The Chronic Disease Index: Analyzing Health Inequalities Over the Lifecycle[*]

Kaveh Danesh[†] **Jonathan Kolstad**[‡] **William Parker**[§]
Johannes Spinnewijn[¶]

October 16, 2025

Abstract

The rich live longer than the poor, but how and when health differences arise over the course of life remains less understood. Leveraging rich administrative data from the Netherlands, we link chronic disease profiles to mortality risk at old-age to construct a health index that is comprehensive, measured repeatedly and at scale. Our index allows us to study the dynamics of health inequality, which translate into mortality differences later in life. We find that about 50% of the health gap between income groups at age 70 has already materialized at age 40. Approximately 60% of the gap is due to low-income individuals developing chronic illness at a faster rate, rather than chronically ill individuals sorting into lower-income groups. We also examine the contributions of a wide range of mediators to the onset of chronic diseases and find that socioeconomic factors play a predominant role.

[*] We thank Rob Alessie, Sandy Black, Zach Bleemer, Raj Chetty, Janet Currie, Liran Einav, Amy Finkelstein, John Friedman, Ingrid Huitfeldt, Henrik Kleven, Camille Landais, Adriana Lleras-Muney, Ellen Meara, Canishk Naik, Carol Propper, Hannes Schwandt, Bram Spinnewijn, Michael Stepner, Frank van Dijk, Winnie Van Dijk and seminar participants at the LSE, IFS, University of Oslo, Bank of Spain, BI Norwegian Business School, CPB Netherlands, Bocconi, Bologna, EIEF, Zurich, Oxford, Royal Holloway, London Business School, NHH Bergen, LSE Health, KULeuven, Tinbergen Amsterdam, Opportunity Insights, Columbia, Princeton, Yale, NYU, and at the CEPR Public Economics symposium, the NBER Aging meeting, the Dutch health econometrics workshop, the lolaHESG conference, and the RFBerlin Economics of Aging 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, Rasmus Duret, Diego Ferreras-Garrucho, and Marco Visentin for excellent research assistance. We thankfully acknowledge financial support from the International Inequalities Institute.

[†] UC Berkeley and UCSF

[‡] UC Berkeley and NBER

[§] London School of Economics

[¶] London School of Economics and CEPR

1 Introduction

Inequality in health is a major source of socioeconomic disparities and a central policy challenge facing many countries. Efforts to 'close the health gap' are a key element of numerous policy agendas (e.g., World Health Organization (1985, 2008, 2017), as well as the EU's recent "Joint Action on Health Inequalities"). A large body of work in public health, epidemiology, sociology and economics has studied health inequalities and delivered detailed insights into how socioeconomic factors associate with self-reported health, diseases and mortality (see for example Cutler, Lleras-Muney, and Vogl 2011; Lleras-Muney, Schwandt, and Wherry 2024). A more recent and rapidly growing literature uses linkages between mortality, tax and other administrative registers to provide a more granular perspective on inequalities in life expectancy and underlying mechanisms (e.g., Chetty et al. 2016; Godøy and Huitfeldt 2020; Finkelstein, Gentzkow, and Williams 2021; Chen, Persson, and Polyakova 2022; Black et al. 2024; Chetty et al. 2025). Yet despite these advances, the powerful data opportunities have not been used to build a comprehensive picture of health inequality across the life course and the resulting socioeconomic differences in mortality. Without this perspective, policy makers lack crucial insights into where inequalities arise, and when interventions could most effectively address them.

In this paper, we leverage novel administrative data from the Netherlands and focus on chronic illness to study how health inequalities develop over the life-cycle. We build a panel for the full Dutch population over a 20 year period, combining multiple administrative registries with detailed data on health and socioeconomic outcomes. Using pharmaceutical data we measure a comprehensive profile of chronic conditions for the approximately 17 million people in our sample, which we link to old-age mortality risk to construct a novel health measure – the *Chronic Disease Index*. While our analysis remains descriptive, its power comes from the index being at the same time (i) comprehensive, (ii) measured repeatedly and (iii) measured at scale. First, the index allows us to explore the evolution of chronic illness as a dynamic marker of health, manifesting itself often much before mortality. Second, as we complement the index with detailed, longitudinal data on income and other socioeconomic factors, we are not only able to document the evolution of population health inequality at different ages, but also to shed light on the different dynamic pathways between health and income. Finally, we also measure a wide range of individual, social and environmental health risk factors, allowing us to analyze how these factors mediate the onset of health disparities over the course of life.

Our data-driven approach, while not without its challenges as we discuss, allows us to document these patterns and provide valuable policy insights with minimal modeling assumptions. 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 since Grossman (1972). Similarly, the observations that chronic conditions can have important health implications and differ in their prevalence across socioeconomic groups throughout life are not new (see, e.g., Loucks et al. 2007; Martinson 2012; Bauer et al. 2014). However, what is less studied is how these chronic conditions develop dynamically and interact with socioeconomic and health risk factors and a comprehensive perspective is lacking. Indeed, data constraints have limited the ability to provide a comprehensive analysis of health and its dynamics and mediators at scale. Instead, researchers have relied on structural assumptions and calibration approaches to model the complex health process and plausible impacts of interventions (see, e.g., Dalgaard and Strulik 2014; Galama and van Kippersluis 2019; Lleras-Muney and Moreau 2022; De Nardi, Pashchenko, and Porapakkarm 2023).

Our data environment and the construction of the Chronic Disease Index (CDI) allow us to relax these constraints. This construction relies on two important building blocks, allowing us to quantify *when* health inequalities arise.

