresponding to this panel is presented in Table B.2. The same notes apply to this figure.

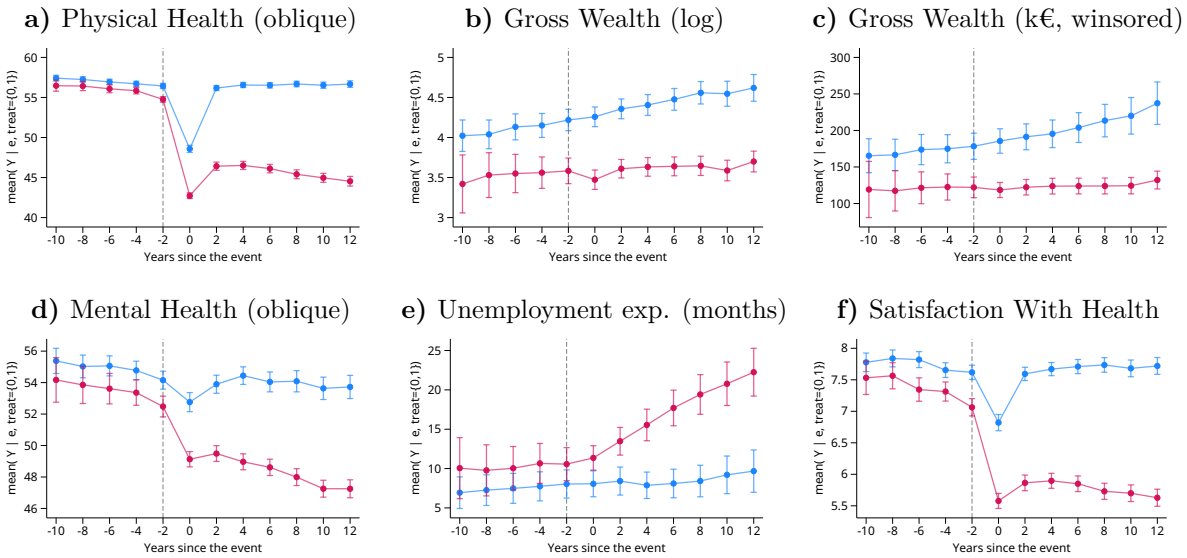


Figure B.3 Development of selected variables from event study (Physical Health Domain)

Panels depict the evolution of selected variables for the treated group (in red) and the untreated (in blue). In panel (a), one can see that, on average, the untreated do experience a strong direct impact at $e = 0$ but it recovers to the previous values in the next period ($e = 2$), while the treated recovers only slightly and carry on on a downward trend. In panel (d), one can see the cross effect on the other health domain. That is, how does the shock in the one health dimension affect the other health dimension. Whiskers depict the 99% confidence interval.

Notes:

- a: The untreated group, by definition, are not assigned a "event year", so their event is set as the year when they experience their worst health outcome, albeit still below the threshold. The actual DiD procedure do not impute any treatment year for the untreated, but for this visualization, one has to set a date or assign randomly and any choice would have some drawbacks. This choice is able depicting the evolution of people on limit of treatment assignment.
- b: The vertical line depict the last period before (imputed) treatment taking place.
- c: To avoid sample composition affecting the interpretation of these figures, the sample is restricted to a balanced panel.

