UNIVERSITÀ POLITECNICA DELLE MARCHE

DIPARTIMENTO DI SCIENZE ECONOMICHE E SOCIALI

RETIREMENT AND HEALTH OUTCOMES IN A META-ANALYTICAL FRAMEWORK

MATTIA FILOMENA AND MATTEO PICCHIO

QUADERNO DI RICERCA n. 458

ISSN: 2279-9575

Luglio 2021

Comitato scientifico:

Giulia Bettin
Marco Gallegati
Stefano Staffolani
Alessandro Sterlacchini

Collana curata da:
Massimo Tamberi

Abstract

This paper presents a meta-analysis on the effects of retirement on health. We select academic papers published between 2000 and 2021 studying the impact of retirement on physical and mental health, self-assessed general health, healthcare utilization and mortality. Among 275 observations from 85 articles, 28% (13%) find positive (negative) effects of retirement on health outcomes. Almost 60% of the observations do not provide statistically significant findings. Using meta-regression analysis, we checked for the presence of publication bias after distinguishing among different journal subject areas and, once correcting for it, we find that the average effect of retirement on health outcomes is small and barely significant. We apply several model averaging techniques to explore possible sources of heterogeneity and our results suggest that the different estimated effects can be explained by the differences in both health measurements and retirement schemes.

JEL Class.: I10; J14; J26.

Keywords: Retirement; health; meta-analysis; meta-regression; publication bias.

Indirizzo:

Mattia Filomena. Department of Economics and Social Sciences, Marche Polytechnic University, Ancona, Italy. E-mail: m.filomena@pm.univpm.it.

Matteo Picchio. Department of Economics and Social Sciences, Marche Polytechnic University, Ancona, Italy; Department of Economics, Ghent University, Ghent, Belgium; IZA – Institute of Labor Economics, Bonn, Germany; GLO – Global Labor Organization. E-mail: m.picchio@staff.univpm.it. Tel.: +39 071 220 7176.

Retirement and health outcomes in a meta-analytical framework[†]

Mattia Filomena and Matteo Picchio

1 Introduction

In recent years, the attention to the effects of retirement on workers' physical and mental health has grown considerably, becoming not only a topic of interest in the medical or psychological field, but also among labour and health economists. For the financial sustainability of the pension systems, in most of the OECD countries the standard retirement age has indeed increased and will keep increasing in the future (OECD, 2019). Understanding the health consequences of retirement is of utmost importance to provide policy-makers with a clearer picture for the design of pension policies, labour market reforms, and healthcare investments that are welfare improving.

The identification of the causal health effects of retirement is the crux of this strand of research and involves methodological issues that are not easy to deal with. Kuhn (2018) provides a clear non-technical summary of these methodological issues. First of all, estimation biases due to reverse causality might arise, because causality not only could run from retirement to health, but it is also likely to go from health to retirement decisions. Second, estimation biases could be due to measurement errors when researchers adopt subjective health measures as outcome variables. Indeed, the decision to retire early might influence the reporting subjective answers of the interviewees, because they could assess their own health differently after retirement. This might happen for example because, when people retire, their reference group changes (Johnston and Lee, 2009). To deliver credibly estimates of the causal impact of retirement on health, more recent studies address endogeneity issues through different methodological strategies, especially using

[†]*Mattia Filomena acknowledges financial support from the Cariverona Foundation Ph.D. research scholarship. We thank Jan van Ours for his comments.*

instrumental variables methods or regression discontinuity design (RDD).

Different identification strategies of the causal health effects of retirement could explain different estimates across studies. However, different findings are also justified by other reasons. For example, some recent reviews of the literature suggest that the heterogeneity in the estimated health effects of retirement depends also on the country or countries involved in the studies or the time span considered by the authors or covered by pension reforms. Furthermore, also the degree of freedom in choosing whether and when to retire matters: Bassanini and Caroli (2015), in reviewing the literature on the effect of working on health, find that both being forced to keep on working while one would like to retire and being forced to retire when one would prefer to continue working have similar adverse effects on health. They also find that voluntary retirement often has a positive effect on mental health. They conclude therefore that different findings among studies may be related to the voluntariness of the retirement decisions.[1] Nishimura et al. (2018) investigate the source of differences among different studies by focusing on the methodological aspect and considering 8 recent papers in the economic literature. They conclude that the key factors in explaining different results are the choice of the estimation method and the surveyed countries. They also find that the results are not sensitive to replacing the definition of retirement. van der Heide et al. (2013) summarize 22 longitudinal studies on the health effects of retirement, describing differences in terms of voluntary, involuntary and regulatory retirement and between blue-collar and white-collar workers. While they find strong evidence for retirement having a positive effect on mental health, their review also reveals that contradictory findings emerge when the studies use perceived general health and physical health as outcome variables. Picchio and van Ours (2020) present a selection of most recent studies focusing on differences in set-up, identification strategy, dependent variables, and heterogeneity of the retirement effects. Pilipiec et al. (2020) investigate the empirical evidence on the effects of increasing the retirement age on the health, well-being, and labour force participation of older workers focusing on 19 studies. They find that the evidence of an increase of the retirement age on health and well-being is scarce and inconclusive, because of the heterogeneity of the retirement effect among different groups of workers and between workers far from retirement and older workers closer to the retirement age. Finally, Zulka et al. (2019) focus on the impact of retirement on cognitive functioning by using a sample of 20 studies. They suggest that different

[1] To study the health effects of retirement, Bassanini and Caroli (2015) refer to 14 studies: 5 of them report negative effects of retirement on health.

effects could be due to different types of prior occupation.

Although detailed, the aforementioned literature reviews focus on single aspects of a multifaceted phenomenon (Kuhn, 2018) and their concluding summaries could be deceptive (Stanley et al., 2013). According to Kuhn (2018), a meta-analysis, i.e. a research methodology used to bring together in a systematic way and with a quantitative perspective all the findings from previous studies on a given issue, has the potential to yield significant insights and improvements into the factors that trigger various health effects of retirement. To the best of our knowledge, only van Mourik (2020) takes up this challenge and proposes a meta-analysis on the effects of retirement on several measures of health by collecting 576 results from 61 manuscripts. However, this meta-analysis does not comply with the MAER-NET guidelines (Stanley et al., 2013; Havránek et al., 2020). The analysis is indeed built on a trinomial outcome instead of effect sizes, revealing that 15% of the studies reported negative health effects of retirement, 35% positive health effects, and 50% statistically insignificant results. Furthermore, it includes not only articles published in scientific journals, but also working papers and Ph.D. dissertations. Also Sewdas et al. (2020) provide a meta-analysis, but with a focus limited to the link between mortality and early and on-time retirement. More in detail, using a sample of 25 studies, they estimate a random-effects model in a meta-regression to identify the pooled effects of retirement and to assess the influence of gender, prior health, and demographics. They conclude that early retirement, compared to continued working, is not associated with higher risk of mortality. However, on-time retirement, compared to continued working, is associated with a higher mortality risk, which might reflect the healthy worker effect, i.e. people in the group of those who work beyond the standard retirement age are on average healthier than those who retire on-time. Finally, both Pabón-Carrasco et al. (2020) and Li et al. (2021) only focus on depressive symptoms:[2] according to the former, the retirees with the highest prevalence of depression are those ones who retire in a mandatory fashion or due to illness; the latter show that the association of involuntary retirement with more depressive symptoms is stronger than voluntary or regulatory retirement, and it is more pronounced in Eastern developed countries.

A rigorous and extensive meta-analysis on the subject is lacking. The main contribution of our article is to fill this gap by a meta-analysis on the evidence of the health effects of retirement which: i) follows the MAER-NET guidelines (Stanley et al., 2013;

[2] Pabón-Carrasco et al. (2020) collect a total of 11 articles, while Li et al. (2021) have a sample of 25 longitudinal studies.

Havránek et al., 2020); ii) is based only on articles published in peer reviewed journals, to reduce the probability that they contain mistakes (Xue et al., 2021), and in English, for the sake of correct interpretability (Vooren et al., 2019); iii) does not focus on a particular measure of health, but it rather considers the most frequently used in the literature, such as self-reported general health, physical and mental health, healthcare utilization, and mortality. Our meta-analysis is carried out on 85 articles. It includes the estimation of FAT-PET meta-regression models which allow us to investigate the issue of publication bias and to look for patterns among different study characteristics after correcting the findings for it. We take into account all the main factors that could lead to different estimates of the effect sizes among studies, such as the institutional context, the research design, the causal effect identification strategy, and other study-related characteristics.

The remainder of the paper is organized as follows. Section 2 focuses on the meta-analytical approach, describing the databases used, the research methods and presenting preliminary and descriptive results of our meta-analysis. Section 3 assesses whether there is publication bias in this kind of empirical literature. Section 4 provides heterogeneity analysis by using meta-regressions with the inclusion of covariates on the basis of Bayesian criteria for model selection. Section 5 concludes. The Appendix reports the full list of the studies included in our meta-analysis and their main characteristics.

2 Meta-dataset

2.1 Search strategy and study selection criteria

The empirical literature does not show clear-cut results on the health effect of retirement. Several reasons could explain different findings: different methodologies of analysis, different identification strategies of the causal effect, different countries, different time spans considered by the studies or covered by pension reforms. As such, a simple comparison of the different studies and of their results could be misleading (Stanley et al., 2013). A rigorous meta-analysis would allow us to systematically review the literature by combining the results of multiple and different studies, so as to identify patterns among diverse study results while taking into account the uncertainty behind each point estimate of the relation of interest and remove bias induced by eventual publication biases. The publication bias (also named 'file drawer problem') is the bias arising from the tendency of editors to publish more easily findings consistent with the conventional view or with statistically

significant results, while studies that find small or no significant effects tend to remain unpublished (Card and Krueger, 1995).

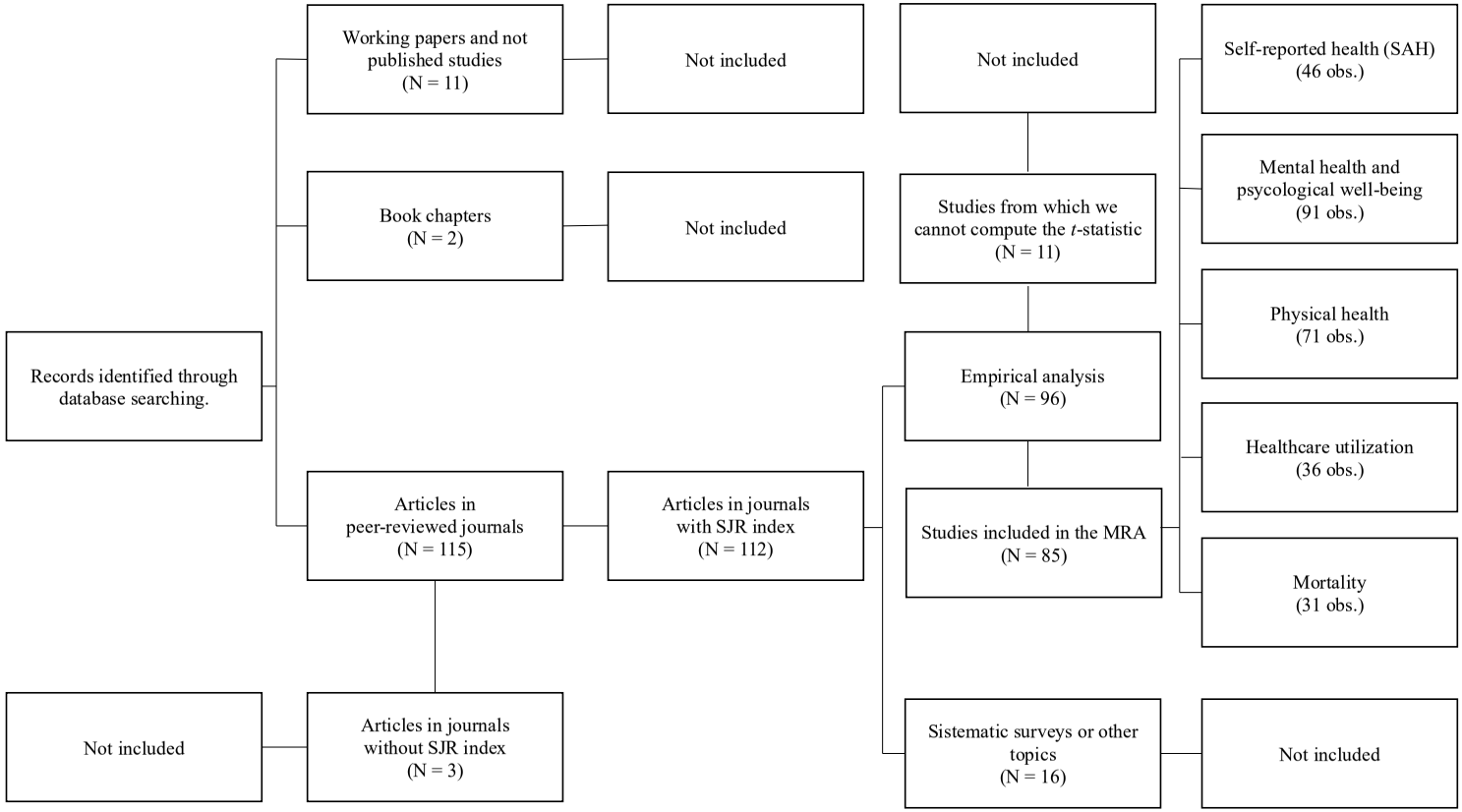
Our search for studies follows the MAER-NET guidelines and was conducted from November 2020 to March 2021 in Ideas/EconPapers, Google Scholar, Scopus and Web of Science by using the following keywords: 'retirement', 'health' and one among 'mental health', 'physical health', 'psychological well-being', 'healthcare' and 'mortality'. We only consider articles published in peer-reviewed journals of health economics, labour economics, social sciences, psychology, and medicine and with the SCImago Journal Rank (SJR) indicator.[3] We excluded theoretical works and studies concerning only cross-partner retirement effects of retiring (Atalay and Zhu, 2018; Bloemen et al., 2019), or general life satisfaction as dependent variable (Abolhassani and Alessie, 2013; Bender, 2012; Horner, 2014; Kesavayuth et al., 2016), or only health behaviours analysis (Evenson et al., 2002; Henkens et al., 2008; Zhao et al., 2017; Motegi et al., 2020). Hence, we selected only micro-level studies on the health effects of retirement. We excluded 11 papers because not published on peer-reviewed journals, i.e. discussion papers (see e.g. Waldron, 2001; Bound and Waidmann, 2007; Coe and Lindeboom, 2008; Lalive and Staubli, 2015; Zulkarnain and Rutledge, 2018), and two book chapters (Charles, 2004; Börsch-Supan and Schuth, 2014). At this point we had 96 articles. Finally, we had to remove 11 articles because they do not contain sufficient information to compute the t-statistic of the estimated retirement effect, on which we will build our meta-regressions.[4] Our final meta-analytic sample is made up of 85 articles, which are listed in Table A.1 in the Appendix. Many studies deal with the retirement effect on multiple health outcomes and some others disaggregate the analysis by gender. In these cases, multiple data points are delivered and our final dataset consists of 275 observations. Figure 1 is a PRISMA flow diagram (Moher et al., 2009): it graphically reports the rules we followed to include/exclude articles in our final sample.