The first building block is grounded in our ability to reliably measure chronic health conditions in the Dutch context. Our procedure builds on mappings in the medical literature using healthcare claims data on dispensed medications to identify chronic conditions in the full population, over time. We directly address concerns regarding under-diagnosis and/or management of chronic conditions, by studying healthcare expenditures and mortality associated with chronic and other medications, as well as concordances with self-reported conditions in survey data. We generally find that the differences across income groups in both diagnostic treatment and mortality associated with chronic conditions are small in the Netherlands; only at the very bottom of the income distribution we find evidence indicative of under-diagnosis or treatment.

The Dutch healthcare system has thus broadly equalized access to *healthcare* at the system level — a stark finding in its own right. Equal access to healthcare is valuable for this work, as we are able to focus on the remaining potential sources of the SES-health gradient.[1] Despite the universal

[1] 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

coverage and broad healthcare access in the Netherlands, we find substantial differences in the prevalence of chronic conditions across income groups, and these differences can explain between 30 and 40% of the gaps in mortality at different ages. This by itself indicates the potential value of a chronic disease index to study differences in health.

The second building block is to translate the chronic conditions in a comprehensive index of chronic disease that is comparable throughout the life-cycle. To construct this index, we focus on the role of chronic disease in mortality as the ultimate health outcome of interest.[2] 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, including multiple lags and interactions.[3] We adapt a Double-Lasso estimation procedure to also flexibly control for a rich set of socioeconomic differences that correlate with chronic conditions and may affect mortality for non-chronic reasons (Belloni, Chernozhukov, and Hansen 2014). Both controlling for comorbidities and socioeconomic confounders matters for the estimated mortality effects, which further underlines the value of comprehensive data for this analysis.

Using the resulting model, we construct the Chronic Disease Index (CDI), which, in essence, re-weights chronic illnesses at any point in life to reflect their eventual contribution to mortality risk at age 70. While the index comes with important caveats, it provides a transparent and consistent way to pinpoint when health disparities emerge over the life course. Although mortality differences are most visible in later life, our analysis reveals that health gaps open much earlier. Nearly half of the chronic disease burden gap between income groups at age 70 is already present by age 40. These differences can also be framed in terms of 'biological age': by age 35, individuals with below-median income have an average CDI equivalent to that of above-median income individuals aged 50.

Having documented *when* health inequalities arise, we then use the CDI to study the question *how* health inequalities arise over the life-cycle.

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).

[2] 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 revisit our method to capture morbidity related to chronic conditions 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 correlated and provide insights into morbidity. We do not focus on this interpretation here, however.

[3] Our findings are robust to alternative definitions of the health outcome, such as mortality at age 65.

First, exploiting the panel structure of the data, we separate to what extent the gap in chronic disease burden at old age is driven by individuals in different income groups 'aging' at different rates — higher incidence of conditions for lower incomes — versus individuals with different chronic illness sorting into different income groups — higher chronic illness levels leading to lower income. We find that the sorting mechanism is present throughout the life-cycle and responsible for the opening of the CDI gap around the start of people's careers. But we also find substantial differences in aging that gradually increase over the life course. That is, chronic disease develops at a faster rate for low-income individuals than for high-income individuals and this difference increases with age. Aggregating these effects over the life course, we find that the aging differences dominate the sorting effects and explain about 40% more of the health gap observed at age 70.

Second, leveraging the granularity of the data, we find a non-linear pattern of the CDI in income with poor health being particularly concentrated at the low end of the income distribution. This corresponds to the non-linearity of life expectancy documented in prior work (e.g., Chetty et al. 2016; Mortensen et al. 2016), but shows that this pattern materializes already earlier in life. Moreover, we find that the non-linearity is driven by sorting effects, while the aging effects are linear across the income distribution. The aging effects are mostly driven by the differential *incidence* of cardiovascular disease and diabetes, especially at older ages. At the same time, we see important differences in the *prevalence* of psychological disorders contributing to the CDI gap, already at young ages.

Third, having separated the differential aging from sorting effects, we can also shed light on the underlying mediating factors over the life-cycle. Here we make use of the rich administrative data and linked survey data to measure a variety of mediating factors jointly for the same individuals, allowing to shed light on the role of biology, geography and work environments as well as socioeconomic factors and individual behaviors. Using Shapley-Owen decompositions, which apportion the common variation to different mediators, we find that socioeconomic factors contribute more than 50% to the explained variation in aging overall. We also find a substantial role for observed health behaviors, including smoking, drinking, physical activity and BMI, particularly at older ages. While this analysis remains descriptive, it provides a unique opportunity to revisit the potential importance of different determinants of health, as they are all measured in the same context and our longitudinal data allows us to measure the incidence rather than the prevalence of disease. In particular, as we find a strong socioeconomic gradient in the commonly measured health behaviors, we also show that we would overestimate their importance for the disparities in health outcomes in more partial analysis that does not control for socioeconomic and other risk factors.

Our results have implications for policy design to address health and inequality, which we further substantiate through counterfactual analysis. As we find that individuals with low socioeconomic status take on chronic illness at a faster rate and this divergence starts early in life, our results point to a focus on earlier life interventions rather than on treatment and access in older ages, the common remit of health policy design. Counterfactual analysis conveys these returns from intervening sufficiently early in the lifespan. 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.

