NBER WORKING PAPER SERIES

THE CHRONIC DISEASE INDEX: ANALYZING HEALTH INEQUALITIES OVER THE LIFECYCLE

Kaveh Danesh
Jonathan T. Kolstad
William D. Parker
Johannes Spinnewijn

Working Paper 32577
http://www.nber.org/papers/w32577

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2024

We thank Rob Alessie, Liran Einav, Amy Finkelstein, Camille Landais, Adriana Lleras-Muney, Canishk Naik, Carol Propper, Bram Spinnewijn, Frank van Dijk, Winnie Van Dijk and seminar participants at the LSE, IFS, University of Oslo, Bank of Spain, BI Norwegian Business School, CPB, Bocconi, Bologna, EIEF, Zurich, Oxford, Royal Holloway and at the CEPR Public Economics symposium, the NBER Aging meeting, the Dutch health econometrics workshop, and the lolaHESG conference for valuable comments and suggestions. We also thank the Dutch Central Bureau of Statistics and Thomas Minten for help with the data and Boris Beyen, Ben Dahmen, Diego Ferreras-Garrucho, and Marco Visentin for excellent research assistance. We thankfully acknowledge financial support from the International Inequalities Institute. Kolstad thanks the London School of Economics for support through the BP Centennial Professorship that he held during work on this paper The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The Chronic Disease Index: Analyzing Health Inequalities Over the Lifecycle
Kaveh Danesh, Jonathan T. Kolstad, William D. Parker, and Johannes Spinnewijn
NBER Working Paper No. 32577
June 2024
JEL No. I1, I12, I14

ABSTRACT

The rich live longer than the poor, but relatively little is known about the evolution of health inequality across the lifecycle. Using rich administrative data from the Netherlands, we develop an index of chronic disease burden based on the projected contribution to old-age mortality. Chronic conditions account for one-third of the mortality gap in old age. Using our index we demonstrate that health inequality arises much earlier in life; by age 35, the bottom half of the income distribution has the same disease burden as those age 50 in the top half. Approximately 60% of the divergence across income groups is due to low-income individuals developing chronic illness at a faster rate, rather than chronically ill individuals sorting into lower-income groups. Using linked health survey data, we then examine the contributions of various mediators to the incidence of chronic diseases over the life cycle. Socioeconomic and geographic factors explain most of the variation, while individual health behaviors play a moderate role. Our findings align with calls to target health policies toward early-life social determinants of health.

Kaveh Danesh
University of California, Berkeley
University of California, San Francisco
Kaveh.Danesh@ucsf.edu

Jonathan T. Kolstad
Haas School of Business
University of California, Berkeley
Berkeley, CA 94720
and NBER
jkolstad@berkeley.edu

William D. Parker
London School of Economics
w.d.parker@lse.ac.uk

Johannes Spinnewijn
Department of Economics
London School of Economics
Houghton Street
London, WC2A 2AE
J.Spinnewijn@lse.ac.uk

1 Introduction

Inequality in health is a major source of socio-economic disparities and a central policy challenge facing many countries to date (e.g., Marmot 2015; Chetty et al. 2016). Efforts to 'close the health gap' continue to be high on the international policy agenda. See for example the sequence of reports by the World Health Organization (1985, 2008, 2017), as well as the EU's recent "Joint Action on Health Inequalities". Despite the importance of the issue and the attention paid in policy and research, the complexity of the health production function in combination with measurement challenges have limited our understanding of the drivers of health inequalities. As Angus Deaton (2002) observed : "there is no general agreement about [its] causes ... [and] what apparent agreement there is is sometimes better supported by repeated assertion than by solid evidence."

In this paper we turn to chronic illness to study how health inequalities develop over the lifecycle. Chronic illness is not only a key contributor to the mortality-income gradient, but also a measurable, dynamic marker of population health. As such, it allows us to track a key measure of health that ultimately contributes to mortality gaps in older age, long before these mortality effects manifest themselves. The idea that health status in later life, when mortality effects are large, is a result of long-run life experience and investments in health is conceptually well developed (Grossman 1972). However, these theoretical insights have been insufficient to guide policy design beyond some simple comparative statics (O'Donnell, Van Doorslaer, and Van Ourti 2015). Similarly, the notion that specific chronic conditions can have important health implications is not new, certainly not to medicine (e.g., Bauer et al. 2014). However, what is less studied is the dynamic development of these chronic conditions and how they interact with socio-economic factors and other determinants of health.

Data constraints have generally limited our ability to provide a comprehensive measure of health and its dynamics at a large scale. As a result, researchers have relied on structural assumptions and calibration approaches to model the complex health process and plausible impacts of interventions (see, e.g., Lleras-Muney and Moreau 2022; De Nardi, Pashchenko, and Porapakkarm 2023). We take a different approach and overcome the data challenges by building a panel for the full Dutch population over a range of 20 years, combining multiple administrative registries with granular and comprehensive data on health and socio-economic outcomes. Using medicine data we can measure the profile of chronic illness for the approximately 18 million people in our

sample. This direct measurement approach allows us to flexibly and comprehensively explore the evolution of chronic illness. By further complementing measurable health with detailed data on income and other socio-economic factors, we are able to empirically document the evolution of population health inequality over the lifecycle and start unpacking the different dynamic pathways between health and income. The approach, while not without its challenges as we discuss, relies on data-driven estimation and allows us to document and measure these effects with minimal assumptions.

Our analysis of health inequalities through the lens of chronic illness relies on two building blocks. The first building block is to measure chronic health conditions. Here we build on prior work (see Huber et al. 2013) and use healthcare claims data on dispensed medication to identify chronic conditions. We directly address concerns regarding unequal under-diagnosis and/or management of chronic conditions, using additional information from survey data and other medicine use. We generally find that it is limited in the Dutch context with the exception of the very bottom of the income distribution. The second building block is to translate the chronic conditions in a comprehensive index of chronic disease that is comparable throughout the lifetime. To construct this index, we focus on the role of chronic disease in mortality. As mortality effects manifest themselves in later life, we develop a transparent model of the relationship between chronic illness and mortality at old age. In particular, we formulate the health to mortality relationship as a prediction problem and use chronic illness for individuals at age 70 to predict their 5-year mortality, flexibly accounting for all measured chronic conditions, using multiple lags and including their interactions.[1] We use a Double-Lasso estimation procedure to also flexibly control for a rich set of socio-economic differences that correlate with chronic conditions and may affect mortality for non-chronic reasons (Belloni, Chernozhukov, and Hansen 2014). Using the resulting model, we construct a Chronic Disease Index (CDI) that, in essence, re-weights chronic illness at any point in an individual's life to capture the eventual effect such an illness will have on mortality at age 70, regardless of socio-economic outcomes.

We then use the chronic disease index to study health inequalities over the life-cycle. The empirical analysis provides three sets of results:

First, despite the universal coverage and broad health care access in the Netherlands, we find

1. We follow much of the literature here in focusing on mortality (e.g., Chetty et al. 2016). Health is not simply a discrete outcome, though, and there are efforts to better measure the utility effects of morbidity. One could also interpret our measure of chronic illness as a measure of morbidity directly though the appropriate weights would need to be developed. To the extent that they map to the eventual mortality effects later in life, our index would be an appropriate measure as is.

substantial gaps in mortality across income groups and show that differences in the prevalence of chronic conditions can explain between 30 and 40% of these gaps. This by itself indicates the potential value of a chronic disease index to study differences in health. Strikingly, we do not find meaningful differences in the mortality associated with these chronic conditions across income groups. This suggests that individuals with chronic conditions are in fact similarly treated across income groups, which we corroborate with more direct evidence using healthcare expenditures. These descriptive findings are important by themselves, but also underscore an advantage of studying this question in the context of the Netherlands. We are able to look at a system where access to care seems equally allocated and can thus focus on the important question of where the remaining, large differences in mortality by income come from.[2]

Second, while mortality differences are only apparent later in life, the CDI allows to measure health earlier in life too. We find that almost half of the gap in the chronic burden between income groups at age 70 has already materialized at age 40. The CDI gap quickly opens up in early adulthood and then increases with age. Between the ages of 20 and 45, the CDI's age profile is remarkably flat for high-income individuals compared to low-income individuals. Exploiting the panel structure of the data, we can separate to what extent this is driven by individuals in different income groups 'aging' at different rates vs. individuals with different chronic illness sorting into different income groups. The sorting mechanism includes the causal pathway from illness to earnings decline. Applying this dynamic decomposition, we find that the sorting mechanism is important throughout the life-cycle and responsible for the opening of the CDI gap at the start of the career. But we also find substantial differences in aging that gradually increase over the lifetime. That is, chronic disease develops at a faster rate for low-income individuals than for high-income individuals and this difference increases with age. The gap in CDI growth is mostly driven by the differential incidence of cardio-vascular disease and diabetes, especially at older ages, while we also see important differences in the *prevalence* of psychological disorders contributing to the gap in CDI levels. Aggregated up, we find that the aging differences dominate the sorting effects and explain about 40% more of the health gap observed at age 70.

Third, having separated two key pathways underlying health inequalities, we can also shed light on the relevant mediating factors for these pathways. We leverage rich administrative data and

2. The income gradient in life expectancy in the Netherlands is about 75% as large as in the US. The difference in life expectancy between $p1$ and $p100$ for Netherlands: Females 7.6, Males 11.6; for the US: Females 10.3, Males 14.8. These measures were created as proceedings of the Fall 2019 NBER conference "Income and Life Expectancy: What Can Be Learned from International Comparisons"; They are constructed to be consistent with definitions and methods in Chetty et al. (2016).

linked survey data to measure a variety of mediating factors for the same individuals. In contrast with prior work, we can thus provide a comprehensive account of their relative importance as we can account for all mediating factors jointly in the same context. Using Shapley-Owen decompositions, which apportion the commonly explained variation by different mediators, we find that both socio-economic status and geography contribute most to the explained variation in CDI growth, accounting for about one third each. In contrast, we find a less important role for observed health behaviors (e.g., smoking), especially at younger ages. While this analysis is descriptive, it provides an important re-calibration of the potential importance of different determinants of health. Moreover, the richness of the data allows us to illustrate the potential for mis-estimation due to data challenges. In particular, as we find a strong socio-economic gradient in health behaviors, we also show that we would over-estimate the importance of the commonly measured health behaviors for the disparities in health outcomes in a more partial analysis and when not accounting for reverse causalities.

The paper also aims to draw policy conclusions from the empirical analysis and demonstrates how better measurement can help tightening the recommendations. First, while disparities in healthcare treatment seem limited in the context of the Netherlands, the health gap remains substantial and to close this gap we should focus our effort on reducing the differential burden of disease across socio-economic backgrounds. Our analysis suggests that this requires a comprehensive approach, not just focusing on the 'usual suspects' like smoking, drinking, and healthy diets, but also focusing on other social and geographic factors that are not directly under individuals' control. Second, we show that individuals with low socio-economic status take on chronic illness at a faster rate and this divergence starts early in life. Counterfactual analysis conveys the returns from intervening sufficiently early in the life-time. We find substantial losses in terms of life-expectancy from intervening too late. For example, stopping the divergence in chronic conditions for individuals with below-median income from age 40 would increase their life-expectancy by more than a year. Waiting until age 60 reduces this gain in life-expectancy to less than half a year. We find further savings in healthcare costs from intervening even earlier in life.

Our paper proceeds as follows. We start by discussing the data and the measurement of chronic conditions in Section 2. Section 3 then establishes the contribution of chronic conditions to the mortality gap. This motivates the construction of a chronic disease index in Section 4. Section 5 then uses the index to study how the health gap evolves over the lifecycle. Throughout the

paper we develop a conceptual framework that guides the empirical analysis and makes the link to policy. Section 6 then presents counterfactual policy analysis and studies the role of different mediators. Section 7 concludes.

Related literature We aim to contribute to three strands of the broader literature on health inequalities. First, a rapidly growing literature in economics has complemented the broad literature on social gradients (see for example Cutler, Deaton, and Lleras-Muney 2006; Cutler, Lleras-Muney, and Vogl 2011) by combining administrative data from mortality registers and tax records. The seminal contribution is by Chetty et al. (2016) for the US, but this has been documented to other countries (Kinge et al. 2019; Mortensen et al. 2016; Chen, Karimi, and Mölken 2020), for different age groups (Kennedy Moulton et al. 2022), across time (Costa 2015; Lleras-Muney and Moreau 2022), across geography (Chetty et al. 2016; Finkelstein, Gentzkow, and Williams 2021), etc. Our contribution is to measure health before death and study how the health gap leading to mortality differences at late age evolves over the lifecycle.

Second, the question of how health evolves of the lifecycle is at the centre of health economics since Grossman (1972), but the empirical evidence on the lifecycle of health inequalities is lagging behind (Lleras-Muney and Moreau 2022). There are some important exceptions documenting life-cycle patterns by socio-economic status using survey data (e.g., Case and Deaton 2005, O'Donnell, Van Doorslaer, and Van Ourti 2015) or focusing on specific health conditions (e.g., Bolt 2022). A number of recent papers also calibrate structural models using survey data to shed light on the dynamics of the health process, including De Nardi, Pashchenko, and Porapakkarm (2023), Hosseini, Kopecky, and Zhao (2021), Hosseini, Kopecky, and Zhao (2022) and Galama and van Kippersluis (2019). Our main focus is on analyzing the difference in health processes across socio-economic groups, leveraging a data-driven approach that uses rich, administrative panel data and relies on minimal assumptions otherwise.

Finally, our paper speaks more generally to the large and influential interdisciplinary literature studying the potential drivers of the health gap (Marmot 2015; Mackenbach 2019; Murray et al. 2020). An important part of this work in the economics literature has focused on specific causal pathways, both from socio-economic shocks to health outcomes (e.g., Lleras-Muney and Moreau 2022; Adda, Banks, and von Gaudecker 2009; Sullivan and von Wachter 2009; Gathmann, Jürges, and Reinhold 2015; Black, Devereux, and Salvanes 2015) and from health shocks to socio-economic outcomes (e.g., Dobkin et al. 2018; Stepner 2019), and on specific mediators of

the health gap within (high income) countries, including access to medical care (e.g., Cutler, Lleras-Muney, and Vogl 2011; Finkelstein and McKnight 2008; Deaton and Paxson 2004), health behaviors (e.g., Pampel, Krueger, and Denney 2010; Darden, Gilleskie, and Strumpf 2018), early life factors (e.g., Case, Fertig, and Paxson 2005; van den Berg, Lindeboom, and Portrait 2006), social structures and stress (e.g., Marmot et al. 1991; Sapolsky 2005). Still, thinking about their general importance, there is still much debate as to the key mechanisms that underlie health inequalities and how they should be addressed. We believe our analysis provides an ideal re-calibration of the potential importance of specific mechanisms, like for example the recent literature uncovering the causal impact of place on health (e.g., Finkelstein, Gentzkow, and Williams 2021; Kulshreshtha, Salm, and Wübker 2022), and as such provides an ideal roadmap for further empirical work.

2 Data and Measurement

This section describes the data sources and our measurement of chronic conditions. We build a panel for the full Dutch population, using micro data from the national statistical agency (Dutch: *Centraal Bureau voor de Statistiek*, CBS). The panel contains data running from 2003 and 2021, and combines multiple administrative registries and survey sources. The panel of individuals observed includes for practical purposes the full resident Dutch population (17.2m in 2016). The basis is the Municipal Register (Dutch: *Gemeentelijke Basisregistratie*, GBA). This includes an individual identifier that forms the linkage between datasets.

2.1 Data Sources

Mortality and Medicines The two main health outcomes we consider are mortality and chronic illness. Mortality is derived from death certificates in the mortality register, completed by either a physician or pathologist and collated by CBS for statistical purposes.[3] To measure chronic conditions, we use prescribed medication as described further in section 2.2. The data for prescription medication is administered by the National Health Care Institute (Dutch: *Zorginstituut Nederland*). This contains medicines dispensed to an individual, outside the hospital

3. The mortality register also includes cause of death, defined as "The disease or injury that initiated the train of morbid events leading directly to death". Instead, we are most interested in the underlying conditions; for instance hyperipidemia, rather than an acute myocardial infarction. Further, cause-of-death coding is often less reliable for more chronic diseases (Harteloh, de Bruin, and Kardaun 2010).

setting. Medicines are classified by the Anatomical Therapeutic Chemical (ATC) code, at ATC3 digit level: for example, the code $N06A$ corresponds to antidepressants. This prescription data is available from 2006 onwards.

Other health data The data on annual healthcare costs is collated by commercial data provider *Vektis*, using the raw information from health insurers. The data provided to CBS relates to costs insurable under the Dutch Health Insurance Act (Dutch: *Zorgverzekeringswet*, ZVW). This includes costs insurable under compulsory standard insurance, which represents 52% of all health and medical care costs in the Netherlands in 2022, excluding long term care. Data are annual totals split by type: for example GP costs, medicine, mental health, and hospital costs are each itemised. These data are available for the resident population from 2009 onwards. We also use information on the frequency and duration of hospitalisations, from the hospital discharge register (Dutch: *Landelijke Medische Registratie*, LMR) for the period 2011-2017.

In addition to the administrative data, representative surveys focusing on self-reported health and health behaviours and merged into the sample at the individual level. This includes large scale 'Health Monitor' surveys (Dutch: *Gezondheidsmonitor*, GEMON), fielded in 2012 and 2016. Data from around 400,000 individuals were collected, as a repeated cross-section. Information collected covers self-reported illness, sensory capacity and mobility, BMI, measures of mental health, and health behaviours: drinking, smoking, rates of physical activity. In addition, smaller scale 'Health Inquiry' surveys (Dutch: *Gezondheidsenquete*, GECON) are included, which report take-up rates of health screening activities, such as blood pressure tests. This is an annual cross-sectional survey, with approximately 9000 individuals per year.

Household Income Income data is collated by CBS from the tax authorities, and is available at the population level from 2003 onwards. In this work, we focus on *Standardized disposable household income*.[4] This measure is constructed by CBS. In line with prior literature, negative or zero disposable income households are omitted: these represent less than 1% of observations. We also follow previous literature (Chetty et al. 2016; Kinge et al. 2019) in using lagged income: We take the mean of $(Y_{t-4}, Y_{t-3}, Y_{t-2})$ to mitigate the reverse causality from health shocks on income and use pre-retirement incomes (Y_{60}, Y_{61}, Y_{62}) for those aged 65 and above. We also consider

4. Disposable income is defined by CBS as all gross income and government insurance/benefit transfers, less insurance premia, and taxes on income and wealth. This measure is standardized by CBS at the household level by dividing the sum of members' disposable income by a 'household equivalence factor', which adjusts for differences in the size and composition of households.

alternative markers of individuals' socio-economic status, including parents' income, education and wealth. We also use the panel structure of the data to shed direct light on the empirical importance of reverse causality, from health to income.

We also follow prior work in using income ranks. That is, we rank lagged incomes within gender, birth cohort, and calendar year. Much of the following analysis calls for a binary classification of relative income. For this, we define those below median income as **Low income**, and above median as **High income**. Notably, the bottom income decile includes some households with high net assets: this suggests some have targeted low incomes perhaps for tax reasons, rather than reflecting overall financial resources. However, it also predominantly includes individuals with very little financial resources, whom we wish to include. Nevertheless, in several aspects of the analysis, we treat the bottom decile separately to account for its distinct composition.

Other Administrative Data The CBS microdata environment is comprehensive in its administrative data collection. Beyond birth year, gender at birth, birth country, it also includes linkages to biological parents, household membership and composition, and residential postcode. Highest education attained is taken from a combination of administrative records for younger cohorts and labour force survey data for older cohorts. Thus education coverage is not comprehensive: around 60% for those aged 40, falling to 20% for 70 year old's. Beyond the demographic data, CBS also provide a linked employee-employer dataset (Dutch: *Banen en lonen op basis van de Polisadministratie*, SPOLIS). This provides information on jobs and earnings for employees at Dutch companies. We use this to construct a within-firm pay rank as a candidate driver of health, building on Marmot et al. (1991).

Descriptive Statistics Selected descriptive statistics on socio-economic and health outcomes are included in Table 1. To illustrate the difference in health outcomes across socio-economic groups we consider the cohort of 55 year old's in 2007, partitioned by their household income, and study the differential in survival rates over 15 years to 2022. Panel A of Figure 1 shows the differential mortality risk faced by those below median income, compared those above median income. Over that time, the cumulative mortality risk is 1.67 times greater for those from poorer households. Similarly, we can partition the 55 year old cohort by income quintile and observe subsequent survival rates, as shown in Panel B of Figure 1: the relation between income and survival probability is clearly monotonic, but the bottom income quintile faces a markedly higher

mortality hazard. By fixing membership of the income group initially, these figures abstract from the interaction between health and earnings processes as the cohort gets older, which we turn to later in our analysis.

2.2 Measurement of Chronic Disease

We follow prior work using the medicine dispensation data from the National Health Care Institute to identify chronic conditions. This approach overcomes challenges of coverage and accuracy faced by alternative approaches. In the Biomedical literature, chronic disease is measured using hospital databases or discharge abstracts, which are available only for the recently hospitalised (see, e.g., Yurkovich et al. 2015 for a review of indices using this data), and has none of the demographic information required to examine inequality in health outcomes. In the Public Health literature, chronic conditions are measured using self-reported information from representative surveys, such as the PSID or NLSY, or HRS in the case of older cohorts. While these data sources contain some demographic information, sample sizes are limited and self-reported health can be subject to non-classical measurement error, leading to biased measures of health inequality.[5]

The medicine dispensation data provides approximately population-wide coverage, from 2006 onwards. These data contain the Anatomical Therapeutic Chemical (ATC) code of all prescribed medication dispensed outside the hospital setting. Several studies have used prescription medication ATC data as indicators of chronic disease. We build on Huber et al. (2013) as our basis to translate medication data into chronic disease indicators.[6] They extended prior work using further medical expertise, to reduce type-I errors by focussing on ATC codes that are used exclusively for the treatment of a given chronic disease. We do, however, make a number of modifications, partly reflecting that our data is resolved to the ATC3 level, whereas their mapping sometimes uses ATC5 resolution. The mapping used for the bulk of our analysis is given in Table 2, with further description of specific refinements given in Appendix D.

5. For instance, Dowd and Todd (2011) found reporting differences by group implies naive estimates of health inequality are downward-biased. Besides this, survey sample sizes mean the analysis can be under-powered for rarer outcomes, such as early life mortality.

6. Huber et al. (2013) apply their mapping to establish new prevalence estimates using Swiss administrative data. Chini et al. (2011) use pharmacy data to identify the prevalence of 20 chronic conditions in the Lazio region of Italy. Examples in the Netherlands are Lamers and van Vliet (2004), constructing a mapping from ATC codes to 22 chronic conditions for the purposes of risk adjustment in the social health insurance sector, and van Ooijen, Alessie, and Knoef (2015), using pharmacy-derived chronic conditions to predict an index of self-reported health status over the lifecycle.

We can test the comprehensiveness of our mapping to a data-driven benchmark: using a LASSO regression of five-year mortality on ATC codes directly, we can see that the 25 non-zero ATC codes that remain are all captured within the refined ATC-CC mapping. We also leverage the GEMON health survey data in which individuals are asked to self-report whether they have either diabetes or high blood pressure. As seen in Figure 2, for Diabetes the concordance is high: of those who self-reported as having diabetes, 85% were detected as taking diabetes medication, conversely of those detected as taking diabetes medication, 95% self-reported having diabetes. The precision is slightly lower for the set of medications indicating cardiovascular disease, as these include conditions beyond just hypertension.

2.3 Under-diagnosis and Under-treatment

Our approach only identifies chronic conditions that are actively treated through medication. This ignores chronic conditions that are not properly diagnosed or managed. Any chronic conditions that we miss as a consequence will make us under-estimate their role for mortality. Moreover, if under-diagnosis is more severe for low-income groups, we would under-estimate the gap in chronic health. We do note, however, that given the Dutch institutional setting with universal access to high-quality healthcare differences in healthcare utilisation, and chronic medications in particular, may be small. For example, only 0.4% of poor households report unmet medical needs, compared to for example 5.1% of poor households in the UK (Eurostat 2023) or even 8.5% of *all* households in the US (National Center for Health Statistics 2022).

To understand the role of under-diagnosis and under-treatment, we present further evidence leveraging both the administrative records and survey data. While not conclusive, the evidence suggests that the scope for under-diagnosis and under-treatment is relatively limited in the Dutch context. We find some differences in under-diagnosis across income groups, so we are likely to provide a lower-bound on the health gap between low-and high-income individuals, but a lower-bound that is arguably tight.

First, not only do we find high concordance between self-reported chronic conditions and the conditions being medicated, we find that is true across the income distribution as shown in Figure 2. Both the precision and sensitivity of our medicine-based measurement of diabetes and cardio-vascular disease remain basically unchanged across income deciles. Hence, individuals who know they have a disease are just as likely to be medicated. Treatment conditional on

diagnosis seems very stable across incomes. In the empirical analysis in Section 3.3, we present more evidence suggestive of equal healthcare treatment for diagnosed chronic diseases, across income groups - both in terms of healthcare expenditures and corresponding survival rates. Of course, this still does not exclude any under-diagnosis of diseases.

Second, to gauge the potential for under-diagnosis, we compare mortality rates for individuals with no measured chronic conditions to those who have some measured chronic illness. The former group could be a mix of truly healthy individuals and individuals that are insufficiently engaged with the healthcare system. We therefore split them out into those who have no measured chronic conditions, but do take some other prescribed medication and those who do not take any prescribed medications. The former group reveals some engagement with primary care and any differential mortality between the former and the latter group would be indicative of the importance of under-diagnosis. Figure 3 presents the results. Up to age 60, the mortality rate among those without measured chronic conditions is the same, independent of whether they take other medication or not, as shown in Panel A. The mortality rates are higher for those with chronic conditions and increasing in the number of measured conditions. At older ages, the mortality rates among those without measured chronic conditions start diverging. Indeed, the mortality rate of the group without any medication even overtakes the mortality rate of groups with one or more chronic conditions. Panel B shows that this group becomes smaller and more selected with age. While in their fifties more than 2 in 10 individuals take no prescribed medication, this falls to less than 1 in 10 individuals in their seventies. Moreover, it is the bottom income decile that becomes heavily over-represented among those individuals without any medication at older ages. This is further illustrated in Appendix Figure C.1.[7]

To investigate the potential for under-diagnosis further, we consider the mortality rates depending on medicine use for different income groups in Panels C to F. The patterns confirm that the bottom income decile jumps out. Already at younger ages, individuals without medication have higher mortality rates than individuals with medication. This pattern is not present for higher income deciles, including the second income decile. For higher income deciles, the mortality rates are the same or even lower for those without medication at younger ages. They then start increasing more rapidly around age 65-70 for those without medication relative to the others,

7. Panel A shows that at age 40, people with lower income are less likely to be without any medication, consistent with them being healthy. This pattern changes drastically at older age with a clear reversal at the 10th percentile, suggestive of the distance to the healthcare system for the bottom income decile. Panel B confirms that that the under-representation of the lowest income decile in the no-medication sample steadily reverts to over-representation between the age of 45 and 70.

while for the bottom decile this divergence happens earlier. Overall, this analysis suggests that during prime ages under-diagnosis is more pronounced in the bottom income decile, but even in this group most individuals are actively connected with the healthcare system. At older ages, under-diagnosis seems to become more widespread across the income distribution, but is probably limited to less than 1 in 10 individuals who are not actively seeking care.

Finally, while the administrative data allows us to focus on the full spectrum of incomes, we can only infer under-diagnosis indirectly. Appendix Figure C.2 presents evidence from the smaller scale *Gezonheidsenquete* survey, which includes the question "*Have you had a* [Cholesterol/Blood pressure/Blood sugar] *test in the past 12 months?*". We consider the reported test rates for those who were not already prescribed the relevant medication. We find that the degree of testing is quite comprehensive; by age 70, around half who are not currently medicated for a condition are being tested for it in a given year. The trajectories of testing rates are quite similar across incomes, with low incomes testing at marginally higher rates across ages and conditions. A priori, higher testing rates for lower incomes suggest under-diagnosis of lower incomes is not an issue, but they may also be at higher risk. We therefore study the number of *newly* prescribed in the year following the survey, and find somewhat higher prescription rates per test for low incomes, which suggest a higher share of positive test results (since Appendix Figure C.2 establishes that positive tests translate equally into filled prescriptions). Hence, lower incomes seem to be somewhat under-tested, and hence under-diagnosed. Still, the test rates are already high and the detection rates are low overall (10-15% for cardiovascular disease and hyperlipidemia, and 2-3% for diabetes), suggesting that it is unlikely to have a first-order impact on our analysis of the prevalence of chronic illness using medicated conditions.[8]

3 Chronic Disease and the Mortality Gap

This section documents the predominant role chronic conditions play in explaining the mortality gap between low and high SES groups. We show that the gap can be mostly explained by differences in the prevalence of chronic conditions and not by differences in treatment of chronic conditions. While mortality differences are only apparent later in life, the descriptive evidence in

8. Naturally it would be of interest to perform this analysis throughout the income distribution, but we are limited by the *Gezonheidsenquete* sample size. In ongoing work, we are using the merged cancer registry and find some differences in the stage of diagnosis across income groups, but conditional on stage and cancer type, the profile of treatment is very similar (Danesh et al., in preparation).

this section demonstrates the value of using chronic conditions to study differences in health earlier in life.

3.1 Static Framework

We start by providing a conceptual framework that guides our empirical analysis and helps highlighting its policy implications. Consider a linear model of mortality for an individual i at age a:

$$M_{i,a} = \alpha_{i,a} + \beta_{i,a} CC_{i,a} \tag{1}$$

where $M_{i,a}$ denotes the mortality rate. The vector $CC_{i,a}$ is short-hand notation for the individual's chronic conditions and the relevant interactions between them. The intercept $\alpha_{i,a}$ captures health (e.g., infectious diseases) and external factors (e.g., accidents) affecting mortality, unrelated to chronic diseases. The slope $\beta_{i,a}$ denotes the linear healthcare technology that converts chronic conditions into mortality. The triple $\{\alpha_{i,a}, \beta_{i,a}, CC_{i,a}\}$ may be different across socio-economic groups. Our focus is on the mortality gap at a given age a between individuals with low vs. high income:

$$\underbrace{M_{L,a} - M_{H,a}}_{\text{Mortality Gap}} = [\alpha_{L,a} - \alpha_{H,a}] + \underbrace{[\beta_{L,a} - \beta_{H,a}]}_{\text{Treatment Gap}} CC_{H,a} + \beta_{L,a} \underbrace{[CC_{L,a} - CC_{H,a}]}_{\text{Prevalence Gap}}, \tag{2}$$

where $Z_{Y,a} \equiv E\left(Z_{i,a} | Y_{i,a} \in Y\right)$ denotes the average outcome for individuals with income $Y_{i,a}$ in income group Y. We can consider this gap at any age a, but mortality differences become only sizeable at older ages. As the baseline marker for socio-economic status, we consider household income, but we extend our analysis using education and parental income too.

The decomposition quantifies the potential importance of three different factors. First, different income groups can be subject to different chronic conditions ($CC_{i,a}$). Potential reasons include differences in genetics, health behaviors, environmental exposure, work conditions, etc across socio-economic groups, but also the reverse channel where individuals' earnings potential depends on their health and ability to work. We refer to these jointly as *prevalence* effects. Second, chronic conditions may have differential health implications ($\beta_{i,a}$) across socio-economic groups, e.g., due to differential access to healthcare, differential treatment of chronic conditions, etc. We refer to these as *treatment* effects. Finally, individuals are exposed to other health and external factors ($\alpha_{i,a}$), which may differ across socio-economic groups. These *residual* effects can also

include differences in under-diagnosis as discussed before.

Despite the descriptive nature of the decomposition, the 'prevalence gap' and the 'treatment gap' point to different policy options for governments in trying to reduce the health gap. The former suggests the value of public health interventions that can reduce the burden of chronic illness for lower SES individuals, while the latter suggests the need for the healthcare system to improve either access or take-up of health treatments for these individuals. Importantly, the respective gaps provide arguably a lower-bound on the impact that interventions on chronic conditions for individuals with low income can have on the health gap. The reason is that as we improve individuals' health, we may improve their socio-economic circumstances, which can further improve their health.[9] In Section 5.1, we explicitly try to separate the causal pathways between health and socio-economic outcomes using the panel structure of our data. Still, we can in principle circumvent this challenge when considering policies that directly intervene on individuals' health and not on individuals' socio-economic status. The prevalence gap provides a lower-bound on how much intervening on chronic conditions can reduce the health gap, granted that the treatment effect β captures the causal effect of the chronic conditions on the individual's health. A key challenge in estimating the treatment effect is that the residual mortality effects differ across socio-economic groups and thus need to be controlled for. We turn to this in Section 4.

3.2 The Prevalence Gap

We first consider the prevalence of chronic conditions and how much it contributes to the mortality gap across income groups. We focus on individuals at age 70, as mortality rates and thus differences across income groups become more apparent then. Figure 4 reports the difference in prevalence for all 22 chronic conditions across the low- and high-income group. We consider men and women separately and pool all observations for years between 2013 and 2021. The overall pattern is very clear as the burden of all common chronic conditions falls more on

9. To illustrate this, consider an intervention that changes the incidence of chronic conditions:

$$\frac{\partial M_{i,a}}{\partial CC_{i,a}} = \beta_{i,a} + \left[\frac{\partial \alpha_{i,a}}{\partial Y_{i,a}} + \frac{\partial \beta_{i,a}}{\partial Y_{i,a}} CC_{i,a} + \beta_{i,a} \frac{\partial CC_{i,a}}{\partial Y_{i,a}}\right] \frac{\partial Y_{i,a}}{\partial CC_{i,a}}.$$

The second term captures how much the improvement in health improves an individual's SES and how this further improves her health. This can be through either one of the three factors in the health production function: the incidence of chronic conditions, its treatment or the residual health part. This indirect effect is arguably positive and would add to the direct effect, i.e., $\frac{\partial M_{i,a}}{\partial CC_{i,a}} \geq \beta_{i,a}$. A similar argument can be made for interventions that improve the treatment of chronic conditions.

the low-income than on the high-income individuals. The differences in prevalence between the high-income and low-income groups are often sizeable. This is also true for the most prevalent chronic conditions, including cardio-vascular disease, diabetes, high cholesterol, respiratory disease and acid-related disorders. For example, 12.8% of high-income men are treated for diabetes, while this share is 17.5% and 20.8% for the respective low-income groups. We split out the below-median income group between the bottom decile and other deciles, acknowledging the potential under-diagnosis for the bottom decile group. The prevalence is higher for the bottom decile for most conditions, but there are a few exceptions, including cardio-vascular disease and high cholesterol.

To evaluate how much the different prevalence of chronic conditions explains the mortality gap, we run linear age-specific regressions of 5-year mortality M_i on income Y and a set of controls X_i that varies by specification:

$$M_{i,a} = \delta_{Y,a} Y_{i,a} + X_{i,a} \gamma_a + \varepsilon_{i,a}. \tag{3}$$

We consider the above-median income group as the reference group and report the mortality for the below-median income group in comparison to the above-median income group. We pool the observations for years between 2013 and 2016.[10] All our specifications control for year and gender and allow for interactions between gender and the health-related variables.

Panel A of Figure 5 shows for the 70-year olds how much δ_Y changes when we include the chronic conditions $CC_{i,a}$ as controls. The average mortality rate in the high-income group is 66‰. For the low-income group, the mortality rate is 44‰ higher (row A1). However, when we control for the 22 chronic conditions observed in the prior year, the difference in mortality rates drops from 44 to 31‰ (row A2). That is, about one third of the gap in mortality between the low- and high-income individuals can be explained by the difference in measurable chronic conditions. Adding further lags of chronic conditions (row B1) hardly changes this, consistent with their persistence over time.

The advantage of the simple regression framework is that we can also compare the role of chronic conditions for the mortality gap to other measurable health information, leveraging our granular data with detailed information on healthcare utilization and diagnoses. Quite strikingly, the mortality gap reduces not much further when adding other health-related controls, once we

10. For individuals aged 65 and older, pre-retirement income is defined as the average income from 60 to 62 years of age. Since data on income is only available starting in 2003, the oldest cohort whose pre-retirement income is available turns 70 year old in 2013. Since we use data on mortality until 2021, the dependent variable, five-year mortality, is only observed until 2016.

have controlled for chronic conditions. Panel A of Figure 5 shows that the estimated gap hardly reduces when adding other prescribed medication (row B2), or when adding comprehensive information on hospital visits, including the number, length and main diagnosis of hospital visits (row C1), or when adding categorized healthcare expenditures in the prior year (row C2), including primary care, specialist care, medicines, mental health care, etc. Interestingly, these variables do strongly increase the explanatory power of our regression model.[11] Including all other health-related variables jointly more than doubles the R-squared (from 0.05 to 0.13) relative to the specification with last year's chronic conditions, but it only reduces the mortality gap from 31‰ to 26‰ (row D).