From most of the articles, we directly extracted the estimated retirement effects ($\widehat{\beta}_i$) along which their standard errors ($SE_i(\widehat{\beta}_i)$) and computed the t-statistics as their ratio. In other cases, we could directly retrieve the t-statistics, because reported among the study

[3] See www.scimagojr.com/SCImagoJournalRank.pdf for details on the calculation of the SJR. The following studies were not included in the final sample because their journals are not indexed in SCImago: Lee and Smith (2009), Fonseca et al. (2014), and Son et al. (2020).

[4] These 11 articles are: Allen and Alpass (2020), Barban et al. (2020), Carlsson et al. (2012), Dufouil et al. (2014), Finkel et al. (2009), Fisher et al. (2014), Kühntopf and Tivig (2012), Mazzonna and Peracchi (2012), Nishimura et al. (2018), Olesen et al. (2014), Rohwedder and Willis (2010).

Figure 1: PRISMA flow diagram

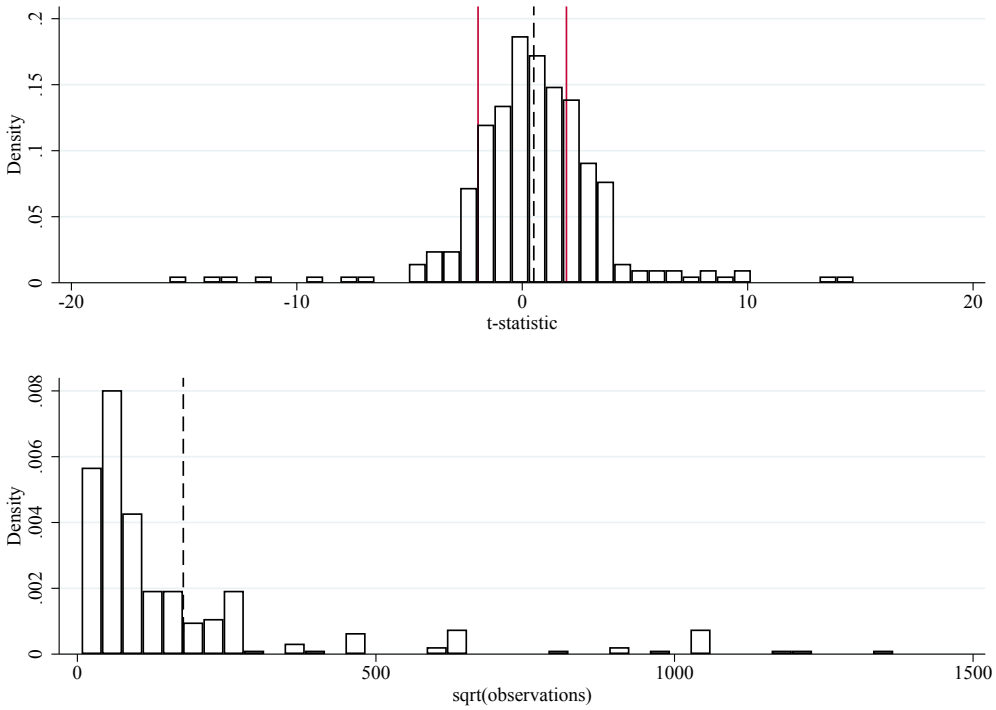


results. Finally, in some studies only the estimated effects and their 95% confidence intervals were displayed. In these cases, we approximated the standard errors in linear models (and then we computed the t-statistics) as follows: $SE_i = (ub - lb)/(2 \times 1.96)$, where ub and lb are the upper bound and the lower bound of the confidence interval, respectively. For studies with non-linear models, such as multinomial logit or Cox proportional hazard models, and reporting only the odds ratio (OR) and its 95% confidence interval, we calculated the standard error as $SE_i = [\ln(ub) - \ln(lb)]/(2 \times 1.96)$ and then the t-statistic as $t_i = [ln(\widehat{\beta}_{1i})\widehat{\beta}_{1i}]/SE_i$.

The health outcomes are quite different between and, sometimes, within studies. In some cases, when the sign of the coefficient of retirement is positive, it means that there is a health improvement, like for general physical health indexes or self-assessed health. In some other cases, it is the negative sign that implies a health improvement, as when mortality or depression are the health outcomes. We manipulated the sign of the t-statistics, so that a "positive" ("negative") sign means a health improvement (deterioration), and all the rest of our analysis is based on this modification of the t-statistics.

Graph a) of Figure 2 shows the distribution of t-statistics, which is quite dispersed, with a minimum of -15.66, a maximum of 14.70, and a standard deviation of 3.27. Most of the findings (58.5%, 161 outcomes) are not significantly different from 0, having a t-statistic smaller than 1.96 in absolute value; in 28.4% (13.1%) of the cases, 78 (36) results, the retirement effect on health is instead significantly positive (negative). Graph b) of Figure 2 plots the distribution of the square root of the observations exploited to estimate the retirement effects. The number of observations is also very heterogeneous with a minimum of 49 and a maximum of 1,866,974. Since in what follows the t-statistics and the number of observations will be used to build a comparable measure of the estimated effect across different studies, the presence of extreme values in these two key variables raises concerns about outliers, especially because the linear models typically used in meta-regressions may be particular sensitive to them (Viechtbauer and Cheung, 2010). As suggested by Xue et al. (2021), who had a similar problem in conducting a meta-analysis on the education effect on health, we moderate the problem by winsorization of t-statistics and number of observations at the top and bottom of their distribution: we replace values that are lower (larger) than the 5th (95th) percentile with the value of the 5th (95th) percentile.

Figure 2: Distribution of t-statistics and observations of study outcomes



Notes: The number of study results is 275. The dashed vertical lines are the sample average of t-statistics in the top graph (0.512) and of the square root of observations in the bottom graph (177.40). The solid vertical lines in the top graph denote the critical values for the 5% significance level in two-tailed tests (± 1.96).

2.2 Descriptive statistics

We provide some basic descriptive statistics of our meta-analytic sample by research findings. We first focus on the publication year and on study quality measures, like the number of citations on average per year and the journal SJR indicator at the time of publication. Table 1 reports the average number of citations per year according to Google Scholar (retrieved on 05/04/2021) and the SJR indicator at the time of publication by research outcome.[5]

Table 1: Descriptive statistics on article citations and SJR

	$t \leq -1.96$	$-1.96 < t < 1.96$	$t \geq 1.96$
a) Number of citations per year on 08/03/2021 (Google scholar)			
Mean	20.247	9.760	11.991
Standard deviation	14.058	9.697	10.613
Minimum	1.000	0.000	1.000
Maximum	49.333	50.600	50.600
b) SJR at the time of publication[(a)]			
Mean	2.260	1.725	1.715
Standard deviation	1.965	1.065	1.150
Minimum	0.365	0.186	0.186
Maximum	7.563	7.563	5.667
Observations	36	161	78

Source: Data retrieved from Google Scholar and Scimago Institutions Rankings on 05/04/2021.
[(a)] At the time of publication, some journals did not have the SJR index yet, either because they were published in too recent years or because the journal was not indexed yet in Scimago. Footnote 5 explains how we deal with these cases of missing information.

The average number of yearly citations is the smallest (9.8) when the null hypothesis of no effect cannot be rejected. It is instead the highest (20.2) when significant negative effects emerge and almost twice as large as the average number of yearly citations of findings supporting significant positive effects (11.9). Differences in the scientific influence of the journals where the articles were published are smaller. In both cases, articles finding negative outcomes display a larger standard deviation. It is noteworthy that statistically insignificant results are not underrepresented in journals of high scientific influence compared to those with more clear-cut findings, but rather they correspond to almost 60% of our sample. This might suggest that, at a first and very descriptive level, the publication bias is not an issue in this research strand.

[5] At the time of publication, some journals did not have the SJR index yet, either because they were published in too recent years or because the journal was not indexed yet in Scimago. In these cases, we assign to the journal the available value of the SJR index which is chronologically closer.

To understand the relation between year of publication and study findings, we report in Table 2 the distribution of the absolute frequencies by year of publication divided in three groups and the research outcomes. Two features are worth of mention: first, the availability of empirical findings has largely increased in the last years; second, study results pointing to negative consequences of retirement on health has become much less frequent in relative terms.

Table 2: Summary statistics on research outcomes over time

	$t \leq -1.96$	$-1.96 < t < 1.96$	$t \geq 1.96$	Total
2000-2009	13	19	16	48
2010-2014	8	33	11	52
2015-2020[(a)]	15	109	51	175

[(a)] In this time frame we also include the first 3 months of 2021.

Table 3 reports descriptive statistics by research outcomes of further variables that we will use as covariates in the meta-regressions to capture the factors underlying the heterogeneous effects in the empirical literature: journal subject area, identification strategy, gender, institutional context, regions and the way in which the t-statistic was calculated. We consider 3 subject areas according to Scimago classification: i) Economics, Econometrics and Finance or Business, Accounting and Management (27% of our observations); ii) Medicine or Psychology (44% of the observations); iii) a residual category containing journals belonging to multiple subject areas (29% of the observations).[6]

Since health is a multidimensional concept, we refer to the main measures analysed in the empirical literature. Among the particular health measures evaluated, positive effect has the largest absolute frequency when we focus on general or self-assessed health. In all the other cases, no statistically significant effect is the prevailing outcome. These different health measures are physical health (26%), mental health (33%),[7] healthcare utilization which includes doctor visits and hospitalization (13%), and mortality (11%).

Identifying the causal effect of retirement on health is not trivial because of several sources of potential endogeneity of the retirement decision, such as reverse causality,

[6]This category comprises even 2 observations by Kalwij et al. (2013), the only article in our sample published in a social-sciences journal.

[7]Physical health includes chronic conditions, mobility, body mass index (BMI), activities of daily living (ADL) and a measure of general physical status. Mental health consists of cognitive functioning, depression or anxiety, and a more general measure which includes general mental health index and psychological well-being (in this case, it also includes happiness as a proxy of well-being).

Table 3: Descriptive statistics of explanatory variables

	$t \leq -1.96$			$-1.96 < t < 1.96$			$t \geq 1.96$		
	Absolute frequencies	Mean	Std. Dev.	Absolute frequencies	Mean	Std. Dev.	Absolute frequencies	Mean	Std. Dev.
Scimago subject areas									
Medicine/Psychology	15	0.417	0.500	72	0.447	0.499	35	0.449	0.501
Economics/Business	12	0.333	0.478	38	0.236	0.426	23	0.295	0.459
Multi area	9	0.250	0.439	51	0.317	0.467	20	0.256	0.439
Health outcomes									
General and self-reported health	6	0.167	0.378	17	0.106	0.308	23	0.295	0.459
Physical health	10	0.278	0.454	47	0.292	0.456	14	0.179	0.386
Mental health	12	0.333	0.478	51	0.317	0.467	28	0.360	0.483
Healthcare utilization	2	0.056	0.232	24	0.149	0.357	10	0.128	0.336
Mortality	6	0.167	0.378	22	0.137	0.345	3	0.038	0.193
Identification strategies									
Regression discontinuity design (RDD)	7	0.194	0.401	33	0.205	0.405	19	0.244	0.432
Instrumental variables (IV)	15	0.417	0.500	66	0.410	0.493	40	0.513	0.503
Difference-in-differences (DiD)	1	0.028	0.167	18	0.112	0.316	4	0.051	0.222
Propensity score matching (PSM)	3	0.083	0.280	14	0.087	0.283	0	0.000	0.000
Fixed-effects/First-differences	7	0.194	0.401	9	0.056	0.230	1	0.013	0.113
Other methods	3	0.083	0.280	21	0.130	0.338	14	0.179	0.386
Institutional contexts									
Mandatory or involuntary retirement	7	0.194	0.401	12	0.074	0.263	5	0.064	0.246
Early retirement	4	0.111	0.319	30	0.186	0.390	8	0.103	0.305
Statutory retirement	22	0.611	0.494	101	0.627	0.485	64	0.820	0.386
Postponed retirement	3	0.083	0.280	18	0.112	0.316	1	0.013	0.113
Geographical areas									
Europe	12	0.333	0.478	78	0.484	0.501	36	0.461	0.502
Extra-European countries	20	0.556	0.504	64	0.398	0.491	32	0.410	0.495
Multi-country analyses	4	0.111	0.319	19	0.118	0.324	10	0.128	0.336
Gender									
Females	6	0.167	0.378	51	0.317	0.467	24	0.308	0.464
Males	12	0.333	0.478	51	0.317	0.467	27	0.346	0.479
Females+Males	18	0.500	0.507	59	0.366	0.483	27	0.346	0.479
Calculation of t-statistic									
t-statistic from $\widehat{\beta}_i / SE_i$	32	0.889	0.319	142	0.882	0.324	63	0.808	0.397
t-statistic from 95% CI or from OR	4	0.111	0.319	19	0.118	0.324	15	0.192	0.397
Observations	36			161			78		

Notes: Both = observations for which authors do not separate estimates for men and women. Other methods = simple OLS regressions or non-linear models, such as multinomial logit and Cox proportional hazard models.

negative self-selection, unobserved heterogeneity,[8] and measurement error.[9] These could affect not only the magnitude, but also the sign of the estimated effect. Hence, we use a set of indicators to control for the methodology used to identify and estimate the impact of retirement on health. The instrumental variables (IV) method is the one that is used more frequently (44%), followed by regression discontinuity design (RDD) (22%). The difference-in-differences (DiD) estimator is mostly used in evaluating policy reforms and represents 8% of our observations. In 14% of the study results, no particular methods is used to tackle the endogeneity of the retirement decision (e.g. linear model, multinomial logit or Cox proportional hazard models).

Some indicator variables are used to capture the institutional context and, in particular, the retirement scheme. The survey of the empirical literature provided by Bassanini and Caroli (2015) highlights the role played by choice vs. constraint in shaping the health impact of work and retirement. They focus on that strand of the literature which studies the voluntariness of retirement and from which evidence of adverse health effects arises when individuals are forced to stop working. In our analysis, we consider both the voluntariness of retirement decisions and its timing: we distinguish among early (15%), on-time (68%), postponed (8%), and mandatory or involuntary retirement (8%).

A further control variable is the gender associated to the estimated effect. The retirement effects could be different for men and women, for example because the career trajectory and the involvement in the labour market are typically different by gender. We will also control for the geographical areas. In particular, we consider results for Europe (46%), for extra-European countries (42%), and from multi-country analyses (12%).

Finally, we also control for the method used to calculate the t-statistics. 86% of our observations are based on t-statistics derived from the ratio between $\widehat{\beta}_i$ and the corresponding standard error. The remaining 14% are derived from 95% confidence intervals or starting from odds ratios (OR).

[8] Omitted variables biases might be induced by differences in unobserved individual characteristics that influence both health and retirement decisions (e.g. subjective life expectancy). Unobserved heterogeneity could be time-constant but also time-varying. To control for unobserved time-constant individual heterogeneity, researchers typically use individual fixed-effects panel data models (Eibich, 2015).

[9] Self-reported health measures are at risk of two kinds of measurement error: i) self-assessed health might not be comparable across individuals ("classical measurement error"); ii) individual who do not work might justify their labour market status by their ill health ("justification bias"). It refers to retirees' tendencies to exaggerate their poor health conditions in order to provide socially acceptable justification for their retirement and observed health would be understated for retirees (Behncke, 2012; Insler, 2014).