Related literature This paper leverages rich, administrative data sources to provide a comprehensive, longitudinal and population-wide analysis of health inequalities and aims to contribute to three strands of the broader literature.[4]

First, a rapidly growing literature in economics studies social gradients by combining administrative data from mortality registers and tax records. Following the seminal contribution by Chetty et al. (2016) in the US, Chetty et al. (2025) are currently studying the mechanisms underlying trends over time following a framework closely related to ours. The income gradients have been extended to other countries (Kinge et al. 2019; Chen, Persson, and Polyakova 2022), for different age groups (Kennedy Moulton et al. 2022), across geography (Chetty et al. 2016; Godøy and Huitfeldt 2020; Finkelstein, Gentzkow, and Williams 2021), etc. Our contribution is to use administrative data to measure health before death and study how the health gap leading to

[4] Preceding the work using administrative data, the longstanding health inequality literature can be categorized by the type of data used: a) longitudinal surveys focused on old-age individuals, including HRS, ELSA, & SHARE (e.g. Smith 2004; Avendano et al. 2009; Banks, Muriel, and Smith 2010; Zaninotto et al. 2020). Most recently and closely related to our work, Russo et al. (2024) construct a frailty index from the HRS data to study health disparities after age 55 in the US. b) cross-sectional population surveys at all ages such as NHIS, BRFSS, & EHIS (e.g. Bleich et al. 2012; Singh et al. 2017) and NHANES & HSE (e.g. Loucks et al. 2007; Seeman et al. 2008; Martinson 2012; Nesson and Robinson 2017; Campbell et al. 2023; Abdalla et al. 2025), which include biomarkers too, c) long-term birth cohort studies such as NSHD & NLSY (e.g. Pais 2014; Khanolkar et al. 2021; Bolt 2022) or d) studies that sample from a electronic health record (EHR) system, often specific to a disease or geographically focused, such as the TOGETHER study (London, UK), the Lifelines study (Groningen, NL), as well as data provided by healthcare networks including Geisinger Health System (PA, US) and Sentara Healthcare (VA/NC, US) (e.g. Dharmayat et al. 2020; Zhu, Dekker, and Mierau 2023; Poulsen et al. 2024; Lu et al. 2024) . Most notably, Shui et al. (2025) build on our work to study health gaps over the life-cycle based on biomarkers observed in the Lifelines study.

mortality differences at late age evolves over the life-cycle.

Second, the question of how health itself evolves dynamically over the life-cycle has been central to health economics since Grossman (1972), but the empirical evidence has been arguably lagging behind. Seminal contributions have documented life-cycle patterns of income gradients in health (Case, Lubotsky, and Paxson 2002; Case and Deaton 2005), but health status has been typically measured as a cross-section, and so cannot describe within-individual health dynamics. While some contemporary work used longitudinal health data (e.g., Currie and Stabile 2003; Case, Fertig, and Paxson 2005), the coarse health measure and modest sample sizes have left room for a more detailed examination of within-individual health dynamics by age and SES. More recent work has calibrated structural models using survey data to shed light on the dynamics of the health process, including Galama and van Kippersluis (2019), Hosseini, Kopecky, and Zhao (2022), De Nardi, Pashchenko, and Porapakkarm (2023), Borella et al. (2024) and Hosseini, Kopecky, and Zhao (2025). For example, Hosseini, Kopecky, and Zhao (2025) use a structural model to study the role of health in lifetime earnings inequality, assuming no impact of earnings on health. Our work is closely related to this literature, but 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 (Cutler, Lleras-Muney, and Vogl 2011; Marmot 2015; Mackenbach 2019; Murray et al. 2020). An important part of this work in the economics literature has focused on specific causal pathways and on specific mediators of the health gap, as reviewed by O'Donnell, Van Doorslaer, and Van Ourti 2015.[5] 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 a valuable recalibration of the potential importance of specific mechanisms over the life-cycle and as such provides an ideal roadmap for further empirical work estimating causal effects.

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] This includes the pathway from socioeconomic determinants to health outcomes (e.g., Currie 2009; Adda, Banks, and von Gaudecker 2009; Sullivan and von Wachter 2009; Black, Devereux, and Salvanes 2015) and from health shocks to socioeconomic outcomes (e.g., Currie 2009; Dobkin et al. 2018; Stepner 2019), as well as specific mediators such as access to medical care (e.g., Finkelstein and McKnight 2008; Deaton and Paxson 2004) obesity (e.g., Bolt 2022)) , 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), etc.

5 then uses the index to study how the health gap evolves over the life-cycle. 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.

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), which 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.[6] To measure chronic conditions, we use prescribed medication as described further in section 2.3. 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 setting. Medicines are classified by the Anatomical Therapeutic Chemical (ATC) code, at ATC3 digit level: for example, the code *N*06*A* 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

[6] 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 hyperlipidemia, 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).

includes costs insurable under compulsory standard insurance, which represented 90% of all hospital, medical practice, and pharmacy costs in the Netherlands from 2009-2021. 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, as well as the primary diagnosis ICD code, 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 behaviors are merged with administrative data at the individual level. These 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 behaviors: drinking, smoking, rates of physical activity. We also link an annual cross section 'Health Inquiry' survey with approximately 9000 individuals per year (Dutch: *Gezondheidsenquete*, GECON). These report take-up rates of health screening activities, such as blood pressure tests.

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*.[7] 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 alternative markers of individuals' socioeconomic status, including parents' income, education and wealth.

To form measures of income we rank lagged incomes within gender, birth cohort, and calendar year. Much of the following analysis calls for a binary classification of relative income. We define those below median income as **Low income**, and above median as **High income**.[8]

[7] 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.

[8] Notably, the bottom income decile includes some households with high net assets: this suggests some have

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. We are able to link around 60% for those aged 40. That rate falls to 20% for 70 year old's. Beyond the demographic data, CBS also provides 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).

Spatial data A number of spatial variables are included using data from the Geoscience and health cohort consortium (GECCO) (Timmermans et al. 2018). Pollution data is included, specifically levels particulate matter ($PM_{10}, PM_{2.5}$), Nitrogen Dioxide (NO_2), and Elemental Carbon. These are imputed on a 25x25 meter grid, then interpolated to the six-digit postcode level. The local food retail environment is observed, specifically the density of fast food retailers, fresh food retailers, supermarkets, restaurants, and convenience stores. Each of these are measured as a kernel density within 1km of each six-digit postcode. Green space density is similarly observed. The pollution, green-space, and food environment data were resolved to the six-digit postcode level, before being aggregated to the neighbourhood level. Alongside the GECCO data, CBS data on the proximity to healthcare facilities (nearest GP, pharmacy and hospital) average property values, and population density are also included at the neighbourhood (Dutch: *Buurt*) level.