The two main results hold when extending the analysis to individuals between 40 and 70, as shown in Panel B of Figure 5. First, chronic conditions explain a substantial part of the difference in the mortality gap between high- and low-income individuals, ranging between 30 and 40 percent. At younger ages, the mortality rate can be twice as high for the low-income group compared to the high-income group. However, it is only half as high once we control for chronic conditions. Second, other measurable health information in our data does not help much in further closing the gap. At younger ages, the further reduction in the mortality gap is not even statistically significant.

In sum, this lends strong support for our focus on measurable chronic conditions to analyze the health gap across income groups.

3.3 The Treatment Gap

We now turn to the difference in treatment of chronic conditions and whether this contributes to the mortality gap across income groups.

Our prior regressions did not allow chronic conditions to differentially affect mortality across income groups. Including interactions between chronic conditions and the income groups in regression equation (3) allows to compare how chronic conditions relate differently to mortality for the two income groups. As discussed, this becomes relevant when chronic conditions are treated differently, but also if under-diagnosis differs across income groups, since any under-diagnosis would bias the estimated effect of chronic conditions downward. To gauge the attenuating effect, we again separately show the bottom income decile for this estimation. Of

11. Without controlling for chronic conditions, each of these health-related variables contribute meaningfully to the mortality gap, albeit less than the chronic conditions (see Appendix Figure C.4).

course, some caution is naturally warranted when interpreting these separate coefficients. The chronic conditions or underlying medicine use can be correlated with other factors affecting mortality. We explicitly address these potential confounders in Section 4.3 when constructing our chronic disease index.

The regression estimates are shown in the bottom panels of Figure 4. The regression controls for all chronic conditions jointly. The differences between the estimates across income groups tend to be small and no clear pattern emerges, especially when we compare this to the difference in prevalence in the top panels of the Figure. For example, diabetes for men is associated with an increase in the mortality rate of 46‰, 51‰ and 43‰ for high income, low-income and bottom deciles respectively. One notable exception is respiratory illness, where the associated mortality rate is significantly higher for low-income individuals than for high-income individuals (71‰ cf. 47‰ for women). We also find that some chronic conditions (or the underlying medication) are associated with lower mortality rates, such as migraine and hyperlipidemia. These protective effects are strongest for the bottom income deciles.

To evaluate how much differential treatment effects contribute to the mortality gap and compare this to the importance of difference in prevalence, we provide a Oaxaca-Blinder decomposition, following equation (2). Overall, the difference in prevalence clearly outweighs the difference in treatment (see Panel A of Appendix Figure C.5). At the age of 70, we find that differences in treatment effects do not contribute to the mortality gap. At younger ages, if anything, the mortality rates for individuals with the same measured chronic conditions are on average higher for low-income than for high-income individuals.

The limited difference in how chronic conditions relate to mortality across income groups is important because it suggests that individuals with chronic conditions receive similar healthcare as individuals without chronic conditions. We can test this more directly by considering healthcare expenditures and studying how these relate to chronic conditions for individuals with low vs. high income. Mirroring Figure 5, Appendix Figure C.3 show that while health expenditures are higher for low-income individuals than for high-income individuals (up to a difference of 989 euros at age 60), this gap is mostly explained by controlling for chronic conditions. This holds again at all ages and is suggestive of equalized treatment by the healthcare system across income groups. We confirm this further through a Oaxaca-Blinder decomposition of the gap in health expenditures (see Panel B of Appendix Figure C.5), where, if anything, the low-income

individuals seem to receive more healthcare for the same measured chronic conditions.[12]

In sum, the difference in treatment effects is small relative to the difference in prevalence. This further confirms that the Dutch context is well-suited to study how gaps in chronic illness as measured through prescribed medicines arise over the life-cycle. However, as allowing for differential mortality rates does not explain a much larger share of the mortality gap, this also implies that at least 50 percent is driven by factors other than measurable chronic conditions. It is thus essential to account for these residual effects when interested in the mortality effects of chronic conditions.

4 Chronic Disease Index

The prior section demonstrated the pre-dominant role of chronic conditions in explaining mortality gaps. We now turn to the construction of a chronic disease index. Our aim is to provide a comprehensive index of health that can be measured throughout the lifetime. To achieve this, we focus on how an individual's chronic illness predicts mortality in old-age, accounting for co-morbidities and interaction effects. We then re-weight chronic illness at any age to reflect the old-age mortality based on *point-in-time* chronic conditions. To provide the mortality interpretation, we need to control for other confounding factors affecting mortality and correlating with chronic conditions. In doing so, we make the chronic disease index comparable both across individuals and within individuals, with the convenient interpretation of capturing health differences that translate into mortality at old age.

12. The higher mortality and healthcare expenditures on low-income individuals with chronic conditions could be driven by their chronic conditions being actually worse. We briefly gauge this by allowing for more lags of chronic conditions and interactions between them. Appendix Figure E.2 shows that this indeed reduces the importance of both differential mortality rates and healthcare treatments. In fact, for the older age groups, the mortality rates associated with chronic conditions are now higher for the high-income group than for the low-income group. However, excluding the bottom decile from the sample makes the contribution of differential treatment effects on mortality more positive again (see Appendix Figure E.3). The contribution is no longer negative at age 65, while it still is at age 70, consistent with under-diagnosis becoming more pervasive in low-income groups beyond the bottom decile. We note that the estimated healthcare treatments for those with chronic conditions seem less affected by this sample restriction, which is consistent with the notion that both under-diagnosed and healthy individuals receive limited healthcare treatments.

4.1 Statistical Framework

To guide our prediction exercise, we impose the following structure on our static model introduced in the prior section:

$$M_{i,a} = \alpha_{i,a} + \beta_{i,a} CC_{i,a} \tag{4}$$

$$= X_{i,a}\gamma_a + \beta_a CC_{i,a} + \varepsilon_{i,a}, \tag{5}$$

with $E\left(\varepsilon_{i,a}|CC_{i,a}, X_{i,a}\right) = 0$. Hence, we assume that chronic conditions have the same mortality effects across individuals, $\beta_{i,a} = \beta_a$, and that non-chronic differences in mortality across individuals can be captured by observables $X_{i,a}$ and independent unobserved heterogeneity $\varepsilon_{i,a}$.

Under these assumptions we can construct a chronic disease index:

$$CDI_{i,a} \equiv \hat{\alpha}_{old} + \hat{\beta}_{old} CC_{i,a}, \tag{6}$$

using an unbiased estimate $\hat{\beta}_{old}$ of the mortality impact of chronic conditions at the old reference age β_{old}, where

$$M_{i,old} = X_{i,old}\gamma_{old} + \beta_{old} CC_{i,old} + \varepsilon_{i,old}. \tag{7}$$

The intercept $\hat{\alpha}_{old}$ of the index captures the average counterfactual mortality rate in the absence of chronic conditions at old age.

We can calculate the chronic disease index $CDI_{i,a}$ for each individual i at *any* age a given their point-in-time chronic conditions $CC_{i,a}$. This then allows to compare the chronic illness across individuals and over the lifecycle, independent of other observable differences, as measured by the expected mortality rate when subject to the point-in-time chronic conditions $CC_{i,a}$ at the old reference age.

In our empirical analysis in Section 3, we have shown the importance of chronic conditions for mortality, but also found that more than half of the mortality gap between income groups cannot be explained. Since the prevalence of chronic conditions is so different across socio-economic groups, this underlines the importance of including socio-economic controls when estimating the mortality effects of chronic conditions. On the other hand, we have found limited evidence for heterogeneity in mortality effects across socio-economic groups, which supports the construction

of a uniform index with the same mortality interpretation.[13]

We can gauge the accuracy of our index by considering residualized mortality at old age, $\tilde{M}_{i,old} = M_{i,old} - X_{i,old}\hat{\gamma}_{old}$, and evaluating how this differs for individuals with difference CDI. Under unbiasedness, we have

$$E(\tilde{M}_{i,old}|CDI_{i,old} = CDI + \Delta) - E(\tilde{M}_{i,old}|CDI_{i,old} = CDI) = \Delta. \tag{8}$$

This can be violated due to unobserved heterogeneity across individuals with different chronic conditions, which would be wrongly attributed to the mortality effect of these chronic conditions. Our focus is on comparing the CDI across socio-economic groups. Hence, if the unobserved heterogeneity and how it correlates with chronic conditions is similar across socio-economic groups, the comparison of differences in CDI across SES groups is still meaningful. In other words, we could relax the conditional independence assumption, as long as confounders apply similarly in different socio-economic groups.[14]

4.2 Double-Selection and Prediction

The empirical task is to estimate the mortality risk M_i at old age from 22 chronic conditions, given a rich set of socioeconomic controls. We continue to focus on the 70-year olds as mortality rates are sizeable at that age. Even with a population-wide sample, both the array of chronic conditions and potential socioeconomic control variables is large enough that including all values, with interactions, to predict mortality for a given age could lead to overfitting and potentially underidentification. Hence it is necessary to use a variable selection step to identify the most relevant variables and/or interactions. Belloni, Chernozhukov, and Hansen (2014) describe a procedure to conduct inference on a focal variable CC_i, with a double selection method to choose relevant control variables X_i in equation (5). Their insight is that estimating this model in a single LASSO step could omit certain relevant controls, if they are highly collinear with the

13. The exception was the bottom income decile where under-diagnosis seems to be important, which would lead us to under-estimate the mortality effects of chronic conditions. Given the limited evidence for under-diagnosis when excluding the bottom income decile until the age of 70, we will estimate our chronic disease index predicting mortality for individuals at age 70 using the 2nd to 10th income decile.

14. Of course, any comparison of how mortality changes with chronic conditions across income groups still depends on the respective treatment effects for different socio-economic groups and thus relies on $\beta_{i,a}$ not to differ across socio-economic groups. With heterogeneity in treatment effects, our chronic condition index reflects the mortality impact of chronic conditions on those who bear them, akin to the *average treatment effect on the treated*, which would complicate any comparison of the index between socio-economic groups as they fundamentally differ in the prevalence of these chronic conditions.

focal variable. Instead they propose a two-step procedure, with separate LASSO estimation for both mortality M_i and the focal variable CC_i and to determine the relevant set of controls to be included.

We build on this procedure, but rather than a single focal variable, we construct an index based on a focal vector of 22 chronic conditions $CC_i = \{c_i^1, ..., c_i^{22}\}$. Hence we run a total of 24 LASSO estimations to select the relevant socioeconomic controls: (i) a first LASSO estimation to establish the socio-economic variables relevant for mortality, (ii) a LASSO estimation for each of the 22 chronic disease indicators, and (iii) a final LASSO estimation to determine the set of relevant interactions between chronic condition types and lags. That is, omitting the a subscripts for convenience,

$$\begin{aligned} M_i &= X_i'\theta_m + \zeta_i \\ c_i^k &= X_i'\theta_c^k + \nu_i^k \qquad \forall \quad k = \{1, \ldots, 22\} \\ M_i &= f(CC_i)'\theta_v + \varepsilon_i \end{aligned} \tag{9}$$

Here M_i denotes an indicator for five-year mortality, c_i^k is an indicator for the kth of 22 chronic conditions. X_i is the set of all potentially relevant socioeconomic information. $f(CC_i)$ is the basis of all potentially relevant chronic condition information. It include three years of chronic condition indicators, within-condition interactions across different lags, and two-way cross-condition interactions for the most recent lag. Socioeconomic variables and chronic condition interaction terms that are selected in *any* of the LASSO estimations are then included as regressors in the final prediction step.

The final estimation equation is as follows:

$$M_i = X_i^{*\prime}\beta_X + f(CC_i)^{*\prime}\beta_{CC} + \zeta_i, \tag{10}$$

X_i^* denotes the set of socioeconomic variables found to be relevant in one or more of the LASSO estimations: $X_i^* = \{x_i : \hat{\theta}_m > 0 \cup x_i : \hat{\theta}_c^k > 0\}$, and $f(CC_i)^*$ denotes the union of the set of chronic conditions and their relevant interactions from the final LASSO estimation: $f(CC_i)^* = \{CC_i \cup f(CC_i) : \hat{\theta}_v > 0\}$. We estimate equation (10) using a linear probability model, by gender and with calendar year fixed effects absorbed. To minimize the impact of under-diagnosis, we exclude individuals who belong to the bottom income decile based on pre-retirement income. We randomly select 50% of the population of 70-year-old individuals to

estimate model (10). The results regarding accuracy and predictive value below are reported for the hold-out sample. We then calculate the CDI for an individual at any age a in the full 2009-2021 period, regardless of socio-economic status, as

$$CDI_{i,a} \equiv \bar{X}^{*\prime}\widehat{\beta}_X + f(CC_i)^{*\prime}\widehat{\beta}_{CC}, \tag{11}$$

where the intercept captures the mean socioeconomic effects.

4.3 Accuracy and Predictive Value

As shown in Section 3.2, chronic conditions are unevenly distributed by income. Hence, the purpose of the LASSO procedure described above is to ensure our estimate of CDI is orthogonal to SES measures, rather than being biased by a correlation between SES and chronic conditions. Appendix Figure C.6A shows this bias to be around 16%.[15] This implies a "naively" estimated CDI without the LASSO-selected SES controls would overstate the chronic disease health gap by 16%.

This bias also has implications on how we test prediction accuracy. We cannot simply compare the observed mortality rates and the predicted mortality rates reflected by the CDI to evaluate its accuracy, so instead we focus on residuals. The "CDI & SES residual" series in Appendix Figure C.6A also shows that, when taking out the socio-economic correction, the conditional mean error $E\left[\hat{\zeta}_i \mid CDI_i\right]$ is close to zero over the entire range of CDI predictions, suggesting that we accurately predict mortality, also at the bottom and the top of the risk distribution. This is expected, given the LASSO procedure ensures the CDI measure is saturated with respect to all relevant chronic conditions and their interactions.

We also want to understand how the CDI performs across incomes. Appendix Figure C.6A compares mean error for the low incomes versus high incomes separately (i.e., $E\left[\hat{\zeta}_i \mid CDI_i, Y_i = Y_L\right]$ vs. $E\left[\hat{\zeta}_i \mid CDI_i, Y_i = Y_H\right]$). As shown, there is no significant divergence of the residuals for these two subpopulations. This suggests our CDI is not biased across incomes, and also that the additive separability assumption is a reasonable one.

Having established the accuracy of the CDI, we can also evaluate its predictive value. We document substantial heterogeneity in the CDI; we find a predicted 5-year mortality rate of

15. This bias is equivalent to an omitted variable bias of $\tilde{\beta}_{CC}$ relative to $\widehat{\beta}_{CC}$, where $\tilde{\beta}_{CC}$ is estimated from $M_i = f(CC_i)^{*\prime}\tilde{\beta}_{CC} + \tilde{\zeta}_i$.

44‰ for the healthiest 10 percent of the 70-year olds in the CDI distribution, which compares to 251‰ for the sickest 10 percent of the sample. The dispersion is comparable for men and women and for different incomes groups separately, but substantially smaller for younger cohorts (see Appendix Figure F.2). Still, even for the 70-year olds, the out-of-sample R-squared when regressing 5-year mortality on the CDI index is only 5.2%. The predictive power of the index is quite modest, but the dependent variable is a binary, random realization of a probability. The low R-squared thus reflects that death remains mostly random.[16] The predictive value increases when allowing for interactions between the chronic conditions and adding more lags, but the additional information from adding lagged conditions quickly tapers off: this is illustrated in Appendix Figure F.1 We find that the in-sample R-squared value (at 5.6%) for the prediction model is only slightly above the out-of-sample R-squared, which indicates limited risk of over-fitting given the large sample size. Overall, the CDI predictions are robust to a number of modelling decisions in their construction. In particular, the choice of lags, the linkage function of the model, the target ages, the estimation sample and the method to choose the λ penalty parameter in the LASSO estimation are each tested and discussed in Appendix F.

We can use the CDI to revisit the marginal effect of separate chronic conditions on mortality, but now accounting for their interactions and lag structure, and controlling for socio-economic factors. Appendix Figure C.7 shows these marginal effects at age 70, defined as

$$\beta^j = E[CDI_{70} \mid c^j_{69} = c^j_{68} = c^j_{67} = 1, c^{-j} = \overline{c^{-j}}] - E[CDI_{70} \mid c^j_{69} = c^j_{68} = c^j_{67} = 0, c^{-j} = \overline{c^{-j}}] \tag{12}$$

for any of the 22 chronic conditions c^j and evaluating all other chronic conditions c^{-j} at their average prevalence. Overall, there is a high correlation between the CDI marginal effects and our earlier estimates regressing mortality on all chronic conditions jointly, shown in Figure 4B and reproduced in Appendix Figure C.7. However, the comparison confirms again that by not accounting for socio-economic factors, we would over-estimate the mortality associated with a specific chronic condition. This correction can be significant and very large. For example, for anemia the estimated mortality rate would increase from 10.7% to 15.7% for women. For

16. While our focus is on mortality, the CDI strongly correlates with self-reported health too ($r = 0.36$). Moreover, while low-income individuals report worse health conditional on CDI, the association with CDI and self-reported health is the same for both groups (see Appendix Figure F.3). An alternative method for assessing the model's predictive capability involves examining the area under the ROC curve (AUC). This illustrates the relationship between the false-positive rate and the true-positive rate, taking into account the chosen threshold for converting predicted mortality probability into a binary outcome. The AUC equals 0.66, which surpasses the minimum of 0.5 expected for a random prediction, but is far from the maximum of 1 indicative of complete accuracy.

dementia, it would be inflated from 24.1% to 38.5%. In a similar spirit, we show how the estimated mortality rates increase further when not controlling for the prevalence of other conditions. These adjustments tend to be even more sizeable as many individuals with chronic illness suffer from multiple conditions. For example, the estimate of the mortality rate associated with anemia would further increase from 15.7% to 23.1%. For cardiovascular disease, it would increase from 2.1% to 4.3%. By controlling for socio-economic factors and co-morbidities in the construction of the CDI, we are arguably getting the estimates closer to the impacts these diseases have on mortality. Still, caution remains warranted when interpreting these estimates causally. Note for example that we continue to find protective effects for migraine and intestinal inflammatory diseases, with presumably the selection into medication offsetting the mortality risk of the disease itself.

5 Chronic Disease over the Lifecycle

This section uses the CDI to study how the health gap arises over the life-cycle. Two key advantages of the CDI are that we can measure health earlier in life and observe it repeatedly for the same individual. As a result, we can evaluate how much of the health gap translating into mortality at old-age already materializes earlier in life. We can also separate how much of the health gap is driven by individuals in different income groups aging at different rates vs. individuals with different health sorting into different income groups.

5.1 Dynamic Framework

We are interested in the evolution of the health gap $\Delta CDI_a \equiv CDI_{L,a} - CDI_{H,a}$ over the life-cycle. A simple comparison of the health gap across ages allows to evaluate when in the life-cycle the health gap opens up and how much of the gap observed at old-age, already materializes earlier in life. The interpretation of the fraction $\Delta CDI_a / \Delta CDI_{70}$ is particularly useful as the CDI captures the predicted mortality at age 70 and thus allows us to capture differences in mortality later in life ΔM_{70} at earlier ages when the observed differences in mortality ΔM_a are not apparent yet. While before we argued that the *static* prevalence gap, captured by ΔCDI_a, helps quantifying the potential for targeted health interventions, any *dynamic* change in this gap sheds light on the desirable timing of these interventions. However, when the SES measure is endogenous to

individuals' health, any comparison of the static gap across ages confounds differential growth in CDI across SES groups and the sorting into different SES groups based on health. Our dynamic measure of health and the panel structure of our data allow us to separate the two forces.

For any given individual we can observe how his or her CDI grows due to the incidence of new chronic conditions. The within-individual change $dCDI_{i,a} \equiv CDI_{i,a+1} - CDI_{i,a}$ thus reflects how an individual's chronic health develops with age, which we can refer to as *biological aging*. We evaluate this aging process separately for different income (or socio-economic) groups, $E\left(dCDI_{i,a}|Y_{i,a} = Y\right)$. Conversely, for any given individual we can also observe how his or her income changes and how this relates to their chronic health. In particular, as the composition of an income group Y changes, we can evaluate how much of the sorting in the income group is related to chronic illness, $E(CDI_{i,a+1}|Y_{i,a+1} = Y) - E(CDI_{i,a+1}|Y_{i,a} = Y)$. Any causal effect of chronic illness on the income process would be reflected in this term, but other underlying factors that affect both the health and the income process may be at play as well.

The observed change in CDI across ages for income group Y can thus be decomposed as follows:

$$\begin{aligned} dCDI_{Y,a} &\equiv E(CDI_{i,a+1}|Y_{i,a+1} = Y) - E(CDI_{i,a}|Y_{i,a} = Y) && (13) \\ &= \underbrace{E\left(dCDI_{i,a}|Y_{i,a} = Y\right)}_{\text{Biological aging}} + \underbrace{E(CDI_{i,a+1}|Y_{i,a+1} = Y) - E(CDI_{i,a+1}|Y_{i,a} = Y)}_{\text{Health-Based Sorting}}. && (14) \end{aligned}$$

By separating out the sorting component, we can meaningfully compare the biological aging across different income groups and evaluate how much differences in the biological aging contribute to the health gap over the lifecycle. Still, we should not interpret these differences as being caused by the differences in income or socio-economic status, but, as argued earlier, they allow us to quantify the potential value of targeting health interventions that reduce the incidence of chronic conditions for a specific socio-economic group. In this spirit, we can re-construct the CDI for a specific socio-economic group as the accumulation of the aging effects between age a_0 and a as

$$CDI^{aging}_{Y,a} = CDI_{Y,a_0} + \sum_{\tilde{a}=a_0}^{a-1} E\left(dCDI_{i,\tilde{a}}|Y_{i,\tilde{a}} = Y\right). \tag{15}$$

Assuming a Markovian process for health with socio-economic status as the relevant state variable, this counterfactual captures the lifecycle of the CDI for individuals who remain in a specific socio-economic group and thus allows evaluating the direct returns from intervening on the differential incidence for that group. As in our discussion of the prevalence gap, this

incidence gap can again be interpreted as a lower-bound for the returns from intervening as the improvements in health may improve their socio-economic status. Interestingly, our estimate of health-based sorting sheds empirical light on whether health can have meaningful impacts on income indeed.

We note that the presented decomposition assumes a balanced panel where all individuals are observed at ages $a+1$ and a. In practice, different cohorts are observed at different ages and thus individuals enter and exit the sample at different ages. We will account for these entry and exit effects in our decomposition (see Appendix G). In particular, individuals can also exit the sample due to death and this will attenuate the age-profile of the average CDI as those with higher CDI face higher mortality rates. We calculate this attrition effect due to death as:

$$E(CDI_{i,a}|Y_{i,a} = Y) - E(CDI_{i,a}|Y_{i,a} = Y, S_{i,a+1}),$$

where $S_{i,a+1}$ denotes survival into age $a+1$.

5.2 Health Gap over the Life-Cycle

To evaluate how the CDI evolves over the life-cycle, we first make a CDI prediction for all individuals in our sample between 10 and 70, based on the lagged chronic conditions at their respective ages. Remember that individuals are divided in the low-income vs. high-income group at age a depending on their average income between ages $a-4$ and $a-2$, except from 65 onwards, we consider average income between ages 60 and 62. Hence, after age 64, the composition of cohorts becomes more skewed towards younger cohorts.

Figure 6 plots the average CDI for the low- and high-income groups at different ages, pooling all observations in our sample for each age. The top panel shows the levels and thus how the CDI gap evolves over the life-cycle. The bottom panel expresses this relative to CDI gap at age 70. The health gap is close to zero until early adulthood with a difference of only .08 percentage points at age 20. That is, the health gap at age 20 would translate into a .08 percentage points difference in 5-year mortality at age 70. In early adulthood, the gap between the CDI's opens up and reaches a difference of more than .3 percentage points by the age of 30. The gap between the two income groups then gradually increases between mid-age and old age up to a difference of 1.3 percentage points at 65. After 65, the divergence seems to stop. As discussed, we know that

under-diagnosis becomes more important at those ages, especially for low-income groups, but we also note the difference in income group definitions and the corresponding change in cohorts.

The CDI allows us to measure health throughout life and thus pick up potential health differences at younger ages. Panel B of Figure 6 shows that the gap in CDI's at age 40 is 48 percent of the gap at 70. That is, about half of the health gap at age 70 has already materialized at age 40. When using mortality rates to evaluate health gaps, the picture is very different. At age 40, the mortality gap is only 7 percent of the mortality gap at 70. Of course, mortality rates become exponentially more important at older ages, but it would be misguided to conclude that differences in health only arise later in life. The CDI allows us to capture the health differences that translate into mortality later in life much earlier.[17]

Separate Chronic Conditions The goal of the CDI is to provide a comprehensive measure of health. However, having established how the CDI gap opens up already during early adulthood and grows steadily during adulthood raises the question which conditions are driving this. To evaluate this, we calculate the gap in prevalence for each chronic condition and scale this gap by the condition's marginal impact on the CDI.[18] Appendix Figure C.8A plots the evolution of the contribution to the health gap for six of the most relevant conditions according to this metric. At younger ages, we find that differences in mental health conditions captured by psychoses and psychological disorders contribute most to the health gap. But as the health gap becomes more important at older ages, we also see that diabetes and cardio-vascular diseases increase in their importance. The role of pain and respiratory disease in explaining the gap remains more stable across ages. Appendix Figure C.9 compares the life-cycle path for these six chronic conditions.

Further Heterogeneity by SES We can extend our analysis to other socio-economic measures and for different sub-groups, but the overall empirical patterns are robust. Appendix Figure C.10 panel A shows that the differences are somewhat larger for men than for women and that the dynamic patterns are different around child-bearing ages. Panel B considers income quintiles and shows that the bottom quintile stands out. The gap at a given age increases linearly when

17. This early age health gap is present in the self-reported health profile, as shown in Appendix Figure F.3B: however the profile is subject to some sampling error, and since self-reported health response is categorical not cardinal, it is difficult to interpret mean gaps in the same way as the CDI. Further, given the available surveys are repeated cross-sections, we cannot look at within-individual effects.

18. The relative contribution is computed as $\kappa^j = (S^j_L - S^j_H) \cdot \beta^j$, where S^j_Y is the share of income group Y with condition j, and β^j is the marginal effect on predicted CDI, as depicted in Appendix Figure C.7.

considering different quintiles, except for the bottom quintile. This non-linear pattern is similar as for the mortality rates considered in Figure 1, but now adds the insight that this patterns already materializes early in life. At age 50, the CDI gap between the bottom and top income deciles already exceeds 2 percent. Panels C and D plot the average CDI by education groups and mother's income respectively. These socio-economic measures are more stable at older ages, but the coverage for both variables decreases with age. We again find very large differences overall. We find somewhat larger differences in CDI already at 20, especially when considering high-school drop outs to the others, and a less dramatic opening of the gap in CDI in early adulthood. Panel E shows the split by net wealth, which is again endogenous over the life-cycle, but allows us to make a meaningful comparison beyond age 78 (which is the oldest age at which we can measure pre-retirement income). Interestingly, like for income, the gap in CDI between wealth groups does not meaningfully increase after retirement ages. Moreover, for both wealth groups, the average CDI decreases for the individuals who survive beyond 85 and older. Finally, Panel F zooms in on specific cohorts, splitting them into low- and high-income groups in 2009 and then showing how the CDI diverges for the respective income groups until 2021. Overall, these panels highlight that the difference in age-gradients between socio-economic groups depends on compositional changes too and thus do not only capture differences in biological aging, which we turn to next.

5.3 Differential Aging over the Life-Cycle

While the average CDI is rapidly increasing for the low-income group during early adulthood, for the high-income group the CDI is almost flat up until mid-age and even slightly decreases around age 30. The vertical gap in CDI in Figure 6 also leads to substantial horizontal differences between the ages at which the same CDI is reached. This horizontal difference is commonly referred to as the 'biological age gap'. For example, already at age 35 the average CDI for the low-income group surpasses the average CDI for the high-income group at age 50.[19] However, any horizontal comparison of the CDI across ages for a given income group mixes biological aging and health-related sorting. Following equation (19), we can separate the two forces. To do this properly, we also account for the unbalanced nature of the sample across ages and thus separate out attrition due to mortality and cohort effects as we move to older ages. The full

19. These biological age gaps increase up to 25 years when comparing the bottom and top income quintiles and even up to 34 years when comparing high-school drop outs and post-graduates (see Appendix Figure C.10)

decomposition is described in Appendix G.

Table 3 shows the results of this dynamic decomposition. Both differential aging and health-based sorting contribute substantially to the CDI gap observed at old age. The role of differential aging, however, is more important, contributing 37 percent more than the health-based sorting to the observed CDI gap at age 70. The difference in aging explains 1.4 percentage points, while health-based sorting results in a gap of 1.0 percentage points. Together this exceeds the observed difference of 1.2 percentage points in the CDI. Indeed, both cohort effects and the attrition due to mortality have a substantial dampening effect on the CDI gap. Low-income individuals are more likely to die and this reduces the CDI of the surviving individuals more in comparison to the high-income individuals.[20]

We can also compare the role of the different forces over the lifecycle, as shown in Table 3. The importance of differential aging increases gradually with age. While the CDI gap increases by 0.10 percentage points between age 20 and 30 due to differential aging, the corresponding increase is 0.47 percentage points between age 60 and 70. Panel A of Figure 7 illustrates this graphically, by adding the aging effects for the respective income groups to the observed CDI at age 10, following Equation (15). The simulated CDI's steadily diverge over the life-cycle, but we no longer observe the sudden opening in the CDI gap during early adulthood. The latter is instead driven by health-based sorting, which is important throughout the career, including the start and the end.[21, 22]. Also, the high-income group ages slowly into adulthood, but the evolution of the CDI is no longer as flat (or even decreasing) when we have taken out sorting effects. The high-income group then starts aging faster after middle age and as the difference in aging reduces, the horizontal age gap slowly decreases too. Panel B in Figure 7 uses these horizontal differences between the simulated CDI's, providing a more proper account of the difference in 'biological ages' between low- and high-income individuals. The Figure shows that this difference increases for young adults, before starting to revert slowly around mid-age. The low-income group reaches the biological age of the high-income 50-year olds at age 40. This was

20. Note that the attrition and cohort effects are particularly important at older ages, and contribute to the stabilization of the CDI gap after 65 in Figure 5. Appendix Figure G.1 plots the four different components as a function of age for the low and high-income group separately.

21. We note that the sorting effects remain as important when controlling for household composition, suggesting that effects in early adulthood are not dominated by e.g. children differentially leaving the parents' household or differential family extensions depending on health. This is shown in Appendix Table G.2.

22. The sorting at the start of the career is consistent with the importance of health for earnings at labor market entry (e.g., O'Donnell, Van Doorslaer, and Van Ourti 2015). Looking at finer age bins in Appendix Figure G.1, we see the importance of health-based sorting at the end of the career, which supports evidence of poor health driving premature exits from the labor markets (e.g., Blundell et al. 2021; Kolsrud et al. 2024).

already at age 35 when using the observed CDI's instead of the simulated CDI's.[23]

Separate Chronic Conditions We now study which chronic conditions contribute most to the differential aging and how that evolves over the life-cycle. To evaluate this, we now consider not the prevalence, but the incidence of new chronic conditions for each individual in an income group and scale this by the condition's marginal impact on the CDI. Panel B of Appendix Figure C.8 plots for the same six chronic conditions as in Panel A how much each contributes to the differential aging between the low- and high-income group. Interestingly, while mental health conditions were key in explaining the gap in CDI levels at young ages, they become less important for explaining the gap in CDI growth. This indicates the importance of the reverse channel where individuals with poor mental health sort into lower income groups. On the other hand, the differential incidence of cardio-vascular disease and diabetes is already apparent at younger ages, as it then explains most of the differential aging, even though they become only dominant in explaining the gap in CDI levels at older ages.

Robustness We consider the decomposition of aging and sorting effects for finer income groups in Appendix Figure G.3. This decomposition shows that the earlier non-linearity in the relation between CDI and income at a given age is mostly driven by sorting effects. Between 20 and 70, the CDI increase due to aging equals 6.6 percentage points for the bottom quintile and 4.7 for the top quintile, and it changes rather linearly for the quintiles in between. The health-based sorting, however, worsens the CDI of the lowest income deciles and improves the CDI of all other income deciles. For the bottom quintile we find a further increase by 1.50 percentage points. For the second to top quintile we find decreases ranging from 0.11 to 0.52 percentage points. Appendix Table G.1 compares the aging effects for income groups and education groups, confirming the importance of differential aging and how it increases with age.

We can also consider the robustness of the decomposition of aging and sorting effects depending on the lagging and the averaging of income for our categorization of income groups (see Appendix Figure G.3B). Our baseline categorization uses average income two to four years prior to being considered. First, the lagging follows prior work in the literature aiming to mitigate reverse causality concerns when considering differences in CDI *levels* across income groups.

23. We note that the aging effects are estimated for the surviving sample at each age. Individuals with worse CDI are more likely to die, which improves the CDI of the surviving sample. Appendix Figure G.2 reflects the differential importance of attrition for the low- and high-income group, especially at older age.

We, however, find that our aging estimates, capturing within-individual CDI *changes*, remain basically unchanged when we instead average income one to three years prior. Second, the averaging over three years aims to capture persistent differences across groups. In the same spirit, we can instead consider individuals who are considered low vs. high-income in two consecutive years (based on the respective 3-year averages). We again find very similar results, supporting that our income definition allows to meaningfully categorize individuals to evaluate differential aging and health-based sorting.

6 Counterfactual and Mediation Analysis

The empirical analysis has documented important socio-economic differences in the burden of chronic conditions already early in life and how those with lower incomes age at a faster rate. This final section draws the work closer to policy relevant discussions. First, a counterfactual exercise asks if a policy could close the gap in aging, what are the implications for life expectancy and healthcare costs? And what are the gains vs. losses from intervening earlier or later? Second, we compare the strength of potential mediators, guided by prior work (e.g. Cutler, Lleras-Muney, and Vogl 2011; Mackenbach 2019). We are describing strength in a correlational sense, rather than a causal sense, but in contrast with most of the prior work we consider various mediators jointly and consider the relation with CDI growth rather than CDI levels.

6.1 Counterfactual Analysis

Our analysis allows us to study the potential impact of health interventions targeted to socio-economic groups and how it depends on the timing of the intervention in the life-cycle. We argued before that equalizing the prevalence of chronic conditions provides a lower-bound on how much we can reduce the health gap. Still, the prevalence gap arises over the life-cycle both because of differential aging and individuals re-sorting across income groups based on health. As we have now separated out the aging effect, we can focus instead on a health intervention that targets the incidence of chronic conditions starting from a specific age. We study the impact on these health interventions on individuals' biological age, but also on life-expectancy and healthcare costs. We provide more detail on our estimations and calculations in Appendix H.[24]

24. Despite our focus on the incidence of chronic conditions, the counterfactual analysis continues to provide a lower bound for the potential health effects, as we again ignore the positive impact improved health can have on socio-economic outcomes and how that can further improve one's health.

We use the simulated CDI's from subsection 5.3 capturing the accumulated aging effects and evaluate the impact of equalizing these aging effects from different ages onwards. We first consider the impact on the 'biological age' of the low-income group relative to the high-income group. Panel B of Figure 7 visualises this and shows clearly that by intervening at 20, we can avoid the otherwise steady increase in the biological age gap in early adulthood. By intervening at 40, we miss this opportunity, but still have sufficient time to have mostly closed the gap by age 70. This is no longer true when we wait even longer before intervening.

While decreasing the chronic burden is valuable by itself, we can also evaluate the corresponding gains in life-expectancy. To do so, we first impute the mortality rates corresponding to the counterfactual CDI's using age-specific regressions of mortality on CDI, while accounting for the residual difference in mortality across income groups.[25] We then aggregate the mortality rates into an estimate of life-expectancy at age 40 for different counterfactual scenario's building on (Chetty et al. 2016).[26] Following this procedure, we find an estimated life expectancy of 81.2 for individuals with below-median income and of 85.3 for individuals with above-median income, as reported in Table 4. Now when equalizing the aging process from age 20 onwards, we would increase the life expectancy of the low-income group by 1.3 years, closing 31 percent of the gap in life-expectancy. This reduction is comparable in magnitude to our earlier findings on the contribution of the differential prevalence of chronic conditions to the mortality gap. We can now evaluate how the life-expectancy gain decreases by intervening later. When equalizing the aging process from age 40, the life expectancy increases to 82.3. Hence, while intervening 20 years later, we are still closing 23 percent of the gap. However, when we wait until age 60, life expectancy increases to 81.6, only closing 8 percent of the gap.

A similar extrapolation and aggregation can be used to estimate the gap in expected healthcare costs and how it depends on the differential aging over the life-cycle. To do so, we again translate the CDI into age-specific healthcare costs for the respective income groups up and aggregate over the life-cycle (after age 40).[27] Table 4 reports the estimated life-time healthcare costs, which

25. Note also that for the estimation of life-expectancy we correct the aging-based simulation of the CDI over the life-cycle for the attrition due to mortality. The reason is that the counterfactual aims to capture how intervening on the biological aging at earlier ages affects mortality of the *surviving* sample at later ages. This is still imperfect as some of the estimated attrition due to mortality is also driven by health-based sorting and/or cohort effects. Our conclusions are robust to changing this assumption as we discuss in Appendix H.