2.3 Comparable effect sizes

The estimated retirement effects on health $\widehat{\beta}_i$ are not easily comparable across the models and the estimation techniques generating them. In this topic, we indeed observe a large heterogeneity in the health measures used as outcome variables. For example, the most frequently used are self-reported general health, physical health indexes, like the body mass index (BMI) or the activities of daily living (ADL), mental health measures, as depression or the 5-item mental health inventory (MHI-5), healthcare utilization, and mortality. Their units of measurement are therefore not comparable. Moreover, even when a similar health outcome is used across studies, different model specifications and/or different estimation methods could alter their comparability. To make the effect estimates comparable we compute the partial correlation coefficient r_i, which is commonly used in meta-analyses in economics, business and social sciences since Doucouliagos (1995). A very recent example is Xue et al. (2021), who exploited it in reviewing the effect of education on health.[10]

The partial correlation coefficient is computed as

$$r_i = \frac{t_i}{\sqrt{t_i^2 + dk_i}}, \tag{1}$$

where dk_i is the degrees of freedom in the model from which the i-th t-statistic is derived. Keef and Roberts (2004) show that the estimate of r_i contains a small positive bias, since it increases as the number of independent variables in the regression model increases, i.e. as the degrees of freedom decreases. However, asymptotically this bias disappears. Moreover, since in our meta-dataset many studies do not report precise information about the number of covariates and we cannot therefore recover the degrees of freedom, we replace in Equation (1) dk_i with the number of observations (minus one). Because the smallest number of observations, after the aforementioned winsorization, is 523, this approximation generates a very mild upward bias which asymptotically disappears. The standard error of the partial correlation coefficient is given by

$$SE(r_i) = \sqrt{\frac{1 - r_i^2}{dk_i}}. \tag{2}$$

[10] See Reed (2020) and the meta-analyses cited therein for other examples of meta-analyses using the partial correlation coefficient as effect size.

It can be shown that $r_i/SE(r_i) = t_i$.

The partial correlation coefficient r is a unitless measure, which takes value between -1 and 1. It enables direct comparisons across the different ways of approaching and measuring health outcomes in the empirical literature and in the diverse literatures (Doucouliagos and Laroche, 2009). The partial correlation coefficient drops as the degrees of freedom or the sample size increase. This implies that nearly similar t-statistics will produce very different partial correlations if the sample sizes are diverse: the larger the sample size, the more the effect size measured by the partial correlation is scaled down.

Table 4 displays summary statistics of partial correlations, t-statistics, and number of observations of the full sample and of the results by the type of health measure. Figure 3 shows the scatter plots of the t-statistics and the partial correlation coefficients r with respect to the (natural logarithm of) observations.[11] Two aspects are worth of mention. First, a simple comparison between the graph at the top and the one at the bottom shows how the standardization modifies the t-statistic into the partial correlation coefficient r. Second, the scatter graph at the bottom is a funnel plot of a measure of precision, the square root of the observations, versus a non-standardized effect, the partial correlation coefficients in our case. In the absence of publication bias, the effect should vary randomly around its mean, which is an estimate of the true effect. Hence, the symmetry of the funnel around the mean effect is of help in graphically visualising an eventual publication bias (Stanley, 2005). In our case, the funnel looks roughly symmetric.

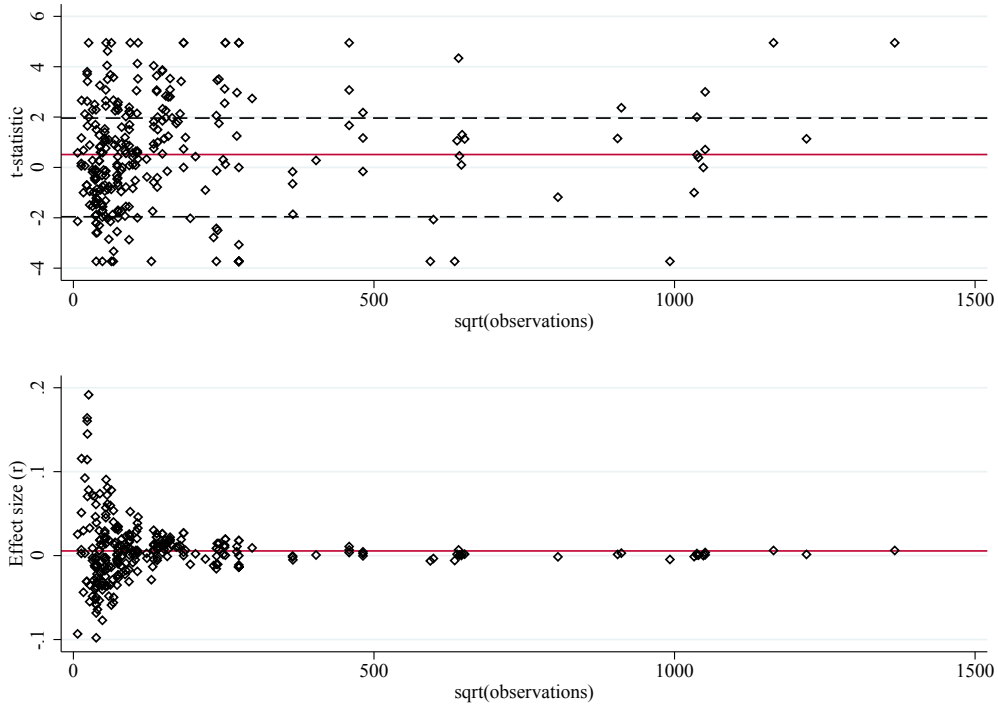
Table 4: Summary statistics of partial correlations, t-statistics, and number of observations by type of health outcome

Outcome variables used as health measures	Absolute frequency	Relative frequency (%)	Average partial correlation (r)	Average t-statistic[(a)]	Average observations[(a)]
Mental health	91	33.1	0.0095	0.6043	13,638
Physical health	71	25.8	0.0058	0.4224	47,394
General and self-reported health	46	16.7	0.0092	1.0967	17,178
Healthcare utilization	36	13.1	-0.0054	0.6828	287,773
Mortality	31	11.3	0.0008	-0.6168	291,287
Total	275	100.0	0.0055	0.5123	90,131

[(a)] These averages are computed before the winsorization.

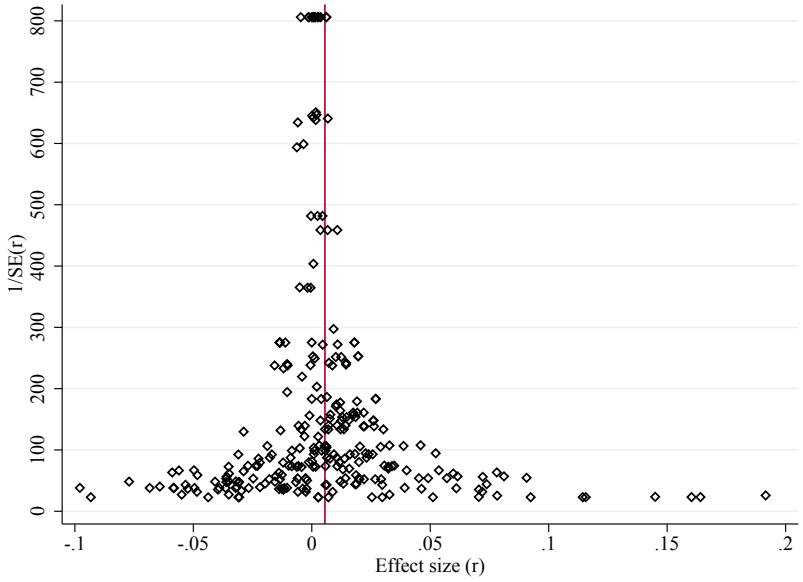
[11] The scatter graph at the top of Figure 3 is named Galbraith plot (Galbraith, 1988). The scatter diagram at the bottom of Figure 3 is known as funnel plot (Light and Pillemer, 1984). The former plots a measure of precision against a standardized effect, the latter against a non-standardized effect.

Figure 3: Scatter plots of t-statistics (top) and partial correlation coefficients r (bottom) versus the square root of observations



Notes: The number of observations is 275. The horizontal continuous red lines are the mean of the t-statistic (0.512) and of partial correlation coefficient r (0.0055). The dashed horizontal lines in the top graph denote the critical values for the 5% significance level in two-tailed tests (± 1.96).

Figure 4: Funnel plot of precision $(1/SE(r))$ versus effect size (r)



Notes: The number of observations is 275. The horizontal line is the mean of the partial correlation coefficient r (0.0055).

To better visualize the eventual presence of publication bias, we plot in Figure 4 the relationship between the partial correlation coefficient and its precision, measured by the inverse of its standard error as defined in Equation (2). The funnel graph shows a mild asymmetry, pointing to a longer tail to the right of the average partial correlation coefficient. It is not easy to arrive to a conclusion about publication bias by way of this graphical approach. Indeed, it relies on the assumption that there is a single 'true' effect common to all empirical studies, so if there is heterogeneity across articles due to different datasets, time spans, countries or methodologies, it might cause the funnel's skewness. In this case, it seems to suggest that there is not an evident publication bias. However, in the next section, on the basis of meta-regression analysis (MRA), we will formally test for the presence of publication bias.

3 Testing for publication bias

To formally assess the relevance of publication bias and to eventually remove it from the estimate of the genuine retirement effect on health, we use the "Funnel Asymmetry

Test – Precision Effect Test" (FAT-PET) (Egger et al., 1997; Stanley, 2005, 2008). It is a standard model to assess the presence of publication bias, used since the end of the 1990s in the economic literature (Card and Krueger, 1995; Ashenfelter et al., 1999; Görg and Strobl, 2001), and based on a simple regression of the i-th effect size on a constant and its standard error:

$$r_i = \gamma_1 + \gamma_0 SE(r_i) + \varepsilon_i, \tag{3}$$

where ε_i is the idiosyncratic error terms and γ_0 will be equal to zero when the effect size r_i varies randomly around the precision effect γ_1, meaning no publication bias. The publication bias is proportional to the inverse of the square root of the sample size, which in turn is proportional to the standard error (Begg and Berlin, 1988). The Funnel Asymmetry Test (FAT) tests the hypothesis of no publication bias (Egger et al., 1997), i.e. $H_0 : \gamma_0 = 0$, and therefore also a test of funnel asymmetry (Sutton et al., 2000). If the null hypothesis is rejected, a publication bias is affecting this strand of the literature, posing a serious problem to the interpretation of the scientific research (Begg and Berlin, 1988). The Precision Effect Test (PET) tests the null hypothesis $H_0 : \gamma_1 = 0$. The rejection of the null hypothesis can be interpreted as the presence of an authentic empirical effect, corrected for publication selection: when the sample size goes to infinity and the standard error goes to 0, the observed effects goes to γ_1 (Stanley, 2008).

Table 5 displays the results of different estimation and specifications of Equation (3). Model (1) reports the ordinary least squares (OLS) estimates of Equation (3), without taking advantage of the known form of heteroskedasticity affecting the distribution of r_i, as seen in Equation (2). This knowledge in instead exploited in Model (2), which displays the results when Equation (3) is estimated by Weighted Least Squares (WLS-FE) using $1/SE(r_i)^2$ as weights. Models (3) and (4) are robustness checks. In Model (3) we replicate our simple FAT-PET estimates by replacing $SE(r_i)$ with the inverse of the square root of the sample size as an alternative precision measure. As the sample size is not subject to estimation error, it avoids errors-in-variables bias that could instead affect $SE(r_i)$. If $SE(r_i)$ is endogenous in Models (1) and (2) because affected by measurement error, we might solve the problem by using an IV approach, instrumenting $SE(r_i)$ with the square root of the number of observations, which is strongly correlated to the standard error but should not be able to explain the estimated effect once we control for the standard error. Finally, in Model (5) we report the results if in Equation (3) we replace $SE(r_i)$ with its square to capture eventual non-linearities: this is the PEESE model which is to

be preferred in correcting for publication bias when a significant publication selection bias is detected (Stanley and Doucouliagos, 2012, 2014). From none of the five models reported in Table 5 we find evidence of publication bias. Furthermore, the FAT-PET point estimates of γ_0, ranging from 0.27 to 0.45, suggest that, if any, the publication bias is positive and small.

Table 5: FAT-PET and PEESE tests for publication bias

	FAT-PET								PEESE	
	(1) OLS		(2) WLS-FE		(3) WLS-FE[a]		(4) FAIVE[b]		(5) WLS-FE	
Publication bias (γ_0)	0.450	(0.486)	0.415	(0.254)	0.416	(0.255)	0.266	(0.319)	11.402	(10.375)
Precision effect (γ_1)	-0.001	(0.005)	0.001	(0.001)	0.001	(0.001)	0.002	(0.002)	0.002*	(0.002)
R^2	0.020		0.021		0.021		0.039		0.009	

Standard errors robust heteroskedasticity and within-study correlation are in parenthesis. The number of observations (studies) is 275 (85).
[a] The inverse of the square root of the sample size is used instead of $SE(r_i)$ as precision measure.
[b] The F-statistic for the power of the excluded instrument is 39.66.

The recognition of publication bias as a threat to the reliability of the scientific knowledge took place in different moments in different disciplines. For example, psychological and medical research has long acknowledged it since the end of the 1950s (Sterling, 1959; Rosenthal, 1979; Begg and Berlin, 1988). The economic research has taken instead some more years, until the 1990s (see e.g. Card and Krueger, 1995; Ashenfelter et al., 1999). Therefore, one might wonder whether researchers and journal editors might have different sensibilities towards the problem across different disciplines, resulting in the publication bias being limited only to some disciplines. To check whether this could be the case, we distinguish the study results in 3 broad subject areas: medicine/psychology, economics/business, and a residual category. Then, we generalize Equation (3) by having one constant per each subject area and the standard error interacted with the subject area indicator and we replicate the estimation of FAT-PET and PEESE models. Equation (4) clarifies how we generalize Equation (3):

$$r_i = \boldsymbol{\gamma}_1 \mathbf{z}_i + \boldsymbol{\gamma}_0 \mathbf{z}_i \times SE(r_i) + \varepsilon_i, \tag{4}$$

where $\mathbf{z}_i$ is a full set of dummies for the subject area of the journal of the i-th study result.[12] Table 6 displays the estimation results of Equation (4). We detect weak evidence for publication bias only in economics/business (the publication bias coefficient is

[12] As such, this equation does not contain the constant term.

significant at the 10%), with the FAIVE estimates in Model (4) suggesting a moderate publication bias in magnitude.

Table 6: FAT-PET and PEESE tests for publication bias by subject area

	FAT-PET								PEESE	
	(1) OLS		(2) WLS-FE		(3) WLS-FE[(a)]		(4) FAIVE		(5) WLS-FE	
Publication bias in economics/business	0.459	(0.366)	0.523	(0.343)	0.524	(0.339)	0.731**	(0.023)	31.030*	(0.089)
Publication bias in medicine/psychology	0.640	(0.563)	0.334	(0.449)	0.336	(0.447)	0.189	(0.794)	8.509	(0.698)
Publication bias in multi-area	0.252	(0.601)	0.233	(0.663)	0.234	(0.661)	0.007	(0.990)	2.093	(0.883)
Precision effect in economics/business	0.002	(0.792)	-0.001	(0.871)	-0.001	(0.863)	-0.002	(0.707)	0.000	(0.956)
Precision effect in medicine/psychology	-0.005	(0.635)	0.001	(0.530)	0.001	(0.546)	0.002	(0.430)	0.001	(0.272)
Precision effect in multi-area	0.001	(0.827)	0.003*	(0.042)	0.003*	(0.044)	0.005*	(0.055)	0.004**	(0.026)
R^2	0.048		0.083		0.083		0.036		0.081	

We report in parentheses wild cluster bootstrap p-values obtained from the wild cluster bootstrap-t procedure proposed by Cameron et al. (2008), with clusters at study level (5,000 bootstraps using the Webb's (2014) six-point distribution as weights). We report wild cluster bootstrap p-values to take into account that, in each subject area, the number of clusters is small (from 16 to 36). The number of observations (studies) is 275 (85), 70 (16) in economics/business, 122 (36) in medicine/psychology, and 83 (33) in the residual multi-area category.