2.2 Descriptive Statistics

Selected descriptive statistics on socioeconomic and health outcomes are included in Appendix Table C.1. To illustrate the difference in health outcomes across socioeconomic 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.

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.

Similarly, we can partition the 55 year old cohort by income quintile and observe subsequent survival rates, as shown in Panel B ofFigure 1: the relation between income and survival probability is clearly monotonic, but the bottom income quintile faces a markedly higher mortality hazard. This corresponds to the pronounced non-linearity in the relation between mortality and income at the bottom of the income distribution (e.g., Chetty et al. 2016). 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.

2.3 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 often measured using self-reported information from representative surveys, as discussed in footnote 4. 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.[9]

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.[10] They extended prior work using further medical expertise, to reduce type-I errors by focusing on ATC codes that are used exclusively for the treatment of a given chronic disease. We make a number of modifications, leveraging our longitudinal data to further reduce type-I errors, but also reflecting that our data

[9] For instance, Dowd and Todd (2011) found reporting differences by group implies naive estimates of health inequality are downward-biased. Conversely, Nesson and Robinson (2017) finds evidence that reporting differences yield an upward bias of health inequality. Besides this, survey sample sizes mean the analysis can be under-powered for rarer outcomes, such as early life mortality.

[10] Huber et al. (2013) apply their mapping to establish new prevalence estimates using Swiss administrative data. 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 life-cycle.

is resolved to the ATC3 level, whereas their mapping sometimes uses ATC5 resolution. Our mapping is given in Table 1, with further description of specific refinements given in Appendix D. Almost all chronic diseases are related to only one ATC3 code, with cardiovascular disease as a notable exception being mapped to several medications. The one chronic disease we are mostly missing is cancer, since only 5% of diagnoses are treated with pharmacy-dispensed medication. Digestive tract and skin cancers dominate this measure: they account for over 60% of the detected diagnoses. In a companion paper (Danesh et al. 2025) we focus on the socioeconomic gradient of cancer using data from the Netherlands Comprehensive Cancer Organization (Dutch: *Integraal Kankercentrum Nederland*, IKNL). While the data is very rich, it captures incidence rather than prevalence, over a five year panel rather than 15, hence we cannot incorporate that data into the index construction.

As a first validation check, we can test the comprehensiveness of our mapping relative to a data-driven benchmark: using a Lasso regression of five-year mortality on ATC codes directly, we find that all 25 selected ATC codes are already captured within the refined mapping to chronic conditions. As a second check, we can use the *Gezondheidsmonitor* 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 lower for the set of medications indicating cardiovascular disease, as these include conditions beyond just hypertension. We address concerns regarding under-identification in more detail next.

2.4 Under-diagnosis and Under-treatment

Our approach only identifies chronic conditions that are actively treated through medication and does not distinguish between the severity of each chronic condition nor the intensity of treatment, which we turn to in Section 3.2. Importantly, our approach also ignores chronic conditions that are not properly diagnosed or not managed at all. If under-diagnosis or treatment is more important for low-income groups, we would under-estimate the gap in chronic health. Moreover, missing chronic conditions will make us under-estimate their role for mortality.

To understand the role of under-diagnosis and under-treatment, we can leverage both the administrative records and survey data. While not conclusive, the evidence suggests that the scope for under-diagnosis and under-treatment is limited in the Dutch context. Given the relative differences in under-diagnosis across income groups, 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. This may all not be as surprising given the Dutch institutional setting with universal access to high-quality healthcare. 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).

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 essentially 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, we present more evidence that indicates 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 (under-diagnosed) 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 share of individuals without any medication for four income groups: the first decile (D1), the second decile (D2), the third to fifth decile (D3-D5) and the fifth to tenth decile (D6-10). The share without any medication 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. Importantly, it is only the bottom income decile (D1) that becomes heavily over-represented among those individuals without any medication at older ages.[11]

To investigate the potential for under-diagnosis further, we consider the mortality rates depending on medicine use for these different income groups in Panels C to F of Figure 3D. 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, while for the bottom decile this divergence happens earlier. Overall, this analysis suggests that during prime ages under-diagnosis is limited to the bottom income decile and 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 still limited to less than 1 in 10 individuals who are not actively seeking care.

3 Chronic Disease and the Mortality Gap

In this section we study the role chronic conditions play in mortality differences between high and low income 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 this section demonstrates the potential value of using chronic conditions to study which differences in health already appear earlier in life.

3.1 Conceptual Framework

We start by providing a conceptual framework that guides our empirical analysis and helps highlighting its policy implications. Here we focus on income, but the framework can be equally

[11] The over-representation of the bottom income decile is further illustrated in Appendix Figure C.1. 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, indicating a lack of engagement with the healthcare system. 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.

applied to other socioeconomic dimensions, such as wealth, education, or parental income. Consider a stylized, linear model of mortality for an individual i at age a:

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

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

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

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.

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 income 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 income 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 income groups. These *residual* effects can also include differences in under-diagnosis as discussed before.

This simple framework points to different policy options for governments in trying to reduce the health gap depending on the relative importance of 'prevalence gap' and the 'treatment gap'. The former suggests the value of public health interventions and programs targeting the social determinants of disease that can reduce the burden of chronic illness for lower income 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. This is due to the fact that if we improve individuals' health, this may improve their income trajectory, which can further improve their health.[12] In Section 5.1, we explicitly try to separate the causal pathways between health and incomes using the panel structure of our data. Still, we can in principle circumvent this challenge when considering policies that directly intervene on individuals' health but not their incomes. 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 income groups and thus need to be controlled for. We turn to this in Section 4.