26. Between ages 40 and 78, we use the imputed mortality rates for the income group and counterfactual scenario of interest. For the ages between 79 and 90, we use a Gompertz extrapolation $\log \tilde{M}_{a,j} = b_{0,j} + b_{1,j}a$ estimated on the mortality rates for the younger age group. For the ages between 91 and 110, we revert to the observed mortality rates, but now for the full sample. We provide more detail and show the extrapolations graphically in the Appendix.

27. Given the poor fit for costs using a Gompertz extrapolation we use the observed age-pattern for the full population and we take a weighted average between the income-specific costs at age 70 and the average population

are only 1.2 percent higher for low-income individuals than for high-income individuals (157.7k and 155.9k EUR resp.). While the costs are significantly higher for low-income individuals at any give age, aggregated over the lifetime the difference is muted because the faster aging for low incomes results into shorter life-expectancy (see also Van Baal et al. 2008).[28] The same offsetting effects are at play when we equalize the aging process. Even though we would substantially reduce healthcare costs at any age for low-income individuals, we also improve their survival rates. This results into cost savings of less than 3 percent, even when intervening already at age 20. Keeping the survival rates unchanged, the cost savings would increase up to 8 percent.

Taken together, our counterfactual analysis shows that when equalizing the incidence of chronic conditions, by starting at middle-age we can still realize most of the life-expectancy gains. By starting earlier we can also reduce the biological age gap throughout the lifetime. Waiting until later ages seems 'too late' on both accounts. The reduction in the life-expectancy gap is much smaller, but also the biological age stays much higher throughout the life-time. The effects on healthcare costs are relatively limited due to the offsetting effects on life-expectancy.

6.2 Mediators of Health over the Lifecycle

Our counterfactual analysis relies on interventions that can be targeted and effectively reduce the incidence of chronic conditions. However, what is driving the onset of chronic conditions and deteriorating health more generally is the larger puzzle, which has been subject to much research and even more debate. A variety of factors including genetic disposition, environmental exposure, health behaviors, physical and mental strain, access to healthcare, etc. have been discussed in the medical and public health literature and these factors may play a different role for different socio-economic groups. Our analysis has already shown how differences in healthcare treatment seem minimal. We harness three key advantages of our setting and data to provide a modest attempt to re-calibrate the importance of the other factors: (i) we can measure health over the life-cycle, (ii) we can study a variety of mediating factors jointly, (iii) we can measure within-individual changes in health and thus focus on the incidence rather than prevalence of chronic disease.

cost at a given age, using weights that change linearly with age so that the estimated costs converge to the population average at age 90. Note that before age 40, average healthcare expenditures are higher for high income women than for low income women at some ages due to different timing of pregnancies.

28. Note that if we instead combined the higher age-specific healthcare costs for low-income individuals with the higher simulated survival of high-income individuals, this would increase life-time healthcare costs by 17% to 182.7k EUR.

To evaluate the potential role of mediating factors, we run simple age-specific linear regressions:

$$dCDI_{i,a} = \sum_{j=1}^{J} X_{i,a}^{j} \gamma_{j,a} + \varepsilon_{i,a} \tag{16}$$

where $dCDI_{i,a} = CDI_{i,a+5} - CDI_{i,a}$ equals the within-individual change in CDI over the next 5 years and $X_{i,a}^{j}$ is a group of mediating factors. The groups we consider are respectively: (i) parental health, including fathers' and mothers' CDI if alive, or their age at death if not, to proxy for genetic disposition, (ii) self-reported health behaviors, including smoking, drinking, physical activity. We also include BMI in the absence of direct information on nutrition, but this of course captures health more broadly and thus makes us potentially over-state the role of health behaviors, (iii) municipality of residence, as a proxy for environmental exposure and other geographic factors, (iv) employment status and occupational sector, proxying for work factors and occupational health, (v) pay-rank within employer, to zoom in on the role of hierarchy and control and its potential effect on stress, and (vi) a rich set of socio-economic variables including income and wealth, education, parental resources and basic demographics (position in household, household composition, whether foreign born, whether parents were foreign born). The sample we consider are 400k respondents to the national health survey (*Gezondheidsmonitor*) for whom we can measure health behaviors and which is linked to the other registers. We provide more details on the sample selection and full list of variables in Appendix I.

To quantify the relative importance of different factors, we decompose the total R-squared of this regression using the Shapley-Owen values. These values calculate the average contribution of each regressor group to the R-squared over all possible sequences of the regressor groups. The Shapley-Owen method is valuable as it allocates any explanatory power that is common to multiple mediators equally. For example, suppose geography alone explains 30 percent of the variation in CDI, health behaviours alone explain 20 percent, and both geography & health behaviours jointly explain 40 percent. This implies 10 percent of the variation is common to both drivers; the Shapley-Owen procedure attributes 5 percent to each.[29]

A few striking patterns emerge, as shown in Figure 8.[30] First, the socio-economic variables play

29. To elaborate on the example, suppose now parental health explains 10 percent of the variation in CDI by itself, and that all three factors (behaviour, geography, parental health) jointly explain 50 percent. This would imply that the parental health variation is fully additional to behaviour and geography. In this way, the method allows us to allocate shares of explainable variation to different factors. In this example, the 50 percent of explainable CDI variation is apportioned 25p.p. to geography, 15p.p. to health behaviours, 10p.p. to parental health.

30. We note the low overall R^2, especially at older ages. This is partly driven by considering coarse age groups (e.g., for 55-year olds only we find an R^2 of 0.18), but also suggests important randomness in the incidence of chronic

a dominant role throughout the lifetime. Together they are responsible for 30.7 percent of the explained variation between ages 20-29; this increases up to 35.8 percent between ages 40-49 to then fall to 25.5 percent between ages 60-69. While a large role of social determinants has been conjectured before in the literature, the one third of explained variation is in addition to the variation explained by the other measured factors, where any commonly explained variation has been equally apportioned. Second, we find a substantial role for local factors as captured by the municipality fixed effects. During early adulthood, the municipality effects even account for 42.4 percent of the explained variation. This gradually declines with age, but remains above 27 percent. This finding complements recent work estimating causal local effects on mortality for elderly (e.g., Finkelstein, Gentzkow, and Williams 2021). Finally, some of the other factors, while notable, are less important quantitatively. Employment factors, including status, sector and ranks, jointly account for between 7 and 11 percent of the explained variation. Parents' measurable health also account for only about 9 percent. Perhaps most strikingly, the role of measured health behaviors is also moderate, especially at younger ages. While at older ages, health behaviors explain up to 24 percent of the explained variation in CDI growth, at younger ages this is as small as 3.7 percent. Its overall importance thus seems limited, especially compared to its central position in the epidemiological literature and public health debate.[31]

The estimated patterns are descriptive and uncovering causal pathways with these mediating factors is challenging. Still, prior work has made strong claims about the importance of specific factors separately, often pointing at the importance of individual health behaviors.[32] A few papers have come up with a comprehensive account of the different factors jointly, as surveyed by Finkelstein, Gentzkow, and Williams (2021) and McGovern (2014).[33] Our analysis provides a new perspective on the key mediators and hopefully can serve as a roadmap for further research identifying the causal health effects of these mediators and how they change over the life-cycle.

disease. Of course, even with our rich data, we cannot exclude our inability to observe other relevant features.

31. The explanation for the limited role at younger ages cannot simply be the coarse or imperfect measurement of individuals' behavior, since at older ages differences in health behavior do capture a more significant part of the variation. Still, one explanation for the low correlation between chronic health and health behaviors at younger ages is that it takes time for these behaviors to convert into chronic illness.

32. For example, using geographic variation in health behaviors and mortality, Cutler (2018) concludes: "Adverse health behaviors account for 40 percent of deaths in the United States. Reduce those deaths and the population can live much longer."

33. For example, McGinnis, Williams-Russo, and Knickman (2002) write: "On a population basis, using the best available estimates, the impacts of various domains on early deaths in the United States distribute roughly as follows: genetic predispositions, about 30 percent; social circumstances, 15 percent; environmental exposures, 5 percent; behavioral patterns, 40 percent; and shortfalls in medical care, 10 percent." In their methodology to develop US county health rankings, Booske Catlin et al. (2010) use the following weights: "Social and economic factors 40 percent, Health behaviors 30 percent, Clinical care 20 percent, and Environmental factors 10 percent."

6.3 (Naive) Contribution of Mediators to the Health Gap

The previous section quantified the share of the overall variation in aging that can be attributed to different mediators, including socio-economic differences. One can easily mis-estimate (or mis-interpret) these contributions due to data challenges.

Let us first consider how much mediators associate with the variation in chronic disease between incomes rather than with all variation in chronic disease. That is, we may want to assess the potential mediators of the health gap itself. To do so, we can project the within-individual CDI growth on income percentiles, while controlling for gender and age, and then use the income-projected CDI growth as the object for a Shapley-Owen decomposition. This is reported in Appendix Figure C.11. Health behaviors, but also employment factors, play a much larger role than when considering the overall variation in CDI growth. For instance, health behaviours explain 25 percent of the cross-income variation at age 40-49, compared to 8.5 percent of the overall explainable variation for the same age bracket. What is driving the overall variation in health, versus the income gradient in health, are two distinct questions. In particular, the contribution of potential mediators to the health gap depends not only on their impact on health, but also on their differential importance across incomes. As is well documented more generally (e.g., Pampel, Krueger, and Denney 2010), health behaviors are particularly correlated with income and this is also true in the Netherlands (see Appendix Figure C.12). The same holds for employment factors and for example the sector of work. Still, it is important to correctly acknowledge the impact of these mediators on health, especially when these mediators are so prevalent in specific socio-economic groups. We can thus easily mis-estimate these impacts due to data constraints, which we briefly illustrate in our context.

Figure 9 compares the estimates for different mediators in our baseline regression in (16), using CDI growth as the dependent variable and controlling for all mediators jointly, including socio-economic factors.[34] We find large municipality effects. Being in the worst decile of municipalities corresponds to CDI growth that is 0.7 percentage points (pp) higher than when in the best decile of municipalities. Health behaviors matter too, but perhaps less than expected. Being a smoker increases CDI growth by 0.2pp and being overweight (obese) increases CDI growth by 0.1pp (0.4pp). The association with sports activities is negligible. We also find small effects for parental

34. This is a similar exercise to Figure 6 of Finkelstein, Gentzkow, and Williams (2021), and Figure 7 of Chetty et al. (2016), but here we regress at the individual level, and we have a broader set of variables in consideration. A breakdown of these results by gender is shown in Appendix Figure I.1. Similarly, results for certain age groups are shown in Appendix Figure I.2.

health and employment factors. While the estimated differences in CDI growth control for all other observable factors, including observable socio-economic differences, we should remain cautious interpreting these magnitudes. This is highlighted by the fact that even in our rich data environment we estimate the role of (self-reported) drinking of alcohol to be protective.[35]

Now the estimates change considerably for more 'naive' approaches that may be used due to data challenges. First, Figure 9 shows that the ranking of the mediating factors substantially changes when using CDI *levels* instead of CDI *growth* as the dependent variable. The former is common in the literature, for example studying self-reported health or the prevalence of medical conditions, but of course more sensitive to the reverse causality of health on potential mediators. We find that the estimated coefficients on health behaviors and employment factors (e.g., being on social assistance) increase in relative magnitude, but this may be picking up some reverse causality from individuals' health on how they behave and what work they can do.[36] Second, the potential for mis-attribution is also greater when focusing on one factor at a time, and not controlling for other mediating factors. This is also illustrated in Figure 9. The estimated relationship between health and health behaviors, including for smoking and BMI, becomes again substantially larger and some estimates even double in size. However, this is now picking up the relation between health and other correlated factors including socio-economic differences.

Many studies in public health and epidemiology have underlined the importance of health behaviors and linked the income gradient in health and the income gradient in behaviors, but some conclusions may thus have been exaggerated due to data challenges.[37] Appendix Figure C.13 illustrates this compellingly by showing that health behaviors by themselves can explain about half of the gap in CDI levels between low- and high-income individuals, even at younger ages. However, the granularity of the data allows us to control for various factors jointly and this shows that while health behaviors are strongly correlated with chronic illness and with income, various other observable factors weigh a lot too and some of them are arguably more important

35. Appendix Figure I.3 compares the outcomes of CDI growth, overall mortality and alcohol related hospitalisations. Moderate drinking is protective across all outcomes; heavy drinking results in a slightly greater risk of overall mortality and related hospitalisations, but this is mainly driven by the low income group, despite heavy drinking being marginally more prevalent among high incomes (Appendix Figure C.12). This is partly consistent with the so-called "Alcohol-Harm Paradox" (Bloomfield 2020). However, current heavy drinking could also be associated with greater underdiagnosis, relative to former drinking.

36. The bottom panel of Appendix Figure C.11 confirms the same patterns by reporting the Shapley-Owen decompositions for CDI levels instead of the CDI changes. It shows the more important role played by employment factors in explaining the variation in CDI levels. Also health behaviors become much more important, but with again a pronounced lifetime gradient, explaining less than 1 percent of the overall CDI variation for ages 20-29 and close to 4 percent between ages 60-69.

37. See also Darden, Gilleskie, and Strumpf 2018 who come to a similar conclusion regarding the estimated mortality effect of smoking.

for explaining the observed variation in health, both across and within income groups.

7 Conclusion

We exploit the richness of our data to directly measure health, income and other relevant factors in one and the same setting to provide a comprehensive and transparent account of health inequalities and how they arise over the life-cycle.

Mackenbach (2019, p. 178) articulated the long-standing knowledge gap: "We know that the explanation of health inequalities involves three basic mechanisms: direct causation, reverse causation, and confounding (due to selection on personal characteristics during social mobility). This was already known when I started to work in this area in the late 1980s, but after decades of research we still do not know what the relative importance of each of these mechanisms is." Our work makes significant progress to fill this gap in our understanding.

We have shown that chronic diseases explain a substantial portion of the income gradient in mortality and healthcare costs. We described the twin roles of differential ageing, versus health-based sorting, at play at different parts of the life course. Differential ageing, that is chronic conditions accruing at different rates, is a consistent process, that builds in magnitude with age and dominates over the life-cycle. This dynamic decomposition contributes to our understanding of the mechanisms behind health inequality and can guide public health interventions that target the incidence of chronic conditions in addressing these health inequalities.

While our analysis is mostly descriptive in nature, our comprehensive approach contrasts with descriptive work in epidemiology, either at the national or global level, often focusing on the mortality rates related to specific health conditions, while marginally accounting for other health conditions, and relating these to one or a few specific risk factors, while marginally accounting for other confounding factors (Wang et al. 2016; Murray et al. 2020). As mentioned, our paper can be seen as a re-calibration of the potential importance of specific mechanisms and as such provides an ideal roadmap for further empirical work.

The Chronic Disease Index we have developed can help in these research endeavours. Our index closely relates to indices aiming to provide a comprehensive account of individuals' health like the Charlson and Elixhauser Indices (Charlson et al. 1987; Elixhauser et al. 1998), but differs in two important ways. The first is that CDI can be constructed at scale and measured repeatedly for

the same individual, as it uses administrative data available in panel data for the full population. The second is that the CDI provides a universal interpretation and is constructed in a robust manner, not confounded by socio-economic differences in mortality due to differences in access to healthcare, differences in communicable or acute disease. Both advantages make the CDI particularly valuable for further work.

References

Adda, Jérôme, James Banks, and Hans-Martin von Gaudecker. 2009. "The Impact of Income Shocks on Health: Evidence from Cohort Data." *Journal of the European Economic Association* 7 (6): 1361–1399. ISSN: 1542-4766, accessed December 18, 2023. JSTOR: 40601206.

Bauer, Ursula E, Peter A Briss, Richard A Goodman, and Barbara A Bowman. 2014. "Prevention of Chronic Disease in the 21st Century: Elimination of the Leading Preventable Causes of Premature Death and Disability in the USA." *The Lancet* 384, no. 9937 (July): 45–52. ISSN: 0140-6736, accessed May 9, 2024. https://doi.org/10.1016/S0140-6736(14)60648-6.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "High-Dimensional Methods and Inference on Structural and Treatment Effects." *Journal of Economic Perspectives* 28, no. 2 (May): 29–50. ISSN: 0895-3309, accessed December 18, 2023. https://doi.org/10.1257/jep.28.2.29.

Black, Sandra E., Paul J. Devereux, and Kjell G. Salvanes. 2015. "Losing Heart? The Effect of Job Displacement on Health." *ILR Review* 68, no. 4 (August): 833–861. ISSN: 0019-7939, accessed April 22, 2024. https://doi.org/10.1177/0019793915586381.

Bloomfield, Kim. 2020. "Understanding the Alcohol-Harm Paradox: What Next?" *The Lancet Public Health* 5, no. 6 (June): e300–e301. ISSN: 2468-2667, accessed May 21, 2024. https://doi.org/10.1016/S2468-2667(20)30119-5.

Blundell, Richard, Jack Britton, Monica Costa Dias, and Eric French. 2021. "The Impact of Health on Labor Supply Near Retirement." *Journal of Human Resources* (January). ISSN: 0022-166X, 1548-8004, accessed April 22, 2024. https://doi.org/10.3368/jhr.58.3.1217-9240R4.

Bolt, Uta. 2022. *What Is the Source of the Health Gradient? The Case of Obesity.* Working Paper. Accessed April 22, 2024.

Booske Catlin, Bridget, Jessica K. Athens, David A. Kindig, Patrick L. Remington, and Hyojun Park. 2010. "Different Perspectives for Assigning Weights to Determinants of Health" (January).

Case, Anne, and Angus S. Deaton. 2005. "Broken Down by Work and Sex: How Our Health Declines." In *Analyses in the Economics of Aging*, 185–212. University of Chicago Press, August. Accessed May 9, 2024.

Case, Anne, Angela Fertig, and Christina Paxson. 2005. "The Lasting Impact of Childhood Health and Circumstance." *Journal of Health Economics* 24, no. 2 (March): 365–389. ISSN: 0167-6296, accessed December 18, 2023. https://doi.org/10.1016/j.jhealeco.2004.09.008.

Charlson, Mary E., Peter Pompei, Kathy L. Ales, and C. Ronald MacKenzie. 1987. "A New Method of Classifying Prognostic Comorbidity in Longitudinal Studies: Development and Validation." *Journal of Chronic Diseases* 40, no. 5 (January): 373–383. ISSN: 0021-9681, accessed December 18, 2023. https://doi.org/10.1016/0021-9681(87)90171-8.

Chen, Yi Hsuan, Milad Karimi, and Maureen P. M. H. Rutten-van Mölken. 2020. "The Disease Burden of Multimorbidity and Its Interaction with Educational Level." *PLOS ONE* 15, no. 12 (December): e0243275. ISSN: 1932-6203, accessed January 9, 2024. https://doi.org/10.1371/journal.pone.0243275.

Chetty, Raj, Michael Stepner, Sarah Abraham, Shelby Lin, Benjamin Scuderi, Nicholas Turner, Augustin Bergeron, and David Cutler. 2016. "The Association Between Income and Life Expectancy in the United States, 2001-2014." *JAMA* 315, no. 16 (April): 1750–1766. ISSN: 0098-7484, accessed December 18, 2023. https://doi.org/10.1001/jama.2016.4226.

Chini, Francesco, Patrizio Pezzotti, Letizia Orzella, Piero Borgia, and Gabriella Guasticchi. 2011. "Can We Use the Pharmacy Data to Estimate the Prevalence of Chronic Conditions? A Comparison of Multiple Data Sources." *BMC Public Health* 11, no. 1 (September): 688. ISSN: 1471-2458, accessed January 8, 2024. https://doi.org/10.1186/1471-2458-11-688.

Costa, Dora L. 2015. "Health and the Economy in the United States from 1750 to the Present." *Journal of Economic Literature* 53, no. 3 (September): 503–570. ISSN: 0022-0515, accessed April 22, 2024. https://doi.org/10.1257/jel.53.3.503.

Cutler, David, Angus Deaton, and Adriana Lleras-Muney. 2006. "The Determinants of Mortality." *Journal of Economic Perspectives* 20, no. 3 (September): 97–120. ISSN: 0895-3309, accessed December 18, 2023. https://doi.org/10.1257/jep.20.3.97.

Cutler, David M. 2018. *The School-First Solution.* https://politi.co/2mnNZ9H, January. Accessed May 23, 2024.

Cutler, David M., Adriana Lleras-Muney, and Tom Vogl. 2011. "Socioeconomic Status and Health: Dimensions and Mechanisms." In *The Oxford Handbook of Health Economics*, edited by Sherry Glied and Peter C. Smith. Oxford University Press, April. ISBN: 978-0-19-923882-8, accessed December 18, 2023. https://doi.org/10.1093/oxfordhb/9780199238828.013.0007.

Danesh, Kaveh, Jonathan Kolstad, William Parker, Johannes Spinnewijn, and Mieke Aarts. In preparation. "Socioeconomic Differences in Cancer Incidence, Stage, and Survival in the Netherlands, 2011-2017: An Observational Study."

Darden, Michael, Donna B. Gilleskie, and Koleman Strumpf. 2018. "Smoking and Mortality: New Evidence from a Long Panel." *International Economic Review* 59 (3): 1571–1619. ISSN: 1468-2354, accessed May 22, 2024. https://doi.org/10.1111/iere.12314.

De Nardi, Mariacristina, Svetlana Pashchenko, and Ponpoje Porapakkarm. 2023. "The Lifetime Costs of Bad Health." *The Review of Economic Studies, accepted.*

Deaton, Angus. 2002. "Policy Implications Of The Gradient Of Health And Wealth." *Health Affairs* 21, no. 2 (March): 13–30. ISSN: 0278-2715, accessed December 18, 2023. https://doi.org/10.1377/hlthaff.21.2.13.

Deaton, Angus S., and Christina Paxson. 2004. "Mortality, Income, and Income Inequality over Time in Britain and the United States." *NBER Chapters*, 247–286. Accessed December 18, 2023.

Dobkin, Carlos, Amy Finkelstein, Raymond Kluender, and Matthew J. Notowidigdo. 2018. "The Economic Consequences of Hospital Admissions." *American Economic Review* 108, no. 2 (February): 308–352. ISSN: 0002-8282, accessed December 18, 2023. https://doi.org/10.1257/aer.20161038.

Dowd, Jennifer Beam, and Megan Todd. 2011. "Does Self-reported Health Bias the Measurement of Health Inequalities in U.S. Adults? Evidence Using Anchoring Vignettes From the Health and Retirement Study." *The Journals of Gerontology: Series B* 66B, no. 4 (July): 478–489. ISSN: 1079-5014, accessed January 8, 2024. https://doi.org/10.1093/geronb/gbr050.

Elixhauser, A., C. Steiner, D. R. Harris, and R. M. Coffey. 1998. "Comorbidity Measures for Use with Administrative Data." *Medical Care* 36, no. 1 (January): 8–27. ISSN: 0025-7079. https://doi.org/10.1097/00005650-199801000-00004.

Eurostat. 2023. *EU Statistics on Income and Living Conditions Microdata, Release 2 in 2023, Data 2004-2022 Version 1.* Accessed April 24, 2024. https://doi.org/10.2907/EUSILC2004-2022V1.

Finkelstein, Amy, Matthew Gentzkow, and Heidi Williams. 2021. "Place-Based Drivers of Mortality: Evidence from Migration." *American Economic Review* 111, no. 8 (August): 2697–2735. ISSN: 0002-8282, accessed December 18, 2023. https://doi.org/10.1257/aer.20190825.

Finkelstein, Amy, and Robin McKnight. 2008. "What Did Medicare Do? The Initial Impact of Medicare on Mortality and out of Pocket Medical Spending." *Journal of Public Economics* 92, no. 7 (July): 1644–1668. ISSN: 0047-2727, accessed December 18, 2023. https://doi.org/10.1016/j.jpubeco.2007.10.005.

Galama, Titus J, and Hans van Kippersluis. 2019. "A Theory of Socio-economic Disparities in Health over the Life Cycle." *The Economic Journal* 129, no. 617 (January): 338–374. ISSN: 0013-0133, accessed May 9, 2024. https://doi.org/10.1111/ecoj.12577.

Gathmann, Christina, Hendrik Jürges, and Steffen Reinhold. 2015. "Compulsory Schooling Reforms, Education and Mortality in Twentieth Century Europe." *Social Science & Medicine,* Special Issue: Educational Attainment and Adult Health: Contextualizing Causality, 127 (February): 74–82. ISSN: 0277-9536, accessed December 18, 2023. https://doi.org/10.1016/j.socscimed.2014.01.037.

Grossman, Michael. 1972. "On the Concept of Health Capital and the Demand for Health." *Journal of Political Economy* 80 (2): 223–255. ISSN: 0022-3808, accessed December 18, 2023. JSTOR: 1830580.

Harteloh, Peter, Kim de Bruin, and Jan Kardaun. 2010. "The Reliability of Cause-of-Death Coding in The Netherlands." *European Journal of Epidemiology* 25, no. 8 (August): 531–538. ISSN: 1573-7284, accessed May 21, 2024. https://doi.org/10.1007/s10654-010-9445-5.

Hosseini, Roozbeh, Karen A. Kopecky, and Kai Zhao. 2021. *How Important Is Health Inequality for Lifetime Earnings Inequality?* SSRN Scholarly Paper, 3829973, Rochester, NY, January. Accessed January 9, 2024. https://doi.org/10.2139/ssrn.3829973.

———. 2022. "The Evolution of Health over the Life Cycle." *Review of Economic Dynamics* 45 (July): 237–263. ISSN: 1094-2025, accessed December 18, 2023. https://doi.org/10.1016/j.red.2021.07.001.

Huber, Carola A., Thomas D. Szucs, Roland Rapold, and Oliver Reich. 2013. "Identifying Patients with Chronic Conditions Using Pharmacy Data in Switzerland: An Updated Mapping Approach to the Classification of Medications." *BMC Public Health* 13, no. 1 (October): 1030. ISSN: 1471-2458, accessed December 18, 2023. https://doi.org/10.1186/1471-2458-13-1030.

Kennedy Moulton, Kate, Sarah Miller, Petra Persson, Maya Rossin Slater, Laura Wherry, and Gloria Aldana. 2022. *Maternal and Infant Health Inequality: New Evidence from Linked Administrative Data.* Working Paper 30693. NBER, November. Accessed December 18, 2023.

Kinge, Jonas Minet, Jørgen Heibø Modalsli, Simon Øverland, Håkon Kristian Gjessing, Mette Christophersen Tollånes, Ann Kristin Knudsen, Vegard Skirbekk, Bjørn Heine Strand, Siri Eldevik Håberg, and Stein Emil Vollset. 2019. "Association of Household Income With Life Expectancy and Cause-Specific Mortality in Norway, 2005-2015." *JAMA* 321, no. 19 (May): 1916–1925. ISSN: 0098-7484, accessed December 18, 2023. https://doi.org/10.1001/jama.2019.4329.

Kolsrud, Jonas, Camille Landais, Daniel Reck, and Johannes Spinnewijn. 2024. "Retirement Consumption and Pension Design." *American Economic Review* 114, no. 1 (January): 89–133. ISSN: 0002-8282, accessed April 22, 2024. https://doi.org/10.1257/aer.20221426.

Kulshreshtha, Shobhit, Martin Salm, and Ansgar Wübker. 2022. "Does Population Sorting through Internal Migration Increase Healthcare Costs and Needs in Peripheral Regions?" *IZA Discussion Papers,* no. 15559 (September). Accessed December 18, 2023.

Lamers, Leida M., and René C. J. A. van Vliet. 2004. "The Pharmacy-based Cost Group Model: Validating and Adjusting the Classification of Medications for Chronic Conditions to the Dutch Situation." *Health Policy* 68, no. 1 (April): 113–121. ISSN: 0168-8510, accessed January 8, 2024. https://doi.org/10.1016/j.healthpol.2003.09.001.

Lleras-Muney, Adriana, and Flavien Moreau. 2022. "A Unified Model of Cohort Mortality." *Demography* 59, no. 6 (December): 2109–2134. ISSN: 0070-3370, accessed December 18, 2023. https://doi.org/10.1215/00703370-10286336.

Mackenbach, Johan P. 2019. *Health Inequalities: Persistence and Change in European Welfare States.* Oxford University Press, August. ISBN: 978-0-19-186911-2, accessed January 9, 2024. https://doi.org/10.1093/oso/9780198831419.001.0001.

Marmot, M. G., S. Stansfeld, C. Patel, F. North, J. Head, I. White, E. Brunner, A. Feeney, M. G. Marmot, and G. Davey Smith. 1991. "Health Inequalities among British Civil Servants: The Whitehall II Study." *The Lancet,* Originally Published as Volume 1, Issue 8754, 337, no. 8754 (June): 1387–1393. ISSN: 0140-6736, accessed December 18, 2023. https://doi.org/10.1016/0140-6736(91)93068-K.

Marmot, Michael. 2015. *The Health Gap: The Challenge of an Unequal World.* 1st edition. New York, NY London Oxford New Delhi Sydney: Bloomsbury Press, November. ISBN: 978-1-63286-078-1.

McGinnis, J. Michael, Pamela Williams-Russo, and James R. Knickman. 2002. "The Case For More Active Policy Attention To Health Promotion." *Health Affairs* 21, no. 2 (March): 78–93. ISSN: 0278-2715, accessed April 22, 2024. https://doi.org/10.1377/hlthaff.21.2.78.

McGovern, Laura. 2014. *The Relative Contribution of Multiple Determinants to Health.* Technical report. Health Affairs, August. Accessed May 28, 2024.

Mortensen, Laust H., Johan Rehnberg, Espen Dahl, Finn Diderichsen, Jon Ivar Elstad, Pekka Martikainen, David Rehkopf, Lasse Tarkiainen, and Johan Fritzell. 2016. "Shape of the Association between Income and Mortality: A Cohort Study of Denmark, Finland, Norway and Sweden in 1995 and 2003." *BMJ Open* 6, no. 12 (December): e010974. ISSN: 2044-6055, 2044-6055, accessed December 18, 2023. https://doi.org/10.1136/bmjopen-2015-010974.

Murray, Christopher J. L., Aleksandr Y. Aravkin, Peng Zheng, Cristiana Abbafati, Kaja M. Abbas, Mohsen Abbasi-Kangevari, Foad Abd-Allah, Ahmed Abdelalim, Mohammad Abdollahi, Ibrahim Abdollahpour, et al. 2020. "Global Burden of 87 Risk Factors in 204 Countries and Territories, 1990–2019: A Systematic Analysis for the Global Burden of Disease Study 2019." *The Lancet* 396, no. 10258 (October): 1223–1249. ISSN: 0140-6736, 1474-547X, accessed January 9, 2024. https://doi.org/10.1016/S0140-6736(20)30752-2.

National Center for Health Statistics. 2022. *Health, United States, Annual Perspective, 2020-2021.* Technical report. National Center for Health Statistics (U.S.) Accessed April 24, 2024. https://doi.org/10.15620/cdc:122044.

O'Donnell, Owen, Eddy Van Doorslaer, and Tom Van Ourti. 2015. "Chapter 17 - Health and Inequality." In *Handbook of Income Distribution*, edited by Anthony B. Atkinson and François Bourguignon, 2:1419–1533. Handbook of Income Distribution. Elsevier, January. Accessed May 21, 2024. https://doi.org/10.1016/B978-0-444-59429-7.00018-2.

Obozinski, Guillaume, Martin J. Wainwright, and Michael I. Jordan. 2011. "Support Union Recovery in High-Dimensional Multivariate Regression." *The Annals of Statistics* 39 (1): 1–47. ISSN: 0090-5364, accessed April 22, 2024. JSTOR: 29783630.

Pampel, Fred C., Patrick M. Krueger, and Justin T. Denney. 2010. "Socioeconomic Disparities in Health Behaviors." *Annual Review of Sociology* 36 (1): 349–370. Accessed December 18, 2023. https://doi.org/10.1146/annurev.soc.012809.102529.

Sapolsky, Robert M. 2005. "The Influence of Social Hierarchy on Primate Health." *Science* 308, no. 5722 (April): 648–652. Accessed December 18, 2023. https://doi.org/10.1126/science.1106477.

Stepner, Michael. 2019. *The Insurance Value of Redistributive Taxes and Transfers.* Working Paper. Accessed April 22, 2024.

Sullivan, Daniel, and Till von Wachter. 2009. "Job Displacement and Mortality: An Analysis Using Administrative Data*." *The Quarterly Journal of Economics* 124, no. 3 (August): 1265–1306. ISSN: 0033-5533, accessed December 18, 2023. https://doi.org/10.1162/qjec.2009.124.3.1265.

Van Baal, Pieter H. M, Johan J Polder, G. Ardine De Wit, Rudolf T Hoogenveen, Talitha L Feenstra, Hendriek C Boshuizen, Peter M Engelfriet, and Werner B. F Brouwer. 2008. "Lifetime Medical Costs of Obesity: Prevention No Cure for Increasing Health Expenditure." Edited by Andrew Prentice. *PLoS Medicine* 5, no. 2 (February): e29. ISSN: 1549-1676, accessed January 10, 2024. https://doi.org/10.1371/journal.pmed.0050029.

van den Berg, Gerard J., Maarten Lindeboom, and France Portrait. 2006. "Economic Conditions Early in Life and Individual Mortality." *American Economic Review* 96, no. 1 (March): 290–302. ISSN: 0002-8282, accessed April 22, 2024. https://doi.org/10.1257/000282806776157740.

van Ooijen, Raun, Rob J. M. Alessie, and Marike Knoef. 2015. *Health Status Over the Life Cycle.* SSRN Scholarly Paper, 2743110, Rochester, NY, October. Accessed January 8, 2024. https://doi.org/10.2139/ssrn.2743110.

Wang, Haidong, Mohsen Naghavi, Christine Allen, Ryan M. Barber, Zulfiqar A. Bhutta, Austin Carter, Daniel C. Casey, Fiona J. Charlson, Alan Zian Chen, Matthew M. Coates, et al. 2016. "Global, Regional, and National Life Expectancy, All-Cause Mortality, and Cause-Specific Mortality for 249 Causes of Death, 1980–2015: A Systematic Analysis for the Global Burden of Disease Study 2015." *The Lancet* 388, no. 10053 (October): 1459–1544. ISSN: 0140-6736, 1474-547X, accessed January 9, 2024. https://doi.org/10.1016/S0140-6736(16)31012-1.

World Health Organization. 1985. "Targets for Health for All: Targets in Support of the European Regional Strategy for Health for All," accessed January 9, 2024.

———. 2008. *Closing the Gap in a Generation: Health Equity through Action on the Social Determinants of Health.* Technical report. Geneva: World Health Organization, August. Accessed May 9, 2024.

———. 2017. *National Health Inequality Monitoring: A Step-by-Step Manual.* World Health Organization. ISBN: 978-92-4-151218-3, accessed May 21, 2024.

Yuan, Ming, and Yi Lin. 2006. "Model Selection and Estimation in Regression with Grouped Variables." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 68, no. 1 (February): 49–67. ISSN: 1369-7412, accessed April 22, 2024. https://doi.org/10.1111/j.1467-9868.2005.00532.x.

Yurkovich, Marko, J. Antonio Avina-Zubieta, Jamie Thomas, Mike Gorenchtein, and Diane Lacaille. 2015. "A Systematic Review Identifies Valid Comorbidity Indices Derived from Administrative Health Data." *Journal of Clinical Epidemiology* 68, no. 1 (January): 3–14. ISSN: 0895-4356, 1878-5921, accessed January 8, 2024. https://doi.org/10.1016/j.jclinepi.2014.09.010.

A Tables

Table 1: Sample Descriptive Statistics

	Full sample	2006	2016	Aged 40	Aged 70
A. Demographics					
Age	41.12	39.31	41.95	40	70
Foreign-born	0.12	0.10	0.12	0.19	0.09
Male	0.50	0.49	0.50	0.50	0.49
Self-employed	0.06	0.05	0.06	0.11	0.02
With partner	0.72	0.74	0.72	0.77	0.73
With kids	0.55	0.56	0.54	0.72	0.07
B. Education					
Less than High School	0.21	0.21	0.22	0.04	0.05
High School	0.26	0.19	0.29	0.29	0.11
College	0.08	0.05	0.09	0.16	0.03
Further Studies	0.04	0.03	0.05	0.10	0.01
Education missing	0.40	0.51	0.35	0.41	0.79
C. Income and Wealth					
Household Income	27,608	21,999	29,736	27,397	27,088
Household Wealth	181,684	-	168,090	100,844	296,894
D. Health and Healthcare					
Chronic cond. count	0.85	0.77	0.87	0.50	2.03
Has chronic conditions	0.39	0.38	0.39	0.31	0.76
Cardiovascular disease	0.18	0.15	0.19	0.05	0.53
Diabetes	0.05	0.04	0.05	0.01	0.14
Respiratory illness	0.09	0.08	0.09	0.07	0.14
Pain	0.06	0.05	0.07	0.05	0.11
Psychological disorders	0.08	0.13	0.08	0.09	0.13
Psychoses	0.02	0.01	0.02	0.02	0.02
Medicines taken	2.67	2.56	2.70	1.97	5.11
Takes medicines	0.67	0.68	0.67	0.65	0.88
Total healthcare spending	2,261	-	2,387	1,539	4,140
Hospitalised	0.11	-	0.11	0.08	0.20
5-year mortality	50.67	48.51	54.41	6.03	112.60
Observations	272,889,744	16,499,473	17,187,337	3,650,513	2,653,596
Individuals	21,159,899	16,499,473	17,187,337	3,650,513	2,653,596

Note: This table provides descriptive statistics for the administrative data, for selected ages and calendar years.