[(a)] The inverse of the square root of the sample size is used instead of $SE(r_i)$ as precision measure.

Finally, the precision coefficient is significant only for multi-area journals and equal to 0.004, suggesting a positive effect of retirement on health only in studies in this category. This effect is however fairly low, considering that, according to Cohen (1988), a partial correlation coefficient of 0.1 is to be considered as "small" and in the analysis of Doucouliagos (2011), who focused on economic results, it should be at least 0.07 to be considered as "small".[13]

In recent years, further techniques have been developed to detect and correct the publication bias. Ioannidis et al. (2017) proposed the Weighted Average Adequately Powered (WAAP) method, which restricts the meta-analysis only to those study results with statistical power of at least 80%, i.e. those estimates with standard errors smaller than the WLS-FE precision effect divided by 2.8. The simulation results in Bom and Rachinger (2019) showed however that the WAAP could be counterproductive if the true effect is very small or publication bias is severe. They proposed the Endogenous Kink (EK) meta-regression model which performs better than WAAP when the true effect is relatively small. The EK method attempts to better fit the non-linearity of the relationship between the estimated effect and its standard error in the presence of publication bias by means of a piecewise linear model instead of a quadratic term, under the assumption that publication selection is triggered only when the standard error exceeds an endogenous cut-off value a. Below this threshold, estimates are sufficiently significant so that marginal increases in standard errors do not induce publication selection. Once determined the cut-off value

[13] In Doucouliagos (2011), 0.17 is the threshold for "moderate" and 0.33 for "large".

a, the EK meta-regression model consists in estimating by WLS the following equation using $\frac{1}{SE(r_i)^2}$ as weights:

$$r_i = \gamma_1 + \delta \cdot [SE(r_i) - a] \cdot \mathbb{1}\{SE(r_i) \geq a\} + \varepsilon_i, \tag{5}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function which returns 1 if the argument is true and 0 otherwise. We estimated the Bom and Rachinger's (2019) EK model, both for the full sample and by subject category. In all the cases we found that $a < 0$. As pointed out by Bom and Rachinger (2019), when this happens, a must be set to 0 and the EK method collapses to the usual FAT-PET estimated by WLS-FE.[14]

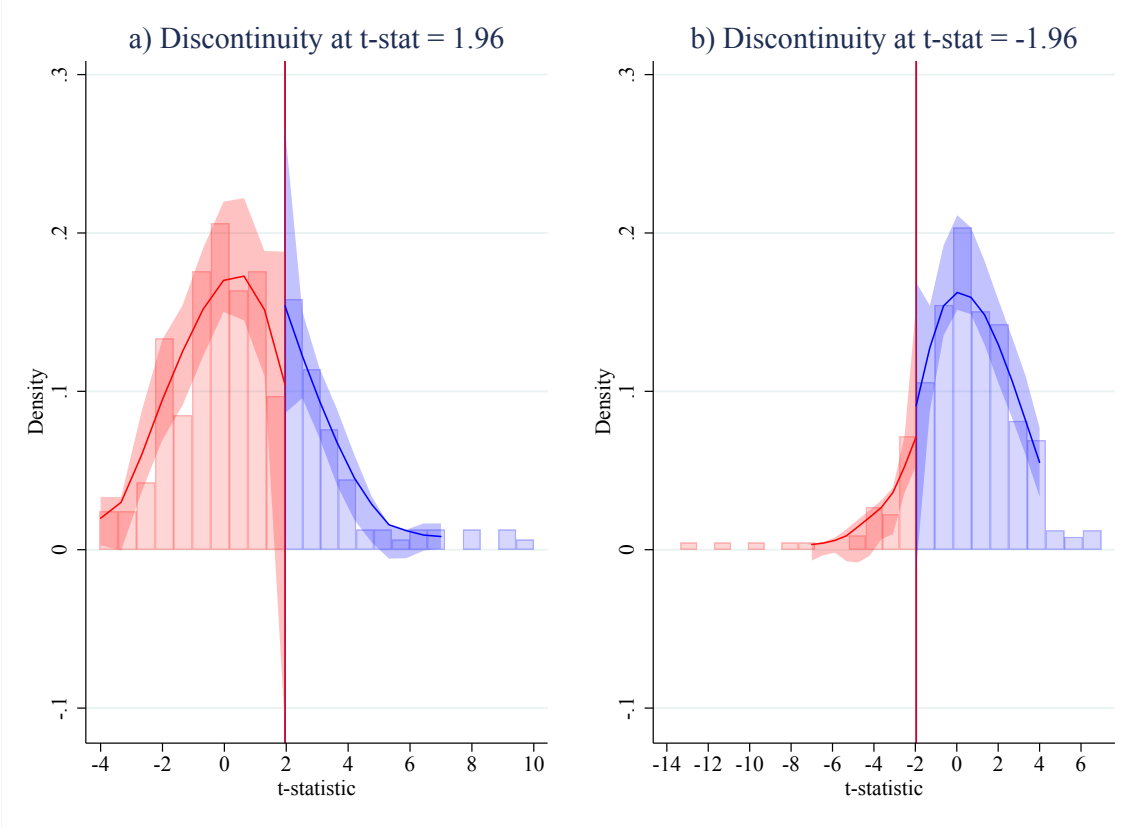
Finally, Andrews and Kasy (2019) suggest to consider the distribution of p-values or t- or z- statistics across published studies. Indeed, if there is no publication bias, the distribution of the t-statistics and p-values across studies should not display discontinuities, especially at critical values like ± 1.96 for the former and 0.05 for the latter. We assess whether the density of the t-statistic is discontinuous at ± 1.96 using the nonparametric local-polynomial density estimator proposed by Cattaneo et al. (2018, 2021). Figure 5 displays the local polynomial density estimates focusing on the discontinuity at $+1.96$ in graph a) and -1.96 in graph b). Although visually the density shows a positive jump at $+1.96$, expected in case of positive publication bias, this is not significantly different from zero. Indeed, the robust bias-corrected test proposed in Cattaneo et al. (2018) cannot reject the null hypothesis of the absence of discontinuity, with a p-value equal to 0.127 in graph a) and 0.608 in graph b). If we split the sample by subject area, we realize that the visual jump at $+1.96$ in graph a) of Figure 5 is induced by studies in medicine and psychology (p-value=0.061).

We check if the distribution of p-values are discontinuous at the 0.05 cutoff by running two-sided binomial tests for the null hypothesis of equal mass in equal size windows around the cutoff. Table 7 reports the result of these binomial tests starting from a window size of 0.005 at each side of the cutoff and by increasing the window length by subsequent steps of 0.005. We cannot reject the null hypothesis, implying that we do not find statistical evidence of publication bias at the p-value cutoff of 0.05.[15]

[14] In our analysis this is due to the fact that the average effect if very close to 0 and we therefore conclude that no particular correction is needed to refine the FAT-PET or FAT-PET-PEESE approaches. Our endogenous computations of the a's are not reported for the sake of brevity. They can be asked upon request from the authors.

[15] We arrive at the same conclusion after splitting the sample by subject area and per each subject area.

Figure 5: Publication bias tests based on t-statistic density discontinuity at $+1.96$ (graph a)) and -1.96 (graph b)



The solid lines are the local polynomial density estimate of the running variable described in Cattaneo et al. (2018) with local polynomial of order 2. The shaded areas around the lines are robust bias-corrected 95% confidence intervals. The vertical lines are ± 1.96 cutoffs. The null-hypothesis of no discontinuity cannot be rejected with a p-value equal to 0.127 in graph a) and 0.608 in graph b).

Table 7: Two-sided binomial tests of equal sample size at each side of p-value=0.05

Window	Two-sided binomial test p-value	Observations ≤ 0.05	Observations > 0.05
[0.045, 0.055]	0.227	8	3
[0.040, 0.060]	0.057	11	3
[0.035, 0.065]	0.263	13	7
[0.030, 0.070]	0.087	19	9

In conclusion, after a battery of tests to detect publication bias, we find some fairly moderate evidence in economics and business when using FAT-PET FAIVE or PEESE meta-regressions. In the next meta-regressions aimed at understanding eventual heterogeneity across studies in the retirement effect on health, we control for publication bias by subject area using the PEESE approach.

4 Multivariate meta-regressions

To detect possible sources of heterogeneous effects of retirement on health, we include into the PEESE specification a series of covariates: measures of health, methods to identify the effect, institutional contexts, geographical areas, gender, year of publication, SJR index, the average number of Google scholar citations per year, and the way in which we derived the t-statistics. We employ the PEESE specification because it has the smallest bias and it easily accommodates systematic heterogeneity and complex and differential publication bias related to study characteristics. The quadratic form of the standard errors of the PEESE approach has been proven to be less biased and often more efficient than the FAT-PET specification when there is a non-zero genuine effect (Stanley and Doucouliagos, 2014).[16]

Formally, we estimate by WLS-FE the following equation for our effect size

$$r_i = \boldsymbol{\gamma}_1 \mathbf{z}_i + \boldsymbol{\gamma}_0 \mathbf{z}_i \times SE(r_i)^2 + \boldsymbol{\beta}_1 \mathbf{x}_i + \varepsilon_i, \tag{6}$$

which is equivalent to estimating by OLS the transformed model

$$\frac{r_i}{SE(r_i)} = \boldsymbol{\gamma}_1 \frac{\mathbf{z}_i}{SE(r_i)} + \boldsymbol{\gamma}_0 \mathbf{z}_i \times SE(r_i) + \boldsymbol{\beta}_1 \frac{\mathbf{x}_i}{SE(r_i)} + \frac{\varepsilon_i}{SE(r_i)}, \tag{7}$$

where $\mathbf{x}_i$ is the vector containing the study characteristics and $\mathbf{z}_i$ is the same set of dummies for the subject area as used in Equation (4).

A problem in estimating Equation (7) is related to the model uncertainty about which variables should be included. We overcome it by employing one of the most commonly used tools in meta-analysis, the Bayesian Model Averaging (BMA). BMA takes into account all possible models by running many regressions with different subsets of control

[16] Table A.2 in the Appendix displays the results of FAT-PET specification. The findings are very similar to the ones from the PEESE model.

variables and computing the weighted averages of the estimated coefficients. The weights are Posterior Model Probabilities (PMP) and are related to the goodness of fit of each model. The sum of PMPs indicates the Posterior Inclusion Probability (PIP) for each regressor, which provides the information on how likely the regressor is to belong to the true specification. A PIP above 0.5 for a given regressor is usually used as a rule of thumb to include it into the final model (Eicher et al., 2011). For each covariate, BMA returns the posterior coefficient distribution, which gives us the posterior mean (PM) of the regression coefficient and the posterior standard deviation (PSD).

We use the BMA estimator discussed by Magnus et al. (2010), who introduced the distinction among two subsets of explanatory variables. The first subset is the set of "focus" regressors, which are those we want in the model because of theoretical (or other) reasons. In our case, the focus variables are those capturing the publication bias and the precision effect by journal subject area. The second subset is the set of "auxiliary" regressors, which are additional covariates that could be relevant in explaining the estimated effect, but this is not certain. Since we have 20 auxiliary covariates, the number of possible models to be considered is 2^{20}. BMA proceeds by applying conventional noninformative priors on the focus variables and the error variance σ^2, and an informative multivariate Gaussian prior on the auxiliary ones.

In a subsequent step, we perform a model-average procedure by using the Weighted Average Least Squares (WALS) (Magnus et al., 2010). WALS is in an intermediate position between the Bayesian approach of BMA and the frequentist model-averaging. It is indeed a Bayesian combination of frequentist estimators (Magnus and De Luca, 2016). WALS uses conventional noninformative priors on the focus regressors and the error variance σ^2 and a distribution with zero mean for the independent and identically distributed elements of the t-ratios associated with linear combinations of the auxiliary regressors.[17] Unlike BMA, WALS relies on preliminary orthogonal transformations of the auxiliary regressors and their parameters, which reduce the computational burden from 2^{20} to 20. For this reason, WALS does not allow to compute the PIPs. An auxiliary covariate is considered to be robustly correlated with the outcome variable if the t-ratio of its coefficient is greater than 1 in absolute value or, equivalently, if the corresponding one-standard error band does not include zero (De Luca and Magnus, 2011). The advantage of WALS over BMA is that it does not impose an *ad hoc* assumption on the prior on the model space (in

[17] The prior distribution of the t ratios can be either a neutral Laplace prior (Magnus et al., 2010), or a neutral Subbotin prior distribution (Einmahl et al., 2011).

general BMA uses a uniform prior assigning equal probability to each model), but it is theoretically based (Magnus and De Luca, 2016).

Finally, as in Havranek et al. (2015) and Xue et al. (2021), we conduct a frequentist check by estimating Equation (7) by OLS after restricting the set of regressors to those with PIP > 0.5 according to BMA. We also run the same frequentist check after the WALS estimates.

Table 8 reports the estimation results. For the BMA, we show the estimated posterior means, the posterior standard deviations and the posterior inclusion probabilities of each regressor. For the WALS, we include the results coming from two different assumptions about the model prior distributions. In the last columns of Table 8, we present the findings from the frequentist checks.

As concerns the focus regressors, while for these variables the Posterior Inclusion Probabilities from BMA model are not informative, both WALS and OLS estimates reveal a barely significant publication bias in economics/business, even after controlling for a set of covariates. According to BMA results, there are 8 auxiliary covariates which are relevant in explaining the heterogeneous effects of retirement on health (PIP > 0.5). In addition to all the different health outcome measures, these variables are: fixed-effects/first-difference estimator, mandatory (or involuntary) retirement, t-statistic calculated from $\widehat{\beta}_i/SE_i$ and year of publication. WALS results are similar, although some further covariates seem to be important: postponed retirement, the SJR indicator, estimates not distinguishing between males and females, RD design, and PSM estimator.

All models reveal that the studies which use general and self-reported health indicators or mental health measures are the most likely to report positive effects of retirement on health. Studies focusing on physical health or healthcare utilization are more likely to find positive effects than those dealing with mortality, although the magnitude of the positive association is smaller. These findings reflect the results from some earlier systematic surveys in this field: as pointed out by Bassanini and Caroli (2015) or suggested by Nishimura et al. (2018) after re-estimating previous analyses, most of the evidence concerning the health effects of retirement move towards a positive impact both on physical and mental dimensions of health, a better self-assessed health, and lower healthcare utilization.