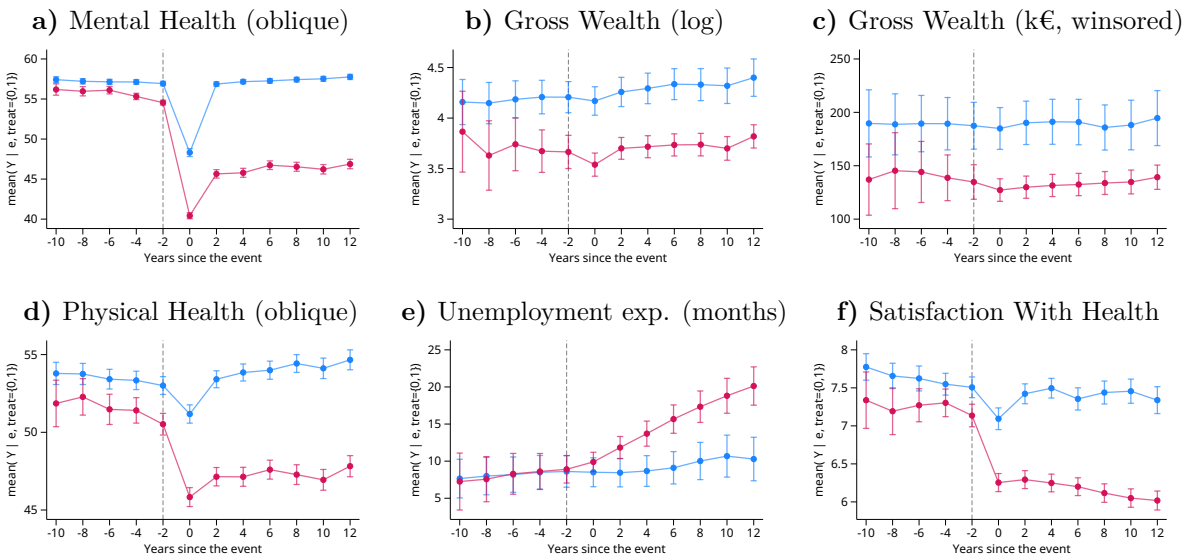


Figure B.4 Development of selected variables from event study (Mental Health Domain)

Notes: points a: b: and c: of Figure B.3 also apply, as well as similar interpretation of each panels. Emphasis here to panel (e), showing a nice example of the effect taking place after not only parallel but equal pre-trend.

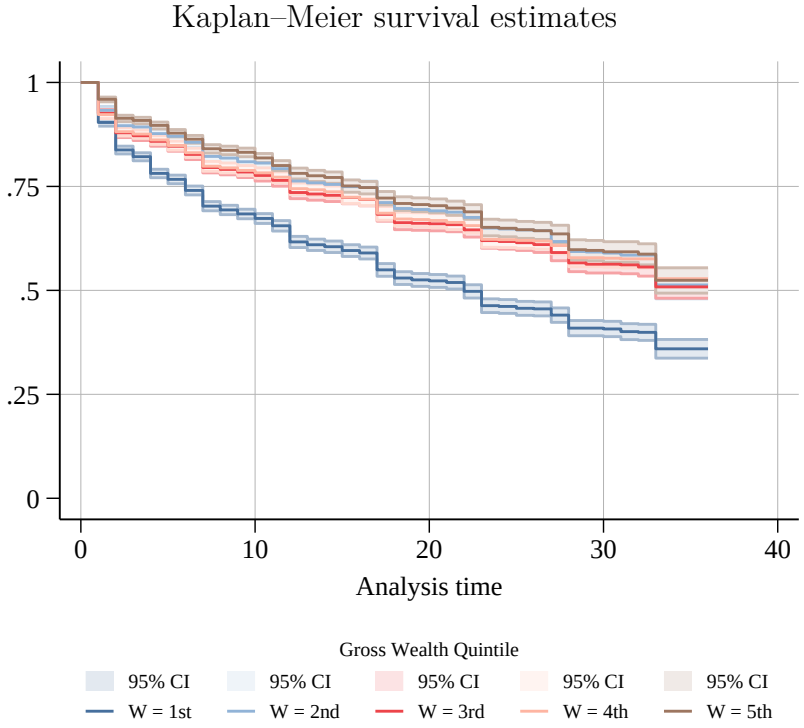


Figure B.5 Survival analysis of SOEP participants by wealth quintiles

Notes: This graph evidentiates the gradient in panel attrition by wealth levels. It depicts an unadjusted survival estimate of participation length for each quintile of age-adjusted gross wealth. While the four upper quintile groups show a similar survival rate, those in the bottom quintile are less likely to remain as long in the SOEP. The graph suggests that the differential attrition rate is stronger in the first 5 years. From then on, the curves evolve parallel to one another. For this analysis, SOEP's entering yeard is used, including if it happened before 2002. Those that dropped out before 2002 (from when on wealth data are available) are not considered for this examination. Data source: SOEPv37.