Table 2: Mapping Pharmaceutical Data to Chronic Disease

Chronic Disease	ATC Code(s)	Medicine Description
Acid related disorders	A02	Drugs for acid related disorders
Bone diseases (osteoporosis)	M05	Drugs for treatment of bone diseases
Cancer*	L01	Antineoplastic agents
Cardiovascular diseases (inc. hypertension)	B01A, C01, C04A, C02, C07, C08, C09	Antithrombotic agents, cardiac therapy, peripheral vasodilators, antihypertensives, beta blocking agents, calcium channel blockers, agents acting on the renin-angiotensin system
Dementia	N06D	Anti-dementia drugs
Diabetes (mellitus)	A10A,A10B,A10X	Insulins and analogues, Blood glucose lowering drugs (excl. insulins), other drugs used in diabetes
Epilepsy	N03	Antiepileptics
Glaucoma	S01E	Antiglaucoma preparations and miotics
Gout, (Hyperuricemia)	M04	Antigout preparations
HIV	J05A	Direct acting antivirals
Hyperlipidemia	C10	Lipid modifying agents
Intestinal (inflammatory) diseases	A07E	Intestinal antiinflammatory agents
(Iron deficiency) anemia	B03A	Iron preparations
Migraines	N02C	Antimigraine preparations
Pain	N02A, N02B	Opioids, other analgesics and antipyretics
Parkinson's disease	N04, N05B, N05C	Anxiolytics, hypnotics and sedatives
Psychological disorders	N06A	Antidepressants
Psychoses	N05A	Antipsychotics
Respiratory illnesses	R03	Drugs for obstructive airway diseases
Rheumatological conditions	L04A	Immunosuppressants
Thyroid disorders	H03	Thyroid therapy
Tuberculosis	J04A	Drugs for treatment of tuberculosis

Note: The table reports concordance or mapping from 3-digit ATC codes to chronic diseases. This was adapted from Huber et al. (2013), with specific refinements described in Appendix D. *Cancer here refers to cancers treated with pharmacy-dispensed medications, which is around 5% of all cancer diagnoses. Digestive tract and skin cancers dominate this measure: they account for over 60% of the diagnoses.

Table 3: Change in CDI gap, x 100

	11-20	21-30	31-40	41-50	51-60	61-70	Life-cycle Aggregate
1. Differential Ageing	0.004	0.101	0.143	0.298	0.374	0.469	1.388
2. Health-Based Sorting	0.000	0.191	0.194	0.226	0.363	0.040	1.014
3. Compositional Effects							
a. Attrition due to death	0.001	-0.003	-0.018	-0.050	-0.162	-0.350	-0.582
b. Cohort effects	-0.004	-0.017	-0.057	-0.139	-0.163	-0.232	-0.610
Total Change	0.001	0.272	0.262	0.335	0.412	-0.073	1.209

Note: The table reports the contribution towards the income gap in the CDI for the aging, sorting and compositional effects as estimated by our lifecycle decomposition. The effects are expressed as the change in the CDI gap for 10-year age bins, multiplied by 100. That is, the numbers in the table are expressed as percentage point changes in the CDI. More detail on the lifecycle decomposition is provided in Appendix Section G.

Table 4: Counterfactual Analysis

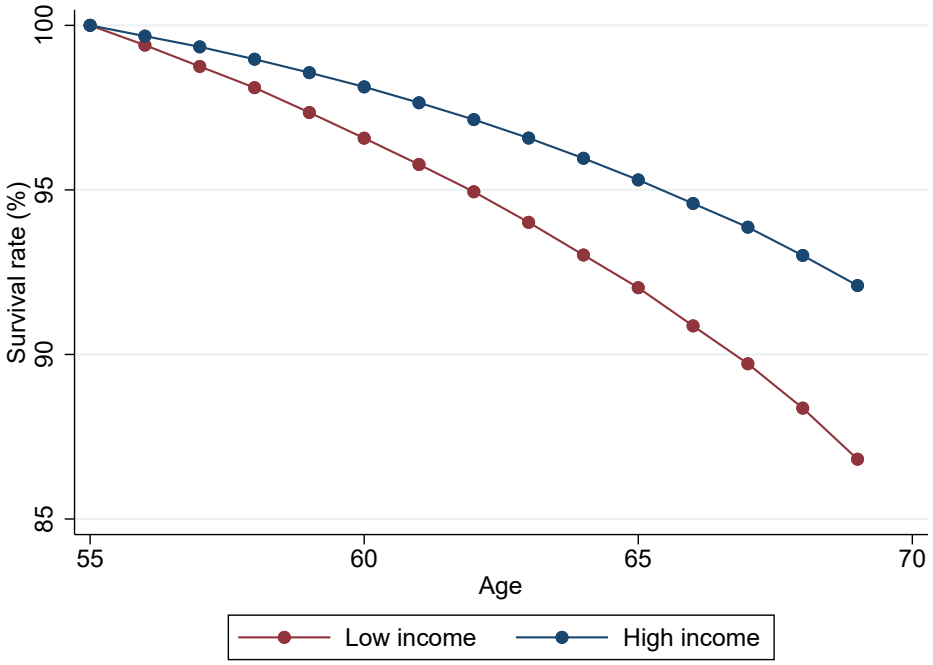
	High Income Y_H	Low Income Y_L			
	Baseline	Baseline	Y_H Ageing		
			From 60	*From 40*	*From 20*
1. Biological Age (relative to Y_H)					
a. at 70	70.0	75.4	73.7	71.2	70.3
b. at 40	40.0	49.3	49.3	49.3	43.3
2. Life Expectancy (at age 40)	85.3	81.2	81.6	82.3	82.5
3. Lifetime Healthcare Costs					
a. Net Effect	155.9k	157.7k	157.6k	155.6k	153.2k
b. Keeping Survival Unchanged	155.9k	157.7k	155.5k	148.9k	145.6k

Note: Simulated biological ages life expectancy and lifetime cost figures are shown. In the biological age calculations, the CDI is simulated over the lifecycle using only aging effects. More specifically, in the baseline calculations, we apply the high and low income aging effects to their respective income groups starting from age 20. In the counterfactual scenarios, the high income aging effects are applied to the low income baseline from different ages onwards. This procedure is visually represented in panel B of Figure 7. More detail on the life expectancy and lifetime costs calculations is provided in Appendix H.

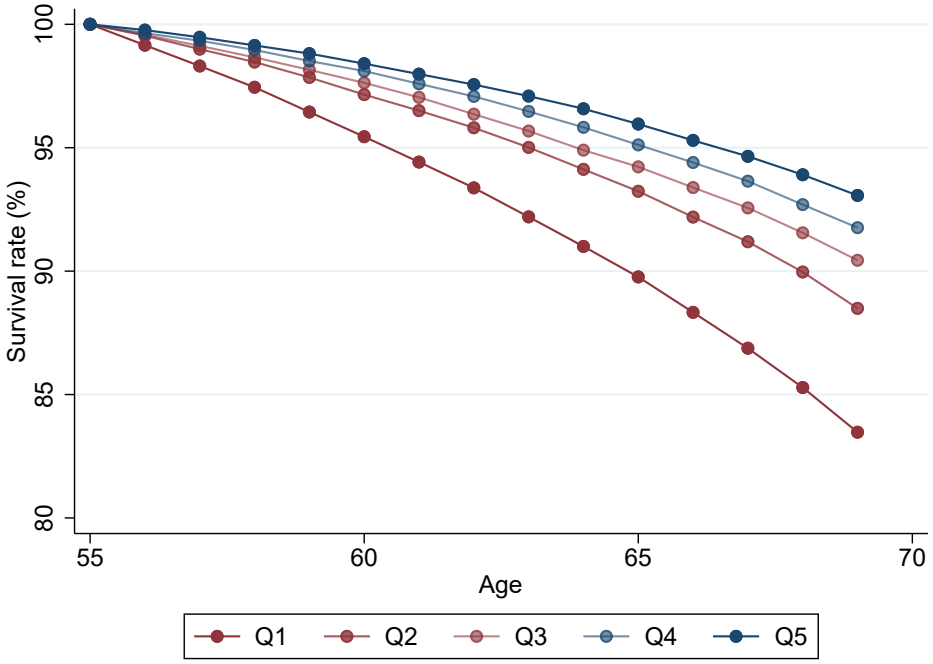
B Figures

Figure 1: SURVIVAL CURVES, 55 YEAR OLD

A. Low and High Income



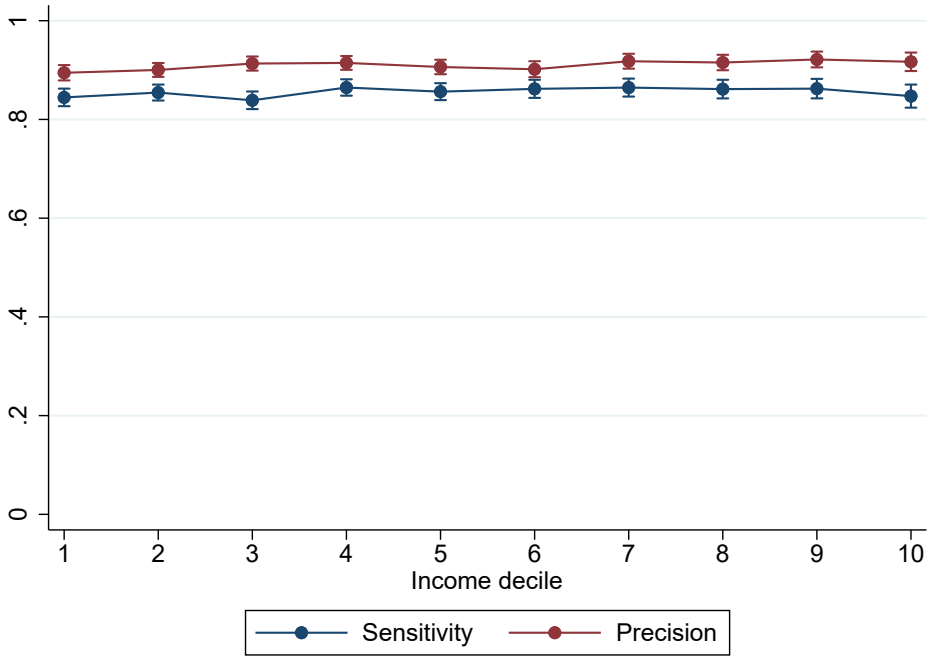
B. By Income Quintile



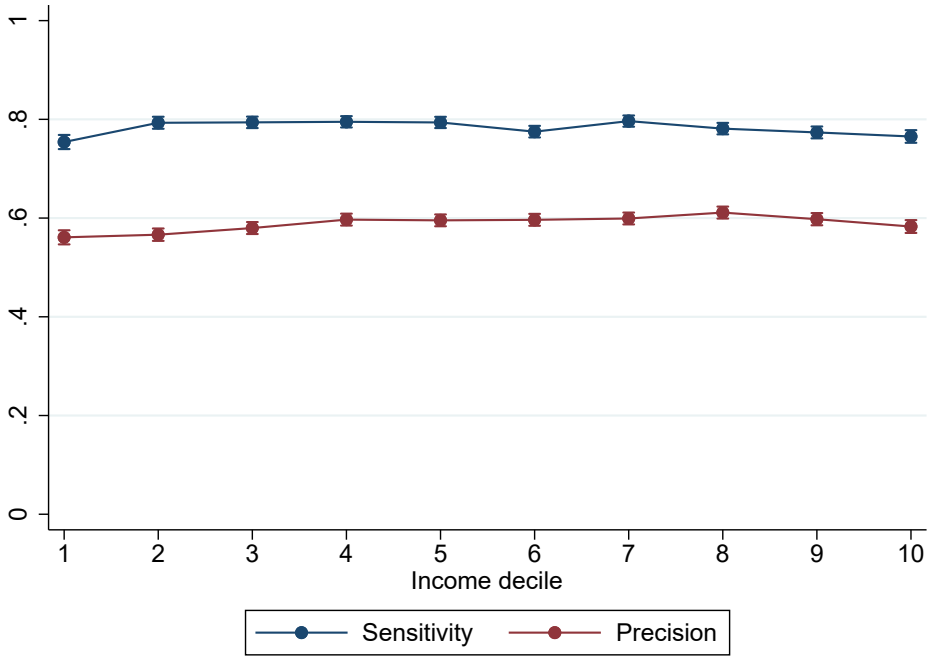
Note: Panel A displays 15-year survival curves for the cohort of 55-year olds in 2007 for low and high income individuals. At each age, the probability of survival until this age, conditional on being observed at 55 in 2007 is presented. Panel B presents 15-year survival curves by income quintile for the same sample of individuals. Income groups are defined within gender in 2007 and kept constant until age 69 (which corresponds to 2021 for this cohort).

Figure 2: COMPARING PHARMACY DATA TO SELF-REPORTED SURVEY RESPONSES BY INCOME

A. Diabetes



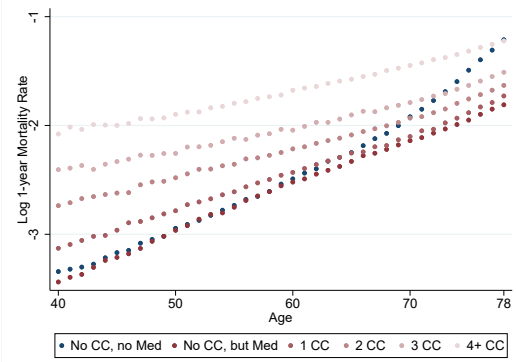
B. Cardiovascular Disease (incl. Hypertension)



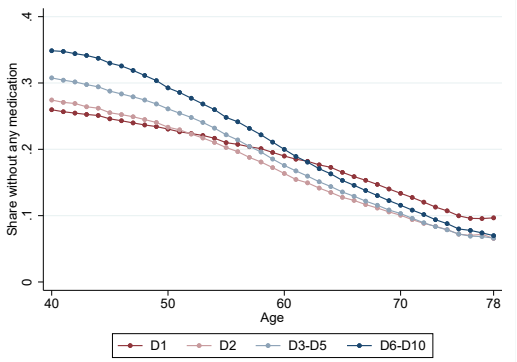
Note: This figure compares the rates of detection of diabetes and cardiovascular disease using our ATC to chronic condition mapping, versus the reporting of diabetes in the *Gezondheidsmonitor* survey data, by income decile. Sensitivity = Pr(condition detected and reported | condition reported in survey), Precision = Pr(condition detected and reported | condition detected in pharmacy data).

Figure 3: UNDER-DIAGNOSIS BY INCOME DECILES

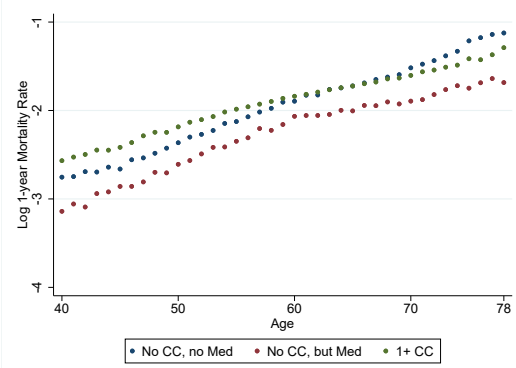
A. Mortality by Chronic Conditions



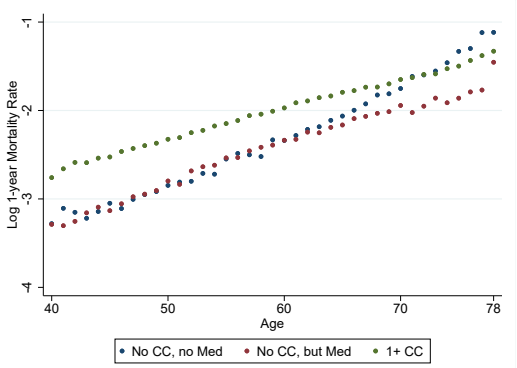
B. Share of no Medication, by Income Deciles



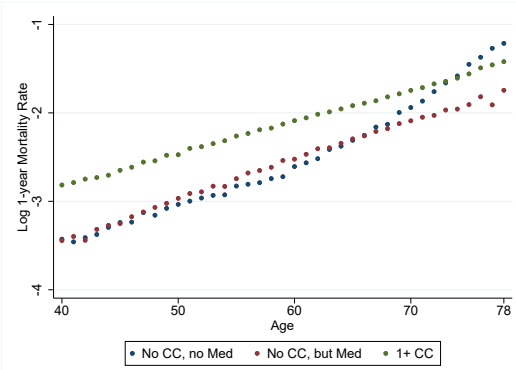
C. Mortality by CC, 1st Income Decile



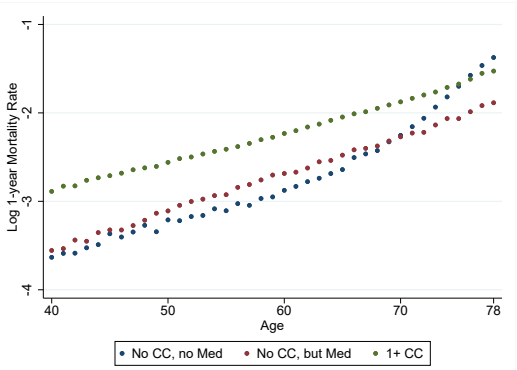
D. Mortality by CC, 2nd Income Decile



E. Mortality by CC, 3rd-5th Income Decile



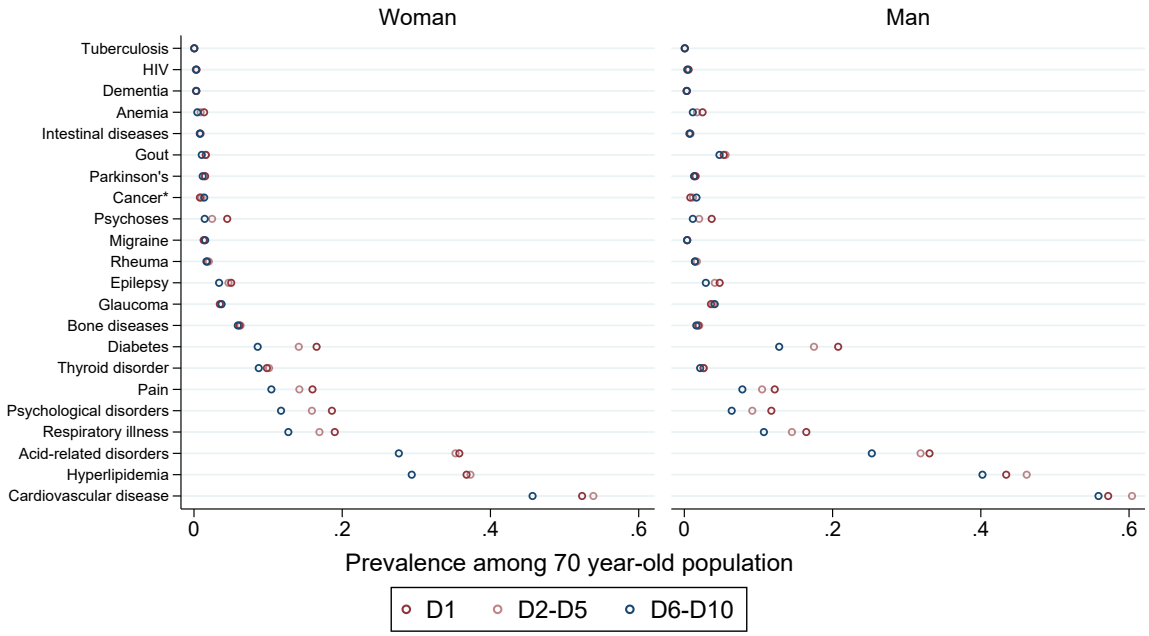
F. Mortality by CC, 6th-10th Income Decile



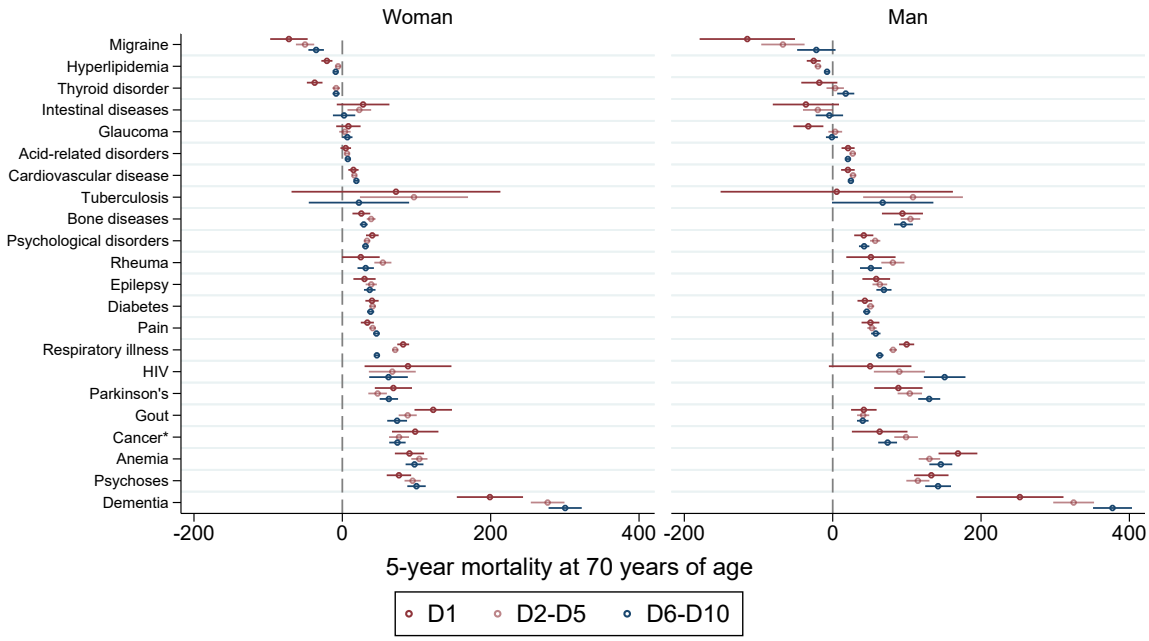
Note: Panel A reports one-year mortality by number of chronic conditions on a logarithmic scale. The category without chronic conditions is divided into a group taking some medication for non-chronic illnesses and a group with individuals taking no medication at all. This panel pools all observations for which prescription data are available. Hence, it includes all individuals between 2006 and 2021. Panel B plots the share of individuals who do not take any medication for different income groups. The income groups considered consist of individuals situated in deciles 1, 2, 3-5 and 6-10, respectively. Income deciles are defined within birth cohort and gender. Panels C-F show 1-year mortality rates for individuals with different medication status for each of these income groups. Panels B-F pool all individuals for which our income measure is defined. Hence, it includes all individuals between 2007 and 2021.

Figure 4: PREVALENCE AND TREATMENT EFFECTS OF CHRONIC CONDITIONS

A. Prevalence of Chronic Conditions at Age 70

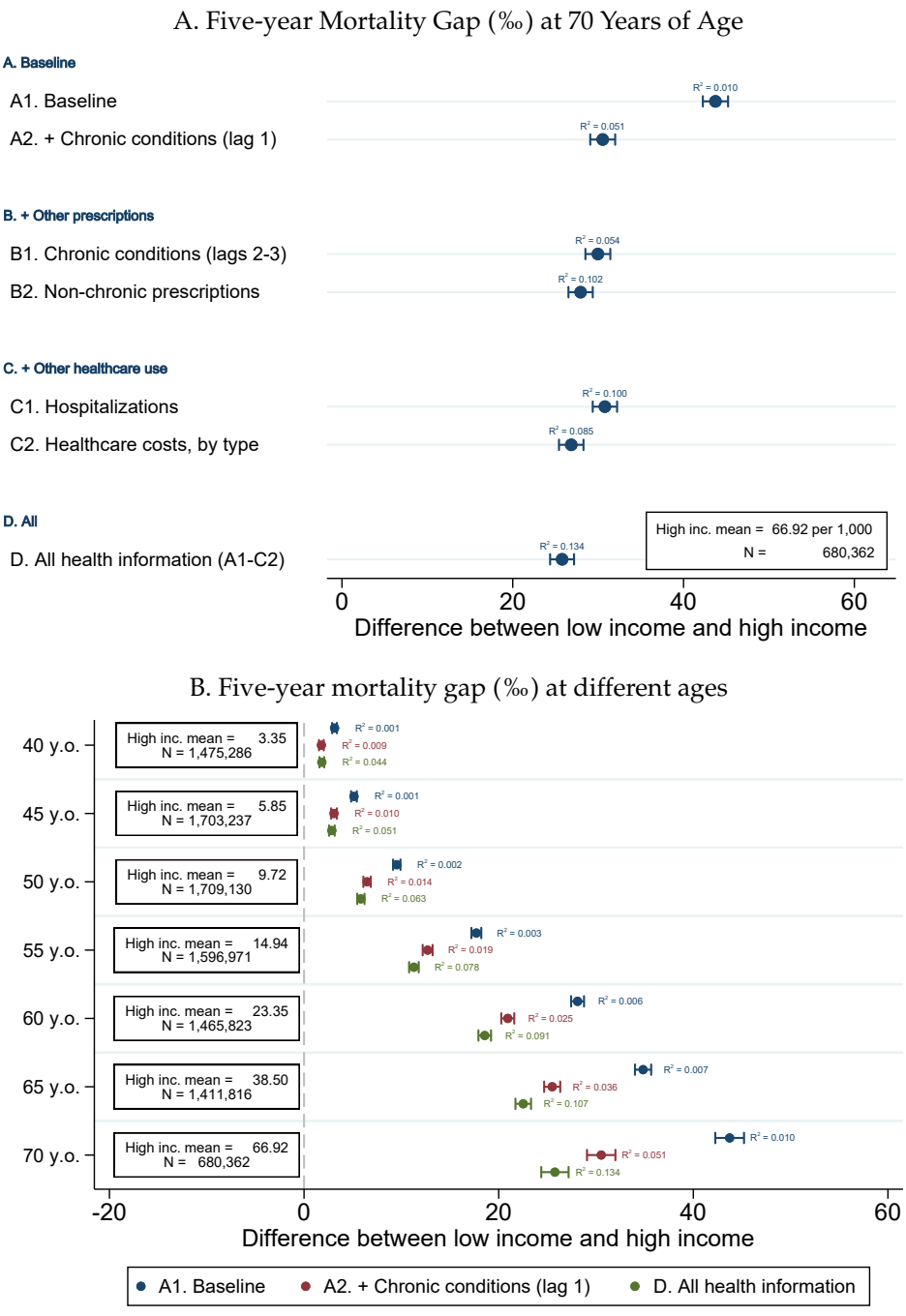


B. Effect of Chronic Conditions on Five-Year Mortality at Age 70 (‰)



Note: Panel A reports the prevalence of each chronic condition among different income strata of the population at age 70, by gender: the bottom income decile (D1); deciles two to five (D2-D5); and deciles six to ten (D6-D10). Confidence intervals are not reported, as they are indiscernible. Panel B reports the coefficient estimates when regressing five-year mortality on all chronic conditions for the different income groups by gender. *Cancer here refers to cancers treated with pharmacy-dispensed medications, which is around 5% of all cancer diagnoses. Digestive tract and skin cancers dominate this measure, they account for over 60% of the diagnoses.

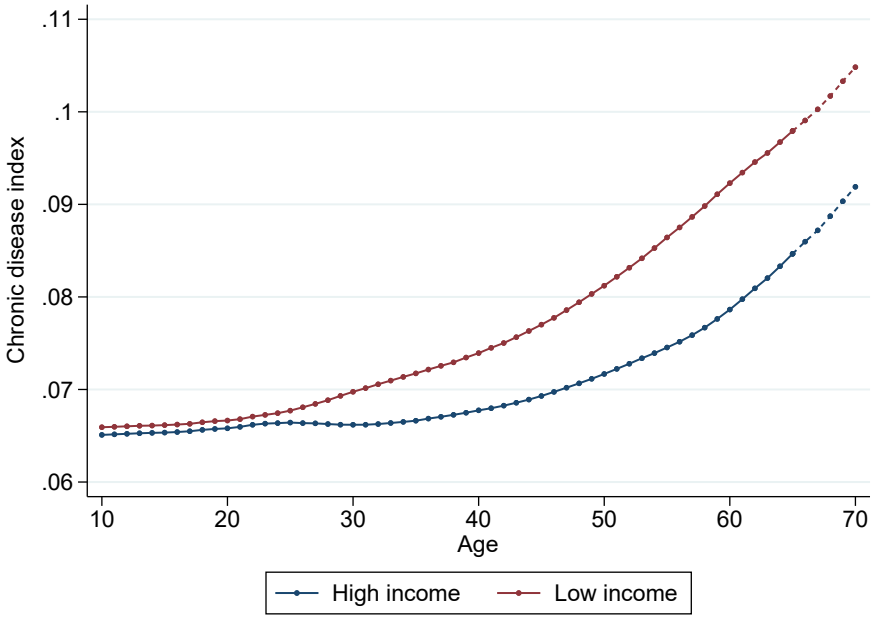
Figure 5: CHRONIC CONDITIONS AND THE MORTALITY GAP



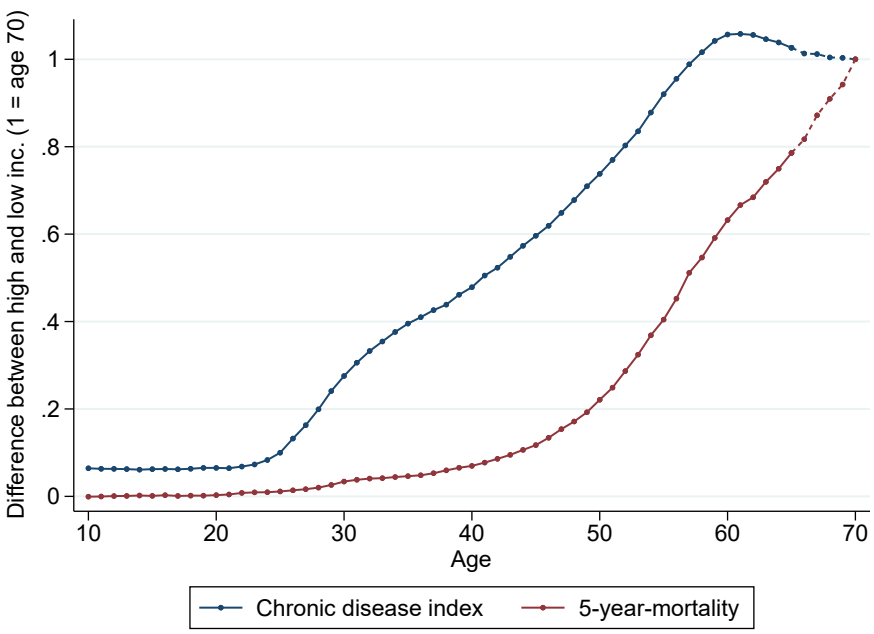
Note: In panel A, each row corresponds to a different regression of five-year mortality (in thousands) on income (defined as low- vs. high-income) and a series of controls identified by the row label on the left. For each specification, the plot shows the estimated coefficient on income. Specifications reported after A2 include all the chronic condition controls used in A2, as well as those listed in the left column. Row D includes all health-related variables jointly. Panel B reports the mortality gap estimates from specifications A1, A2, and D at different ages. For more information, refer to Appendix Section E.

Figure 6: HEALTH GAP OVER THE LIFECYCLE

A. Average CDI by Income Group



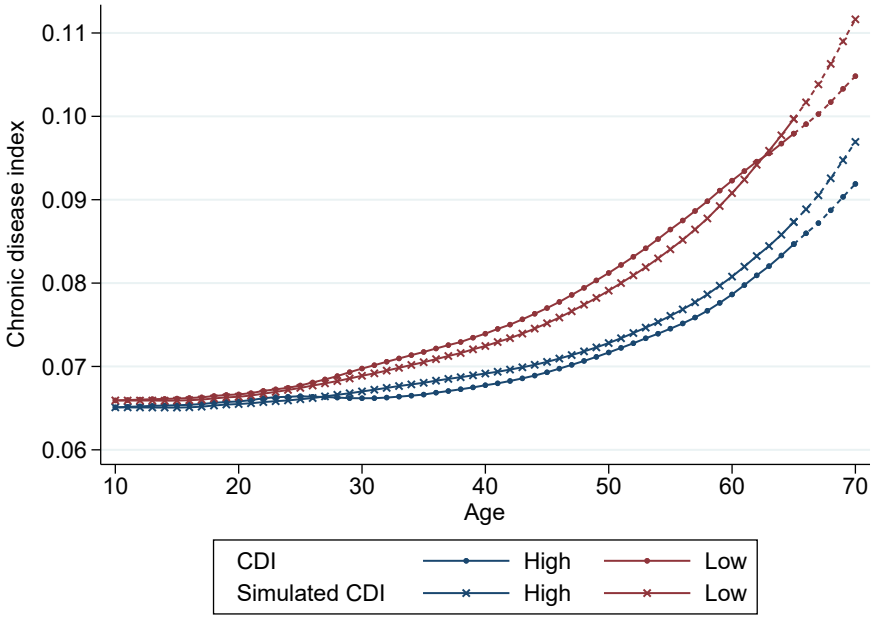
B. CDI and Mortality Gap over the Lifecycle



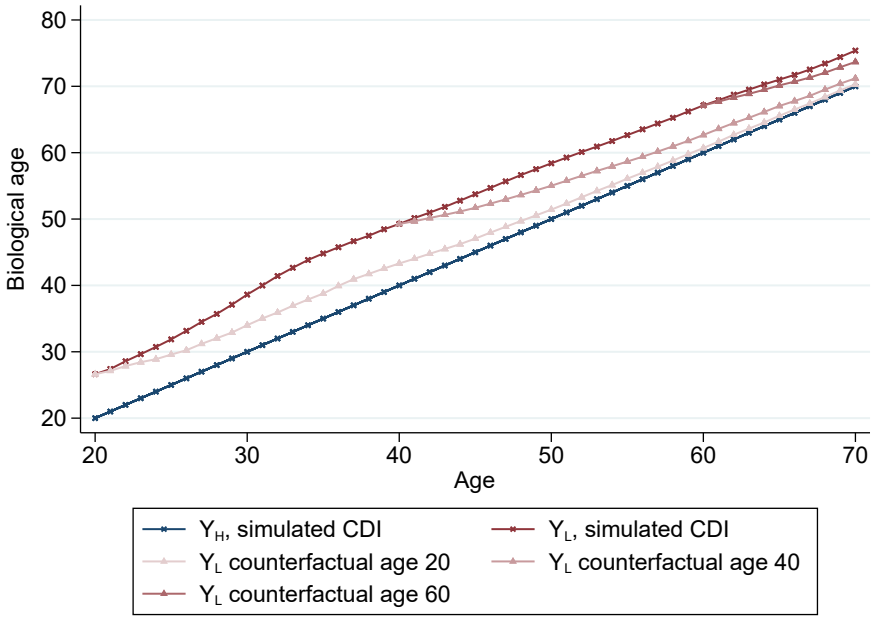
Note: Panel A documents the evolution of the average chronic disease index by income group over the lifecycle. That is, at each age, the average CDI is shown for the relevant income group. Individuals are ranked on the mean of $(Y_{t-4}, Y_{t-3}, Y_{t-2})$ within year, age and gender. High income is defined as above median income, and low income as below median. From age 65 onwards, we fix income as the mean of (Y_{60}, Y_{61}, Y_{62}). This is represented by the dashed lines in the figure from age 65 onwards. Panel B shows the difference in the CDI between both income groups, along with the difference in observed 5-year mortality. Both gaps in panel B are shown relative to age 70, which is set to 1. We pool all observations between 2009 and 2021 in both panels.