The results for the identification strategy suggest that the heterogeneity across this dimension is not so important in explaining different findings. We find that only those studies using a fixed-effects or a first-differences approach are more likely to report neg-

Table 8: Heterogeneity in the estimated effects of retirement on health

	BMA[a]			WALS ($q=1$)[b]		WALS ($q=0.5$)[b]		OLS check after BMA[c]			OLS check after WALS[d]		
	PM	PSD	PIP	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.		p-value	Coeff.		p-value
Focus regressors													
Scimago subject areas													
Publication bias in economics/business	**33.610**	**15.984**	**1.000**	**31.931**	**16.139**	**31.230**	**16.378**	31.299	*	0.059	28.398		0.108
Publication bias in medicine/psychology	**9.509**	**10.191**	**1.000**	8.793	10.042	9.072	10.142	8.686		0.743	11.635		0.625
Publication bias in multiarea	**0.634**	**10.770**	**1.000**	0.956	10.660	0.624	10.705	-0.749		0.951	2.481		0.830
Precision effect in economics/business	**0.005**	**0.004**	**1.000**	**0.015**	**0.006**	**0.014**	**0.006**	0.005		0.142	0.012		0.112
Precision effect in medicine/psychology	**0.006**	**0.004**	**1.000**	**0.016**	**0.005**	**0.016**	**0.005**	0.005		0.167	0.012	**	0.032
Precision effect in multiarea	**0.008**	**0.004**	**1.000**	**0.017**	**0.005**	**0.017**	**0.005**	0.007	*	0.057	0.012	**	0.014
Auxiliary regressors													
Google scholar citations per year	0.000	0.000	0.100	0.000	0.000	0.000	0.000	–		–	–		–
Scimago Journal Ranking	0.000	0.000	0.110	**-0.002**	**0.001**	**-0.002**	**0.001**	–		–	-0.002		0.259
Year of publication	**0.001**	**0.000**	**0.930**	**0.001**	**0.000**	**0.001**	**0.000**	0.001	**	0.043	0.001	**	0.015
Health outcomes (reference category: Mortality)													
General and self-reported health	**0.013**	**0.003**	**1.000**	**0.011**	**0.002**	**0.011**	**0.002**	0.015	***	0.000	0.015	***	0.000
Physical health	**0.002**	**0.002**	**0.530**	**0.003**	**0.001**	**0.003**	**0.001**	0.004	**	0.018	0.004	**	0.041
Mental health	**0.010**	**0.003**	**1.000**	**0.007**	**0.002**	**0.007**	**0.002**	0.011	***	0.000	0.011	***	0.000
Healthcare utilization	**0.003**	**0.002**	**0.810**	**0.003**	**0.001**	**0.003**	**0.001**	0.004	***	0.004	0.004	**	0.040
Identification strategies (reference category: Other methods)													
Regression discontinuity design (RDD)	0.000	0.001	0.110	**-0.007**	**0.005**	**-0.007**	**0.005**	–		–	-0.002		0.503
Instrumental variables (IV)	0.000	0.001	0.110	-0.005	0.005	-0.004	0.005	–		–	–		–
Difference-in-differences (DiD)	0.000	0.002	0.130	-0.002	0.004	-0.001	0.004	–		–	–		–
Propensity score matching (PSM)	-0.001	0.004	0.080	**-0.012**	**0.009**	**-0.011**	**0.010**	–		–	-0.012		0.293
Fixed-effects/First-differences	**-0.008**	**0.005**	**0.830**	**-0.013**	**0.005**	**-0.013**	**0.005**	-0.010	*	0.093	-0.010	**	0.015
Institutional contexts (reference category: Statutory retirement)													
Mandatory or involuntary retirement	**-0.027**	**0.009**	**0.960**	**-0.021**	**0.007**	**-0.022**	**0.008**	-0.028	**	0.021	-0.028	**	0.020
Early retirement	0.000	0.001	0.110	-0.001	0.002	-0.001	0.002	–		–	–		–
Postponed retirement	-0.001	0.003	0.270	**-0.006**	**0.002**	**-0.006**	**0.002**	–		–	-0.005	**	0.046
Geographical areas (reference category: Multi-country analyses)													
Europe	0.000	0.001	0.060	-0.002	0.003	-0.002	0.003	–		–	–		–
Extra-European countries	0.000	0.001	0.070	-0.002	0.003	-0.002	0.003	–		–	–		–
Sex (reference category: Males)													
Females	0.000	0.001	0.150	0.001	0.001	0.001	0.001	–		–	–		–
Males+Females	-0.001	0.002	0.310	**-0.004**	**0.003**	**-0.005**	**0.003**	–		–	-0.004	*	0.076
Calculation of t-statistic (reference category: from 95% CI or from OR)													
t-statistic from $\widehat{\beta}_i / SE_i$	**-0.006**	**0.004**	**0.780**	**-0.005**	**0.004**	**-0.005**	**0.004**	-0.007		0.100	-0.008		0.119

Notes: The results are from the PEESE specification by using the inverse of the SE_i^2 as weights. PM = Posterior Mean of the coefficient; PSD = Posterior Standard Deviation; PIP = Posterior Inclusion Probability. The number of observations (studies) is 275 (85), 70 (16) in economics/business, 122 (36) in medicine/psychology, and 83 (33) in the residual multi-area category. Auxiliary variables for which the PIP is above 0.5 in BMA or the corresponding one-standard error band does not include zero in WALS are in bold. *** Significant at 1%, ** significant at 5%, * significant at 10%.

[a] In the BMA, we use the uniform distribution for model priors, the Zellner's g prior for the distributions of the coefficients and a Markov Chain Monte Carlo algorithm to search over the model space, by distinguishing between focus and auxiliary regressors.

[b] $q = 1$ indicates the Laplace model prior distribution; $q = 0.5$ implies the Subbotin model prior distribution.

[c] The model specification under "OLS" includes those variables which have a PIP > 0.5 in BMA ($R^2 = 0.35$).

[d] The second model specification under "OLS" includes those variables which are relevant according to WALS ($R^2 = 0.38$). For both OLS checks, we report wild cluster bootstrap p-values obtained from the wild cluster bootstrap-t procedure proposed by Cameron et al. (2008), with clusters at study level (5,000 bootstraps using the Webb's (2014) six-point distribution as weights).

ative effects on health. This finding contrasts with the one in Nishimura et al. (2018), who instead showed that the choice of the estimation strategy is one of the key factors in explaining why the estimated results of the retirement effect on health differ.

One of the most relevant factors in explaining heterogeneous estimated effects of retirement on health is the institutional context and the retirement scheme: mandatory retirement has a PIP close to 1 and the greatest negative effect in magnitude. Following the WALS results and although with a lower magnitude, also studies focusing on postponed retirement are associated with a lower chance of detecting positive retirement effects than studies dealing with early or statutory retirement. These findings confirm the conclusions in Bassanini and Caroli (2015), who showed that being forced to work while preferring to retire and, symmetrically, being forced to stop working because of workers have no control on the retirement and work decisions have a health damaging effect. Similar results are provided by Pabón-Carrasco et al. (2020) and Li et al. (2021), but limited to the effects on depressive symptoms. Moreover, the negative impact of postponed retirement on health, compared to statutory retirement, could reflect the consequences of being stuck in employment while one had planned to retire, for example due to pension reforms which increase the retirement age or the length of the contribution period required to be entitled to pension (see e.g. Blake and Garrouste, 2019; Shai, 2018).

About the publication year, we find that the estimated effects of retirement on health tend to be more and more over time: the year of publication presents a PIP $= 0.93$ and a positive and significant coefficient. As concerns study-quality measures (the average number of citations per year retrieved from Google Scholar and the SJR indicator) and the way in which the t-statistic was calculated, they do not play significant roles in explaining result heterogeneity. Finally, the health effects of retirement are independent on geographical areas and sex, although the studies which do not distinguish between males and females tend to find a slightly smaller estimated effect if we look at WALS results.

The results presented in Table 8 suggest sources of heterogeneity in the study results. However, it is not easy to visualise from it if for particular combinations of study feature the expected retirement effect is significantly positive or significantly negative. To be more informative from this point of view, we use the OLS estimates from the frequentist check after BMA and computed the expected partial correlation coefficients for particular combinations of the covariates, after fixing the publication year to the median (2017) and setting to zero γ_0, so as to mimic the absence of publication bias.

Table 9 displays the expected partial correlation coefficients for interesting combina-

tions of the explanatory variables. In panel a) we use the 8 most frequent combinations of covariates, sum the corresponding OLS estimated coefficients, and test if this sum is zero. These 8 most frequent combinations of covariates encompass 181 observations (out of 275), i.e. 66% of our sample. We find that for these combinations, when the analyzed outcome variable is mental health, independently on the subject area of the study, retirement has a positive and highly significant impact, with a partial correlation coefficient between 0.008 and 0.010. According to the classifications in Cohen (1988) or Doucouliagos (2011), which set to 0.1 and 0.07 the size of the partial correlation coefficient to be considered as "small", the detected magnitude is very modest. The largest impact is detected in studies published in multi-area journals when they focus on general and self-reported health (0.014). Finally, when it comes to physical health, the predicted average effect for the chosen combinations of covariates is still positive, but smaller and not significant at the usual 5% statistical level.

Table 9: Expected partial correlation coefficients of the health effect of retirement for particular combinations of covariates

	Coeff.		Std. Err.	*p*-value	Frequencies Abs.	Frequencies Rel. (%)
a) Most frequent combinations of covariates						
Economic/business + mental health + t from $\widehat{\beta}_i/SE_i$	0.008	***	0.002	0.001	24	8.73
Economic/business + physical health + t from $\widehat{\beta}_i/SE_i$	0.001		0.002	0.494	25	9.09
Medicine/psychology + mental health + t from $\widehat{\beta}_i/SE_i$	0.008	***	0.002	0.001	35	12.73
Medicine/psychology + healthcare utilization + t from $\widehat{\beta}_i/SE_i$	0.001		0.001	0.108	23	8.36
Medicine/psychology + physical health + t from $\widehat{\beta}_i/SE_i$	0.002		0.002	0.348	17	6.18
Multi-area + mental health + t from $\widehat{\beta}_i/SE_i$	0.010	***	0.002	0.000	21	7.64
Multi-area + general and self-reported health + t from $\widehat{\beta}_i/SE_i$	0.014	***	0.002	0.000	16	5.82
Multi-area + physical health + t from $\widehat{\beta}_i/SE_i$	0.004	*	0.002	0.097	20	7.27
b)						
Economic/business + mental health + mandatory or involuntary retirement + t from $\widehat{\beta}_i/SE_i$	-0.020	**	0.003	0.024	5	1.82
Economic/business + physical health + fixed-effects/first-differences + t from $\widehat{\beta}_i/SE_i$	-0.009	***	0.009	0.003	4	1.45
Economic/business + mental health + fixed-effects/first-differences + t from $\widehat{\beta}_i/SE_i$	-0.002		0.003	0.505	2	0.73
Medicine/psychology + mental health + mandatory or involuntary retirement + t from $\widehat{\beta}_i/SE_i$	-0.020	**	0.009	0.025	3	1.09
Medicine/psychology + mental health + fixed-effects/first-differences + t from $\widehat{\beta}_i/SE_i$	-0.002		0.004	0.551	2	0.73
Medicine/psychology + healthcare utilization + mandatory or involuntary retirement + t from $\widehat{\beta}_i/SE_i$	-0.026	***	0.008	0.002	4	1.45
Multi-area + mental health + fixed-effects/first-differences + t from $\widehat{\beta}_i/SE_i$	0.000		0.004	0.977	2	0.73
Multi-area + general and self-reported health + fixed-effects/first differences + t from $\widehat{\beta}_i/SE_i$	0.004		0.004	0.318	2	0.73
Multi-area + physical health + fixed-effects/first differences + t from $\widehat{\beta}_i/SE_i$	-0.007	**	0.003	0.049	2	0.73
Multi-area + physical health + mandatory or involuntary retirement + t from $\widehat{\beta}_i/SE_i$	-0.024	***	0.009	0.007	2	0.73

Notes: *** Significant at 1%, ** significant at 5%, * significant at 10%. Year of publication is normalized at its median value (2017) and γ_0 is set to zero.

Panel b) of Table 9 shows the predicted partial correlation coefficients for the same combinations of covariates in panel a), with the difference that: i) we also focus on the cases in which retirement is mandatory or involuntary or, alternatively, a fixed-effects/first differences estimation strategy was used; ii) we report only predictions of covariate com-

binations for which in the actual dataset we observe at least 2 observations, to avoid out of sample extrapolations. These further combinations consist in 28 observations (10% of the sample). We find that, regardless the journal subject area and the health measurement, when a study focuses on mandatory or involuntary retirement we predict an expected negative effect between -0.020 and -0.026. When fixed-effects or first-differences estimators were used, we predict an expected partial correlation coefficient still negative, but lower in magnitude and statistically significant in economics/business and multi-area journals only when the outcome variable is physical health.

5 Conclusions

To the best of our knowledge, this article is the first attempt to systematically and analytically summarize the empirical findings on the impact of retirement on health outcomes by following the MAER-NET guidelines (Stanley et al., 2013; Havránek et al., 2020). Our meta-sample was made up of 275 observations from 85 articles published on peer-reviewed journals in the period 2000-2021. Among these findings, 28% supported the hypothesis according to which retirement improves health, 59% provided no statistically significant effects, and only 13% reported evidence in favour of a worsening health status after retirement.

In a first step, using a battery of meta-regression techniques, we checked for the presence of publication bias. After distinguishing the study results among journal subject areas according to the Scimago classification, we detected weak evidence of publication bias only in economics and business. After correcting for publication bias, we found that the estimated average retirement effect on health is positive, significant only at the 10% level, and very little, considering the figures suggested by Cohen (1988) or Doucouliagos (2011) to value a partial correlation coefficient as "small".

Then, we used model averaging strategies to explore possible sources of effect heterogeneity across several study characteristics, like research design, estimation strategy, and institutional context. Our results suggest that the different reported estimates are linked to the differences in health outcomes used across studies. The identification/estimation strategy does not appear to be so relevant in explaining heterogeneous findings, although studies which opted for fixed-effects or first-differences tend to report more negative estimated effects. Finally, a further source of finding heterogeneity is the type of retirement scheme: compared to on-time retirement, mandatory or involuntary retirement and,

although lower significance, also postponed retirement are associated to more negative health outcomes.

These findings have important implications for public policy, especially because many countries are considering rising further their retirement age (OECD, 2019). Although our analysis suggests that retirement improves several health dimensions, having no choice about the timing of retirement, being involuntarily retired or being forced to continue working due to policy reforms which postpone the time of retirement might have health damaging implications. For these reasons, policy-makers should consider not only the financial sustainability of the pension system, but also the raising healthcare spending due to the negative impact of mandatory, involuntary, or postponed retirement. Welfare optimal pension policies should ensure workers a greater degree of freedom in choosing whether to retire and its timing, rather than increasing the retirement age or the requirements to be entitled to pension benefits. In summary, a trade-off seems to arise: while a greater voluntariness about retirement and its timing has to be in line with public budget constraints, at the same time the financial sustainability of the pension system cannot ignore that retirement appears to be health improving in some cases.

Finally, as suggested by Kuhn (2018), there are reasons to suspect that the health effects of retirement could be heterogeneous across dimensions, such as different types of prior occupation (e.g. blue vs. white collar workers), different types of physically/mentally demanding previous jobs, time horizons or health behaviours, which we did not investigate in our meta-analysis. The studies in our meta-analysis indeed only rarely focused on effect heterogeneity across these additional dimensions. Future research should take them into account to have a clearer picture on the multifaceted nature of the retirement effects on health.

References

Abolhassani, M. and R. Alessie (2013). Subjective well-being around retirement. *De Economist 161*(3), 349–366.