Appendix C

Replications based on the SF-12 method

Table C.1 Table of coefficients of models based on the SF-12 methodology

	Physical Health				Mental Health			
	(neg)log		level		(neg)log		level	
	gross (%) ($P1_{sf12}$)	net (%)[1] ($P2_{sf12}$)	gross ($P3_{sf12}$)	net ($P4_{sf12}$)	gross (%) ($M1_{sf12}$)	net (%)[1] ($M2_{sf12}$)	gross ($M3_{sf12}$)	net ($M4_{sf12}$)
SimpleATT	−6.12**	−5.78	−8.39***	−7.47***	−10.94***	−14.56***	−5.85*	−5.50**
	(2.10)	(2.88)	(2.27)	(2.03)	(1.93)	(2.45)	(2.29)	(2.06)
Pre average	3.81	4.61	4.65	4.34	9.25*	11.62*	6.61	5.30
	(3.55)	(4.88)	(3.61)	(3.10)	(4.11)	(5.34)	(4.40)	(3.53)
Post average	−8.55**	−7.75*	−10.77***	−9.69***	−12.58***	−17.18***	−7.29*	−6.72*
	(2.69)	(3.43)	(2.93)	(2.59)	(2.40)	(2.99)	(3.10)	(2.66)
$\hat{\theta}_{es}(-10)$	1.63	3.27	7.50	6.36	24.71*	19.85	15.82	15.35*
	(8.42)	(10.85)	(8.31)	(6.69)	(10.76)	(13.79)	(9.54)	(7.41)
$\hat{\theta}_{es}(-8)$	9.54	10.56	5.30	5.86	8.29	15.61	7.26	4.74
	(6.03)	(7.97)	(5.40)	(4.52)	(5.83)	(8.57)	(6.80)	(5.98)
$\hat{\theta}_{es}(-6)$	2.53	2.56	3.09	2.99	2.36	6.34	0.59	−1.09
	(3.35)	(4.61)	(3.28)	(2.80)	(3.39)	(4.48)	(3.85)	(3.21)
$\hat{\theta}_{es}(-4)$	1.74	2.28	2.70*	2.15	3.04	5.35**	2.74	2.20
	(1.57)	(2.28)	(1.22)	(1.18)	(1.63)	(2.06)	(1.46)	(1.30)
$\hat{\theta}_{es}(0)$	−0.76	−1.21	−2.53**	−2.48**	−3.31**	−5.12***	−1.90	−1.64
	(1.16)	(1.72)	(0.87)	(0.79)	(1.05)	(1.48)	(1.00)	(0.93)
$\hat{\theta}_{es}(2)$	−1.04	−1.60	−5.01**	−4.57**	−6.07***	−7.59***	−3.44	−2.96
	(1.86)	(2.64)	(1.68)	(1.60)	(1.62)	(2.21)	(1.92)	(1.75)
$\hat{\theta}_{es}(4)$	−5.91*	−6.89	−8.65***	−7.47***	−12.91***	−15.48***	−6.84**	−7.12**
	(2.50)	(3.60)	(2.48)	(2.22)	(2.24)	(2.81)	(2.65)	(2.39)
$\hat{\theta}_{es}(6)$	−8.04*	−7.62	−9.62**	−8.11**	−16.46***	−21.73***	−7.07*	−7.56*
	(3.04)	(4.09)	(3.27)	(2.91)	(2.55)	(3.38)	(3.39)	(3.09)
$\hat{\theta}_{es}(8)$	−10.67**	−7.57	−11.22**	−9.51**	−18.27***	−22.44***	−7.45	−6.66
	(3.88)	(4.98)	(3.76)	(3.51)	(3.24)	(4.06)	(4.29)	(3.95)
$\hat{\theta}_{es}(10)$	−16.84***	−13.48*	−18.96***	−16.75***	−15.39***	−22.25***	−9.75	−8.58
	(4.20)	(5.78)	(5.17)	(4.70)	(4.01)	(4.78)	(5.42)	(4.84)
$\hat{\theta}_{es}(12)$	−15.30**	−14.97*	−19.40**	−18.95**	−14.62**	−23.61***	−14.56*	−12.53*
	(4.99)	(6.79)	(6.71)	(5.88)	(5.06)	(5.89)	(6.77)	(5.98)
N	92,239	92,239	92,239	92,239	89,112	89,112	89,112	89,112
Unique N	17,943	17,943	17,943	17,943	17,663	17,663	17,663	17,663
Pretrend χ^2 (df)	27.5 (22)	24.5 (22)	37.8 (22)	37.8 (22)	32.8 (22)	27.1 (22)	17.9 (22)	17.6 (22)
Pretrend p-value	0.193	0.322	0.019	0.019	0.065	0.208	0.709	0.732

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This results are a replication of Table 5.1 based on the SF-12 methodology for construction PCS and MCS. Wild Bootstrap standard error in parenthesis. A visual representation of this results are depicted in Figure C.3
[1]Coefficients (and standard errors) of log specifications are transformed to represent the effect in percentage terms, but such interpretation of the neglog transformation might be biased (see Appendix A.3)