Figure 7: BIOLOGICAL AGING

A. Differential Aging

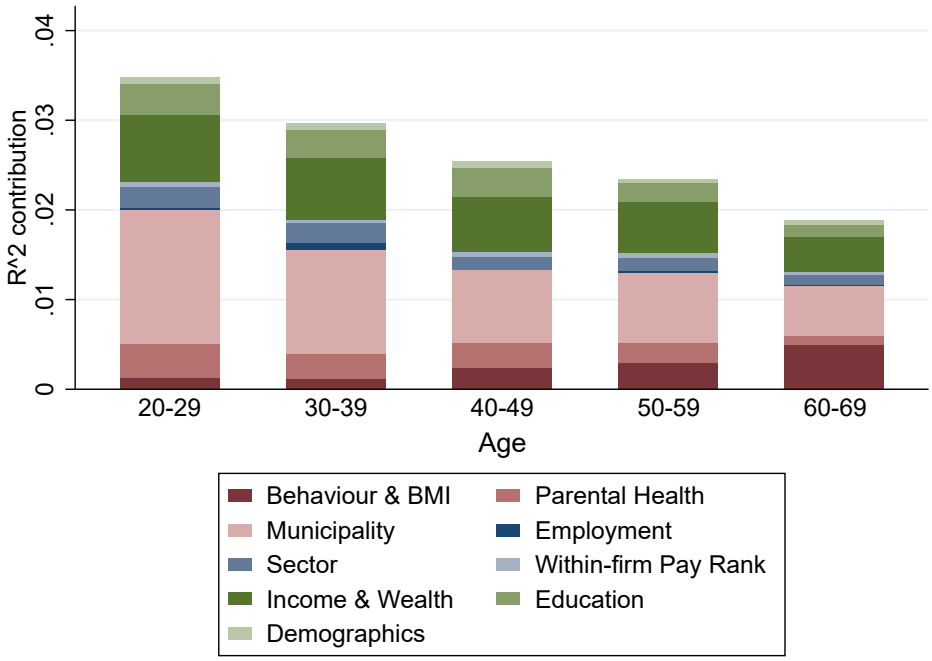


B. Counterfactual Biological Age



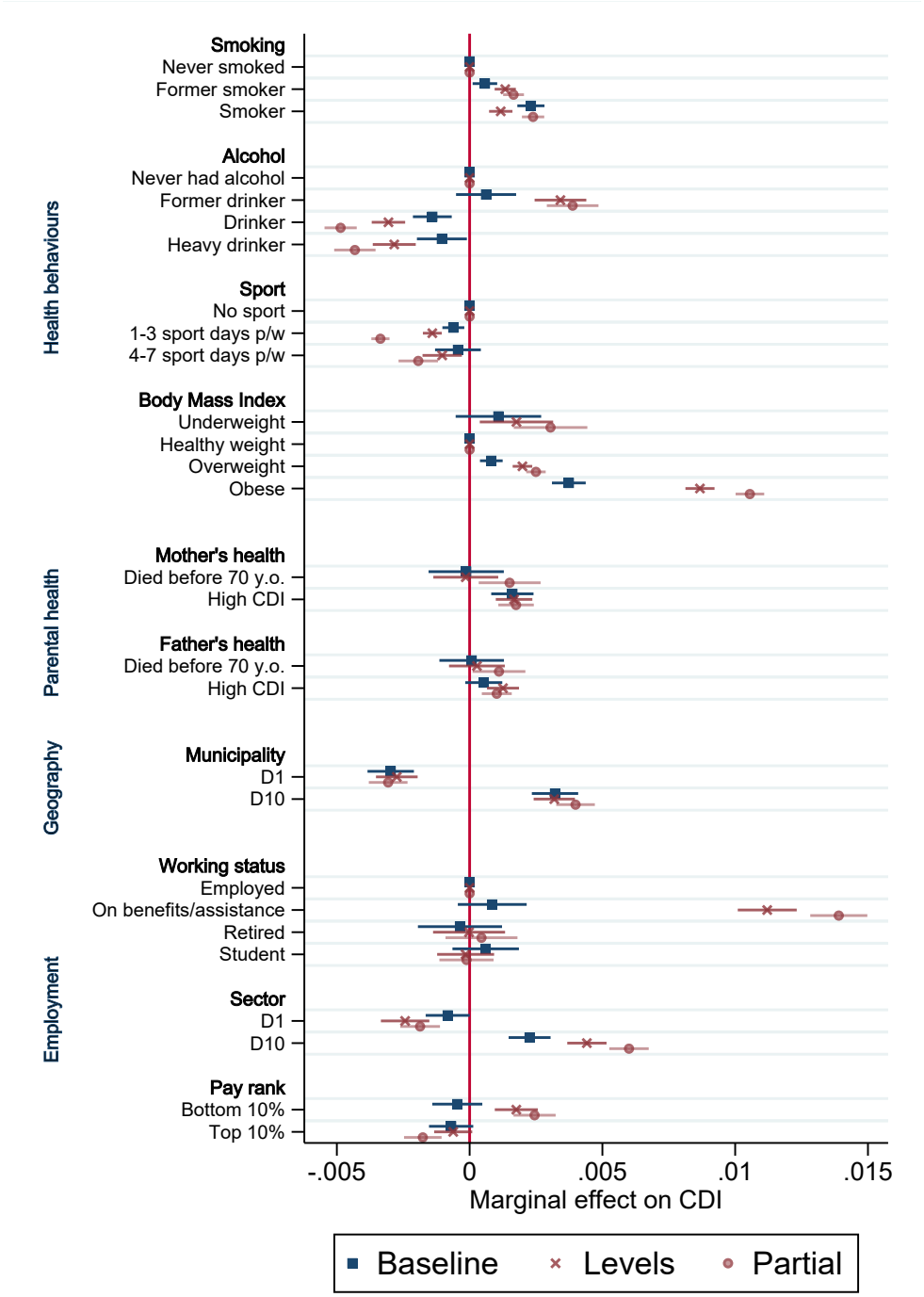
Note: Panel A shows the evolution of the CDI when it is simulated using only aging effects, along with the average CDI's, which are the same as those shown in Panel A of Figure 6. The simulated CDI's start from the observed CDI at age 10, and consider only aging effects, defined in equation (14), to simulate the CDI at all later ages. Panel B shows biological ages for different scenarios. In the baseline scenario, the high and low income CDI are simulated based on their respective estimated aging effects. In the counterfactual scenario's, the high income aging effect is used to simulate the Low Income CDI from different ages onwards. Table 4 shows the impact of these counterfactuals on life expectancy and lifetime health expenditures.

Figure 8: SHAPLEY-OWEN DECOMPOSITION OF THE WITHIN-INDIVIDUAL FIVE-YEAR DIFFERENCE IN THE CDI



Note: The figure shows the results of a Shapley-Owen decomposition of five-year log CDI growth on age, gender, and the sets of predictors reported in the legend, based on equation (16). Separate decompositions are carried out for each age bin. The stacked bars represent the contribution of each set of predictors to the overall R-squared. Detailed information on the predictors in each group is provided in section 6.2. Information on the sample coverage is available in Appendix I.

Figure 9: MEDIATORS OF THE CDI

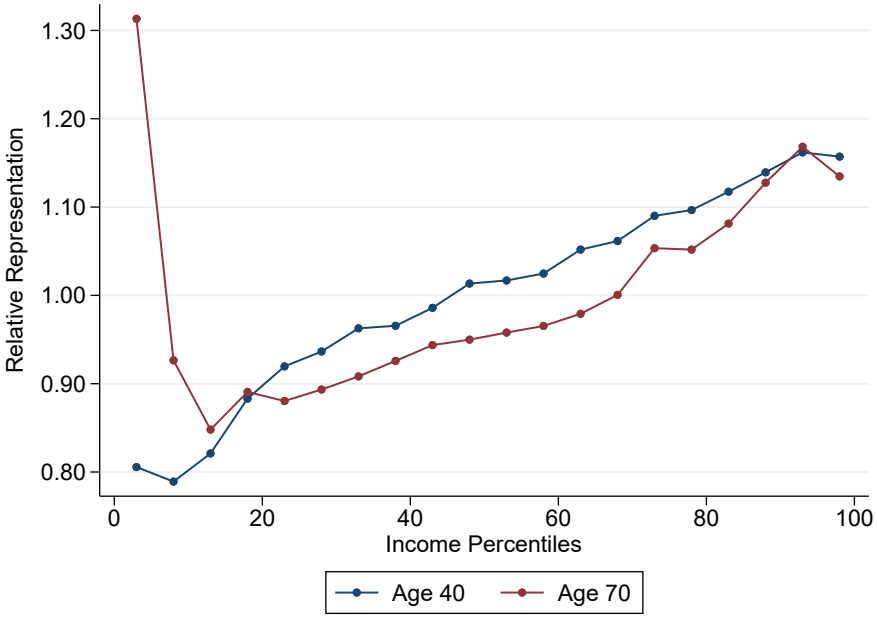


Note: This figure reports coefficients and confidence intervals from regressions of the CDI on mediators. Specification "Baseline" regresses five-year CDI growth ($dCDI$) on the comprehensive set of controls used in the Shapley-Owen Decomposition. These include health behaviors, parental health, geography, and employment predictors, as reported in the figure, but also income, wealth, education, demographics, and parental income and wealth, which are not reported. Specification "Levels" uses the CDI level (CDI) rather than CDI growth ($dCDI$) as dependent variable and uses the same comprehensive set of controls. Finally, specification "Partial" shows the coefficients from separate regressions of the CDI level on each predictor separately. All regressions control for age and gender fixed effects.

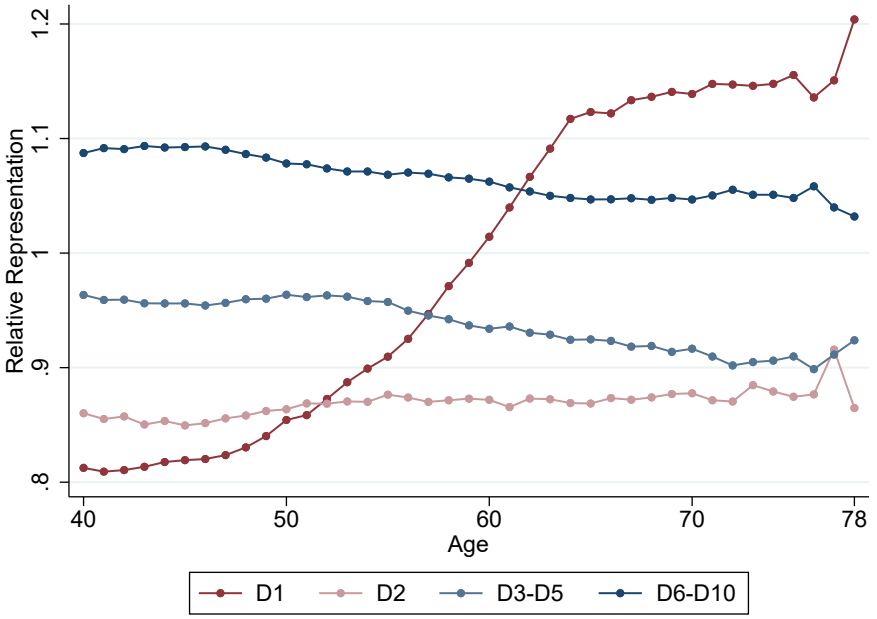
C Appendix: Additional Figures

Figure C.1: UNDER-DIAGNOSIS IN FIRST INCOME DECILE

A. Relative Representation in No Medication Group

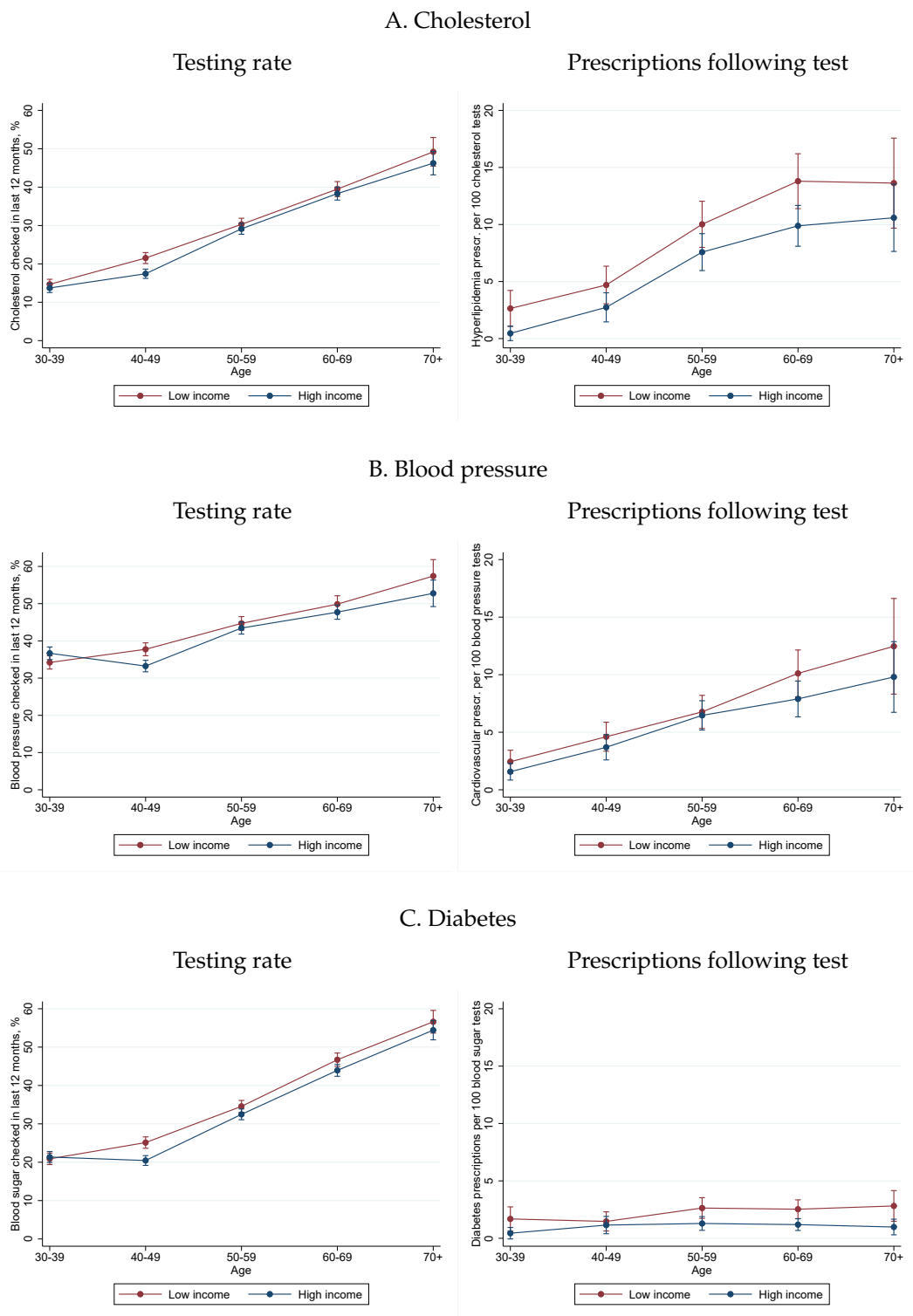


B. Representation in No Medication Sample



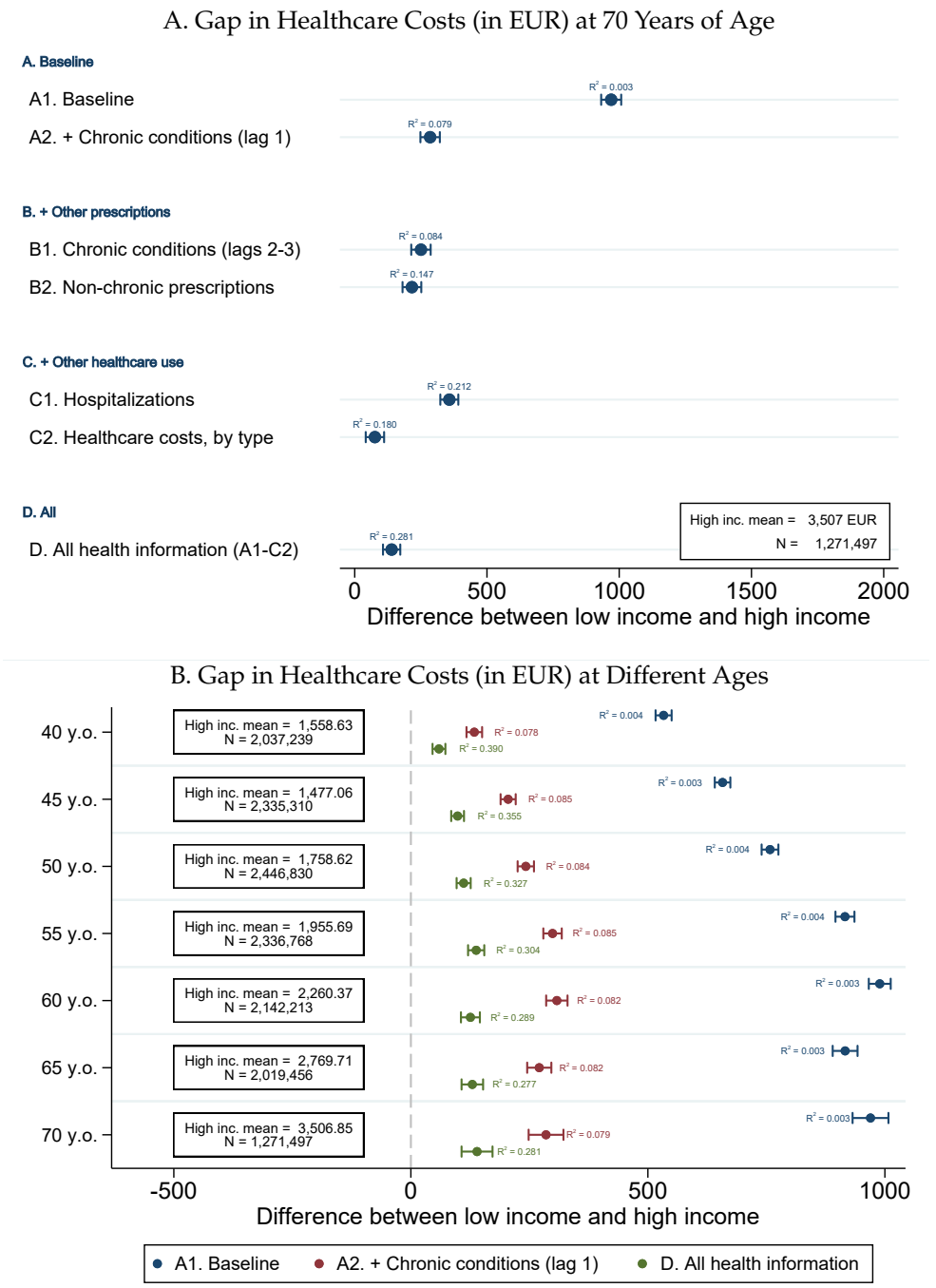
Note: This figure presents evidence for under-diagnosis among very low incomes. Panel A reports relative representation in the sample of people without any medication by income ventile at age 40 and age 70. Relative representation is defined as the share of people within the no medication sample who also belong to a certain income group, relative to the share of this income group in the full sample. Panel B shows the relative representation in the sample of people not taking any medication at ages 40 to 78. The income groups considered consist of individuals situated in decile 1, 2, 3-5 and 6-10, respectively.

Figure C.2: TESTING AND PRESCRIPTIONS FOR SPECIFIC CONDITIONS



Note: This figure compares the rates of screening and subsequent prescription across incomes in the *Gezondheidsenquete* survey data. In the left-hand panels, we focus on the subset of people who are not currently prescribed the relevant medication, and reports responses to the question "*Have you had a* [Cholesterol/Blood pressure/Blood sugar] *test in the past 12 months?*". The right-hand panels then report the share of those tested that were prescribed with the relevant medication in the following year.

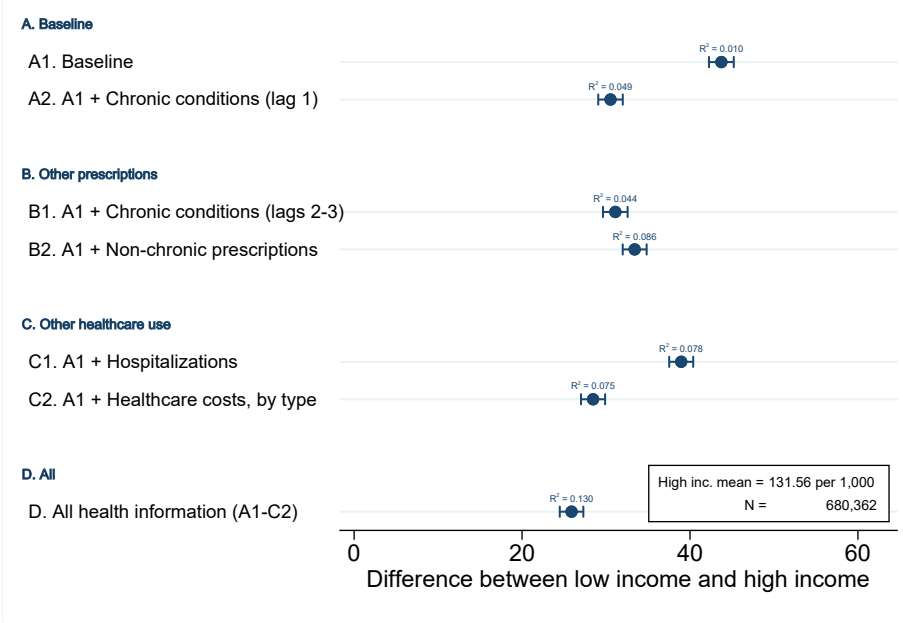
Figure C.3: GAP IN TOTAL HEALTHCARE COSTS BETWEEN HIGH- AND LOW-INCOME INDIVIDUALS



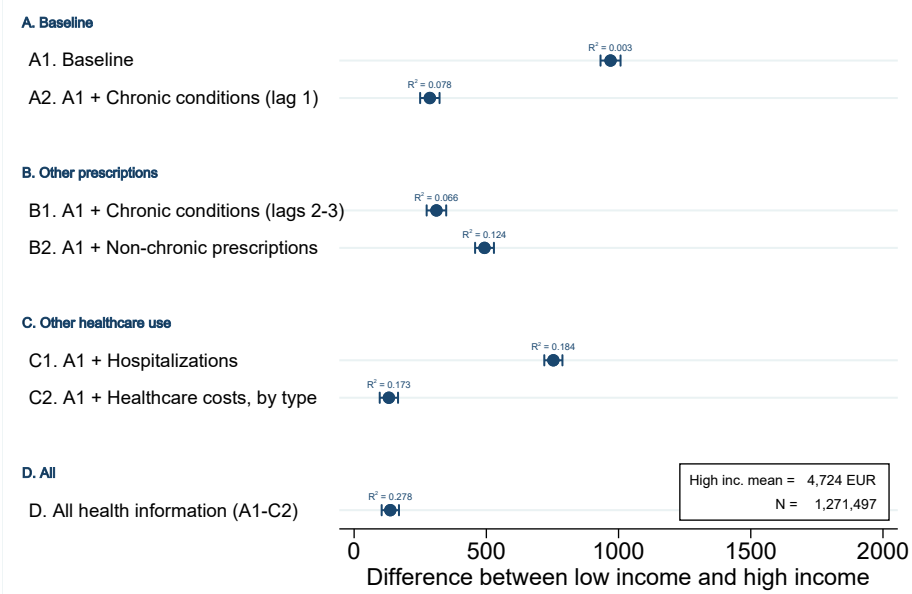
Note: In panel A, each row corresponds to a different regression of total healthcare costs (in euros) on income (defined as low- vs. high-income) and a series of controls identified by the row label on the left. For each specification, the plot shows the estimated coefficient on income. Specifications reported after A2 include all the controls used in A2, as well as those listed on the left. Panel B reports the healthcare costs gap estimates from specifications A1, A2, and D at different ages. For more information, refer to Appendix Section E.

Figure C.4: MORTALITY AND HEALTHCARE COST GAP, SEPARATE CONTROLS

A. Five-year Mortality Gap (‰) at 70 Years of Age



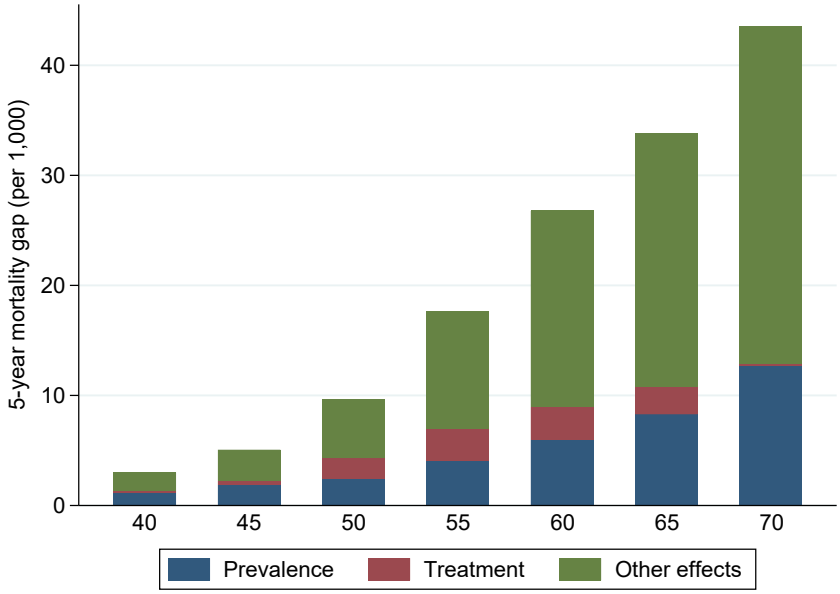
B. Healthcare Cost Gap (EUR) at 70 Years of Age



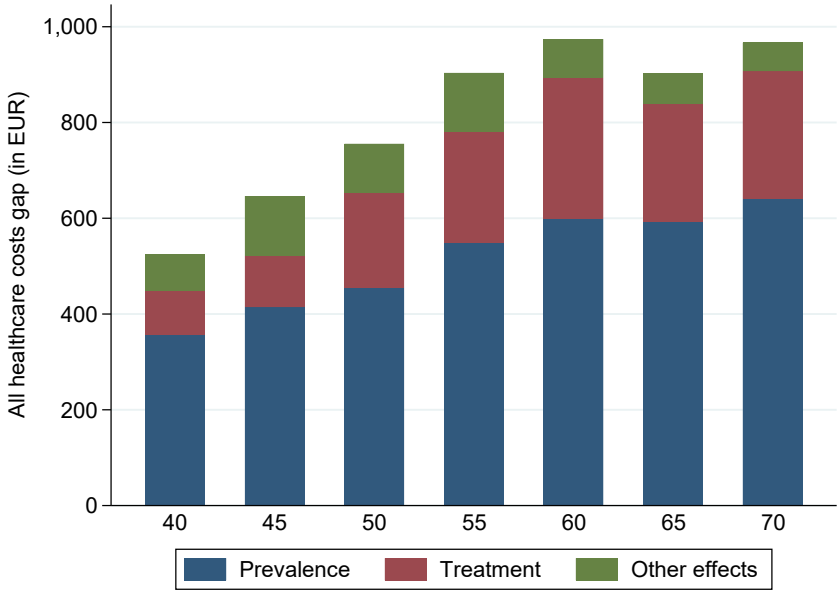
Note: Panel A shows a variation of Figure 5A where specifications B1, B2, C1, and C2 do not control for the first lag of chronic conditions. The coefficients reported thus illustrate how each set of factors reported in the row label affects the estimated income gap in 5-year mortality. Similarly, panel B reports a variation of Appendix Figure C.3A where specifications B1, B2, C1, and C2 do not control for the first lag of chronic conditions. The coefB1, B2, C1, and C2 do not control for the first lag of chronic conditions. The coefficients reported thus illustrate how each set of factors reported in the row label affects the estimated income gapficients reported thus illustrate how each set of factors reported in the row label affects the estimated income gap in healthcare costs. For more information, refer to Appendix Section E.

Figure C.5: OAXACA-BLINDER DECOMPOSITION OF FIVE-YEAR MORTALITY AND HEALTHCARE COSTS

A. Five-year mortality



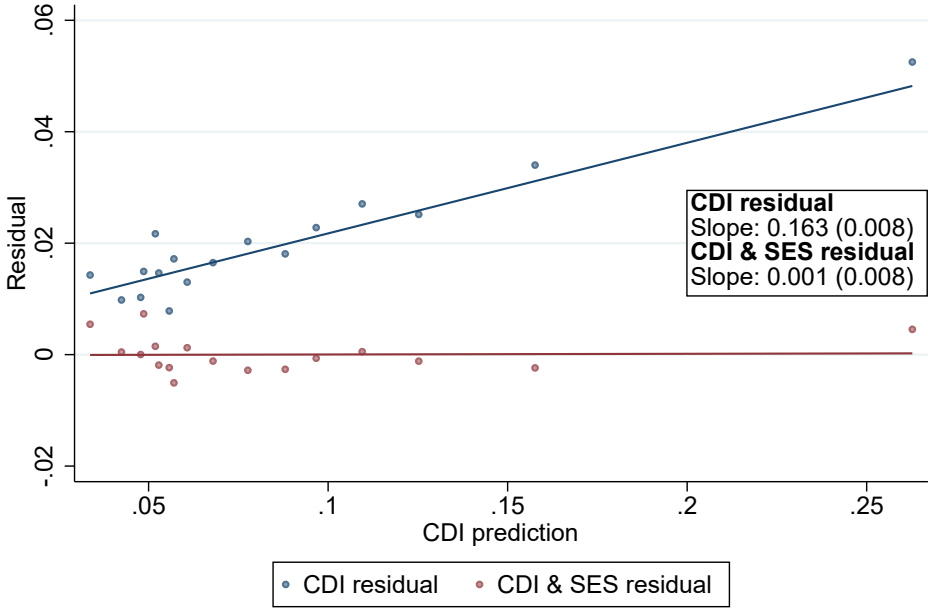
B. Healthcare costs



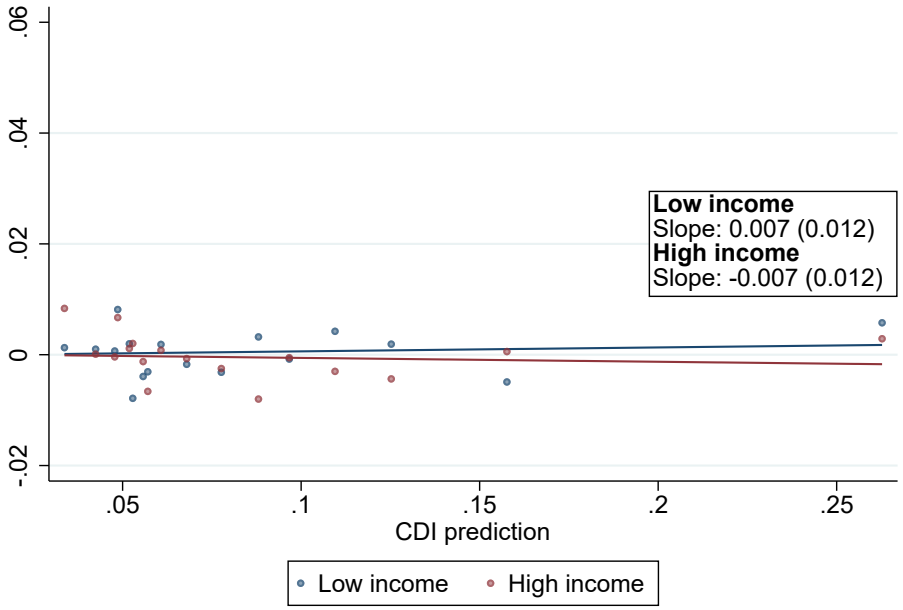
Note: The figure reports the results of a threeway Oaxaca-Blinder decomposition of 5-year mortality (in panel A) and of total healthcare costs (in panel B), using as predictors lagged chronic condition indicators from the previous years. The two groups considered are low-income and high-income individuals, using as threshold the median standardised household income. The "Prevalence" component is given by the part of the difference in means explained by intergroup difference in chronic condition endowments; the "Access / Treatment" component is given by the part explained by intergroup differences in coefficients, excluding the constant term; the "Other effects" component is given by the part explained by intergroup differences in the estimated constant term.

Figure C.6: DIAGNOSTIC BINNED SCATTERPLOTS OF THE CDI

A. Residual mortality risk, controlling for CDI alone or CDI & SES

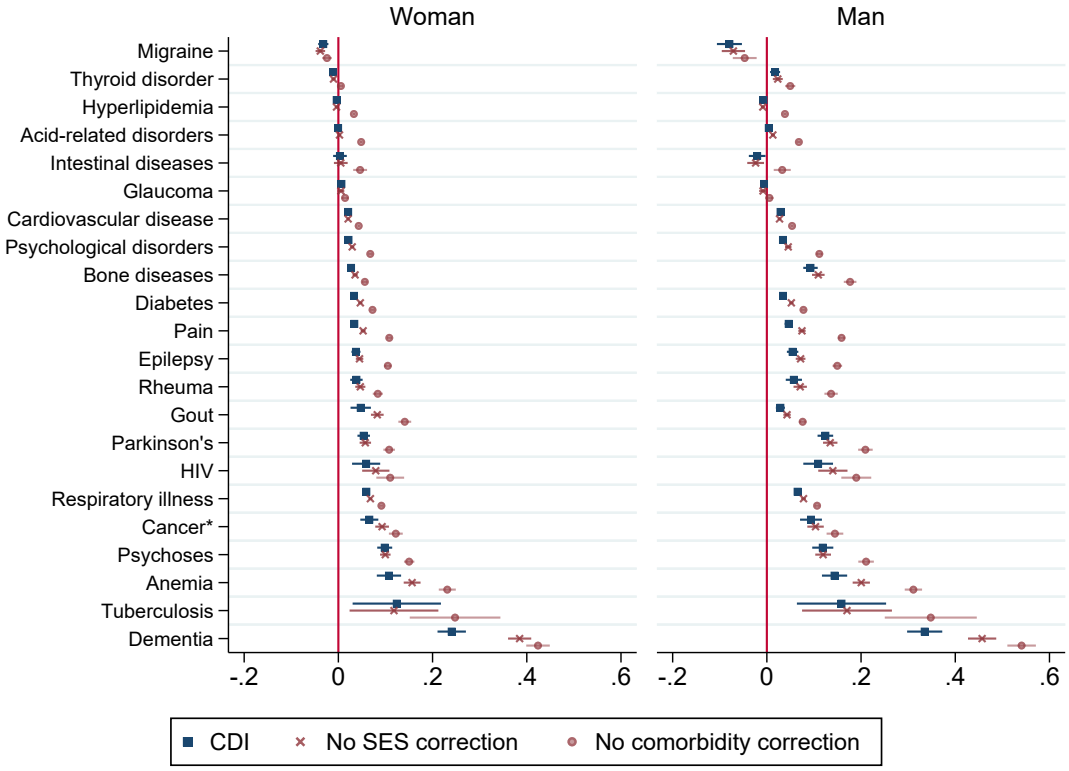


B. CDI & SES residual, by income



Note: Panel A depicts two series, the CDI residual $E\left[m_i - CDI_i \mid CDI_i\right]$, and the CDI & SES residual, which is equivalent to the fitted error term in Equation (10): $E\left[\hat{\zeta}_i \mid CDI_i\right]$. The wedge of 0.163 represents the bias the double-selection procedure excludes. The bias is due to contamination by a correlation between SES and chronic conditions. Panel B depicts the CDI & SES residual, separately for low income and high income $E\left[\hat{\zeta}_i \mid CDI_i, Y_i = Y_L\right]$, $E\left[\hat{\zeta}_i \mid CDI_i, Y_i = Y_H\right]$.