3.2 Prevalence vs. Treatment

We first consider the prevalence of chronic conditions and how much they contribute 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 incomes. The overall pattern is very clear as the burden of all common chronic conditions falls more on low income than on high income individuals, in aggregate and by specific condition.[13]

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 R_i that varies by specification:

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

[12] To illustrate this, consider an intervention that changes the incidence of chronic conditions:

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

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.

[13] 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.

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. 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 5-year mortality rate in the high-income group is 66 per 1000. For the low-income group, the mortality rate is 44 per 1000 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 per 1000 (row A2). That is, about one third of the gap in mortality between the income groups can be explained by the difference in measurable chronic conditions. Adding further lags of chronic conditions (row B1) does not change the explanatory power, 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 with other measures of healthcare utilization and diagnoses. The mortality gap is virtually unchanged when adding other health-related controls, once we have controlled for chronic conditions. Panel A of Figure 5 shows that the estimated gap hardly reduces when adding other prescribed medications (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 healthcare and other related measures. Interestingly, these variables do strongly increase the explanatory power of our regression model. However, their limited contribution to the mortality gap further motivates our focus on measurable chronic conditions in the empirical analysis.[14]

These findings continue to 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 across incomes, 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, the gap is almost halved once we control for chronic conditions. Second, other health information in our data does not help much in further closing the gap. At younger ages, the

[14] Including all other health-related variables jointly more than doubles the R^2 (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 per 1000 to 26 per 1000 (row D). Of course, without controlling for chronic conditions, the other health-related variables do contribute meaningfully to the mortality gap, albeit less than the chronic conditions (see Appendix Figure C.4).

further reduction in the mortality gap is not statistically significant.

The prior regressions do 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 us to compare how chronic conditions carry different mortality risk for different income groups. As discussed, this becomes relevant when chronic conditions differ in severity or are treated differently. This also matters 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. 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.[15]

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). Appendix Figure C.5 shows clearly that the difference in prevalence outweighs the difference in treatment. At the age of 70, we find that differences in treatment effects do not contribute at all to the mortality gap. We can analyse 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 shows 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 indicative of equalized treatment by the healthcare system across income groups.[16]

Taken together, these results provide clear evidence that 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.

[15] For example, diabetes for men is associated with an increase in the mortality rate of 46, 51 and 43 per 1000 for high income, low-income and bottom deciles respectively. Of course, some caution remains warranted when interpreting these separate coefficients. The chronic conditions or underlying medicine use can be correlated with other factors affecting mortality. We explicitly address this in Section 4.3 when constructing our chronic disease index.

[16] 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.

4 Chronic Disease Index

Chronic illness plays a key role in older age mortality, but it can occur much earlier in life. Our aim is to provide a comprehensive index of health that can be measured throughout the life course, but ultimately contributes to mortality. To achieve this, we focus on how an individual's chronic illness predicts mortality in old-age, accounting for co-morbidities and interaction effects, while controlling for other confounding factors. We then re-weight chronic illness at any age to reflect the old-age mortality based on *point-in-time* chronic conditions.

4.1 Empirical Model

To guide our prediction exercise, we impose the following structure on our static model introduced in section 3:

$$\begin{aligned} M_{i,a} &= \alpha_{i,a} + CC_{i,a}\beta_{i,a} + \epsilon_{i,a} && (4) \\ &\equiv X_{i,a}\gamma_a + CC_{i,a}\beta_a + \varepsilon_{i,a}, && (5) \end{aligned}$$

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 $\alpha_{i,a}$ 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 as follows:

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

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

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

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 allows to compare the chronic illness across individuals and over the life-cycle, 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 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 socioeconomic groups, this underlines the importance of including socioeconomic controls when estimating the mortality effects of chronic conditions. On the other hand, we have found limited evidence for heterogeneity in mortality effects across socioeconomic groups, which supports the construction of a uniform index with the same mortality interpretation. 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. Hence, we estimate our chronic disease index predicting mortality for individuals at age 70 using the 2nd to 10th income decile.

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.

We build on Belloni, Chernozhukov, and Hansen (2014) who propose 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 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. 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.

Our estimation proceeds in three steps: (i) a first Lasso estimation to establish the socioeconomic 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,

$$\begin{aligned} M_i &= X_i'\theta_m + \zeta_i \\ c_i^k &= X_i'\theta_c^k + \nu_i^k \qquad \forall \quad k = \{1, \ldots, 22\} \\ M_i &= f(CC_i)'\theta_v + \varepsilon_i, \end{aligned} \tag{8}$$

where M_i denotes an indicator for five-year mortality, c_i^k is an indicator for the kth of 22 chronic conditions, and X_i is the set of all potentially relevant socioeconomic information. The mapping $f(CC_i)$ is the basis of all potentially relevant chronic condition information. It includes three years of chronic condition indicators, within-condition interactions across different lags, and two-way cross-condition interactions for the most recent lag. The 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. This results in $X_i^* = \{x_i : \hat{\theta}_m > 0 \cup x_i : \hat{\theta}_c^k > 0\}$, denoting the set of socioeconomic variables found to be relevant in one or more of the Lasso estimations, and in $f(CC_i)^* = \{CC_i \cup f(CC_i) : \hat{\theta}_v > 0\}$, denoting the union of the set of chronic conditions and their relevant interactions from the corresponding Lasso estimation.

The final estimation is then as follows:

$$M_i = X_i^{*\prime}\beta_X + f(CC_i)^{*\prime}\beta_{CC} + \zeta_i, \tag{9}$$

which we estimate using a linear probability model, by gender and with calendar year fixed effects absorbed. We randomly select 50% of the population of 70-year-old individuals for the estimation. 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 socioeconomic status, as

$$CDI_{i,a} \equiv \bar{X}^{*\prime}\widehat{\beta}_X + f(CC_{i,a})^{*\prime}\widehat{\beta}_{CC} \tag{10}$$

where the intercept captures the mean socioeconomic effects.