Allen, J. and F. M. Alpass (2020). Trajectories of material living standards, physical health and mental health under a universal pension. *Journal of Epidemiology and Community Health 74*(4), 362–368.

Andrews, I. and M. Kasy (2019). Identification of and correction for publication bias. *American Economic Review 109*(8), 2766–94.

Apouey, B. H., C. Guven, and C. Senik (2019). Retirement and unexpected health shocks. *Economics & Human Biology 33*, 116–123.

Ardito, C., R. Leombruni, D. Blane, and A. d'Errico (2020). To work or not to work? The effect of higher pension age on cardiovascular health. *Industrial Relations: A Journal of Economy and Society 59*(3), 399–434.

Ashenfelter, O., C. Harmon, and H. Oosterbeek (1999). A review of estimates of the schooling/earnings relationship, with tests for publication bias. *Labour Economics 6*(4), 453–470.

Atalay, K. and G. F. Barrett (2014). The causal effect of retirement on health: new evidence from Australian pension reform. *Economics Letters 125*(3), 392–395.

Atalay, K., G. F. Barrett, and A. Staneva (2019). The effect of retirement on elderly cognitive functioning. *Journal of Health Economics 66*, 37–53.

Atalay, K. and R. Zhu (2018). The effect of a wife's retirement on her husband's mental health. *Applied Economics 50*(43), 4606–4616.

Bamia, C., A. Trichopoulou, and D. Trichopoulos (2008). Age at retirement and mortality in a general population sample: the Greek EPIC study. *American Journal of Epidemiology 167*(5), 561–569.

Barban, N., X. De Luna, E. Lundholm, I. Svensson, and F. C. Billari (2020). Causal effects of the timing of life-course events: age at retirement and subsequent health. *Sociological Methods & Research 49*(1), 216–249.

Barrett, G. F. and M. Kecmanovic (2013). Changes in subjective well-being with retirement: assessing savings adequacy. *Applied Economics 45*(35), 4883–4893.

Bassanini, A. and E. Caroli (2015). Is work bad for health? The role of constraint versus choice. *Annals of Economics and Statistics 119/120*, 13–37.

Bauer, A. B. and R. Eichenberger (2018). Worsening workers' health by lowering retirement age: the malign consequences of a benign reform. *Journal of the Economics of Ageing 18*, 100296.

Begg, C. and J. Berlin (1988). Publication bias: a problem in interpreting medical data. *Journal of the Royal Statistical Society. Series A (Statistics in Society) 151*(3), 419–463.

Behncke, S. (2012). Does retirement trigger ill health? *Health Economics 21*(3), 282–300.

Belloni, M., E. Meschi, and G. Pasini (2016). The effect on mental health of retiring during the economic crisis. *Health Economics 25*(S2), 126–140.

Bender, K. A. (2012). An analysis of well-being in retirement: the role of pensions, health, and 'voluntariness' of retirement. *The Journal of Socio-Economics 41*(4), 424–433.

Bertoni, M. and G. Brunello (2017). Pappa ante portas: the effect of the husband's retirement on the wife's mental health in Japan. *Social Science & Medicine 175*, 135–142.

Bianchini, L. and M. Borella (2016). Retirement and memory in Europe. *Ageing & Society 36*(7), 1434–1458.

Binh Tran, D. and V. Zikos (2019). The causal effect of retirement on health: understanding the mechanisms. *Australian Economic Review 52*(4), 427–446.

Blake, H. and C. Garrouste (2019). Collateral effects of a pension reform in France. *Annals of Economics and Statistics 133*, 57–86.

Bloemen, H., S. Hochguertel, and J. Zweerink (2017). The causal effect of retirement on mortality: Evidence from targeted incentives to retire early. *Health Economics 26*(12), e204–e218.

Bloemen, H., S. Hochguertel, and J. Zweerink (2019). The effect of incentive-induced retirement on spousal retirement rates: evidence from a natural experiment. *Economic Inquiry 57*(2), 910–930.

Bom, P. R. and H. Rachinger (2019). A kinked meta-regression model for publication bias correction. *Research synthesis methods 10*(4), 497–514.

Bonsang, E., S. Adam, and S. Perelman (2012). Does retirement affect cognitive functioning? *Journal of Health Economics 31*(3), 490–501.

Bonsang, E. and T. J. Klein (2012). Retirement and subjective well-being. *Journal of Economic Behavior & Organization 83*(3), 311–329.

Börsch-Supan, A. and M. Schuth (2014). Early retirement, mental health, and social networks. In *Discoveries in the Economics of Aging*, NBER Chapters, pp. 225–250. University of Chicago Press.

Bound, J. and T. Waidmann (2007). Estimating the health effects of retirement. Working Papers wp168, University of Michigan, Michigan Retirement Research Center.

Bozio, A., C. Garrouste, and E. Perdrix (2021). Impact of later retirement on mortality: evidence from France. *Health Economics 30*(5), 1178–1199.

Brockmann, H., R. Müller, and U. Helmert (2009). Time to retire–time to die? A prospective cohort study of the effects of early retirement on long-term survival. *Social Science & Medicine 69*(2), 160–164.

Butterworth, P., S. C. Gill, B. Rodgers, K. J. Anstey, E. Villamil, and D. Melzer (2006). Retirement and mental health: analysis of the Australian national survey of mental health and well-being. *Social Science & Medicine 62*(5), 1179–1191.

Calvo, E., N. Sarkisian, and C. R. Tamborini (2013). Causal effects of retirement timing on subjective physical and emotional health. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences 68*(1), 73–84.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics 90*(3), 414–427.

Card, D. and A. B. Krueger (1995). Time-series minimum-wage studies: a meta-analysis. *The American Economic Review 85*(2), 238–243.

Carlsson, S., T. Andersson, K. Michaëlsson, D. Vågerö, and A. Ahlbom (2012). Late retirement is not associated with increased mortality, results based on all Swedish retirements 1991–2007. *European Journal of Epidemiology 27*(6), 483–486.

Carrino, L., K. Glaser, and M. Avendano (2020). Later retirement, job strain, and health: evidence from the new state pension age in the United Kingdom. *Health Economics 29*(8), 891–912.

Cattaneo, M., M. Jansson, and X. Ma (2018). Manipulation testing based on density discontinuity. *Stata Journal 18*(1), 234–261.

Cattaneo, M., M. Jansson, and X. Ma (2021). lpdensity: Local polynomial density estimation and inference. *Journal of Statistical Software*, forthcoming.

Celidoni, M., C. Dal Bianco, and G. Weber (2017). Retirement and cognitive decline. A longitudinal analysis using SHARE data. *Journal of Health Economics 56*, 113–125.

Celidoni, M. and V. Rebba (2017). Healthier lifestyles after retirement in Europe? Evidence from SHARE. *The European Journal of Health Economics 18*(7), 805–830.

Charles, K. K. (2004). Is retirement depressing?: Labor force inactivity and psychological well-being in later life. In S. E. Polachek (Ed.), *Accounting for worker well-being*, Volume 23 of *Research in Labor Economics*, pp. 269–299. Emerald Group Publishing Limited.

Che, Y. and X. Li (2018). Retirement and health: evidence from China. *China Economic Review 49*, 84–95.

Chung, S., M. E. Domino, and S. C. Stearns (2009). The effect of retirement on weight. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences 64*(5), 656–665.

Coe, N. B. and M. Lindeboom (2008). Does retirement kill you? Evidence from early retirement windows. IZA Discussion Paper No. 3817, Institute of Labor Economics (IZA), Bonn, Germany.

Coe, N. B., H.-M. von Gaudecker, M. Lindeboom, and J. Maurer (2012). The effect of retirement on cognitive functioning. *Health Economics 21*(8), 913–927.

Coe, N. B. and G. Zamarro (2011). Retirement effects on health in Europe. *Journal of Health Economics 30*(1), 77–86.

Cohen, J. (1988). *Statistical Power Analysis in the Behavioral Sciences*. Hillsdale: Lawrence Erlbaum Associates.

Dave, D., R. I. Rashad, and J. Spasojevic (2008). The effects of retirement on physical and mental health outcomes. *Southern Economic Journal 75*(2), 497–523.

Dayaram, K. and A. McGuire (2019). Retirement reforms: occupational strain and health. *Industrial Relations: A Journal of Economy and Society 58*(3), 522–542.

De Luca, G. and J. R. Magnus (2011). Bayesian model averaging and weighted-average least squares: equivariance, stability, and numerical issues. *Stata Journal 11*(4), 518–544(27).

Doucouliagos, C. (1995). Worker participation and productivity in labor-managed and participatory capitalist firms: a meta-analysis. *ILR Review 49*(1), 58–77.

Doucouliagos, H. (2011). How large is large? Preliminary and relative guidelines for interpreting partial correlations in economics. Working Papers SWP 2011/5, Deakin University, Department of Economics.

Doucouliagos, H. and P. Laroche (2009). Unions and profits: a meta-regression analysis. *Industrial Relations: A Journal of Economy and Society 48*(1), 146–184.

Dufouil, C., E. Pereira, G. Chêne, M. M. Glymour, A. Alpérovitch, E. Saubusse, M. Risse-Fleury, B. Heuls, J.-C. Salord, M.-A. Brieu, et al. (2014). Older age at retirement is associated with decreased risk of dementia. *European Journal of Epidemiology 29*(5), 353–361.

Egger, M., G. D. Smith, M. Schneider, and C. Minder (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal 315*(7109), 629–634.

Eibich, P. (2015). Understanding the effect of retirement on health: mechanisms and heterogeneity. *Journal of Health Economics 43*, 1–12.

Eicher, T. S., C. Papageorgiou, and A. E. Raftery (2011). Default priors and predictive performance in Bayesian model averaging, with application to growth determinants. *Journal of Applied Econometrics 26*(1), 30–55.

Einmahl, J., J. Magnus, and K. Kumar (2011). On the choice of prior in Bayesian model averaging. Discussion Paper 2011-003, Tilburg University, Center for Economic Research.

Evenson, K. R., W. D. Rosamond, J. Cai, A. V. Diez-Roux, and F. L. Brancati for the Atherosclerosis Risk in Communities Study Investigators (2002). Influence of retirement on leisure-time physical activity: the atherosclerosis risk in communities study. *American Journal of Epidemiology 155*(8), 692–699.

Eyjólfsdóttir, H. S., I. Baumann, N. Agahi, J. Fritzell, and C. Lennartsson (2019). Prolongation of working life and its effect on mortality and health in older adults: propensity score matching. *Social Science & Medicine 226*, 77–86.

Fé, E. and B. Hollingsworth (2016). Short-and long-run estimates of the local effects of retirement on health. *Journal of the Royal Statistical Society. Series A (Statistics in Society) 179*(4), 1051–1067.

Feng, J., Q. Li, and J. P. Smith (2020). Retirement effect on health status and health behaviors in urban China. *World Development 126*, 104702.

Finkel, D., R. Andel, M. Gatz, and N. L. Pedersen (2009). The role of occupational complexity in trajectories of cognitive aging before and after retirement. *Psychology and Aging 24*(3), 563–573.

Fisher, G. G., A. Stachowski, F. J. Infurna, J. D. Faul, J. Grosch, and L. E. Tetrick (2014). Mental work demands, retirement, and longitudinal trajectories of cognitive functioning. *Journal of Occupational Health Psychology 19*(2), 231–242.

Fitzpatrick, M. D. and T. J. Moore (2018). The mortality effects of retirement: evidence from social security eligibility at age 62. *Journal of Public Economics 157*, 121–137.

Fonseca, R., A. Kapteyn, J. Lee, G. Zamarro, and K. Feeney (2014). A longitudinal study of well-being of older Europeans: does retirement matter? *Journal of Population Ageing 7*(1), 21–41.

Frimmel, W. and G. J. Pruckner (2020). Retirement and healthcare utilization. *Journal of Public Economics 184*, 104146.

Galbraith, R. F. (1988). A note on graphical presentation of estimated odds ratios from several clinical trials. *Statistics in Medicine 7*(8), 889–894.

Gill, S. C., P. Butterworth, B. Rodgers, K. J. Anstey, E. Villamil, and D. Melzer (2006). Mental health and the timing of men's retirement. *Social Psychiatry and Psychiatric Epidemiology 41*(7), 515–522.

Godard, M. (2016). Gaining weight through retirement? Results from the SHARE survey. *Journal of Health Economics 45*, 27–46.

Gorry, A., D. Gorry, and S. N. Slavov (2018). Does retirement improve health and life satisfaction? *Health Economics 27*(12), 2067–2086.

Görg, H. and E. Strobl (2001). Multinational companies and productivity spillovers: a meta-analysis. *The Economic Journal 111*(475), F723–F739.

Grip, A. d., M. Lindeboom, and R. Montizaan (2012). Shattered dreams: the effects of changing the pension system late in the game. *Economic Journal 122*(559), 1–25.

Grøtting, M. W. and O. S. Lillebø (2020). Health effects of retirement: evidence from survey and register data. *Journal of Population Economics 33*(2), 671–704.

Hagen, J. (2018). The effects of increasing the normal retirement age on health care utilization and mortality. *Journal of Population Economics 31*(1), 193–234.

Hallberg, D., P. Johansson, and M. Josephson (2015). Is an early retirement offer good for your health? Quasi-experimental evidence from the army. *Journal of Health Economics 44*, 274–285.

Havranek, T., R. Horvath, Z. Irsova, and M. Rusnak (2015). Cross-country heterogeneity in intertemporal substitution. *Journal of International Economics 96*(1), 100–118.

Havránek, T., T. D. Stanley, H. Doucouliagos, P. Bom, J. Geyer-Klingeberg, I. Iwasaki, W. R. Reed, K. Rost, and R. C. M. van Aert (2020). Reporting guidelines for meta-analysis in economics. *Journal of Economic Surveys 34*(3), 469–475.

Heller-Sahlgren, G. (2017). Retirement blues. *Journal of Health Economics 54*, 66–78.

Henkens, K., H. van Solinge, and W. T. Gallo (2008). Effects of retirement voluntariness on changes in smoking, drinking and physical activity among Dutch older workers. *The European Journal of Public Health 18*(6), 644–649.

Hernaes, E., S. Markussen, J. Piggott, and O. L. Vestad (2013). Does retirement age impact mortality? *Journal of Health Economics 32*(3), 586–598.

Hessel, P. (2016). Does retirement (really) lead to worse health among European men and women across all educational levels? *Social Science & Medicine 151*, 19–26.

Horner, E. M. (2014). Subjective well-being and retirement: analysis and policy recommendations. *Journal of Happiness Studies 15*(1), 125–144.

Horner, E. M. and M. R. Cullen (2016). The impact of retirement on health: quasi-experimental methods using administrative data. *BMC Health Services Research 16*(1), 1–9.

Hult, C., M. Stattin, U. Janlert, and B. Järvholm (2010). Timing of retirement and mortality-a cohort study of Swedish construction workers. *Social Science & Medicine 70*(10), 1480–1486.

Insler, M. (2014). The health consequences of retirement. *Journal of Human Resources 49*(1), 195–233.

Ioannidis, J. P., T. Stanley, and H. Doucouliagos (2017). The power of bias in economics research. *The Economic Journal 127*(605), F236–F265.

Johnston, D. W. and W.-S. Lee (2009). Retiring to the good life? The short-term effects of retirement on health. *Economics Letters 103*(1), 8–11.