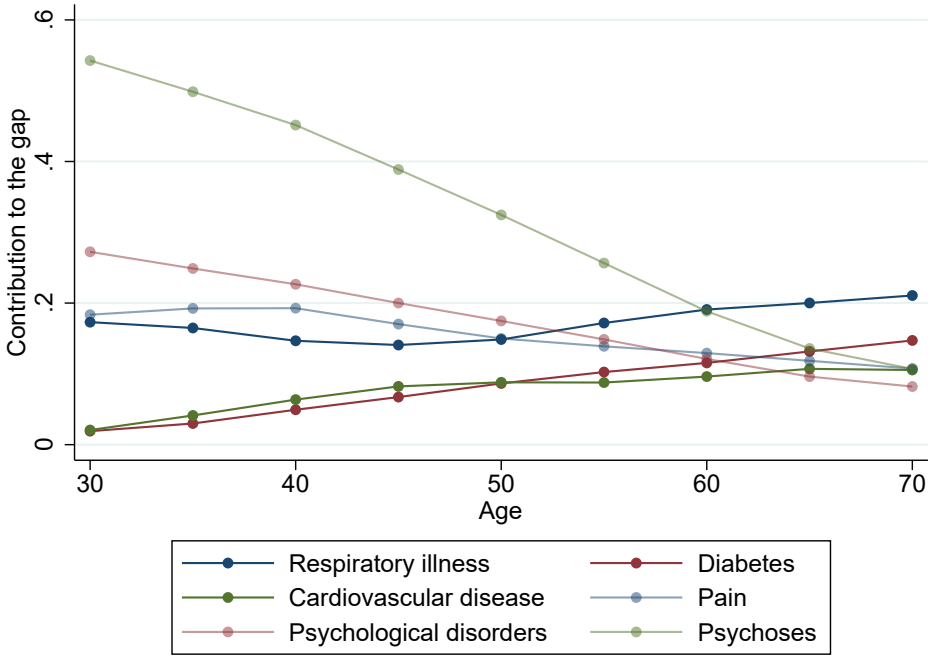
Figure C.7: MARGINAL EFFECTS AND TREATMENT EFFECTS ON PREDICTED CDI, BY CHRONIC CONDITION



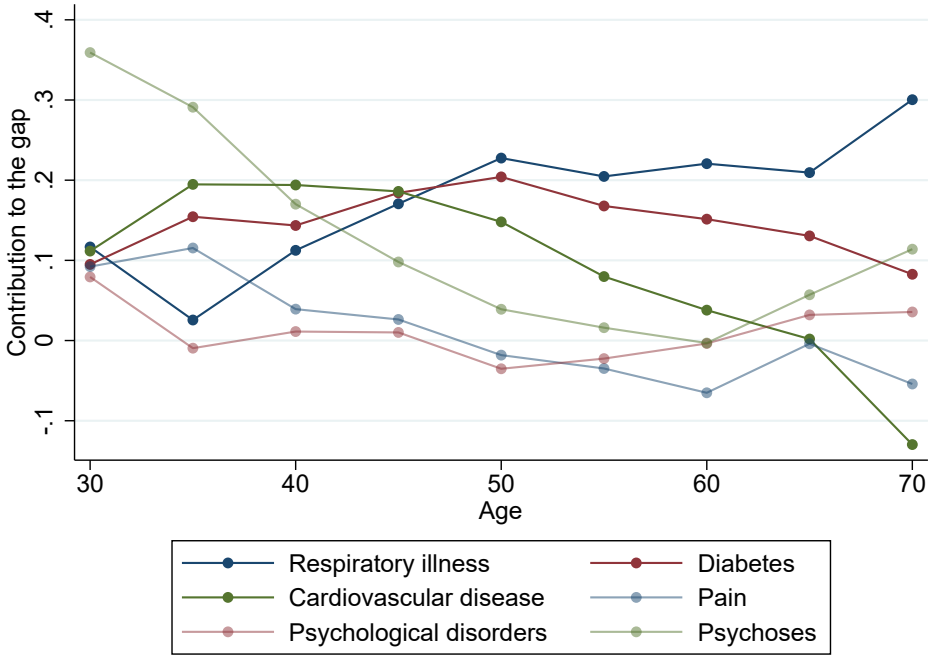
Note: This figure presents the marginal effects of each chronic condition on predicted CDI, by gender. The marginal effects are defined for each gender as follows:
$\beta^j = E[CDI_{70} \mid c^j_{69} = c^j_{68} = c^j_{67} = 1, c^{-j} = \overline{c^{-j}}] - E[CDI_{70} \mid c^j_{69} = c^j_{68} = c^j_{67} = 0, c^{-j} = \overline{c^{-j}}] \quad \forall j = 1, ..., 22.$
Similarly, it presents the treatment effect of each chronic condition from multivariate and univariate regressions of the predicted CDI on lagged chronic conditions. Multivariate regressions estimate the effect of each chronic condition simultaneously. For consistency with the definition of the marginal effects, the displayed treatment effect of a given chronic condition is given by the sum of the coefficients for each of the three lags (one to three) of that chronic condition. *Cancer here refers to cancers treated with pharmacy-dispensed medications, which is around 5% of all cancer diagnoses. Digestive tract and skin cancers dominate this measure, they account for over 60% of the diagnoses.

Figure C.8: CONTRIBUTION OF CHRONIC CONDITIONS TO THE CDI GAPS

A. Contribution to the gap in CDI levels

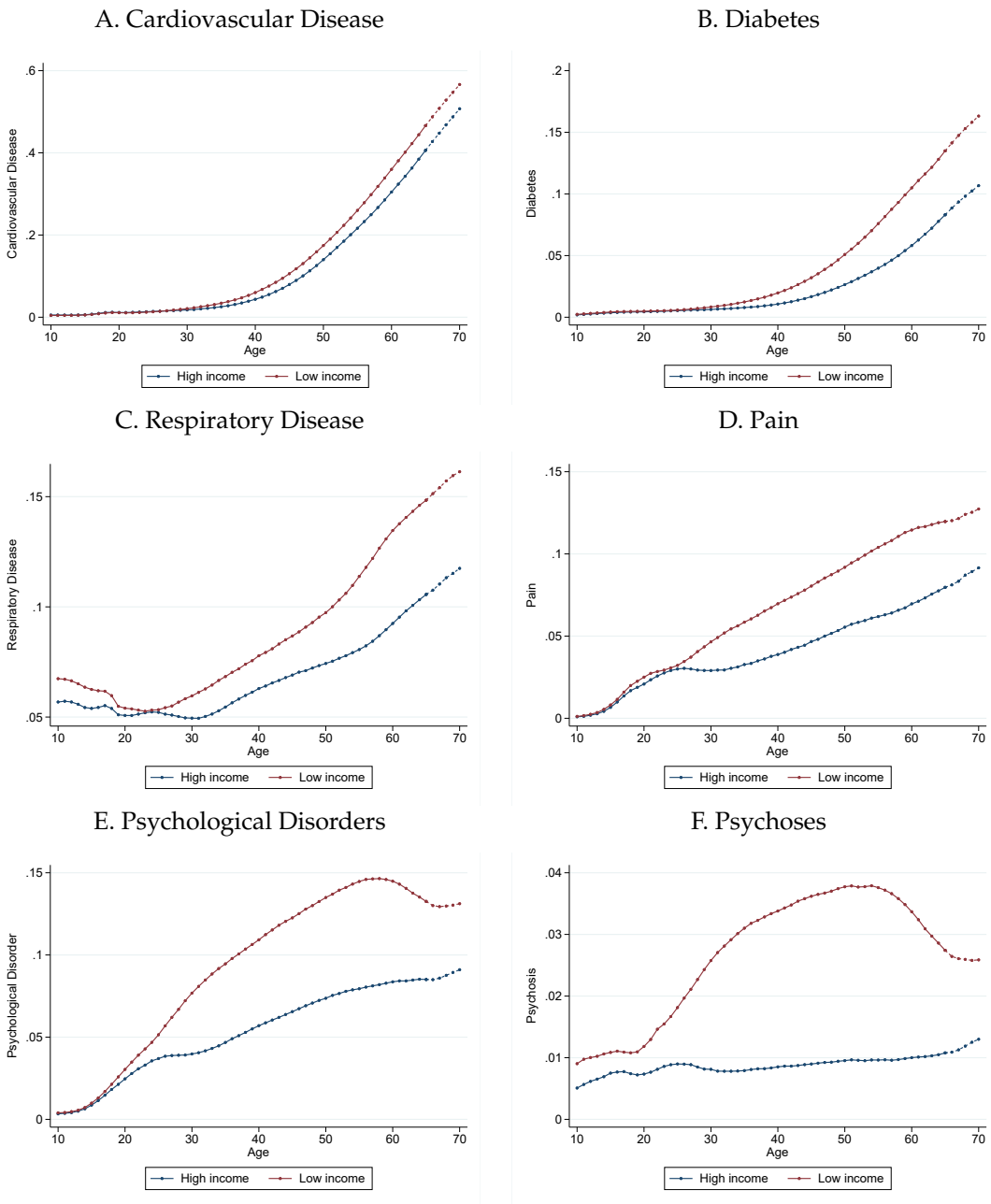


B. Contribution to the gap in five-year CDI growth



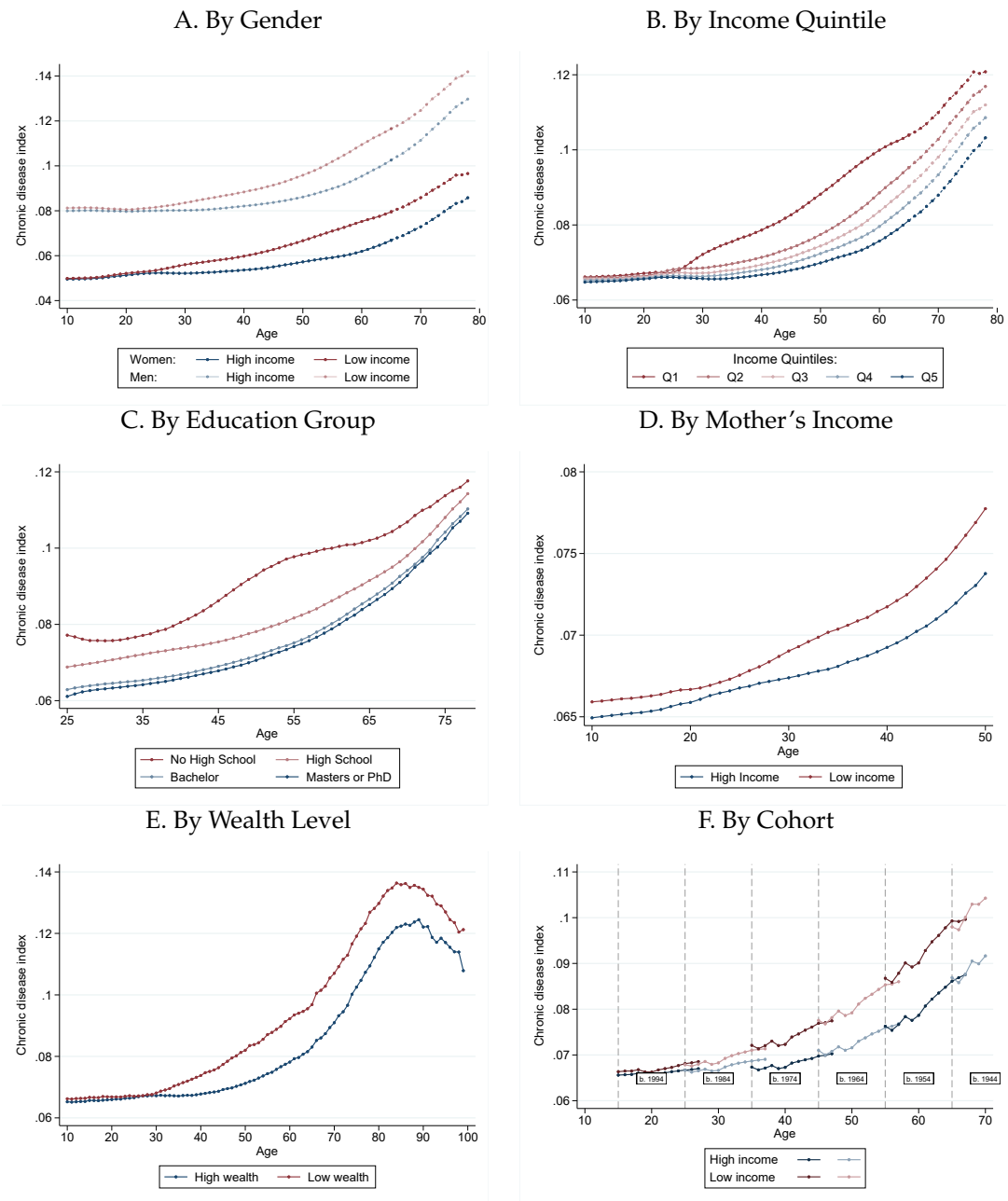
Note: Panel A shows the contribution of a selection of six chronic conditions to the chronic disease index over the life-cycle. Panel B shows the contribution of those chronic conditions to the within-individual five-year change in the chronic disease index. The relative contribution is computed as $\kappa^j = (S^j_L - S^j_H) \cdot \beta^j$, where S^j_Y is the share of income group Y with condition j, and β^j is the marginal effect on predicted CDI, as depicted in Figure C.7.

Figure C.9: LIFECYCLE PREVALENCE OF SPECIFIC CONDITIONS



Note: All panels show the lifecycle prevalence of individual chronic diseases. At each age between 10 and 70, the percentage of individuals taking medication for the specific chronic disease is shown by income group. All panels pool all observations in the period 2009-2021.

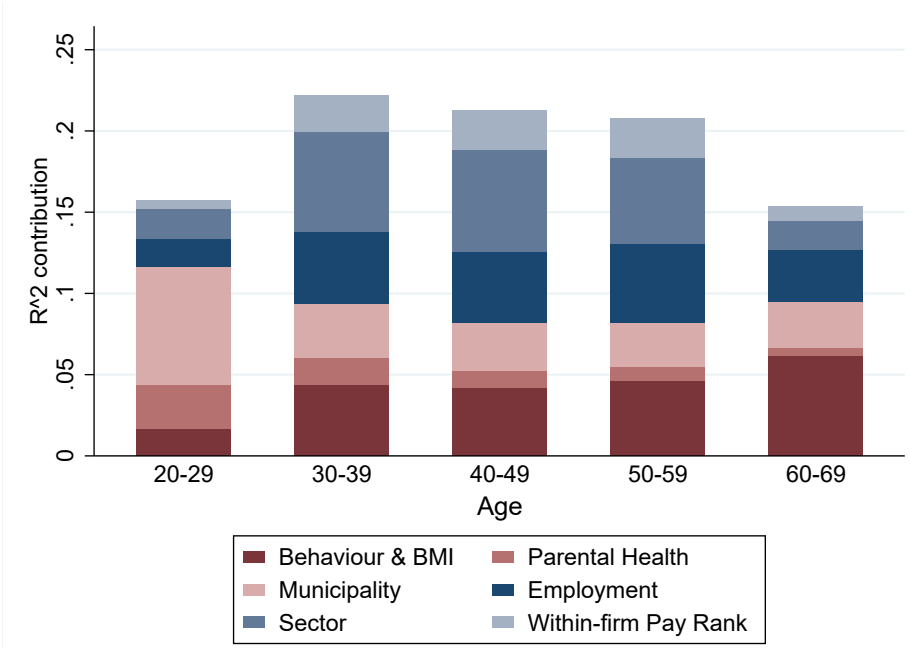
Figure C.10: LIFECYCLE CDI ACROSS SUBGROUPS



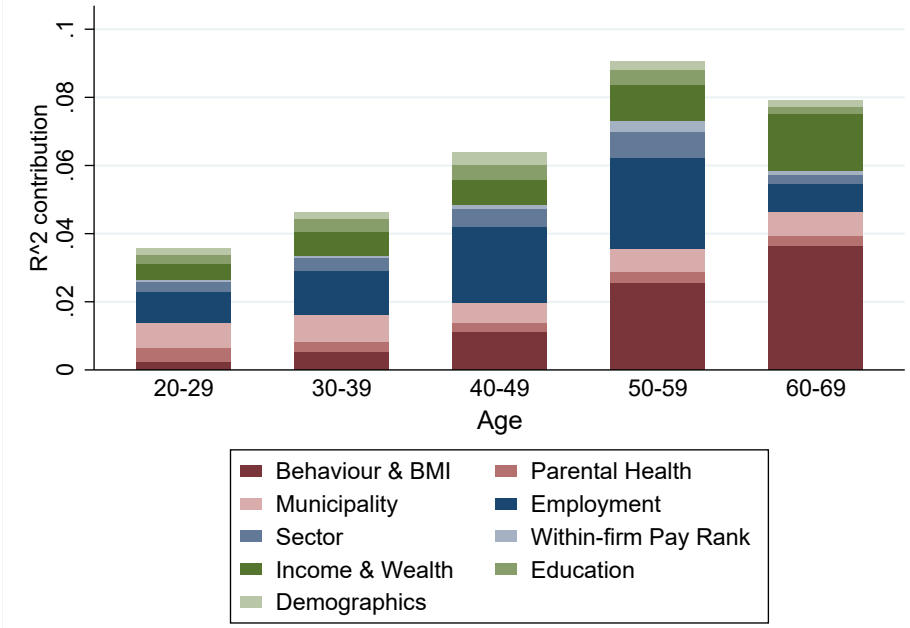
Note: This figure and shows the evolution of the CDI across different subgroups and socio-economic outcomes, similar to Figure 6, which shows the same evolution for high and low income individuals. At each age, the average CDI for the relevant subgroup is shown. Panel A splits by gender and income group. Panel B shows the CDI for 5 income quintiles. Panel C splits by obtained level of education, and panel D splits by income group of the individual's mother. Panel E reports average CDI by above/below median household net wealth. Panels A-E pool all observations in the period 2009-2021. Panel F reports how the average CDI evolves for a selection of birth cohorts. For each cohort, the average CDI's are shown for 13 consecutive years. The earliest age corresponds to the CDI as it is observed in 2009 for each cohort, and the latest age to the CDI as it is observed in 2021. For this analysis, the income groups are defined in 2009 and kept constant until 2021.

Figure C.11: ADDITIONAL SHAPLEY-OWEN DECOMPOSITIONS OF THE CDI

A. Using the Difference in Income-Projected $CDI_{t+5} - CDI_t$

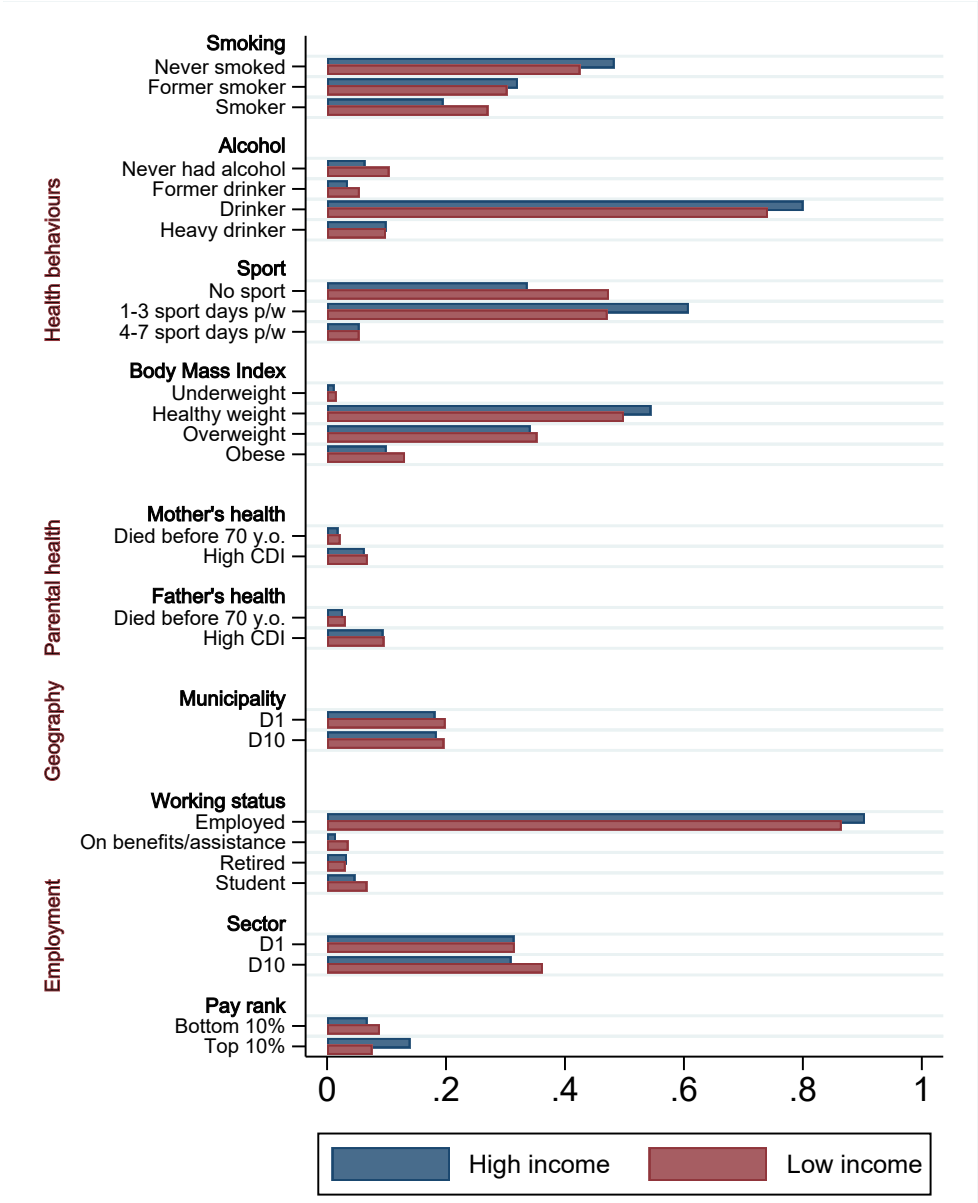


B. Using the CDI level at time t



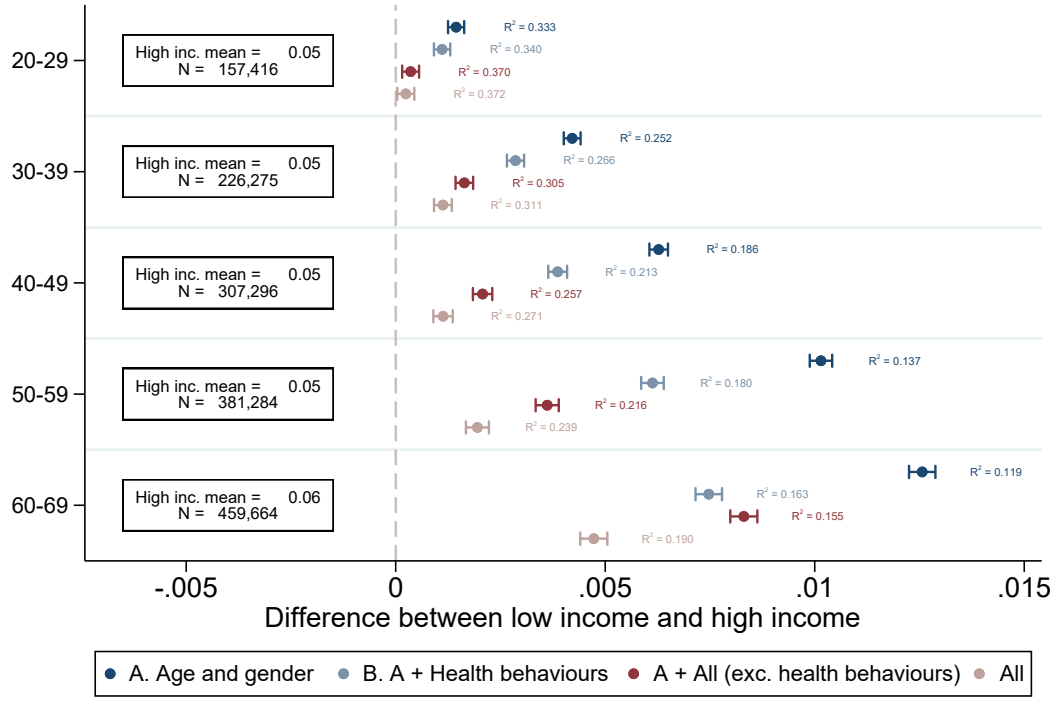
Note: Panel A shows the results of a Shapley-Owen decomposition of the fitted values from a projection of the log five-year CDI growth on income percentile indicators, age, and gender; the predictors used in the decomposition are reported in the legend. Panel B shows the results of a Shapley-Owen decomposition of the CDI on age, gender, and the set of predictors reported in the legend. For both panels, separate decompositions are carried out for each age bin. The stacked bars represent the contribution of each set of predictors to the overall R-squared. Predictors are treated as indicators. Detailed information on the predictors in each group is provided in section 6.2. Information on the sample coverage is available in Appendix I.

Figure C.12: PREVALENCE OF CDI MEDIATORS ACROSS DIFFERENT INCOME GROUPS



Note: The figure reports the prevalence of the CDI mediators within the sample used for regression "Baseline" in Figure 9, separately for individuals with above- and below- median income.

Figure C.13: RELATIONSHIP BETWEEN INCOME GRADIENT IN HEALTH AND BEHAVIOR



Note: The figure reports, for each age shown in the y-axis, multiple estimates of the gap in the chronic disease index between low- and high-income individuals. Each gap estimate is given by the coefficient for a low-income dummy in a regression of the chronic index on a set of predictors identified in the legend. Regressions (A) only control for age and gender when estimating the income gap. Regressions (B) also control for the same health behaviours considered in the Shapley-Owen decompositions (see Figure 8). The third set of regressions controls for all the predictors used in the Shapley-Owen decompositions, except for the health behaviours. The last set of regressions controls for the union of all the predictors used in the previous regressions.

D Refinements to the ATC-Chronic Condition mapping

We use Huber et al. (2013) as our basis to translate medication data into chronic disease indicators. We do however make a number of modifications, as described below.

- **Cardiovascular disease**: Huber et al. (2013) use B01AA (vitamin K antagonists) and B01AC (Platelet aggregation inhibitors excl. heparin), among others. We use B01A (antithrombotic agents). To reduce the number of false positives due to anti-blood-clot medication after an operation, we only consider that the person had cardiovascular disease if she/he took any of the medications in this group for at least two years in a row.

- **HIV**: Huber et al. (2013) use J05AE (protease inhibitors), J05AG (Non-nucleoside reverse transcriptase inhibitors) and J05AR (Antivirals for treatment of HIV infections, combinations). We use J05A (direct acting antivirals). To reduce the number of false positives due to antivirals for acute conditions, we only consider that the person had HIV if she/he took J05A medication for at least two years in a row.

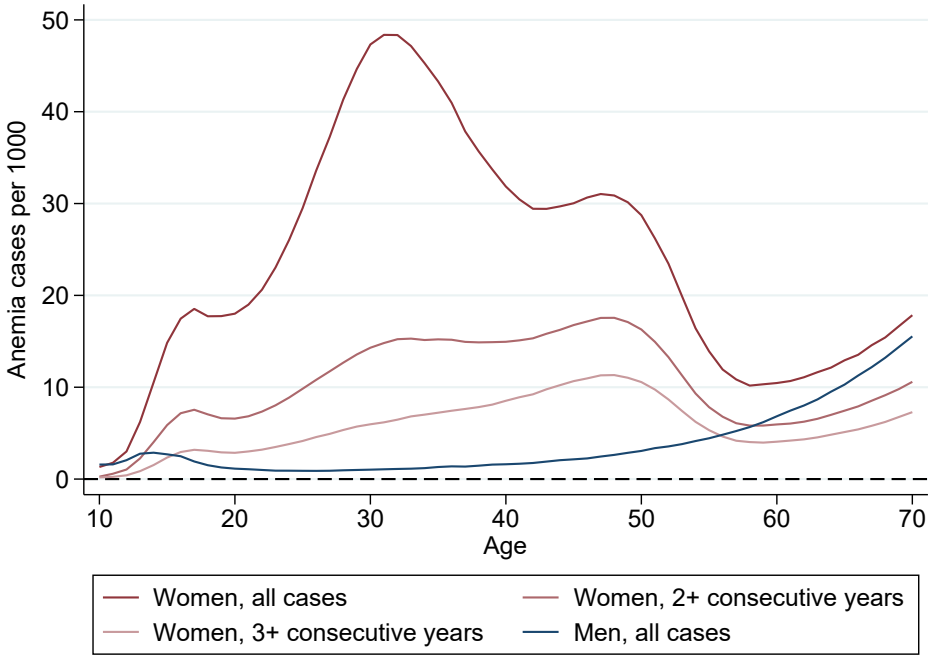
- **Intestinal inflammatory diseases**: Huber et al. (2013) use A07EA (Corticosteroids acting locally) and A07EC (Aminosalicylic acid and similar agents), while we use A07E (intestinal anti-inflammatory agents).

- **Iron deficiency anemia**: Huber et al. (2013) use B03AA (Iron bivalent, oral preparations), B03AB (Iron trivalent, oral preparations) and B03AC (Iron, parenteral preparations). We use B03A (iron preparations). To reduce the number of false positives due to pregnancy related anemia, we only consider that a woman had chronic anemia if she took B03A medication for at least three years in a row. This restriction was informed by diagnostics of the prevalence of medication use by age and gender, as shown in Appendix Figure D.1. B03A medication is predominantly used by women around childbearing age, but this peak is removed when we filter for three consecutive years of use.

- **Rheumatic conditions**: Huber et al. (2013) use M01 (Antiflammatory and antirheumatic products), M02 (topical products for joint and muscle pain), L04AA (selective immunosuppressants) and L04AB (tumor necrosis factor alpha inhibitors), among others. We use the upper group L04A (immunosuppressants), but omit M01 and M02: as shown in Appendix Figure D.2 these are more prevalent than any other chronic conditions at younger ages,

and are associated higher levels of self-reported exercise in the *Gezondheidsmonitor* survey data, suggesting they are being used predominantly for sport injuries at younger ages.

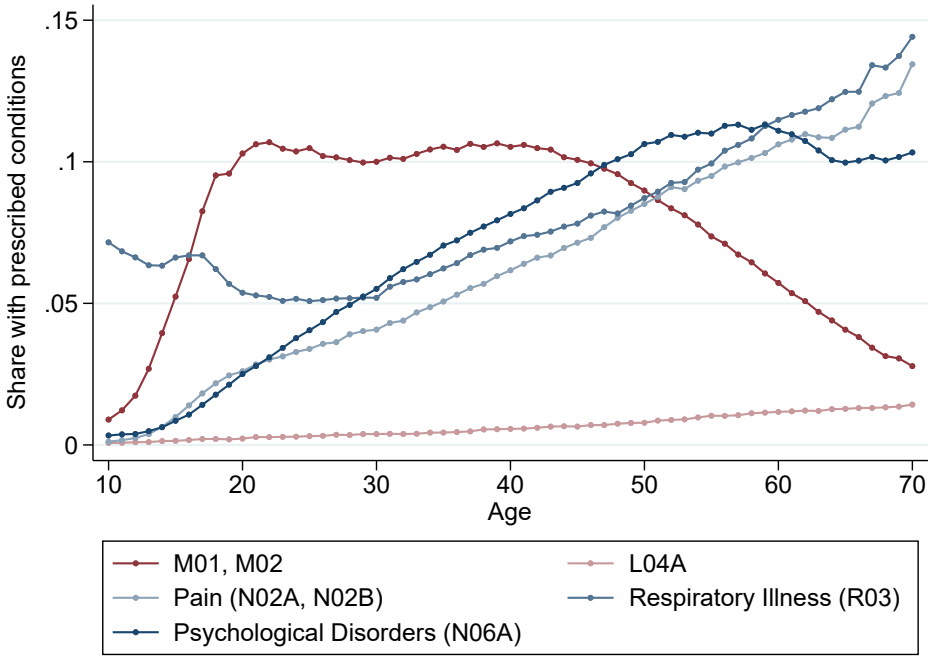
Figure D.1: ANEMIA PREVALENCE OVER THE LIFECYCLE



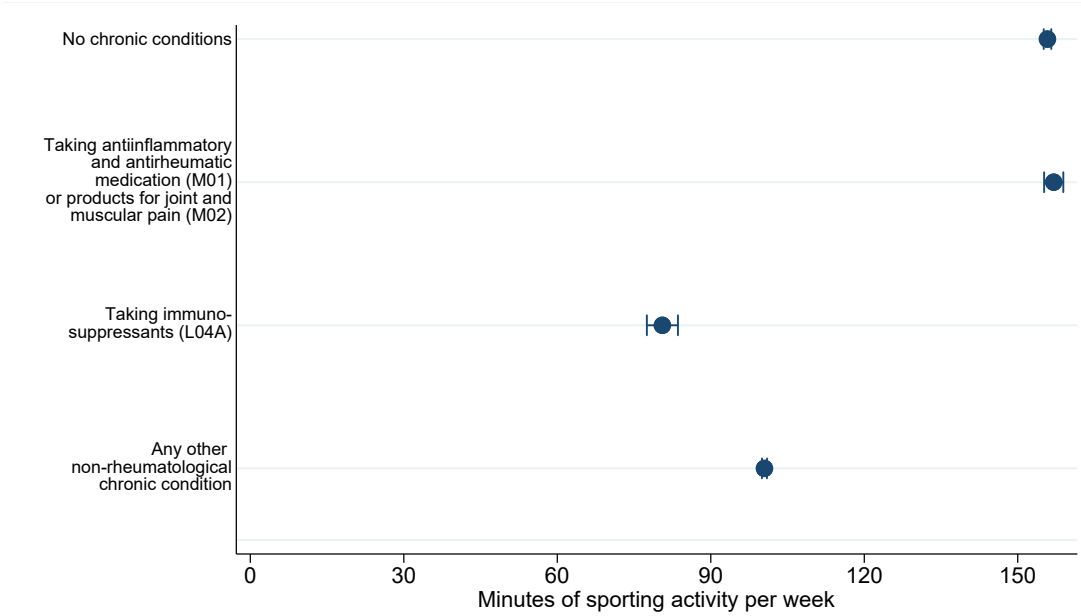
Note: This figure shows the number of Anemia cases per 1000 for men and women between 2011-2021. Furthermore, it shows how the evolution of anemia for women changes when restricted to 2 or 3 consecutive years of anemia.

Figure D.2: MEDICINES FOR RHEUMATOLOGICAL CONDITIONS

A. Share of Prescribed Medicines by Age



B. Sporting Activity by Medication Groups



Note: Panel A shows the share of individuals taking different types of medicines for Rheumatic conditions in 2012. Panel B displays the minutes of sporting activity of different groups. L04A: immunosuppressants excl. corticosteroids. M01: Anti-inflammatory and antirheumatic products. M02: Topical products for joint and muscular pain. Both M01 and M02 were included in Huber et al. (2013), but excluded in our analysis.

E Chronic disease and the mortality gap

Figure 5 and Appendix Figure C.3 estimate the income gap in five-year mortality and total healthcare costs, respectively. Panel A of both Figures reports a coefficient plot which shows the coefficient on a "low income" dummy from regressions of the relevant outcome (mortality or costs) on that dummy and differing sets of controls. The resulting gradient enables us to assess both the raw gap and how much of that gap is captured by other related factors. The dummy takes value one if an individual's income is below the median, zero if it is below.

Both Panel A figures report results from the same specifications. Specification A1 (Baseline) regresses the relevant outcome on the low income dummy, age dummies, and gender. All independent variables are fully interacted with gender in all the specifications. Specification A2 (Chronic conditions (lag 1)) also controls for the prevalence of chronic conditions in the previous year. Specification B1 (Chronic conditions (lags 2-3) adds 2-year and 3-year lags of chronic condition prevalence to the set of independent variables of A2. Specification B2, instead, adds to A2 indicators for the use of non-chronic condition-related medicines. The set of medicines is the union of those selected with separate LASSOs for women and men with the dependent variable being five-year mortality. For computational reasons, the LASSO penalization parameters are chosen to select twenty medicine indicators for each gender-specific regression. Most of the selected medicines are common among the genders, resulting in a union set of 24 medicines. The coefficients from the LASSOs are reported in Appendix Figure E.1.

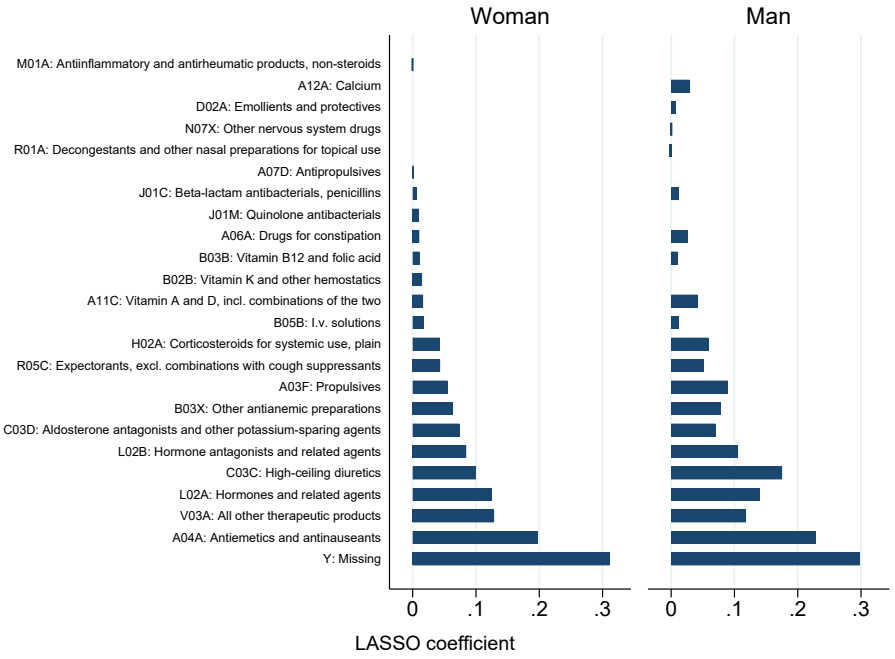
Specification C1 (Hospitalizations) adds to A2 information on hospitalizations: one-year lagged ICD codes (International Classification of Diseases); fourth degree polynomials of the number of previous-year hospitalisations, of the number of previous-year hospitalized nights, and of previous-year hospitalization costs. Specification C2 (Healthcare costs, by type) adds to A2 fourth degree polynomials of previous-year GP costs, medicine costs, mental health costs, and a miscellaneous other healthcare costs variable. Finally, Specification D (All health information (A1-C2)) regresses the relevant outcome on the low income dummy controlling for the union of all the independent variables employed in all the previous specifications.