4.3 Accuracy and Predictive Value

We briefly evaluate the accuracy and predictive value of the prediction model. 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. This bias is around 16%.[17] That is, a "naively" estimated CDI without the Lasso-selected SES controls would overstate the chronic disease health gap by 16%. The 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, but we need to residualize first using the SES controls. Appendix Figure C.6A shows that, when taking out the socioeconomic 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. Appendix Figure C.6B compares the conditional 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.

We document substantial heterogeneity in the CDI, Appendix Figure F.2 depicts the breadth of the CDI distribution within age, gender, and bivariate income. We find a predicted 5-year mortality rate of 44 per 1000 for the healthiest 10 percent of the 70-year olds in the CDI distribution, which compares to 251 per 1000 for the sickest 10 percent. The dispersion is comparable for men and women and for different incomes groups separately, but substantially smaller for younger cohorts. Still, even for the 70-year olds, the out-of-sample R^2 when regressing 5-year mortality on the CDI index is only 5.2%. While the dependent variable is a binary, random realization of a probability, the predictive power of the index is relatively modest. This can in principle be increased using alternative measures of health. As noted before, adding other health-related variables in our data almost triples the predictive power, but does not explain much more of the mortality gap between low and high-income individuals (Figure 5).[18]

Finally, we can use the CDI to revisit the marginal effect of separate chronic conditions on

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

[18] We find that 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. Also, the in-sample R^2 value (at 5.6%) for the prediction model is only slightly above the out-of-sample R^2, which indicates limited risk of over-fitting given the large sample size.

mortality, but now accounting for their interactions and lag structure, and controlling for socioeconomic factors. 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 socioeconomic factors, we would over-estimate the mortality associated with a specific chronic condition. This correction can be significant and very large.[19] 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.[20]

4.4 Further Caveats

The CDI enables meaningful and transparent comparisons of individual health at different ages and across socioeconomic groups. In documenting the measurement of chronic disease and the construction of the CDI, we have taken care to validate our approach along the way. In particular, Section 2.4 demonstrates that inferring disease prevalence from dispensed medication appears robust to concerns about differential under-diagnosis. Moreover, the absence of differential treatment effects (Section 3.2) and the debiasing method discussed (Section 4.2) help ensure that the CDI remains equally valid across the income distribution. Nonetheless, some important caveats remain.

First, the CDI should not be interpreted as a causal measure. We observe seemingly protective effects for conditions such as migraine and intestinal inflammatory diseases - likely due to selection into medication offsetting the underlying mortality risk. Still, by controlling for socioeconomic status and comorbidities, the CDI plausibly captures estimates that are closer to the causal impact of these conditions on mortality.

Second, the CDI focuses on mortality and does not capture all relevant dimensions of health. Most notably, it omits cancer prevalence due to measurement limitations as discussed earlier. It also does not explicitly account for differences in frailty or physical functioning. Yet, the CDI correlates strongly with self-reported health ($r = 0.36$), and while low-income individuals report

[19] For example, for anemia the estimated mortality rate would increase from 10.7% to 15.7% for women. For dementia, it would be inflated from 24.1% to 38.5%.

[20] 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%.

worse health at any given CDI level, the relationship between CDI and self-reported health is similar across income groups (Appendix Figure F.3).

Third, the mortality weights for chronic conditions are estimated at age 70, and these weights are likely to be different at young ages: health stocks are different at younger ages, individual conditions have different implications at younger ages and the sample would include individuals who do not survive until age 70. While this is an inevitable trade-off, we make some refinements to mitigate age-related mis-classification, as discussed in Appendix Section D.[21] Importantly, our aim is not to predict age-70 mortality as precisely as possible using early-life information, but rather to construct a consistent measure of chronic disease burden that is available and comparable across ages, subject to the limited horizon of panel data. Even so, the CDI predictions are robust to a range of modeling choices. This includes the choice of lags of chronic conditions and of target ages, but also the linkage function, the estimation sample and the Lasso penalization, which are all tested and discussed in Appendix F.

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 that translates into old-age mortality 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 briefly revisit our conceptual framework from Section 3 to consider the dynamic evolution of the health gap. In a first step, we are interested in how the health gap between income groups $\Delta CDI_a \equiv CDI_{L,a} - CDI_{H,a}$ compares at different ages a. This simple comparison allows us to evaluate when in the life-cycle the health gap opens up, and how much of the gap observed at old-age is already present 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

[21] For instance, anemia is only counted if treatment extends over three years, excluding cases linked to pregnancy-related iron deficiency.

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 to quantify the potential for targeted health interventions, any *dynamic* change in this gap sheds light on the desirable timing of these interventions. However, since income is endogenous to individuals' health, any comparison of the static gap across ages confounds differential growth in CDI across income groups and the sorting into different income groups based on health.

Our dynamic measure of health and the panel structure of our data allow us to separate the two forces. Indeed, for any given individual, we can observe how his or her CDI grows, capturing the *incidence* of new chronic conditions rather than their *prevalence*. The growth $dCDI_{i,a} \equiv CDI_{i,a+1} - CDI_{i,a}$ reflects how an individual's health develops with age, which we refer to as *biological aging*, or aging in short.[22] We evaluate this aging process separately for the individuals in different income 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 his or her 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. Of course, other underlying factors that affect both the health and the income process would also be captured by this term. The observed change in CDI across ages for income group Y can thus be decomposed as follows:

$$\begin{aligned} dCDI_{Y,a} &\equiv E(CDI_{i,a+1}|Y_{i,a+1} = Y) - E(CDI_{i,a}|Y_{i,a} = Y) \qquad (11) \\ &= \underbrace{E\left(dCDI_{i,a}|Y_{i,a} = Y\right)}_{\text{Biological aging}} + \underbrace{E(CDI_{i,a+1}|Y_{i,a+1} = Y) - E(CDI_{i,a+1}|Y_{i,a} = Y)}_{\text{Health-Based Sorting}}. \qquad (12) \end{aligned}$$

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 life-cycle. In this spirit, we can re-construct the CDI for a specific income group as the accumulation of the aging effects between age a_0 and a as

$$CDI_{Y,a}^{aging} = CDI_{Y,a_0} + \sum_{\tilde{a}=a_0}^{a-1} E\left(dCDI_{i,\tilde{a}}|Y_{i,\tilde{a}} = Y\right). \qquad (13)$$

[22] As our measure of health is a mortality-weighted index of chronic disease, our approach is perhaps more aligned to the conceptual models of health deficit accumulation (e.g. Dalgaard and Strulik (2014)) rather than health capital accumulation (e.g., Grossman (1972).