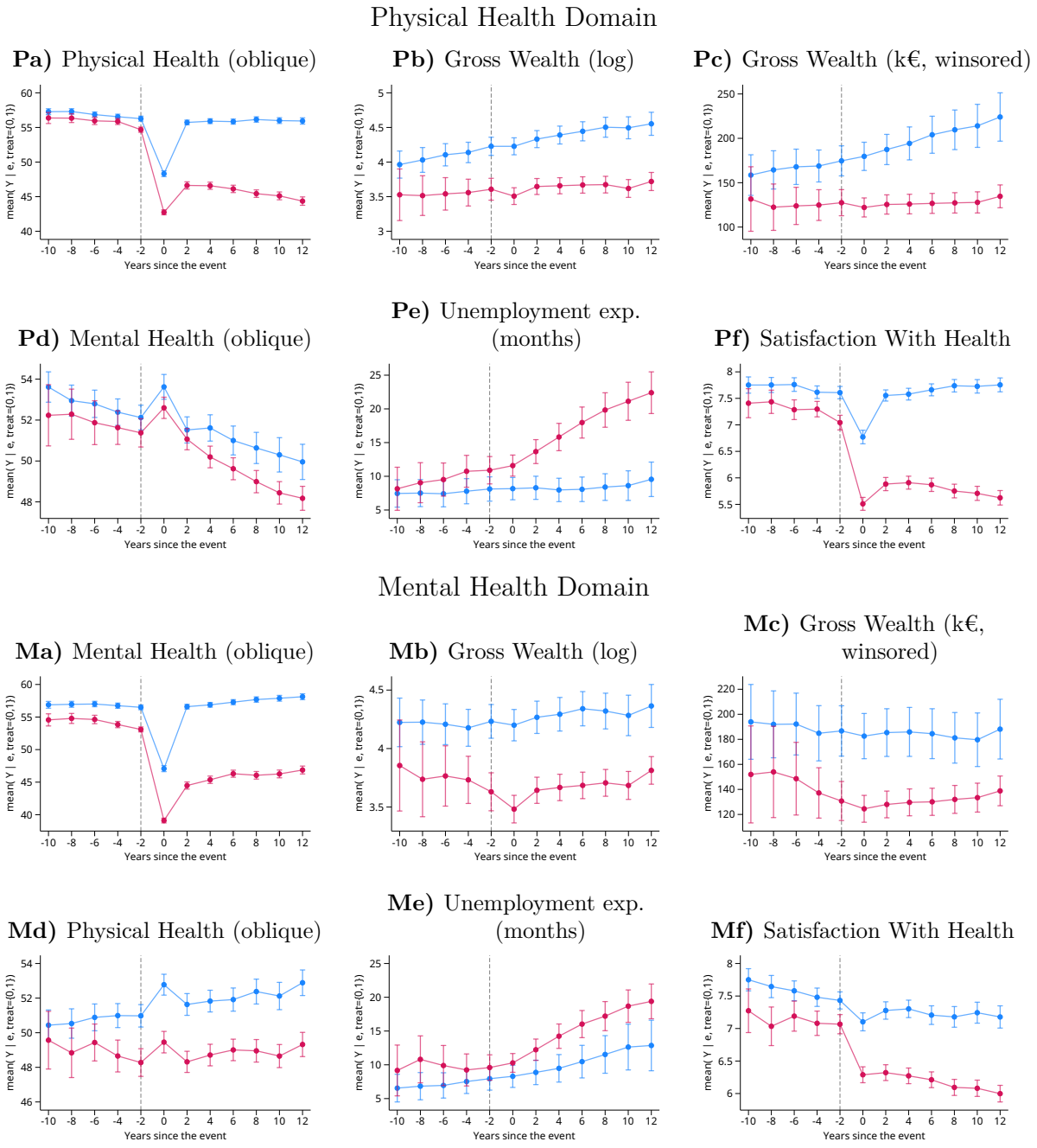


Figure C.1 Development of selected variables based on SF-12 method

Notes: Replication of Figures B.3 and B.4 using PCS and MCS following the SF-12 method. Here one can observe the negative cross-effect in Pd and Md where the adverse outcome in one dimension impacts the other dimension in the opposite direction. The pre-treatment trend, specially in Pe and Me, are less supportive of the parallel trends assumption than in the alternative method. Same considerations regarding the "treatment date" of the untreated stated in the notes of Figure B.3 apply here as well.