Jokela, M., J. E. Ferrie, D. Gimeno, T. Chandola, M. J. Shipley, J. Head, J. Vahtera, H. Westerlund, M. G. Marmot, and M. Kivimäki (2010). From midlife to early old age: health trajectories associated with retirement. *Epidemiology (Cambridge, Mass.) 21*(3), 284–290.

Kajitani, S., K. Sakata, and C. McKenzie (2017). Occupation, retirement and cognitive functioning. *Ageing and Society 37*(8), 1568–1596.

Kalwij, A., R. Alessie, and M. Knoef (2013). Pathways to retirement and mortality risk in the Netherlands. *European Journal of Population/Revue Européenne de Démographie 29*(2), 221–238.

Keef, S. P. and L. A. Roberts (2004). The meta-analysis of partial effect sizes. *British Journal of Mathematical and Statistical Psychology 57*(1), 97–129.

Kesavayuth, D., R. E. Rosenman, and V. Zikos (2016). Retirement, personality, and well-being. *Economic Inquiry 54*(2), 733–750.

Kim, J. E. and P. Moen (2002). Retirement transitions, gender, and psychological well-being: a life-course, ecological model. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 57*(3), P212–P222.

Kim, S. and K. Koh (2020). Does early access to pension wealth improve health? *Economic Inquiry 58*(4), 1783–1794.

Kolodziej, I. W. and P. García-Gómez (2019). Saved by retirement: beyond the mean effect on mental health. *Social Science & Medicine 225*, 85–97.

Kuhn, A. (2018). The complex effects of retirement on health. *IZA World of Labor 430*.

Kuhn, A., S. Staubli, J.-P. Wuellrich, and J. Zweimüller (2020). Fatal attraction? Extended unemployment benefits, labor force exits, and mortality. *Journal of Public Economics 191*, 104087.

Kühntopf, S. and T. Tivig (2012). Early retirement and mortality in Germany. *European Journal of Epidemiology 27*(2), 85–89.

Kuusi, T., P. Martikainen, and T. Valkonen (2020). The influence of old-age retirement on health: causal evidence from the Finnish register data. *Journal of the Economics of Ageing 17*, 100257.

Lalive, R. and S. Staubli (2015). How does raising women's full retirement age affect labor supply, income, and mortality? Evidence from Switzerland. Working Paper NB 14–09, NBER Retirement Research Center.

Latif, E. (2011). The impact of retirement on psychological well-being in Canada. *Journal of Socio-Economics 40*(4), 373–380.

Lee, J. and J. P. Smith (2009). Work, retirement, and depression. *Journal of Population Ageing 2*(1-2), 57–71.

Lei, X. and H. Liu (2018). Gender difference in the impact of retirement on cognitive abilities: evidence from urban China. *Journal of Comparative Economics 46*(4), 1425–1446.

Li, W., X. Ye, D. Zhu, and P. He (2021). The longitudinal association between retirement and depression: a systematic review and meta-analysis. *American Journal of Epidemiology*.

Light, R. and D. Pillemer (1984). *Summing up. The science of reviewing research*. Cambridge, MA: Harvard University Press.

Litwin, H. (2007). Does early retirement lead to longer life? *Ageing and Society 27*(5), 739.

Lucifora, C. and D. Vigani (2018). Health care utilization at retirement: the role of the opportunity cost of time. *Health Economics 27*(12), 2030–2050.

Magnus, J. R. and G. De Luca (2016). Weighted-average least squares (WALS): a survey. *Journal of Economic Surveys 30*(1), 117–148.

Magnus, J. R., O. Powell, and P. Prüfer (2010). A comparison of two model averaging techniques with an application to growth empirics. *Journal of Econometrics 154*(2), 139–153.

Mandal, B. and B. E. Roe (2007). Job loss, retirement and the mental health of older Americans. *Journal of Mental Health Policy and Economics 11*(4), 167–176.

Mazzonna, F. and F. Peracchi (2012). Ageing, cognitive abilities and retirement. *European Economic Review 56*(4), 691–710.

Mazzonna, F. and F. Peracchi (2017). Unhealthy retirement? *Journal of Human Resources 52*(1), 128–151.

Mein, G., P. Martikainen, H. Hemingway, S. Stansfeld, and M. Marmot (2003). Is retirement good or bad for mental and physical health functioning? Whitehall II longitudinal study of civil servants. *Journal of Epidemiology and Community Health 57*(1), 46–49.

Messe, P.-J. and F.-C. Wolff (2019a). Healthier when retiring earlier? Evidence from France. *Applied Economics 51*(47), 5122–5143.

Messe, P.-J. and F.-C. Wolff (2019b). The short-term effects of retirement on health within couples: evidence from France. *Social Science & Medicine 221*, 27–39.

Moher, D., A. Liberati, J. Tetzlaff, D. G. Altman, and T. P. Group (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLOS Medicine 6*(7), 1–6.

Mojon-Azzi, S., A. Sousa-Poza, and R. Widmer (2007). The effect of retirement on health: a panel analysis using data from the Swiss Household Panel. *Swiss Medical Weekly 137*(4142), 581–585.

Mosca, I. and A. Barrett (2016). The impact of voluntary and involuntary retirement on mental health: Evidence from older Irish adults. *Journal of Mental Health Policy and Economics 19*(1), 33–44.

Motegi, H., Y. Nishimura, and M. Oikawa (2020). Retirement and health investment behaviors: an international comparison. *Journal of the Economics of Ageing 16*, 100267.

Müller, T. and M. Shaikh (2018). Your retirement and my health behavior: evidence on retirement externalities from a fuzzy regression discontinuity design. *Journal of Health Economics 57*, 45–59.

Neuman, K. (2008). Quit your job and get healthier? The effect of retirement on health. *Journal of Labor Research 29*(2), 177–201.

Nielsen, N. F. (2019). Sick of retirement? *Journal of Health Economics 65*, 133–152.

Nishimura, Y., M. Oikawa, and H. Motegi (2018). What explains the difference in the effect of retirement on health? Evidence from global aging data. *Journal of Economic Surveys 32*(3), 792–847.

OECD (2019). *Pensions at a Glance 2019: OECD and G20 Indicators*. Paris: OECD Publishing.

Oksanen, T., J. Vahtera, H. Westerlund, J. Pentti, N. Sjösten, M. Virtanen, I. Kawachi, and M. Kivimäki (2011). Is retirement beneficial for mental health? Antidepressant use before and after retirement. *Epidemiology 22*(4), 553–559.

Olesen, K., R. Rugulies, N. H. Rod, and J. P. Bonde (2014). Does retirement reduce the risk of myocardial infarction? A prospective registry linkage study of 617 511 Danish workers. *International Journal of Epidemiology 43*(1), 160–167.

Oshio, T. and M. Kan (2017). The dynamic impact of retirement on health: evidence from a nationwide ten-year panel survey in Japan. *Preventive Medicine 100*, 287–293.

Pabón-Carrasco, M., L. Ramirez-Baena, R. López Sánchez, I. Rodríguez-Gallego, N. Suleiman-Martos, and J. L. Gómez-Urquiza (2020). Prevalence of depression in retirees: a meta-analysis. *Healthcare 8*(3).

Picchio, M. and J. C. van Ours (2020). Mental health effects of retirement. *De Economist 168*(3), 419–452.

Pilipiec, P., W. Groot, and M. Pavlova (2020). The effect of an increase of the retirement age on the health, well-being, and labor force participation of older workers: a systematic literature review. *Journal of Population Ageing*, forthcoming.

Quaade, T., G. Engholm, A. M. T. Johansen, and H. Møller (2002). Mortality in relation to early retirement in Denmark: a population-based study. *Scandinavian Journal of Public Health 30*(3), 216–222.

Reed, W. (2020). A note on the use of partial correlation coefficients in meta-analyses. Working paper no. 8/2020, Department of Economics and Finance, University of Canterbury, New Zealand.

Rijs, K. J., R. Cozijnsen, and D. J. Deeg (2012). The effect of retirement and age at retirement on self-perceived health after three years of follow-up in Dutch 55-64-year-olds. *Ageing and Society 32*(2), 281–306.

Roberts, B. A., R. Fuhrer, M. Marmot, and M. Richards (2011). Does retirement influence cognitive performance? The Whitehall II study. *Journal of Epidemiology and Community Health 65*(11), 958–963.

Rohwedder, S. and R. J. Willis (2010). Mental retirement. *Journal of Economic Perspectives 24*(1), 119–38.

Rose, L. (2020). Retirement and health: evidence from England. *Journal of Health Economics 73*, 102352.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin 86*(3), 638–641.

Sewdas, R., A. De Wind, S. Stenholm, P. Coenen, I. Louwerse, C. Boot, and A. Van Der Beek (2020). Association between retirement and mortality: working longer, living longer? A systematic review and meta-analysis. *Journal of Epidemiology and Community Health 74*(5), 473–480.

Shai, O. (2018). Is retirement good for men's health? Evidence using a change in the retirement age in Israel. *Journal of Health Economics 57*, 15–30.

Son, C.-H. et al. (2020). The effects of retirement on health-related quality of life of retirement-aged adults for four marital status subgroups. *International Journal of Economic Sciences 9*(1), 179–201.

Stanley, T. (2008). Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection. *Oxford Bulletin of Economics and Statistics 70*(1), 103–127.

Stanley, T. D. (2005). Beyond publication bias. *Journal of Economic Surveys 19*(3), 309–345.

Stanley, T. D. and H. Doucouliagos (2012). *Meta-regression analysis in economics and business*. Abingdon: Routledge.

Stanley, T. D. and H. Doucouliagos (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods 5*(1), 60–78.

Stanley, T. D., H. Doucouliagos, M. Giles, J. H. Heckemeyer, R. J. Johnston, P. Laroche, J. P. Nelson, M. Paldam, J. Poot, G. Pugh, et al. (2013). Meta-analysis of economics research reporting guidelines. *Journal of Economic Surveys 27*(2), 390–394.

Sterling, T. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance–or vice versa. *Journal of the American Statistical Association 54*(285), 30–34.

Sutton, A. J., K. R. Abrams, D. R. Jones, D. R. Jones, T. A. Sheldon, and F. Song (2000). *Methods for meta-analysis in medical research*. Chichester: Wiley.

Syse, A., M. Veenstra, T. Furunes, R. J. Mykletun, and P. E. Solem (2017). Changes in health and health behavior associated with retirement. *Journal of Aging and Health 29*(1), 99–127.

Tsai, S. P., J. K. Wendt, R. P. Donnelly, G. De Jong, and F. S. Ahmed (2005). Age at retirement and long term survival of an industrial population: prospective cohort study. *British Medical Journal 331*(7523), 995.

van der Heide, I., R. M. van Rijn, S. J. Robroek, A. Burdorf, and K. I. Proper (2013). Is retirement good for your health? A systematic review of longitudinal studies. *BMC Public Health 13*(1), 1–11.

van Mourik, C. (2020). A meta-analysis into the causal effect of retirement on health. Netspar Academic Series, Thesis for the Master Public Administration, Leiden University.

Viechtbauer, W. and M. Cheung (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Method 1*(2), 112–125.

Vooren, M., C. Haelermans, W. Groot, and H. van den Brink (2019). The effectiveness of active labor market policies: a meta-analysis. *Journal of Economic Surveys 33*(1), 125–149.

Waldron, H. (2001). Links between early retirement and mortality. ORES Working Paper Series No. 93, Social Security Administration, Office of Policy, Washington.

Webb, M. D. (2014). Reworking wild bootstrap based inference for clustered errors. Queen's Economics Department Working Paper No. 1315, Kingston, Canada.

Westerlund, H., M. Kivimäki, A. Singh-Manoux, M. Melchior, J. E. Ferrie, J. Pentti, M. Jokela, C. Leineweber, M. Goldberg, M. Zins, et al. (2009). Self-rated health before and after retirement in France (GAZEL): a cohort study. *The Lancet 374*(9705), 1889–1896.

Wu, C., M. C. Odden, G. G. Fisher, and R. S. Stawski (2016). Association of retirement age with mortality: a population-based longitudinal study among older adults in the USA. *Journal of Epidemiology and Community Health 70*(9), 917–923.

Xue, X., M. Cheng, and W. Zhang (2021). Does education really improve health? A meta-analysis. *Journal of Economic Surveys 35*(1), 71–105.

Zhang, Y., M. Salm, and A. van Soest (2018). The effect of retirement on healthcare utilization: evidence from China. *Journal of Health Economics 62*, 165–177.

Zhao, M., Y. Konishi, and H. Noguchi (2017). Retiring for better health? Evidence from health investment behaviors in Japan. *Japan and the World Economy 42*, 56–63.

Zhu, R. (2016). Retirement and its consequences for women's health in Australia. *Social Science & Medicine 163*, 117–125.

Zulka, L. E., I. Hansson, and L. B. Hassing (2019). Impact of retirement on cognitive function: a literature review. *Journal of Gerontopsychology and Geriatric Psychiatry 32*(4), 187–203.

Zulkarnain, A. and M. S. Rutledge (2018). How does delayed retirement affect mortality and health? CRR Working Paper 2018-11, Center for Retirement Research at Boston College.

Appendix: Articles used in the meta-analysis and further estimation results

Table A.1: Articles included in the meta-analysis ($N = 85$)