We note the parallel between this simulated ageing $CDI^{aging}_{Y,a}$ and the period life expectancy (e.g., Chetty et al. 2016), which aggregates income-specific mortalility rates at different ages to obtain income-specific life expectancies. Instead of aggregating mortality rates, the simulated ageing aggregates the income-specific incidence of chronic conditions at different ages. Assuming a Markovian process for health with an individual's income group as the relevant state variable, this counterfactual captures the life-cycle of the CDI for individuals who remain in a specific income group.

This decomposition approach can also be applied to other measures of socioeconomic status, such as education, parental income, or wealth. Still, we should not interpret these differences as being caused by the differences in socioeconomic factors. Instead, differences in biological aging allow us to quantify the potential value of targeting health interventions that reduce the incidence of chronic conditions for a specific group. As in our discussion of the prevalence gap in Section 3, this incidence gap can again be interpreted as a lower-bound for the returns from intervening as the improvements in health may improve their income. Interestingly, our estimate of health-based sorting sheds empirical light on whether health does indeed have meaningful impacts on income.

We also 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 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 attenuates 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 household income between ages $a-4$ and $a-2$, except from 65 onwards, we consider average household 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 the CDI gap at age 70. The health gap is close to zero until early adulthood with a difference of only 0.08 percentage points at age 20. That is, the health gap at age 20 would translate into a 0.08 percentage point 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 0.30 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 in that age range, especially for low-income groups, but we also note the difference in income group definitions at age 65 and the corresponding change in cohorts for older ages.

The CDI allows us to measure health throughout life and has the potential to pick up 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 incorrect 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.[23] Note that 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 by age 35 the average CDI for the low-income group surpasses the average CDI for the high-income group at age 50.[24]

[23] This early age health gap is also present in the self-reported health profile, as shown in Appendix Figure F.3B. However, the profile is subject to sampling error and does not allow for a natural interpretation as for the CDI, since self-reported health response is categorical not cardinal. Furthermore, given the available surveys are repeated cross-sections, we cannot look at within-individual effects.

[24] 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)

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 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.[25] 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 of the prevalence of these six chronic conditions.

Non-linearity The gap at a given age increases linearly when considering different quintiles, except for the bottom quintile. Appendix Figure C.10 panel A considers income quintiles and shows that the bottom quintile indeed stands out. This non-linear pattern is similar to the mortality rates considered in Appendix Figure 1 and in prior work (e.g., Chetty et al. 2016), but now adds the insight that this pattern materializes early in life. For example, at age 50, the CDI gap between the bottom and top income quintiles exceeds 2 percentage points. In relative terms, the chronic disease burden for bottom-quintile incomes by age 50 is nearly 30% greater than those in the top-quintile.

Further Heterogeneity We can extend our analysis to other socioeconomic measures and for different sub-groups, but the overall empirical patterns are robust. Appendix Figure C.10 panel B shows that the differences are somewhat larger for men than for women with different dynamic patterns around child-bearing ages. Panels C and D plot the average CDI by education groups and mother's income respectively. These socioeconomic 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,

[25] 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.

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 socioeconomic groups partly depends on compositional changes, and thus do not only capture differences in biological aging.

5.3 Differential Aging over the Life-Cycle

While the average CDI increases rapidly with age 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. However, these age gradients mix biological aging and health-related sorting: we can separate these two components, following equation (17). Of course, 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 decomposition is described in Appendix G.

Table 2 shows the age-specific and aggregated results of this dynamic decomposition. Both differential aging and health-based sorting contribute substantially to the CDI gap observed at old age. Aggregated over the course of life, differential aging is more important, contributing 37 percent more than 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 with high CDI's are more likely to die, and this reduces the CDI of the surviving individuals more in comparison to the high-income individuals.[26]

Figure 7 graphically illustrates the role of differential aging over the life-cycle and also compares it to health-based sorting. In Panel A we simulate the mean CDI for each income group when only the aging effects are included, following Equation (13), and when both aging and health-

[26] 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.

based sorting effects are included. When only the aging component is included, the high-income group ages slowly into adulthood, but the evolution of the CDI is no longer as flat since we have excluded the sorting effect. The gap between the CDI for high versus low income steadily grows over the life-cycle. While the gap increases by 0.10 percentage points between age 20 and 30, the corresponding increase is 0.47 percentage points between age 60 and 70 (see Table 2). Once we include the health-based sorting effects, the more rapid opening in the CDI gap during early adulthood becomes apparent.[27, 28]. Individuals with lower CDI then continue to sort into the high-income group and individuals with higher CDI continue to sort into the low-income group over the life-cycle. Panel B in Figure 7 uses the horizontal differences between the simulated CDI's using only aging effects to provide a better 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. Recall that this threshold was reached at age 35 when using the observed CDI instead of the simulated CDI.[29]

Separate Chronic Conditions Which chronic conditions contribute most to the differential aging and how does that evolves over the life-cycle? To evaluate this, we focus on the incidence of new chronic conditions (not the prevalence) 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 condition 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 present at younger ages and explains most of the differential aging in middle age. Consequently, these two conditions are dominant in explaining the gap in CDI levels at older ages.