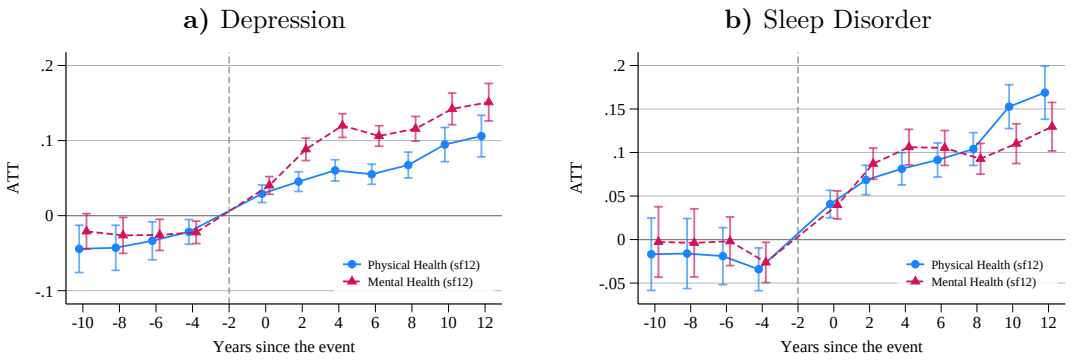


Figure C.2 Replication of validation of mental health diagnoses with orthogonal health scores

Notes: This replication aims to show that also using orthogonal health scores, the effect follows a similar in both domains, similar to using oblique scores. Response variables are binary indicators of being ever diagnosed with the respective condition. Panels have different scale on the y-axis. Whiskers depict the 95% confidence interval

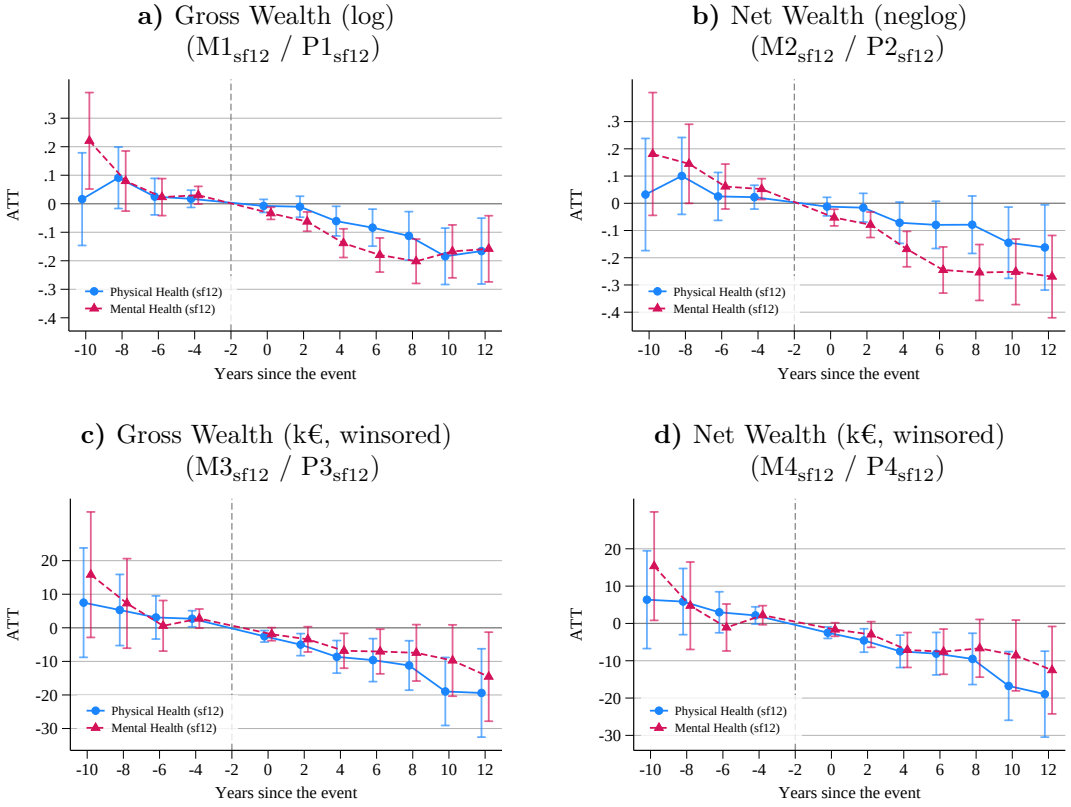


Figure C.3 Replication of main analysis with SF-12 scores

Notes: These panels are a visual representation of the table of coefficients presented in Table C.1. Post-treatment, the effects are in line with those obtained in the main analysis. Pre-treatment, in contrast, seems to be less supportive of the parallel trends assumptions, including in the primary specifications with log-transformed gross wealth ($P1_{sf12}$ and $M1_{sf12}$). To what extend this is driven by the particular treatment rule or whether an artifact from the sf12 or from the alternative scores computes could be a matter of further inquiry. To what extend should be, a priori, expected that the trends remain parallel 8 or 10 years prior to treatment is also worth considering. Whiskers depict the 95% confidence interval.

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Berlin, February 07, 2024

.....

(Marcelo Rainho Avila)