Authors	Citations	Outcome(s)	Country	Data	Time Span	Id. Strategy	Effects	Heterogeneity
Apouey et al. (2019)	7	SAH, PH, MH	Australia	HILDA	2001-2014	FE	+, 0	No
Ardito et al. (2020)	1	HC	Italy	WHIP	2001-2014	IV	–	O, Ph, I
Atalay and Barrett (2014)	36	SAH, PH, MH	Australia	NHSs	1995-2008	IV	0 (SAH), 0/+ (PH, MH)	G
Atalay et al. (2019)	12	MH	Australia	HILDA	2012-2016	FD-IV	0	No
Bamia et al. (2008)	92	M	Greece	EPIC	1994-2006	Other	–	No
Barrett and Kecmanovic (2013)	35	MH	Australia	HILDA	2007	Other	0	V
Bauer and Eichenberger (2018)	1	SAH, PH	Switzerland	Swiss LFS	2004-2015	DiD	0	No
Behncke (2012)	247	SAH, PH, MH	England	ELSA	2002-2007	PSM, IV	– (SAH, PH), 0 (MH)	No
Belloni et al. (2016)	31	MH	10 EU	SHARE	2004-2013	FE-IV	0	O
Bertoni and Brunello (2017)	23	MH	Japan	JPSS	2008-2013	IV	–	No
Bianchini and Borella (2016)	16	MH	10 EU	SHARE	2004-2012	FE-IV	0, + (RD)	RD
Binh Tran and Zikos (2019)	8	SAH, PH, MH	Australia	HILDA	2002-2015	FE-IV	+	No
Blake and Garrouste (2019)	9	SAH, PH, MH	France	Health Barometer	1994-2003	DiD	0	E
Bloemen et al. (2017)	55	M	Netherlands	Administrative Data	2000-2005	FE-IV	0	No
Bonsang et al. (2012)	444	MH	USA	HRS	1998-2008	FE-IV	–	No
Bonsang and Klein (2012)	134	SAH	Germany	GSOEP	1995-2010	FE-IV	+, 0 (MI)	V
Bozio et al. (2021)	2	M	France	Administrative Data	2004-2017	IV	0	No
Brockmann et al. (2009)	104	M	Germany	Gmunder Ersatzkasse	1990-2004	Other	+ (M), 0 (F)	G, Ph, I
Butterworth et al. (2006)	256	MH	Australia	NSMHWB	1997	Other	0	O, AC
Calvo et al. (2013)	152	PH, MH	USA	HRS	1992-2010	FE-IV	–	T
Carrino et al. (2020)	1	SAH, PH, MH	UK	Understanding Society	2009-2016	DiD	0	O
Celidoni et al. (2017)	55	MH	10 EU	SHARE	2004-2012	IV	+	T
Celidoni and Rebba (2017)	51	HC	10 EU	SHARE	2004-2012	FE-IV	0	MS
Che and Li (2018)	10	SAH	China	CHNS	1991-2006	IV	+	No
Chung et al. (2009)	64	PH	USA	HRS	1992-2002	FE-IV	–	AC, I, O
Coe et al. (2012)	153	MH	USA	HRS	1996-2008	IV	+, 0	O
Coe and Zamarro (2011)	506	SAH, MH	11 EU	SHARE	2004-2007	IV	+ (SAH), 0 (MH)	No
Dave et al. (2008)	485	SAH, PH, MH	USA	HRS	1992-2005	FE	–	MS, V
Dayaram and McGuire (2019)	1	PH, MH	Australia	HILDA	2003-2015	PSM	0	No
Eibich (2015)	268	SAH, PH, MH, HC	Germany	GSOEP	2002-2009	RDD	+ (SAH, MH, HC), 0 (PH)	E
Eyjólfsdóttir et al. (2019)	6	PH, M	Sweden	LNU, LISA, SWEOLD	2004-2014	PSM	0	No
Fé and Hollingsworth (2016)	14	SAH, MH	UK	BHPS	1991-2005	RDD	– (SAH), + (MH)	No
Feng et al. (2020)	12	PH	China	CHARLS	2001-2015	RDD	– (M), 0 (F)	E, G
Fitzpatrick and Moore (2018)	84	M	USA	MCOD, SSDMF	1979-2012	RDD	– (M), 0 (F)	G, E
Frimmel and Pruckner (2020)	7	HC	Austria	ASSD	1998-2012	FE-IV	+, 0 (F)	O, G
Gill et al. (2006)	95	MH	Australia	HILDA	2002-2003	Other	0	No
Godard (2016)	90	PH	8 EU	SHARE	2004-2011	FE-IV	0	O
Gorry et al. (2018)	84	SAH, MH, PH, HC	USA	HRS	1992-2014	IV	+, 0 (MH)	No
Grip et al. (2012)	121	SAH, MH, HC	Netherlands	Administrative Data	1997-2006	RDD	– (MH), 0 (HC, SAH)	O, I
Grøtting and Lillebø (2020)	5	PH, HC, M	Norway	NORLAG	2002-2012	RDD	0, + (M PH)	No
Hagen (2018)	40	HC, M	Sweden	LOUISE	1987-2010	DiD	0	No
Hallberg et al. (2015)	64	M	Sweden	Administrative Data	1985-2010	DiD	+, 0	No
Heller-Sahlgren (2017)	67	MH	10 EU	SHARE	2004-2012	FE-IV	– (M), 0 (F)	E, G, O
Hernaes et al. (2013)	111	M	Norway	Administrative Data	1992-2010	IV	0	No
Hessel (2016)	54	SAH, PH	12 EU	EU-SILC	2009-2012	RE-IV	+, 0 (Chronic)	No

(continued on next page)

Table A.1: *Continued from previous page*

Authors	Citations	Outcome(s)	Country	Data	Time Span	Id. Strategy	Effects	Heterogeneity
Horner and Cullen (2016)	23	PH, MH, HC	USA	Administrative Data	1997-2009	IV	0, + (Chronic)	No
Hult et al. (2010)	57	M	Sweden	Health Monitoring	1971-1993	Other	0	Ph
Insler (2014)	198	SAH	USA	HRS	1992-2010	FE-IV	+	No
Johnston and Lee (2009)	149	SAH, PH, MH	UK	HSE	1997-2005	RDD	+, 0 (PH)	No
Jokela et al. (2010)	161	PH, MH	UK	WhiteHall II Cohort Study	1991-2006	Other	+ (MH), 0 (PH)	T, RD
Kajitani et al. (2017)	35	MH	Japan	NSJE	1987-2002	IV	0	O
Kalwij et al. (2013)	11	M	Netherlands	IPO	1996-2010	Other	0	I
Kim and Koh (2020)	1	SAH	Singapore	SLP	2015-2019	RDD	+	No
Kim and Moen (2002)	364	MH	USA	Cornell Retirement Study	1994-1999	Other	0	G, MS, Ph
Kolodziej and García-Gómez (2019)	20	MH	11 EU	SHARE	2004-2013	IV	+ (F), 0 (M)	MS, Ph, G
Kuhn et al. (2020)	19	M	Austria	ASSD	1972-2017	IV	0 (F), – (M)	Ph, O, G
Kuusi et al. (2020)	4	MH, HC	Finland	Statistics Finland	2000-2012	FE-IV	+, 0 (HC)	G, O, I
Latif (2011)	59	MH	Canada	CNPHS	1994-2006	FE-IV	+	AC, MS
Lei and Liu (2018)	14	MH	China	CHARLS	2011-2015	FE-IV	+, 0 (M), 0, – (F)	G, O
Litwin (2007)	51	M	Israel	NHS	1997-2004	Other	0	No
Lucifora and Vigani (2018)	16	HC	10 EU	SHARE	2004-2006	FE-IV	0	G, O
Mandal and Roe (2007)	117	MH	USA	HRS	1992-2002	IV	+, –	No
Mazzonna and Peracchi (2017)	138	SAH, MH	10 EU	SHARE	2004-2006	FD-IV	–, 0 (F SAH)	O
Mein et al. (2003)	348	PH, MH	UK	WhiteHall II Cohort Study	1991-1995	Other	+, 0 (PH F)	No
Messe and Wolff (2019b)	9	SAH, PH	France	LFS	2013-2016	DiD	+ (M), 0 (F PH)	G, O
Messe and Wolff (2019a)	7	SAH, PH	France	INSEE	2012	IV	0	No
Mojon-Azzi et al. (2007)	80	SAH, PH, MH	Switzerland	SHP	1999-2003	Other	+	No
Mosca and Barrett (2016)	17	MH	Ireland	TILDA	2009-2013	FD	0	V
Müller and Shaikh (2018)	49	SAH	19 C.	SHARE	2004-2013	RDD	+	No
Neuman (2008)	235	SAH, PH, MH	USA	HRS	1992-2004	IV	0, + (SAH)	No
Nielsen (2019)	13	PH, HC, M	Denmark	Administrative Data	1980-2010	IV, RDD	0 (IV), 0, + (HC)	G, AC, T
Oksanen et al. (2011)	103	MH	Finland	National Records	1995-2004	Other	+	No
Oshio and Kan (2017)	28	SAH, MH	Japan	PSMOA	2005-2014	FE-IV	+, 0 (F SAH)	G
Picchio and van Ours (2020)	12	MH	Netherlands	LISS	2007-2018	RDD	+ (M), 0 (F)	G, MS
Quaade et al. (2002)	63	M	Denmark	Administrative Data	1986-1996	Other	–	No
Rijs et al. (2012)	41	SAH	Netherlands	LASA	1995-2009	Other	0	E, AC
Roberts et al. (2011)	51	MH	UK	WhiteHall II Cohort Study	1985-1988	Other	0	No
Rose (2020)	4	SAH, PH, MH, HC, M	England	ELSA, BHPS	1990-2011	RDD, FE-IV	+, 0, – (M M)	G, E, MS
Shai (2018)	44	SAH, HC	Israel	IHS, SHARE	1997-2013	DiD	0 (HC), –	E
Syse et al. (2017)	35	PH, MH	Norway	NORLAG	2002-2007	Other	+, 0	No
Tsai et al. (2005)	116	M	USA	Shell Oil	1973-2003	Other	0	No
Westerlund et al. (2009)	340	SAH	France	GAZEL	1990-2006	Other	+	O
Wu et al. (2016)	49	M	USA	HRS	1992-2010	Other	+	Ph
Zhang et al. (2018)	27	SAH, HC	China	CHARLS	2011-2013	RDD	0, – (F)	G
Zhu (2016)	51	SAH, PH, MH	Australia	HILDA	2001-2011	FE-IV	+	No

The sign of the effect is based on the value of t-stat: "–" means $t \leq -1.96$; "+" is for $t \geq 1.96$; "0" when $-1.96 < t < 1.96$;
Identification Strategy: PSM = Propensity Score Matching; IV = Instrumental Variables; DiD = Difference-in-differences; FE = Fixed Effects; RDD = Regression Discontinuity Design;
Other methods = Ordinary Least Square, Duration models, Multinomial logit, Growth-mixture models.
Outcome: MH = Mental health; SAH = Self-assessed / General health; PH = Physical health; HC = Healthcare utilization; M = Mortality.
Effects: M = Males; F = Females; B = Both; MI = Mandatory/Involuntary; ER = Early retirement; PP = Postponed; SR = Short-run; LR = Long-run; RD = Retirement duration.
Heterogeneity: G = Gender; O = Occupation; Ph = Previous health; MS = Marital status; E = Educational attainment; I = Income; V = Voluntariness; AC = Age cohort; T = Timing; C = Country; LS = Living standard.

Table A.2: Heterogeneity in the estimated effects of retirement on health (FAT-PET model)

	BMA[a]			WALS ($q=1$)[b]		WALS ($q=0.5$)[b]		OLS check after BMA[c]			OLS check after WALS[d]		
	PM	PSD	PIP	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.		p-value	Coeff.		p-value
Focus Regressors													
Scimago subject areas													
Publication bias in economics/business	**0.811**	**0.343**	**1.000**	**0.875**	**0.357**	**0.898**	**0.356**	0.856	**	0.040	0.637		0.133
Publication bias in medicine/psychology	**0.180**	**0.273**	**1.000**	0.129	0.272	0.156	0.276	0.201		0.683	0.245		0.597
Publication bias in multiarea	**0.183**	**0.329**	**1.000**	0.163	0.334	0.159	0.336	0.232		0.628	0.224		0.616
Precision effect in economics/business	**0.004**	**0.004**	**1.000**	**0.014**	**0.006**	**0.014**	**0.006**	0.003		0.362	0.012		0.149
Precision effect in medicine/psychology	**0.006**	**0.004**	**1.000**	**0.017**	**0.005**	**0.017**	**0.006**	0.004		0.163	0.012	*	0.056
Precision effect in multiarea	**0.007**	**0.004**	**1.000**	**0.017**	**0.005**	**0.018**	**0.005**	0.006	**	0.048	0.012	**	0.018
Auxiliary Regressors													
Google scholar citations per year	0.000	0.000	0.100	0.000	0.000	0.000	0.000	–		–	–		–
Scimago Journal Ranking	0.000	0.000	0.090	**-0.001**	**0.001**	**-0.001**	**0.001**	–		–	-0.002		0.293
Year of publication	**0.001**	**0.000**	**0.950**	**0.001**	**0.000**	**0.001**	**0.000**	0.001	*	0.054	0.001	**	0.014
Health outcomes (reference category: Mortality)													
General and self-reported health	**0.012**	**0.003**	**1.000**	**0.010**	**0.002**	**0.01**	**0.002**	0.012	***	0.000	0.014	***	0.000
Physical health	0.001	0.002	0.380	**0.003**	**0.001**	**0.003**	**0.002**	–		–	0.004	*	0.070
Mental health	**0.009**	**0.003**	**0.970**	**0.007**	**0.002**	**0.007**	**0.002**	0.009	***	0.002	0.010	***	0.000
Healthcare utilization	**0.003**	**0.002**	**0.750**	**0.003**	**0.000**	**0.003**	**0.001**	0.003	**	0.036	0.004	**	0.043
Identification strategies (reference category: Other methods)													
Regression discontinuity design (RDD)	0.000	0.001	0.120	**-0.009**	**0.005**	**-0.009**	**0.005**	–		–	-0.004		0.694
Instrumental variables (IV)	0.000	0.001	0.120	**-0.007**	**0.005**	**-0.006**	**0.006**	–		–	-0.003		0.691
Difference-in-differences (DiD)	-0.001	0.002	0.150	-0.003	0.004	-0.002	0.005	–		–	–		–
Propensity score matching (PSM)	-0.001	0.005	0.110	**-0.018**	**0.010**	**-0.019**	**0.010**	–		–	-0.015		0.210
Fixed-effects/First-differences	**-0.009**	**0.005**	**0.840**	**-0.014**	**0.005**	**-0.014**	**0.005**	-0.010		0.141	-0.013	*	0.077
Institutional contexts (reference category: Statutory retirement)													
Mandatory or involuntary retirement	**-0.026**	**0.010**	**0.950**	**-0.022**	**0.007**	**-0.024**	**0.008**	-0.028	**	0.022	-0.028	**	0.026
Early retirement	-0.001	0.002	0.180	-0.002	0.002	-0.002	0.002	–		–	–		–
Postponed retirement	-0.002	0.003	0.360	**-0.006**	**0.002**	**-0.007**	**0.002**	–		–	-0.006		0.172
Geographical areas (reference category: Multi-country analyses)													
Europe	0.000	0.001	0.060	-0.002	0.003	-0.002	0.003	–		–	–		–
Extra-European countries	0.000	0.001	0.080	**-0.003**	**0.003**	-0.003	0.003	–		–	–		–
Sex (reference category: Males)													
Females	0.000	0.001	0.150	0.001	0.001	0.001	0.000	–		–	–		–
Males+Females	-0.001	0.002	0.330	**-0.004**	**0.003**	**-0.005**	**0.003**	–		–	-0.004		0.102
Calculation of t-statistic (reference category: from 95% CI or from OR)													
t-statistic from $\widehat{\beta}_i / SE_i$	**-0.006**	**0.004**	**0.780**	**-0.004**	**0.004**	**-0.005**	**0.004**	-0.006		0.149	-0.006		0.427

Notes: The results are from the FAT-PET specification by using the inverse of the SE_i^2 as weights. PM = Posterior Mean of the coefficient; PSD = Posterior Standard Deviation; PIP = Posterior Inclusion Probability. The number of observations (studies) is 275 (85), 70 (16) in economics/business, 122 (36) in medicine/psychology, and 83 (33) in the residual multi-area category. Auxiliary variables for which the PIP is above 0.5 in BMA or the corresponding one-standard error band does not include zero in WALS are in bold. *** Significant at 1%, ** significant at 5%, * significant at 10%.

[a] In the BMA, we use the uniform distribution for model priors, the Zellner's g prior for the distributions of the coefficients and a Markov Chain Monte Carlo algorithm to search over the model space, by distinguishing between focus and auxiliary regressors.

[b] $q = 1$ indicates the Laplace model prior distribution; $q = 0.5$ implies the Subbotin model prior distribution.

[c] The model specification under "OLS" includes those variables which have a PIP > 0.5 in BMA ($R^2 = 0.35$). We report wild cluster bootstrap p-values obtained from the wild cluster bootstrap-t procedure proposed by Cameron et al. (2008), with clusters at study level (5,000 bootstraps using the Webb's (2014) six-point distribution as weights).

[d] The second model specification under "OLS" includes those variables which are relevant according to WALS ($R^2 = 0.38$). For both OLS checks, we report wild cluster bootstrap p-values obtained from the wild cluster bootstrap-t procedure proposed by Cameron et al. (2008), with clusters at study level (5,000 bootstraps using the Webb's (2014) six-point distribution as weights).