Appendix Figure C.4 reports variations of Figure 5A and Appendix Figure C.3A where specifications B1, B2, C1, and C2 do not control for the first lag of chronic conditions. The coefficients reported thus illustrate how each set of factors reported in the row label affects the estimated income gap in the outcome variable (5-year mortality or healthcare costs).

Appendix Figure C.5 reports Oaxaca-Blinder decompositions of five-year mortality and healthcare costs that separate the effects of differential chronic condition prevalence and differential treatment, for different age groups. Chronic conditions are measured with a one-year lag, so that the outcome in period t is regressed on chronic condition indicators in period $t-1$, at the individual level.

Appendix Figure E.2 uses chronic condition indicators from periods $t-3$, $t-2$, and $t-1$. Moreover, based on the distribution of chronic conditions in $t-1$, the most frequent twoway interactions between chronic conditions are retrieved and added to the Oaxaca-Blinder regression for all three lags of chronic conditions considered. Finally, Appendix Figure E.3 replicates Appendix Figure E.2 excluding the bottom decile of income, to assess the robustness of the results to potential differential underdiagnosis of chronic conditions across the income spectrum.

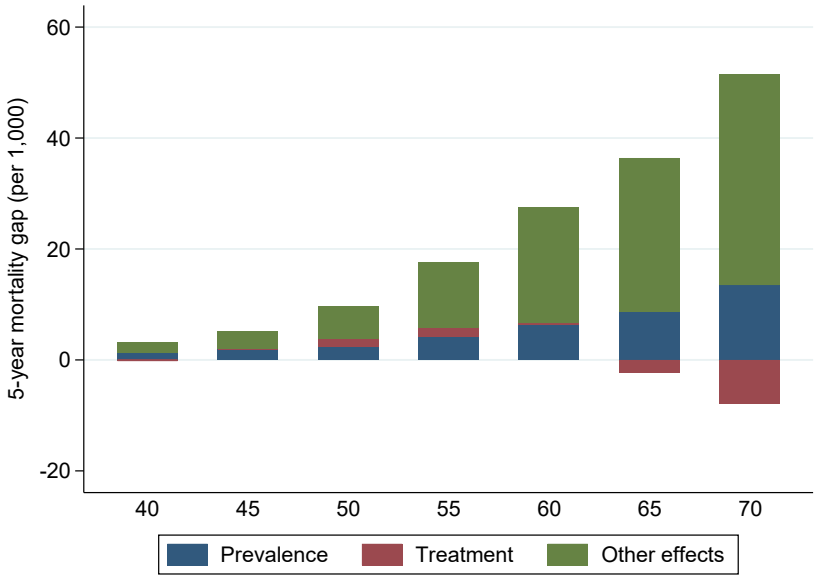
Figure E.1: COEFFICIENTS FROM LASSO REGRESSIONS OF FIVE-YEAR MORTALITY ON SELECTED MEDICINE INDICATORS



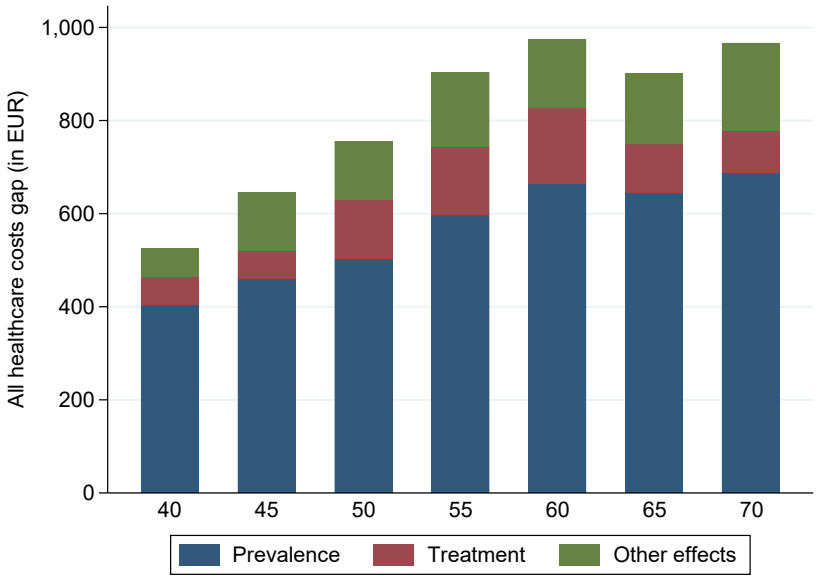
Note: This figure shows the coefficients of selected medicine indicators from gender-specific LASSO regressions of five-year mortality on the full set of non-chronic condition-related medicine indicators.

Figure E.2: OAXACA-BLINDER DECOMPOSITION, USING MORE LAGGED AND INTERACTED CHRONIC CONDITIONS

A. Five-Year Mortality



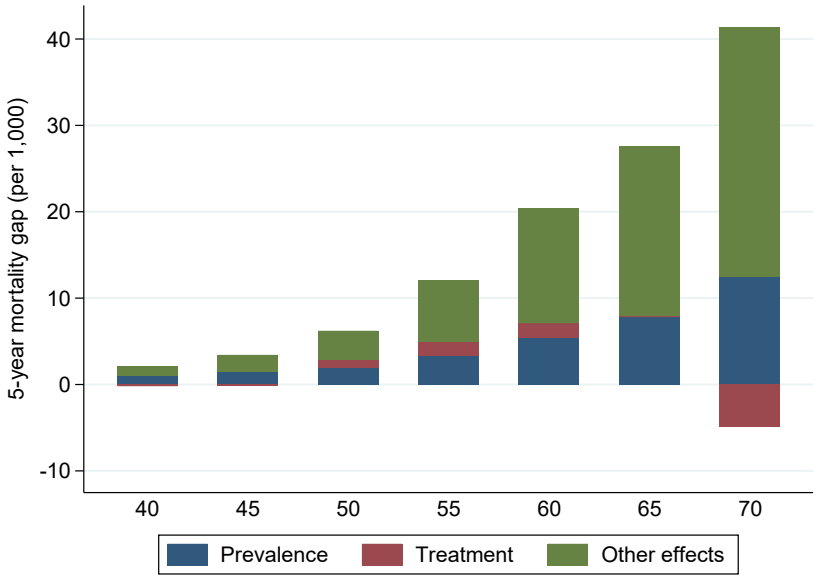
B. Healthcare Costs



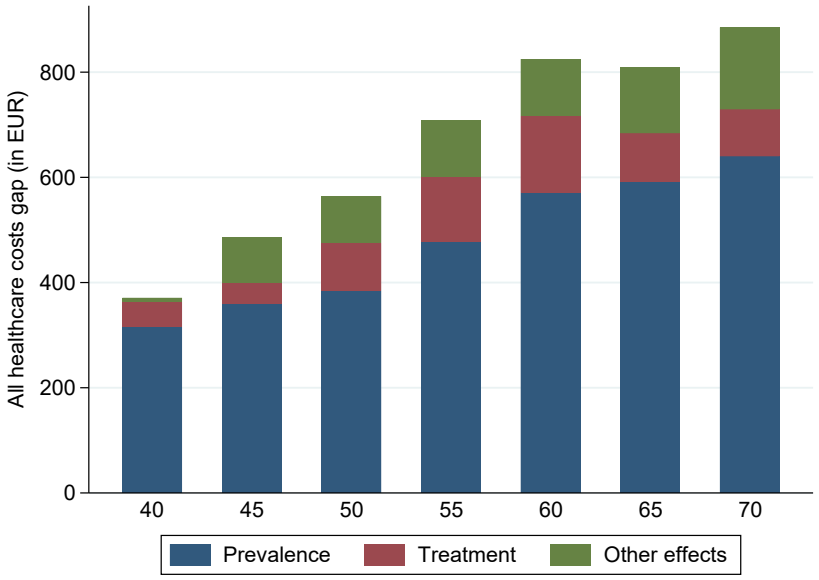
Note: The figure reports the results of a threeway Oaxaca-Blinder decomposition of 5-year mortality (in panel A) and of total healthcare costs (in panel B), using as predictors lagged chronic condition indicators from the previous three years. The ten most frequent within-period chronic condition interactions are also included. The two groups considered are low-income and high-income individuals, using as threshold the median of the main income variable. The "Prevalence" component is given by the part of the difference in means explained by integroup difference in chronic condition endowments; the "Access / Treatment" component is given by the part explained by intergroup differences in coefficients, excluding the constant term; the "Other effects" component is given by the part explained by intergroup differences in the estimated constant term.

Figure E.3: OAXACA-BLINDER DECOMPOSITION, USING MORE LAGGED AND INTERACTED CHRONIC CONDITIONS, EXCLUDING THE BOTTOM DECILE OF INCOME

A. Five-year mortality



B. Healthcare costs



Note: The figure reports the results of a threeway Oaxaca-Blinder decomposition of 5-year mortality (in panel A) and of total healthcare costs (in panel B), using as predictors lagged chronic condition indicators from the previous three years. The ten most frequent within-period chronic condition interactions are also included. The two groups considered are low-income and high-income individuals, using as threshold the median of the main income variable and excluding the bottom income decile. The "Prevalence" component is given by the part of the difference in means explained by integroup difference in chronic condition endowments; the "Access / Treatment" component is given by the part explained by intergroup differences in coefficients, excluding the constant term; the "Other effects" component is given by the part explained by intergroup differences in the estimated constant term.

F Prediction Model Performance

Table F.1: Predictors Included in the CDI

Socioeconomic Status Predictors	
A. Individual Variables	
Gender	Household Composition*
Percentile of Household Disposable Income	Foreign Parents
Number of Household members with Income*	Household Main Source of Income*
Position in Household*	Work Status*
Percentile of Wealth Income**	Percentile of Household Assets**
Main source of household income*	Percentile of Household Savings
Calendar Year	House Owner*
Foreign Born	Percentile of Home value
Number of Household Members*	Percentile of Personal Primary Income
B. Interaction Terms	
Percentile of disposable Income x Main Income source	Percentile of Primary Income x House Owner
Percentile of Primary Income x Household Composition	Percentile of Personal Net Income x Work Status
Percentile of Personal Net Income x Percentile of Disposable Income	Percentile of Gross Income x Main Income Source
Percentile of Gross Income x Main Income Source	Percentile of Primary Income x Main Income Source
Percentile of Disposable Income x Percentile of Household Assets	Percentile of Disposable Income x Main source of Income
Percentile of Wealth Income x Percentile of Household Assets	Personal Net Income x Work Status
Percentile of Disposable Income x Percentile of Wealth Income	Percentile of Disposable Income x Main Income Source
Percentile of Personal Gross Income x Work Status	Percentile of Wealth Income x Percentile of Household Assets

Note: This tables presents the list of socioeconomic status variables included through the LASSO selection procedure. Variables for which multiple lags are included are denoted with $*$. Variables for which multiple lags and higher-order terms are included are indicated with $**$.

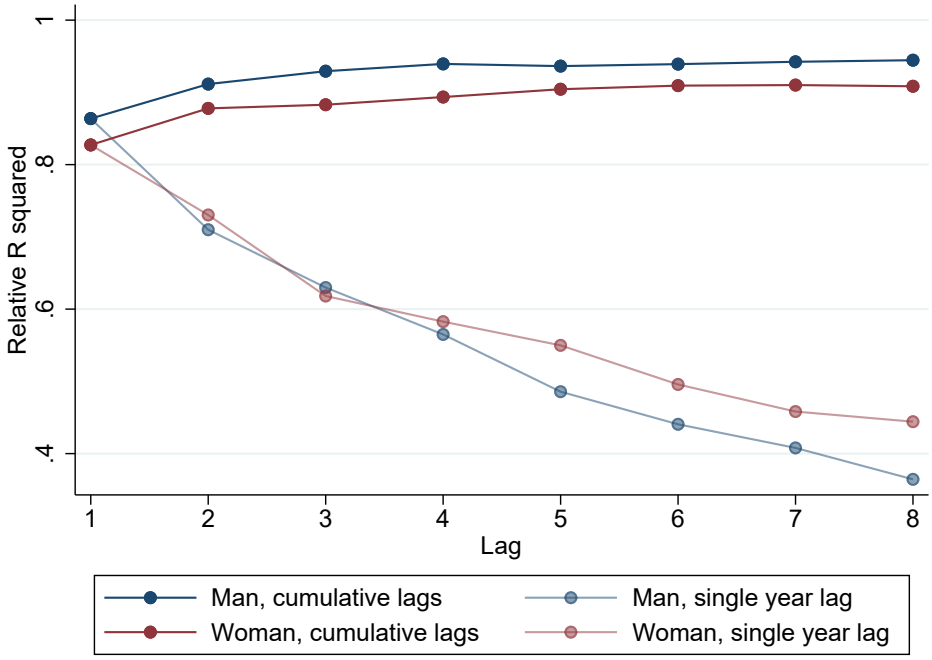
F.1 CDI model robustness

This section describes a series of considerations to understand the robustness or otherwise sensitivity of the CDI predictions to various modelling choices. Overall, the CDI is highly robust along these dimensions.

Lag Structure: One of the main modelling choices in constructing the CDI is selecting the lag sequence of chronic conditions. We choose the set $t = \{-1, -2, -3\}$: longer lags potentially contain more information, but would preclude certain cohorts from the sample. As shown in Appendix Figure F.1, the amount of additional information included the fourth and higher lags starts to taper off. This is despite those higher lags being highly predictive in themselves - due to the persistent nature of the conditions in question.

Linkage functions: For tractability, the CDI is estimated as a linear probability model. However, provided the separability is maintained between the set of chronic conditions and the socioeconomic variables, other linkage functions may be used, for example a logistic or Gompertz function, which naturally bound the prediction and have been used elsewhere in the literature. We have tested these two alternatives, but they do not yield any increases in predictive power,

Figure F.1: PREDICTIVE POWER OF CHRONIC CONDITIONS, BY LAG LENGTH



Note: This figure plots the predictive power of varying lags of chronic conditions on five-year mortality. Predictive power is measured as a relative R-squared statistic: $R^2_{s,t}/R^2_{(1,3),Int}$. The numerator $R^2_{s,t}$ is computed from a regression of five year mortality on the set of chronic condition lags between s and t, without interactions. $R^2_{(1,3),Int}$ is computed from our preferred specification, using lags 1-3 of chronic conditions with interactions.

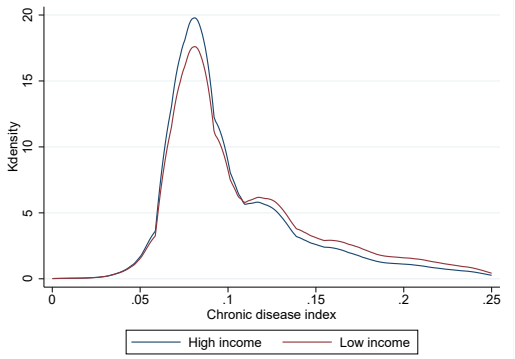
Table F.2: Summary statistics on alternative CDI specifications

	Baseline CDI	Estimated at 65	Estimated on positive medication subsample	Logistic regression
A. Model performance (test sample)				
Test R squared	0.052	0.035	0.056	0.052
Test AUC	0.663	0.654	0.679	0.666
Estimation sample size	402,500	554,519	353,643	402,500
B. Predicted CDI distribution at 70 (whole population)				
10th percentile	0.048	0.034	0.047	0.045
Median	0.077	0.051	0.077	0.073
90th percentile	0.172	0.114	0.171	0.157
C. Explained gradient at 70 (whole population)				
Explained 5-year mortality gap	0.297	0.192	0.293	0.273
Explained healthcare costs gap	0.555	0.492	0.554	0.541

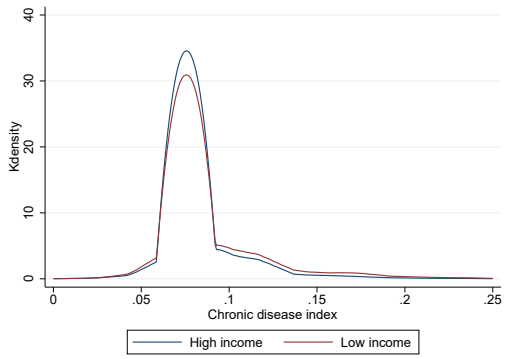
Note: The table displays three sets of statistics on the baseline CDI (the chronic disease index introduced in Section 4), and alternative versions estimated as robustness checks. One version was estimated using a sample of 65 year-olds (instead of 70 year-olds as is the case for the baseline CDI); another one used a sample limited to individuals reported to be taking at least one medication; finally, a version models the relationship between five-year mortality and its predictors using a logistic regression.

Figure F.2: DISTRIBUTION OF THE CDI BY INCOME GROUPS AND GENDER

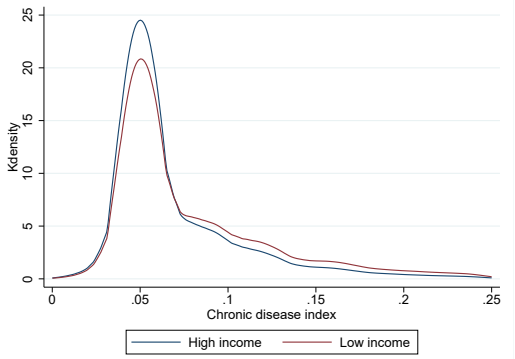
A. Men, at Age 70 (High v. Low Income)



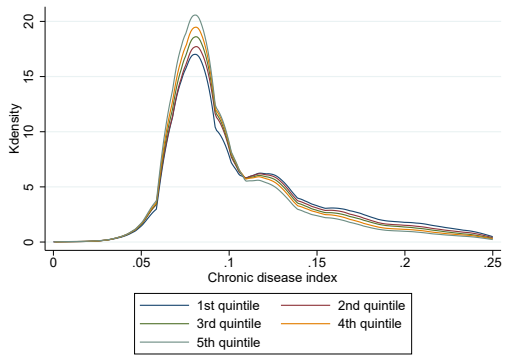
B. Men, at Age 40 (High v. Low Income)



C. Women, at Age 70 (High v. Low Income)



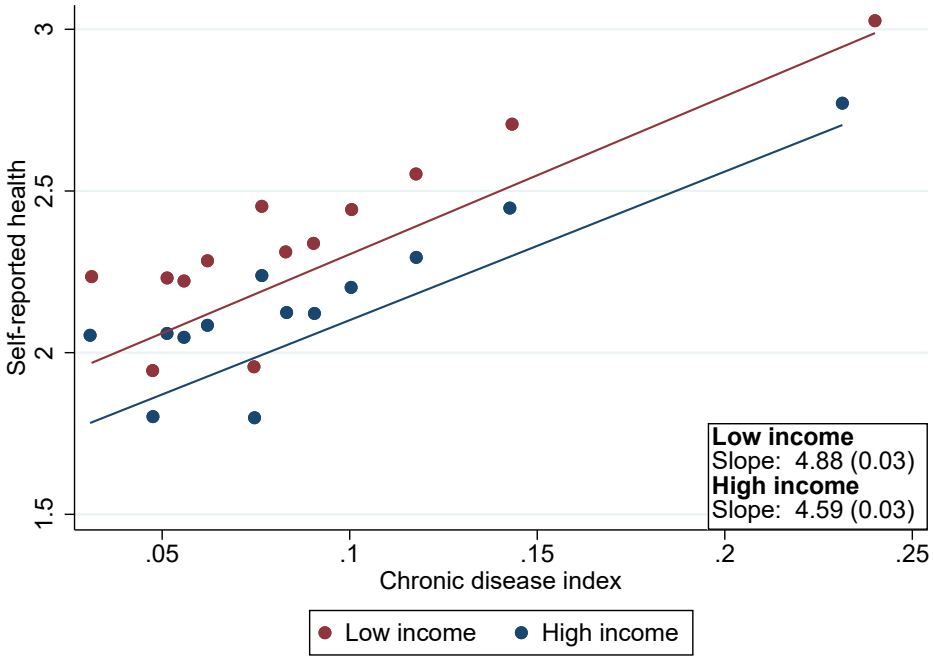
D. Men, at Age 70 (by Income Quintile)



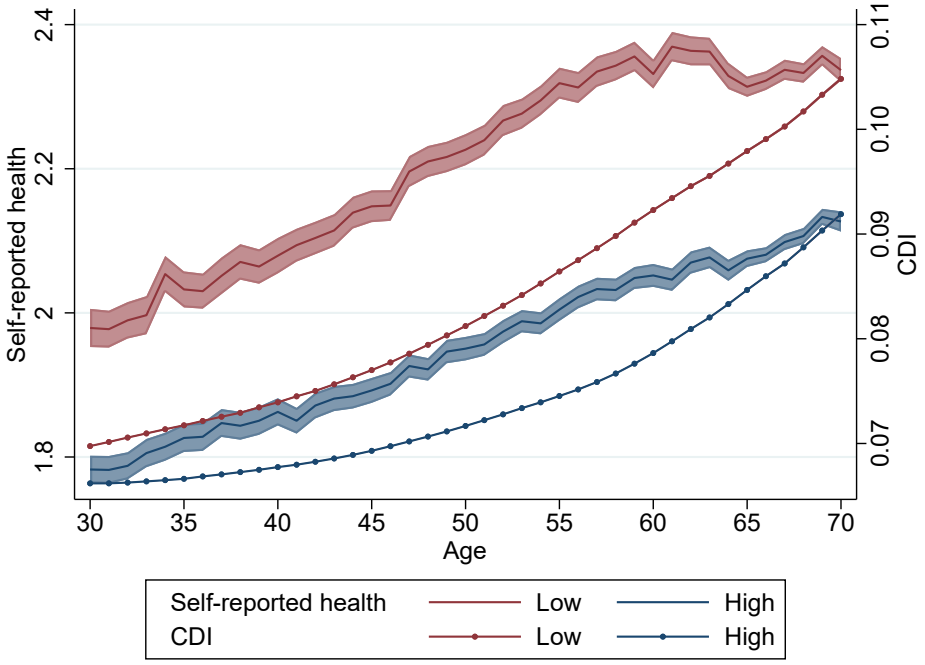
Note: The histograms report the kernel density of the chronic disease index at different ages and for different income splits for both men and women. The range of the x-axis is limited to the interval [0, 0.25] to avoid showing the low-density tails of the distribution, which are composed of outliers.

Figure F.3: DIAGNOSTIC BINNED SCATTERPLOTS OF SELF-REPORTED HEALTH

A. Self-reported health and CDI, by income



B. Self-reported health and CDI over the lifecycle



Note: Panel A shows a binned scatter plot showing the chronic disease index on the x-axis and average self-reported health as reported in the *Gezondheidsmonitor* survey data on the y-axis. A greater value denotes worse health. Panel B plots the evolution of self-reported health and the CDI over the lifecycle, along with 95% confidence intervals. The CDI series pool all observations in the period 2009-2021 and are identical to those reported in Figure 6. The CDI confidence interval is within the thickness of the connector line.

and given the specifications chosen by a LASSO the model is sufficiently regularised such that there are minimal predicted values outside the [0,1] bounds.

Linear model: As a benchmark, we construct $\widehat{M}_i^+$, including chronic conditions additively, without any interactions. This is shown in Equation (18) below. The CDI outperforms this benchmark in terms of R-squared (+10%), while the AUC statistic is a marginal improvement.

$$M_i^+ = CC_i^{*\prime}\beta_{CC}^+ + X_i^{*\prime}\beta_X^+ + \zeta_i^+ \tag{17}$$

$$\widehat{M}_i^+ = CC_i^{*\prime}\widehat{\beta}_{CC}^+ + \overline{X}_i^{*\prime}\widehat{\beta}_X^+ \tag{18}$$

Target ages: As described in Section 4.2, we select 70 to be the reference age for the CDI estimation. This balances a number of considerations: 70yo's are not overly positively selected in terms of survivorship bias, the under-diagnosis issue is not too acute, but the mortality risk is sufficiently high that the dependent variable has a meaningful amount of variation. We test alternative age ranges, including 65, and 65-75. These do not qualitatively change the CDI estimates, although the explained share of the health gap is slightly diminished.

Variable selection: The power of a LASSO framework is that modelling decisions on specification and functional form can be data-driven, rather than based on ad-hoc decisions. however, there are inevitably some decisions that could affect the estimation outcome. First, the candidate set of variables for the LASSO estimation: we could in principle choose any interaction set from the basis of variables documented in Section 2. In practice that would not be feasible given computation constraints. Since the set of relevant variables is less than half the candidate set, the risk of an error of omission from the candidate set is taken to be negligible. Second, the penalty parameter λ is chosen using the default `GLMNET` criteria: it provides the most regularized model such that the cross-validated error is within one standard error of the minimum mean squared error. An alternative criteria is that λ is chosen to minimise mean squared error, resulting in a greater set of relevant variables. Since this choice does not markedly alter the MSE of the the model, the CDI predictions, and subsequent findings will also not vary dramatically. Third, we conduct a separate LASSO estimation per chronic condition, to establish the set of relevant socioeconomic variables. Alternatively, we can conduct a group LASSO estimation exercise, as described in Yuan and Lin (2006). This would choose $f(C_i)^*$ in one step, similar to a 'seemingly unrelated regression' framework. In a theoretical paper, Obozinski, Wainwright, and Jordan (2011) show that if the dimensions of the CC_i are highly correlated, it is superior to use separate

LASSO steps for each, rather than combine all, akin to a variance inflation from multicollinearity argument. Given the degree of correlation between chronic conditions, this supports the decision to perform each LASSO step separately.

G Lifecycle Decomposition

In Figure 6, the average CDI for both income group is shown over the lifecycle. The slope of these curves, i.e. the difference in the CDI between two consecutive ages for a given income group, can be denoted as $E_{a+1}[CDI_{i,a+1}|Y_{a+1}] - E_{a-1}[CDI_{i,a}|Y_a]$, where Y_a denotes the set of individuals who belong to income group Y at age a. The subscript on the expectation operator indicates that we are taking expectations over those observed at age a.

We can decompose the slope of these curves into several terms:

$$\begin{aligned} E_{a+1}[CDI_{i,a+1}|Y_{a+1}] - E_a[CDI_{i,a}|Y_a] = & [E_{a+1}(CDI_{i,a+1}|Y_{a+1}) - E_{a+1,a}(CDI_{i,a+1}|Y_{a+1})] \\ & + [E_{a+1,a}(CDI_{i,a+1}|Y_{a+1}) - E_{a+1,a}(CDI_{i,a+1}|Y_a)] \\ & + [E_{a+1,a}(CDI_{i,a+1} - CDI_{i,a}|Y_a)] \\ & + [E_{a+1,a}(CDI_{i,a}|Y_a) - E_a(CDI_{i,a}|Y_a, S_{a+1})] \\ & + [E_a(CDI_{i,a}|Y_a, S_a + 1) - E_a(CDI_{i,a}|Y_a)] \end{aligned} \tag{19}$$

Below, we describe the interpretation for each of these terms.

1. **Aging**: for individuals in Y_a observed during both periods, we can calculate the average change in their outcome measure between a and $a+1$. We call this the Aging effect:

$$Aging = E_{a+1,a}(CDI_{i,a+1} - CDI_{i,a}|Y_a) \tag{20}$$

   Note that $E_{a+1,a}$ denotes the mean outcome for individuals who were in the sample both at age $a+1$ and a.

2. **Health-based Sorting**: over the lifecycle, people move between different income groups. Conditional on observing people in both periods, we will see two types of transitions: some people who were in Y_a will not be in Y_{a+1}, and some people who were not in Y_a will now be in Y_{a+1}. The (net) sorting effect is just the difference in mean outcome at age a between the members of Y_{a+1} and the members of Y_a:

$$Sorting = E_{a+1,a}(CDI_{i,a+1}|Y_{a+1}) - E_{a+1,a}(CDI_{i,a+1}|Y_a) \tag{21}$$

3. **Attrition due to Death**: some individuals who were in Y_a died at some point during that

year. Call the set of people who survived until age $a+1$, S_{a+1}. The attrition due to mortality is the difference in mean outcome at age a between those individuals in Y_a who survived until age $a+1$ and the mean outcome at age $a-1$.

$$Attrition = E_a(CDI_{i,a}|Y_a, S_{a+1}) - E_a(CDI_{i,a}|Y_a) \tag{22}$$

4. **Cohort Effect**: This is composed of exit and entry effects.

   First, some individuals who were in Y_a and survived into age $a+1$ are no longer in the sample at age $a+1$. This could be because they emigrated or because they aged out of the sample period . The exit effect is the difference in mean outcome at age a between those individuals in Y_a who stayed in the sample and all those who survived. (In other words, this is the expected CDI in a for all who left the sample for reasons other than death).

$$Exit = E_{a+1,a}(CDI_{i,a}|Y_a) - E_a(CDI_{i,a}|Y_a, S_{a+1}) \tag{23}$$

   Second, individuals who are in Y_a but were not in the sample at age $a-1$. This could be because they immigrated, were born or aged into the sample period. What we call "entry" effect is the difference in mean outcome at time a between the full set of individuals in Y_a and those who were also observed at time $a-1$ (In other words, this is the expected CDI in period t for all individuals who were not observed in $a-1$).

$$Entry = E_{a+1}(CDI_{i,a+1}|Y_{a+1}) - E_{a+1,a}(CDI_{i,a+1}|Y_{a+1}) \tag{24}$$

   The exit and entry effects are then combined into the so-called "Cohort Effects", which include includes both cohort, time and migration effects.

$$Cohort = Exit + Entry \tag{25}$$

In our main lifecycle decomposition, we estimate those effects for the low (below median) and high (above median) income group, pooling all observations in the period 2009-2021. The result of this decomposition is shown in Appendix Figure G.1. The aging effects increase steadily over the lifecycle for both income groups, while the sorting effects are most important around labor market entry and exit. Attrition due to death effects become relevant at later ages and

are stronger for the low income group, as low income individuals die at higher rates than high income individuals. Table 3 reports the difference between both panels of Appendix Figure G.1 for each effect.

The estimated effects can be used to simulate counterfactual CDI evolutions. Figure 7 applies aging effects only to simulate the CDI for both income groups. Appendix Figure G.2 instead also accounts for attrition due to death effects. Because those attrition effects are more strongly negative for the low income group, the gap in counterfactual CDI's is smaller when we account for them. Similarly, the biological age gaps in Panel B of Appendix Figure G.2 are somewhat smaller than those in Panel B of Figure 7.

Using this lifecycle decomposition framework, it is also possible to consider more groups based on income or other observable characteristics. Panel A of Appendix Figure G.3 shows cumulative aging and sorting effects when the decomposition is performed by income quintile. Aging effects are strongest for quintile 1 and decrease monotonically for higher income quintiles. Sorting effects, on the other hand, are positive for the first income quintile and negative for the four other quintiles. This shows that a substantial share individuals who fall ill (and see their CDI increasing) move into the lowest income quintile, which worsens the average health of this quintile. Table G.1 summarizes the aging effects for two alternative decompositions. The first uses income quintiles instead of the usual income split and shows that aging effects are monotonically increasing in income quintile for each bin. The second alternative decomposes the CDI using groups based on education level. The results show that more highly educated individuals 'age' slower than lower educated individuals at similar ages.

Furthermore, we also test our decomposition by imposing two robustness checks on the timing of our income variable. First, we restrict the sample to only consider individuals who have been in an age group for at least two years and then run the decomposition again. This means that at age a, only individuals who were in the same income group at a and at $a-1$ are included. Because those individuals are more fixedly in the relevant income group, we might obtain a more 'pure' aging effect. Under this so-called 'Markov Restriction', the aging effect can be written as:

$$Aging_{MR} = E_{a+1,a,a-1}(CDI_{i,a+1} - CDI_{i,a}|Y_a, Y_{a-1}) \tag{26}$$

And sorting is:

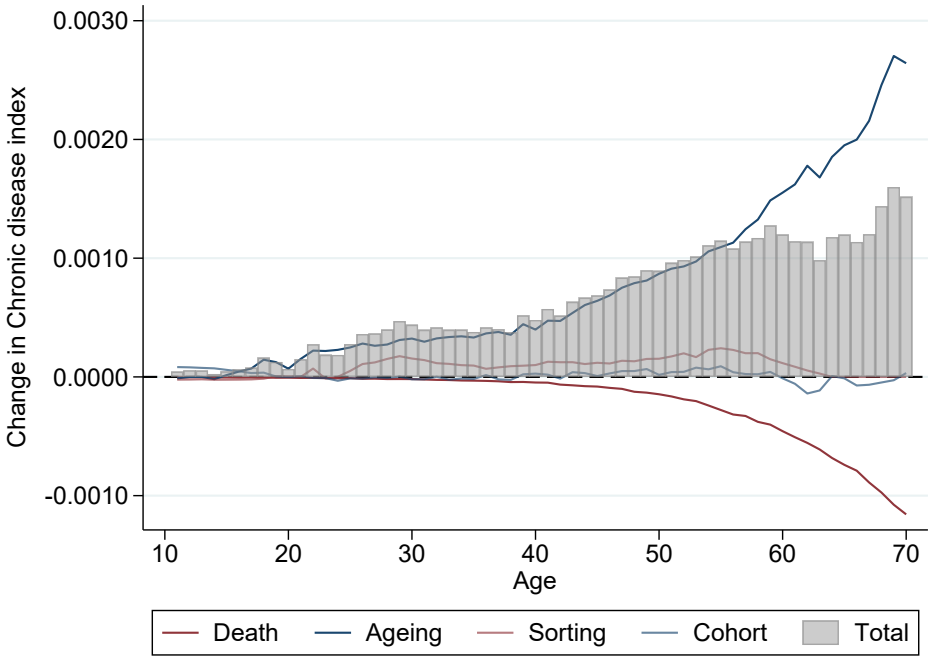
$$Sorting_{MR} = E_{a+1,a,a-1}(CDI_{i,a+1}|Y_{a+1}) - E_{a+1,a,a-1}(CDI_{i,a+1}|Y_a, Y_{a-1}) \tag{27}$$

In the second robustness check, we adapt our income definition and use a rolling average of Y_{a-3}, Y_{a-2} and Y_{a-1}. In this definition, we use income at the same ages for which chronic condition indicators are used to predict the CDI. Therefore, we call the second alternative 'Contemporaneous Income'. Panel B of Appendix Figure G.3 summarizes aging and sorting effects for both robustness checks, along with the original decomposition. The decomposition results are robust to different income definitions, as the obtained effects are very close to those of the original decomposition.

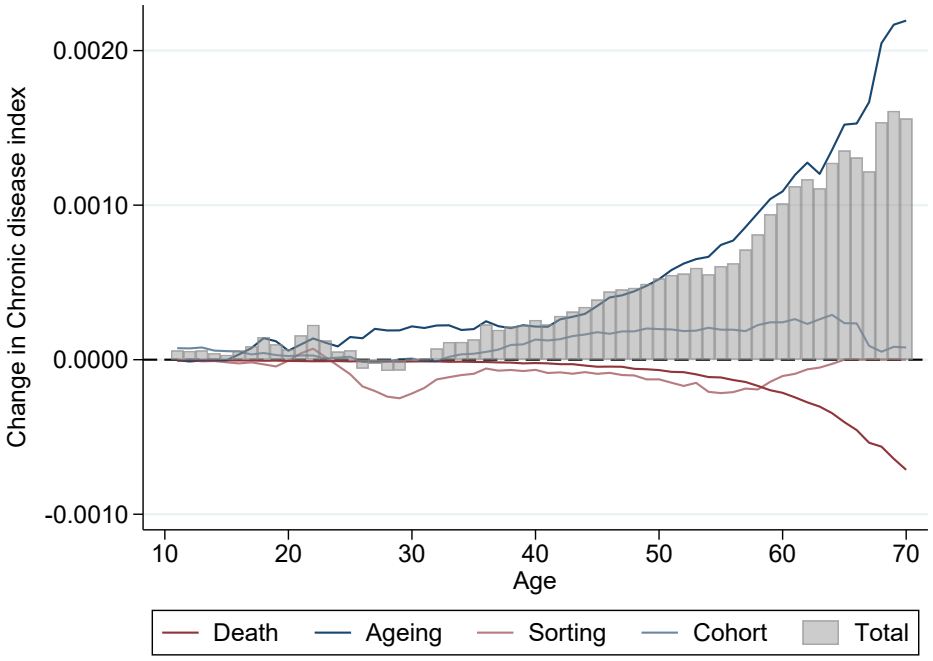
Lastly, we perform a robustness check on our estimated sorting effects. Sorting effects could be driven by individuals who move into a new household composition, affecting the household income in the process. To do so, we estimate the sorting effects separately for household with a changing composition and households which have the same composition. Table G.2 reports the results of this robustness checks, and shows that sorting is present for both changing and constant households. Furthermore, the sorting gap considering only non-changing households is very close to the gap sorting gap in our main decomposition, reported in Table 3. This provides evidence that the sorting effect is not driven by individuals who move into a new household.

Figure G.1: DECOMPOSITION BY INCOME GROUP

A. Low Income



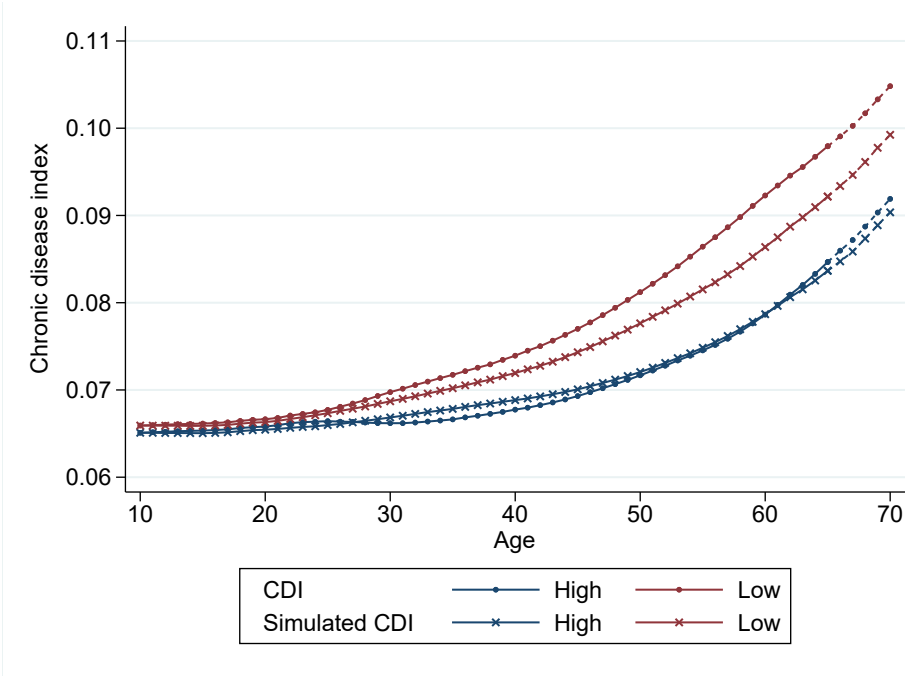
B. High Income



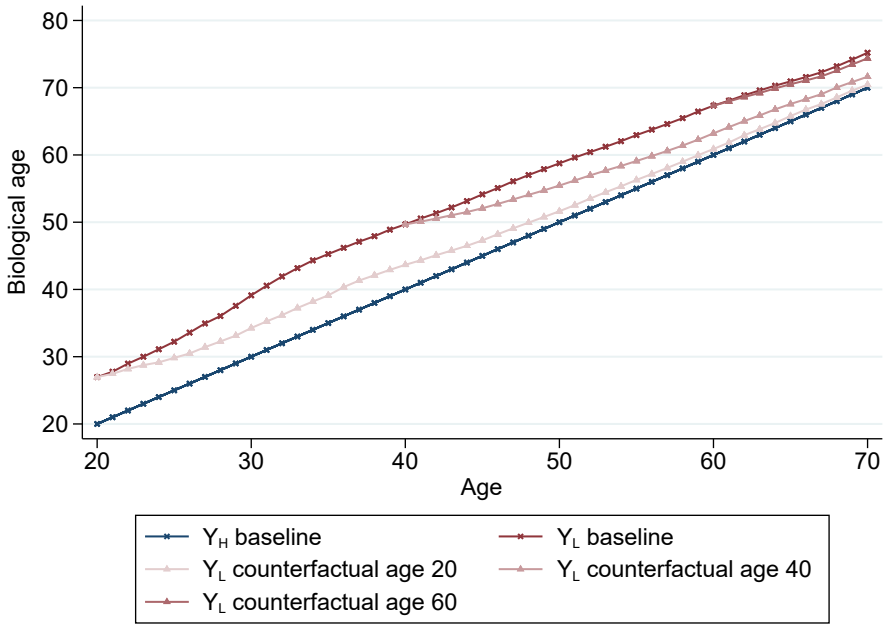
Note: This figure presents the full decomposition of the chronic disease index. Panel A shows the decomposition for the Low Income group, while panel B shows the High Income Decomposition. The total change between age a and $a-1$ is shown for both income groups, along with its decomposition into attrition due to death, aging, sorting and cohort effects. The decomposition pools all observations in the period 2009-2021. The gap between both income groups is summarized for each of the effects in Table 3.

Figure G.2: BIOLOGICAL AGING, ACCOUNTING FOR ATTRITION DUE TO DEATH

A. Differential aging and attrition



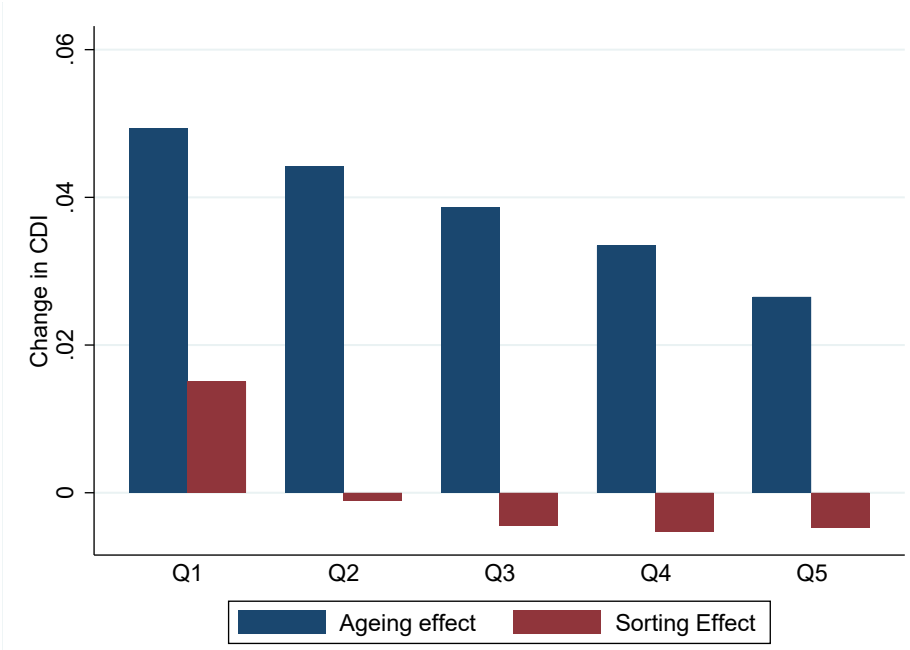
B. Counterfactual biological age



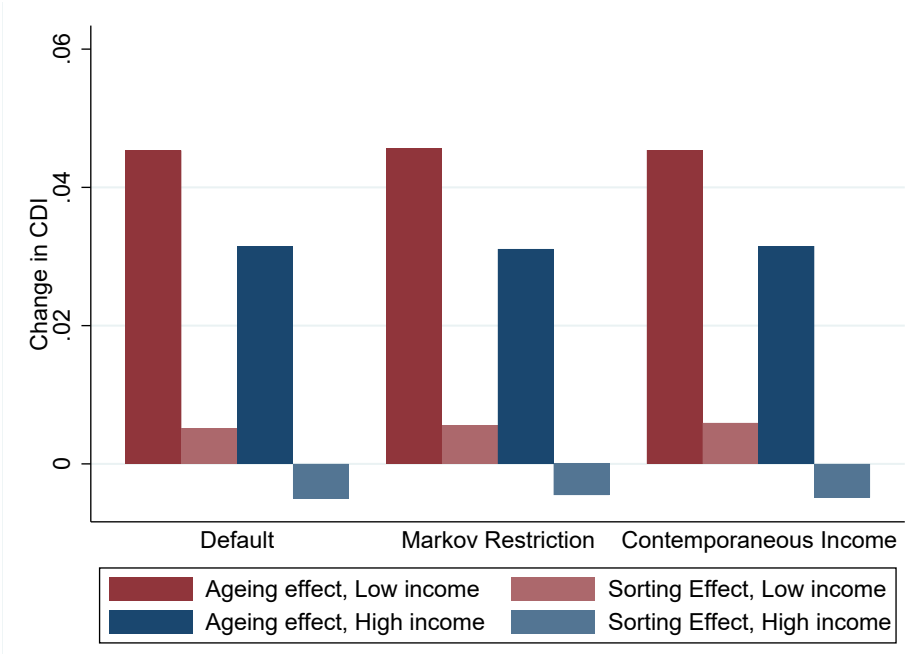
Note: Both panels are equivalent to Figure 7, but use both aging and attrition due to death effects. Panel A shows the evolution of the CDI by income group when it is simulated using aging and attrition due to death effects. Furthermore, 'observed' CDI's are shown, which are the same as those shown in Panel A of Figure 6. Panel B shows biological ages for different scenarios. In the baseline scenario, the high and low income CDI are simulated based on their respective estimated aging and attrition due to death effects effects. In the counterfactual scenario's, the high income aging and attrition effects are used to simulate the low income CDI from different ages onwards.

Figure G.3: AGGREGATE AGING AND SORTING EFFECTS

A. BY INCOME QUINTILE



B. ROBUSTNESS CHECKS



Note: Panel A reports the results of the lifecycle decomposition by income quintile. More specifically, the cumulative aging and sorting effects from ages 20 to 70 are shown for each income quintile. Panel B shows cumulative aging and sorting effects for two alternative decompositions. The Markov restriction robustness check only considers individuals at age a who were in the same income group at a and $a-1$ and repeats the composition for this selected sample of individuals. The contemporaneous income alternative uses the average of Y_{a-3}, Y_{a-2} and Y_{a-1} to rank individuals' incomes. This is the same range of ages for which we use chronic condition information in the CDI prediction model. All decompositions pool all observations in the period 2009-2021.

Table G.1: Aging effect by income quintile and education level, x 100

	11-20	21-30	31-40	41-50	51-60	61-70	Life-cycle Effect
A.By Income Quintile							
Q1	0.057	0.248	0.425	0.785	1.270	2.196	4.982
Q2	0.038	0.266	0.323	0.623	1.143	2.061	4.454
Q3	0.032	0.215	0.280	0.487	0.988	1.889	3.891
Q4	0.030	0.165	0.227	0.407	0.859	1.685	3.373
Q5	0.052	0.114	0.177	0.272	0.659	1.424	2.698
B. By Education level							
No High School	-	0.374	0.681	1.145	1.472	2.233	5.911
High School	-	0.269	0.421	0.711	1.185	1.926	4.511
Bachelor	-	0.094	0.182	0.372	0.779	1.563	2.990
Master or PhD	-	0.0809	0.1529	0.260	0.6239	1.4294	2.548

Note: This table reports aging effects for 5 income quintiles and 4 education groups separately. Effects are reported as the total contribution to the change in CDI of the aging effect for 10-year age groups, multiplied by 100. More detail on the methodology used to perform the lifecycle decomposition is provided in Appendix Section G.

Table G.2: Sorting effect by household composition, x 100

	11-20	21-30	31-40	41-50	51-60	61-70	Life-cycle Aggregate
1. High income							
Ageing	0.04	0.15	0.22	0.37	0.80	1.62	3.18
Sorting, new household composition	-0.07	-0.16	-0.05	-0.10	-0.16	-0.01	-0.55
Sorting, constant household composition	-0.02	-0.11	-0.11	-0.10	-0.18	-0.03	-0.54
2. Low income							
Ageing	0.04	0.25	0.36	0.66	1.17	2.08	4.57
Sorting, new household composition	-0.12	0.16	0.26	0.14	0.36	0.18	0.98
Sorting, constant household composition	-0.01	0.07	0.09	0.13	0.17	0.01	0.46
3. Gap							
Ageing	0.00	0.10	0.14	0.30	0.37	0.47	1.39
Sorting, new household composition	0.05	0.32	0.30	0.24	0.52	0.19	1.52
Sorting, constant household composition	0.01	0.18	0.20	0.23	0.35	0.03	0.99

Note: The table reports the contribution towards the Chronic Disease Index for the aging and sorting effects. The sorting effects are estimated separately for households which change their composition between $a+1$ and a and those which stay the same. The effects are expressed as the change in the CDI for 10-year age bins, multiplied by 100. That is, the numbers in the table are expressed as percentage points change in the CDI. More detail on the lifecycle decomposition is provided in Appendix Section G.

H Life Expectancy and Lifetime Costs

H.1 Life Expectancy

In Section 6, we perform a counterfactual analysis which calculates a range of counterfactual life expectancy estimations. In this appendix, we explain the methodology lying behind those calculations.

We observe income-specific mortality rates until age 78. Therefore, we run the following age- and gender-specific regressions relating mortality to our Chronic Disease Index:

$$M_{i,a,Y} = \alpha_{a,Y} + \beta_a CDI_{i,a,Y} + \varepsilon_{i,a,Y} \tag{28}$$

where age $a \in [40, 78]$, $CDI_{i,a}$ is our index for individual i based on lagged chronic conditions and $h_{i,a,Y}$ is same-year mortality. Based on the estimation of these age-, gender- and income-specific coefficients, we predict same-year mortality for the observed average CDI by age, gender and income group.

To construct the counterfactuals shown in Table 4, we simulate alternative evolutions of the CDI based on the aging and attrition effects estimated in the lifecycle decomposition, explained in Appendix G. More specifically, we compute a baseline CDI simulation applying the aging and attrition due to death effect for the relevant income group from age 20. That is, we start from the observed CDI at age 20 for each income group and then let the CDI evolve according to the estimated aging and attrition due to death effects only. Then, we simulate different counterfactuals which let the low income CDI evolve at the aging rate of high income individuals from different ages (20, 40 & 60) onwards. Using these simulated CDI's, we predict same-year mortality rates using equation (28) for each baseline and counterfactual series of the CDI.

The above procedure yields income-specific mortality rates for each counterfactual until age 78 for each alternative. To estimate life expectancy figures, however, we need a full set of same-year mortality rates for group of interest j. We estimate the mortality rates at later ages as follows:

- For ages 79 to 90, we use a Gompertz extrapolation to predict mortality. That is, we linearly extrapolate log one-year mortality rates to estimate counterfactual-specific mortality rates between ages 79 and 90. This means that we estimate $\log \tilde{M}_{a,j} = b_{0,j} + b_{1,j}a$, where $\tilde{M}_{a,j}$ are

the one-year mortality rates estimated using equation (28) for ages 40-78. Then, we predict $\log \tilde{M}_{a,j}$ until age 90.

- For ages between 91 and 110, we rely on the (gender-specific) full-sample one-year mortality rates and set the hazard rate to 1 at age 110.

Once we have a full set of mortality rates, life expectancy at 40 is computed as:

$$\mathbb{E}[A|A \geq 40] \approx \sum_{a=40}^{110} Pr(A = a|A \geq 40) \cdot a \tag{29}$$

where A is age at death. Using this framework, we first compute baseline life expectancy for both income groups, based on the observed CDI averages.

The resulting life expectancy at age 40 is reported in Table 4. Panel A of Appendix Figure H.1 visually shows the survival probabilities for the high and low income baseline, and the counterfactual applying high income aging effects from age 20 onwards.

H.2 Lifetime Costs

Apart from life expectancy estimations, we also calculate counterfactual lifetime healthcare costs. First, we start with the following version of equation (28) :

$$k_{i,a,Y} = \gamma_{a,Y} + \delta_a CDI_{i,a,Y} + u_{i,a,Y} \tag{30}$$

Where age $a \in [40, 70]$ and $k_{i,a,Y}$ is logged, detrended healthcare costs for individual i who belongs to income group Y at age a.

Then, we use the same counterfactual CDI evolutions described in Section H.1 above to estimate healthcare costs at each age between 40 and 70. Again, a baseline CDI evolution for each income group using the respective aging and attrition due to death effects are used, along with counterfactuals which apply high income aging and death effects to the low income CDI from age 20, 40 and 60 onwards.

We then estimate cost at later ages as follows:

- Between 71 and 90, we calculate yearly costs in two steps. First, we impose the empirical high- and low-income costs to grow at the same rate as the full population costs. Then, we

compute a weighted average of both, with linearly increasing wights on the full population costs. This procedure is visually represented in Panel B of Appendix Figure H.1 for the high and low income baselines.

- Between 91 and 110, we revert to the overall cost rates $\bar{k}^a$ (not income specific), computed on the full sample, for all sets of costs.

Once we have a full set of mortality rates, life-time expected cost at age 40 is computed as:

$$\mathbb{E}\left[\sum_{a=40}^{\infty} K_a \mid A \geq 40\right] \approx \sum_{a=40}^{110} S_a \cdot K_a \tag{31}$$

Where A is age at death and S_a is the survival probability, $S_a := Pr(A \geq a|A \geq 40) = S(a|A \geq 40)$.

We use two alternative approaches with respect to the survival probabilities in our counterfactuals. In the first approach, we compute counterfactual healthcare costs when intervening at age 20, 40 or 60, but use the baseline low income survival probabilities computed in the life expectancy calculations. That is, we allow costs to be affected but not survival probabilities by the hypothetical intervention. This approach corresponds to row 3.a in Table 4.

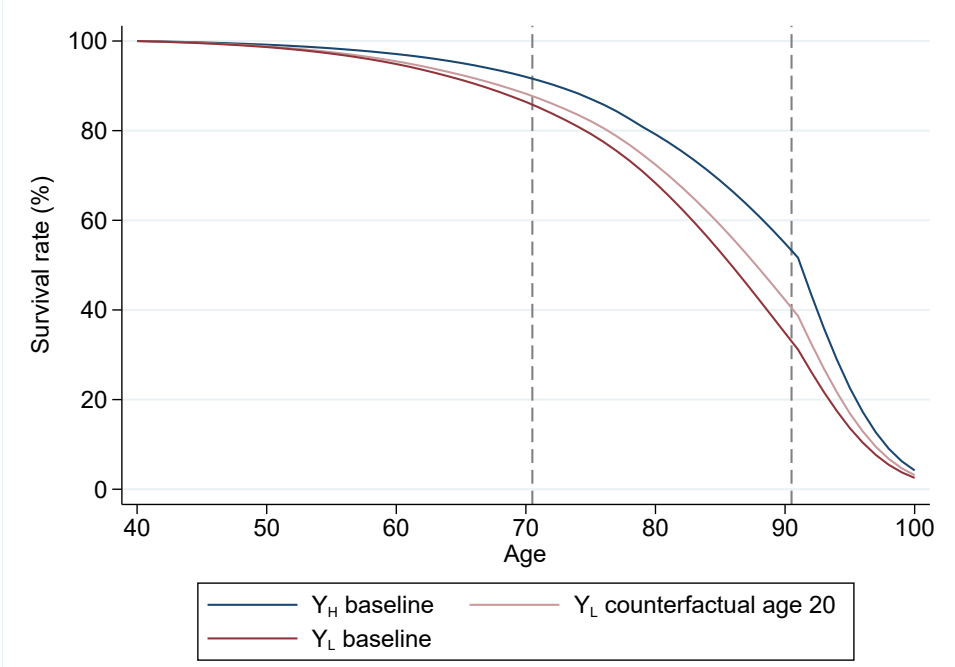
In the second approach, survival probabilities are adjsuted for each counterfactual. More specifically, they are taken from the corresponding alternative estimated in the life expectancy calculations, using equation (29). That is, we allow both costs and survival probabilities to be affected by the hypothetical intervention. This approach corresponds to row 3.b in Table 4.

H.3 Alternative Estimates

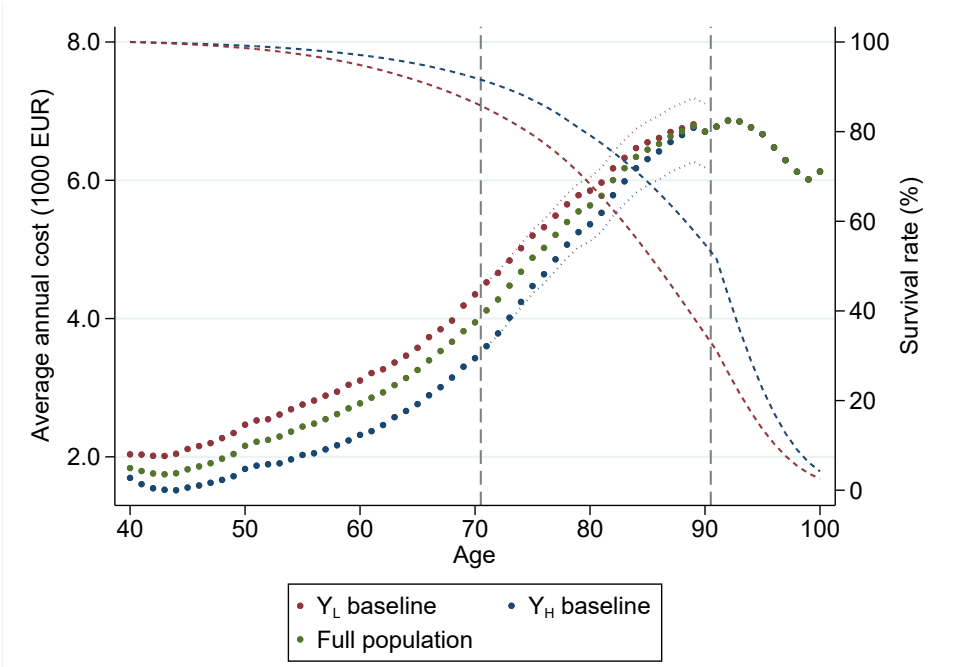
In our counterfactual analysis, we use simulated CDI's based on aging effects. We can also apply the methodology described in Sections H.1 and H.2 to estimate life expectancy and expected lifetime costs using average CDI's for both income groups. Table H.1 shows the resulting estimates when we use average CDI's for low and high income, and how those estimates change when we assign high income CDI's or survival rates to the low income group.

Figure H.1: SURVIVAL RATES AND COSTS OVER THE LIFECYCLE

A. Survival curves, women



B. Weighted costs, women



Note: This figure illustrates the procedure used for our life expectancy and lifetime costs estimations. Both panels show the procedure for women. Panel A displays the survival rates in the baseline scenario for both income groups. Furthermore, the counterfactual scenario where high income aging effects are applied from age 20 is shown. Between ages 40 and 78, one-year mortality rates are observed for both income groups. Between ages 78 and 90, a Gompertz extrapolation is performed to estimate one-year mortality rates. Between ages 91 and 100, full sample mortality rates are assigned to each group. Panel B shows average costs healthcare costs over the lifecycle. Between ages 40 and 70, annual costs are observed by income group. Between ages 71 and 90, income-specific healthcare costs are imposed to grow at the same rate as full sample healthcare costs. Then, a weighted sum of the income-specific and full-sample costs is applied, with linearly increasing weights on the full-sample costs. Above age 90, full-sample costs are applied to all individuals.

Table H.1: Estimates using average CDI's

	High Income	Low Income		
	Baseline	Baseline	Y_H Survival	Y_H CDI
1. Life Expectancy	85.2	80.8	85.2	84.3
2. Lifetime Costs	159.0 k	163.9k	171.4k	147.8k

Note: This table shows additional life expectancy and lifetime cost estimations. The first two columns use CDI averages by age to estimate life expectancy and expected life time costs. The third column assigns the observed high income survival rates to the low income group. Column 4 assigns high income CDI averages to the low income groups. Each alternative estimate applies the methodology described in Appendix H to estimate costs and surival rates at higher ages.

I Mediators Analysis

Figure 8 reports the Shapley-Owen values for regression equation (16), separately for each 10-year age bin from 20-29 to 60-69 years of age. The dependent variable is the within-individual five-year growth of the log of the CDI from 2013 to 2018. Using 2013 as the base year allows to use previous-year lagged behavioural predictors from the 2012 wave of the GEMON survey.

To allow for enough flexibility, all predictors (listed in Section 6.2) are treated as binary indicators (e.g., there are 100 binary indicators corresponding to income percentiles, one of which is omitted). In addition to the predictor groups already listed and shown in the legend, which are considered for the Shapley-Owen decomposition, each specification controls for age and gender indicators.

Some of the variables used in the decomposition have poor coverage. It is the case, as discussed in Section 2, of those related to education. The same holds for the parental chronic disease index, as many parents are not observed; and for sector and pay rank, in particular at older ages, as they are not defined after retirement. Table I.1 reports summary counts of variable coverage for the Shapley-Owen decompositions reported in Figure 8.

In Panel A of Appendix Figure C.11, instead, the dependent variable is the projection on income percentiles and gender of the within-individual five-year growth of the log of the CDI. In Panel B, the dependent variable is the log of the CDI. Again, in both cases, the base year used is 2013.

Figure 9 reports selected coefficients from a number of different linear regressions. In specification "Baseline", the outcome is the within-individual five-year difference in the CDI. In addition to the dependent variables shown in Figure (smoking, alcohol consumption, sport, Body Mass

Index, maternal and paternal health, municipality, working status, sector, and pay rank), the specification controls for age and gender indicators, as well as for percentile indicators of income, wealth, parental income, and parental wealth; the education attained and the field; the position in the household, the household composition; indicators for being foreign and for having foreign parents. In specification "Levels", the dependent variable is the same-year CDI, and the same independent variables are used. Finally, specification "Partial" aggregates the results of several regressions, whose dependent variable is the same-year CDI. Each regression has as independent variable one of the factors shown in Figure (e.g., smoking, alcohol, etc.) and controls for age and gender indicators. Appendix Figures I.1 and I.2 report the results of specification "Baseline" for specific subgroups of the population.

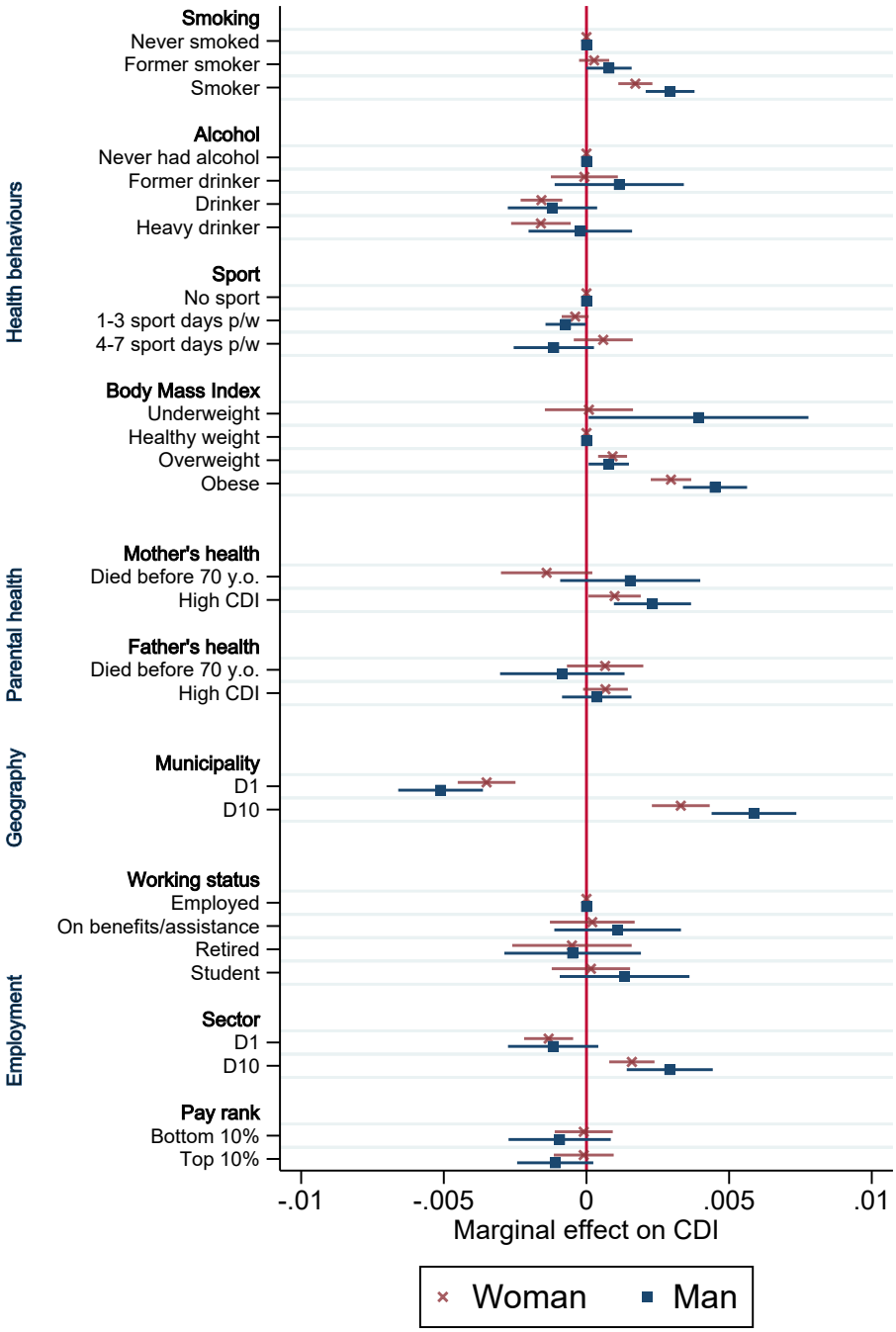
The Figures report coefficients for sector and municipality aggregated into deciles, the top and bottom of which are shown. This results from an ex-post categorization, conducted as follows. The regressions use separate indicators for each municipality and sector, which are then aggregated into deciles based on the cumulative distribution of their effects on the dependent variable, weighting each sector and municipality by its respective population. The coefficients reported are the average of those for the sectors and municipalities in a given decile, weighting for the population, and subtracting the weighted average of the coefficients around the median (in percentiles 45 to 55 of the effect size).

Table I.1: COVERAGE OF THE SAMPLE USED FOR THE SHAPLEY-OWEN DECOMPOSITIONS

	20-29	30-39	40-49	50-59	60-69
Observations					
Observations	31,429	35,043	50,694	58,356	79,419
Sample used	25,654	28,938	41,533	47,565	63,572
Used, no filling in	15,314	11,775	8,299	2,285	122
Filled in values					
Education level	529	7,031	17,776	27,113	46,518
Education field	1,005	7,697	18,655	28,117	47,503
Foreign parents	2,195	3,111	3,793	3,541	3,278
Maternal CDI	1,227	3,783	9,793	24,033	54,998
Paternal CDI	2,412	5,538	16,472	35,307	61,473
Sector	2,897	4,338	8,357	12,424	46,945
Pay rank vigintile	6,913	8,345	14,170	17,546	50,430

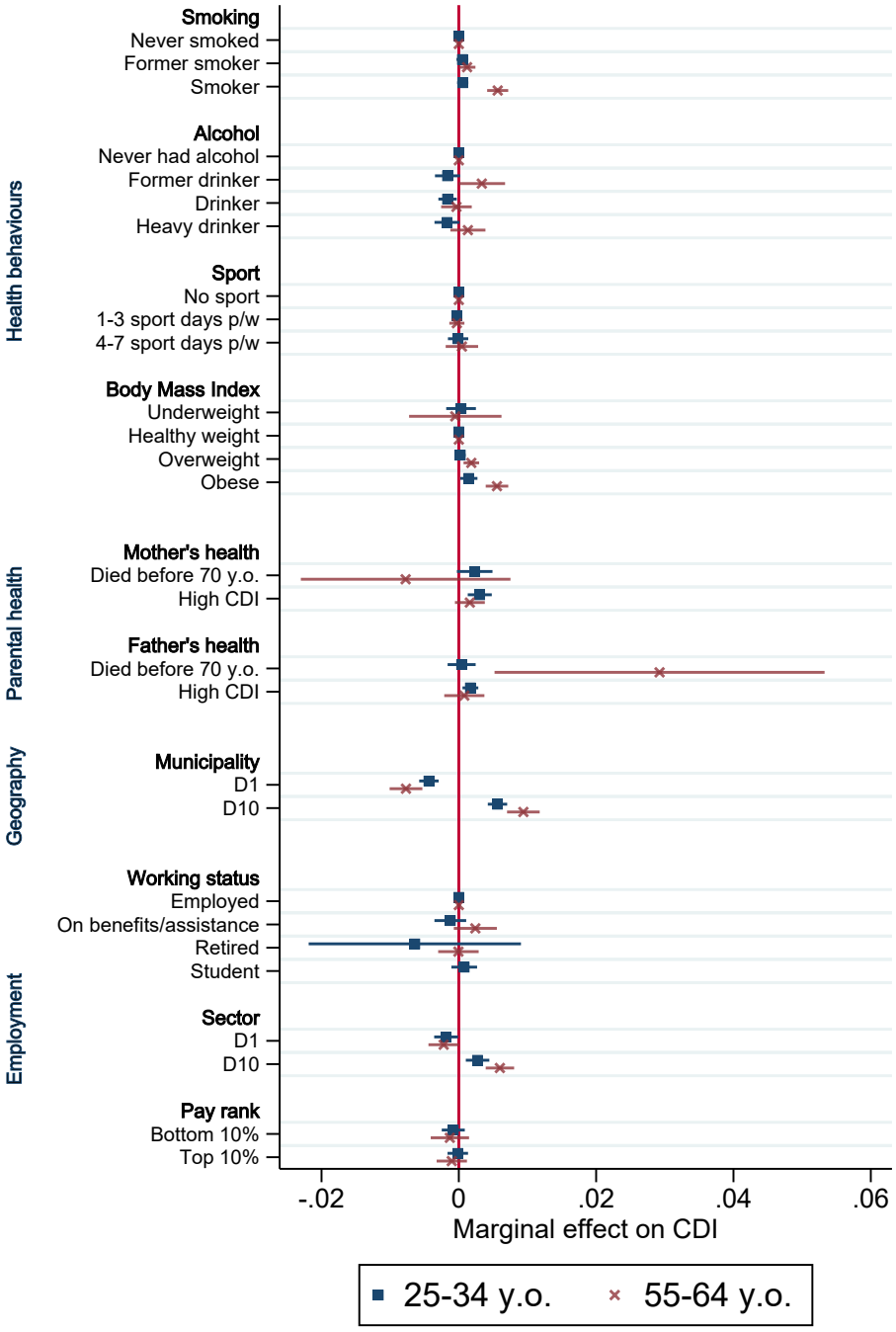
Note: Row "Observations" report the number of observations of the dependent variable (the within-individual five-year CDI growth) in 2013 for the age group in column. Row "Sample used" reports the number of observations actually used in the regressions. Row "Used, no filling in" reports the number of observations which were not supplemented by a "missing value" indicator to avoid dropping variables, out of those used in the regressions. Part "Filled in values" shows how many values were "filled in" for each of the variables that required it, in each specification.

Figure I.1: MEDIATORS OF THE CDI, BY GENDER



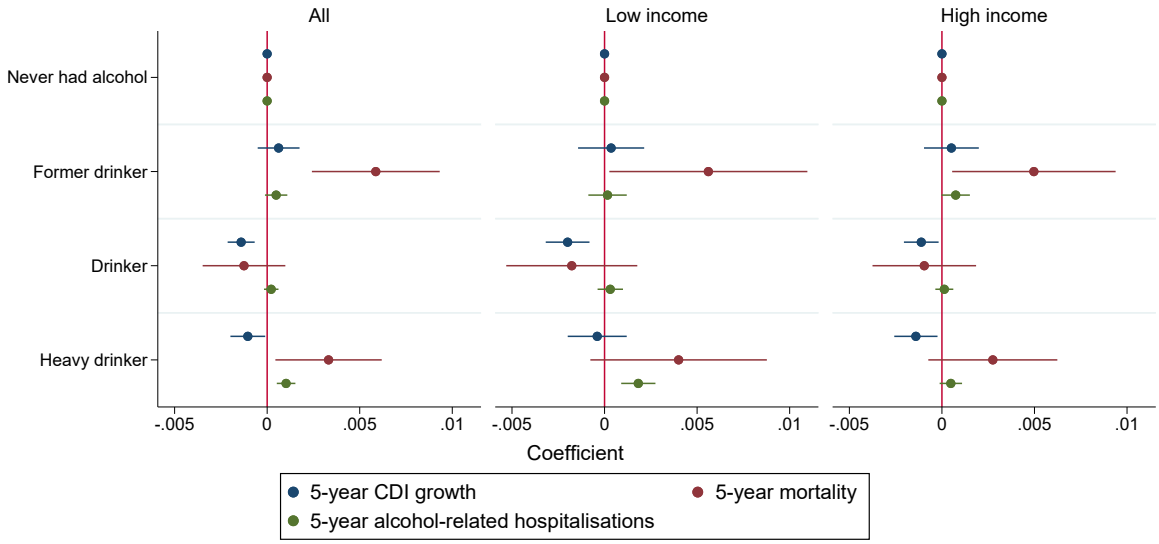
Note: This figure reports coefficients and confidence intervals from regressions of the CDI on mediators, separately by gender. Both gender-specific regressions use specification "Baseline" from Figure 9.

Figure I.2: MEDIATORS OF THE CDI, BY AGE



Note: This figure reports coefficients and confidence intervals from regressions of the CDI on mediators, separately for individuals aged 25-34 and 55-64. Both age-specific regressions use specification "Baseline" from Figure 9.

Figure I.3: CDI GROWTH, MORTALITY AND HOSPITALISATION RISK BY ALCOHOL CONSUMPTION



Note: This figure reports coefficients and confidence intervals from regressions of surveyed alcohol consumption on CDI growth, 5-year all-cause mortality, and hospitalisation due to alcohol-related liver disease, or other alcohol-related disorders. All regressions use the same set of controls as "Baseline" from Figure 9.