[27] 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.

[28] 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).

[29] 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.

Non-linearity We can consider the decomposition of aging and sorting effects for finer income groups. The decomposition in Appendix Figure G.3 shows that the earlier non-linearity in the relation between CDI and income at a given age is driven by sorting effects. Between 25 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 relatively linearly for the quintiles in between. The health-based sorting, however, worsens the CDI of the bottom income quintile and improves the CDI of all other incomes. That is, if poor health lowers your income rank, it will push you all the way to the bottom quintile. The Q1 sorting effect is most acute between ages 25-30, as cohorts are forming their own households. Over the lifecycle, for Q1 the CDI is increased by 1.50 percentage points due to health-based sorting. For Q2-Q5 we find decreases ranging from 0.11 to 0.52 percentage points. Appendix Table G.1 compares aging effects for different education groups instead of income quintiles, confirming the important, but gradual relation between aging and socio-economic status.

Robustness Our baseline categorization of income groups uses average income two to four years prior to being considered. This follows prior work in the literature (e.g., Chetty et al. 2016) where the lagging is aimed to mitigate reverse causality concerns. Of course, this addresses only health-based sorting based on the most recent health shocks. Our decomposition has shown that persistent health-based compositional changes among the low and high-income groups are important and thus the lagging is by itself not sufficient to fully address the reverse causality. Still, we can test the robustness of the dynamic decomposition results and in particular the estimated differences in aging when categorizing income group using different lags and moving average lengths (see Appendix Figure G.4). First, we find that our aging estimates, capturing differences in CDI growth, remain essentially unchanged when we instead average income one to three years prior. The same is true for the sorting effects. Second, the averaging over three years aims to capture persistent differences in income. To gauge whether this is the relevant state variable of Markovian health process, 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. We discuss this in more detail in Appendix Section G.

6 Counterfactual and Mediation Analysis

The empirical analysis has documented important socioeconomic differences in the burden of chronic conditions early in life and how those with lower socioeconomic status age at a faster rate. This final section draws the work closer to policy discussions on approaches to tackling health inequality. 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 versus 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 - capturing the incidence instead of the prevalence of chronic disease.

6.1 Counterfactual Analysis

Our analysis allows us to study the potential impact of health interventions targeting socioeconomic groups depending on their timing. We argued in Section 3 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 use the simulated CDI's from Section 5.3 capturing the accumulated aging effects and evaluate the impact of equalizing these aging effects from different ages onwards. We provide more detail on our estimations and calculations in Appendix H.[30]

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 shows 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 disease burden is valuable by itself, we can also evaluate the corresponding gains in life expectancy.

[30] 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 socioeconomic outcomes and how that can further improve one's health.

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.[31] We then aggregate mortality rates into an estimate of period life expectancy at age 40 for different counterfactual scenario's building on Chetty et al. 2016.[32] 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 3. If we could equalize the aging process from age 20 onwards, the life expectancy of the low-income group would increase by 1.3 years, closing 31 percent of the gap. This reduction is comparable in magnitude to our earlier findings on the contribution of the differential prevalence of chronic conditions to the mortality gap. If we instead equalized the aging process from age 40, low-income life expectancy would increase to 82.3. Hence, intervening 20 years later still closes 23 percent of the gap. However, waiting 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, and aggregate over the life-cycle (after age 40).[33] Table 3 reports the estimated lifetime healthcare costs, which 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 given age, the lifetime difference is muted because the faster aging for low incomes results in shorter life expectancy (see also Van Baal et al. 2008).[34] The same countervailing effects are at play when we equalize the aging process. Even though we substantially reduce healthcare costs at any age for low-income individuals, we also improve survival rates. This results in

[31] 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.

[32] 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.

[33] 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 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.

[34] 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 lifetime healthcare costs by 17% to 182.7k EUR.

cost savings of less than 3 percent, even when intervening at age 20. Keeping survival rates unchanged, the cost savings would increase to 8 percent.

Taken together, our counterfactual analysis shows that when equalizing the aging process, we can still realize most of the life expectancy gains by starting at middle-age. By starting earlier we can also reduce the biological age gap throughout the life-cycle. Waiting until later ages seems too late on both dimensions. The reduction in the life expectancy gap is much smaller, but also the biological age stays much higher throughout the life course. The effects on healthcare costs are relatively limited due to the countervailing 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 overall health is an even larger puzzle, which has been subject to much research and 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. These factors may differ across socioeconomic groups, but socioeconomic status may be of importance beyond the observable risk factors too. We harness three key advantages of our setting and data to provide an attempt to calibrate the importance of each factor: (i) we can measure health over the life-cycle, (ii) we can study a wide range 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.

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

$$dCDI_{i,a} = \sum_{j=1}^{J} X_{i,a}^{j} \gamma_{j,a} + \varepsilon_{i,a} \tag{14}$$

where $dCDI_{i,a} = CDI_{i,a+5} - CDI_{i,a}$ equals the CDI growth over the next 5 years and $X_{i,a}^{j}$ is a group of mediating factors. We calculate the Shapley-Owen values, which represent the average contribution of each regressor group to the R^2 over all possible combinations of the regressor groups. This 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 behaviors alone explain 20 percent, and both geography & health behaviors 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.[35]

Registry versus Survey Data In the registry data, we consider the following the groups of mediators: (i) parental health, consisting of fathers' and mothers' CDI if alive, or their age at death if not, to proxy for genetic disposition, (ii) a set of spatial variables , comprising pollution exposure, green-space, food retail quality, healthcare proximity, population density, and mean residential property value,[36] (iii) employment status and occupational sector, proxying for work factors and occupational health, (iv) pay-rank within employer, to zoom in on the role of hierarchy and control and its potential effect on stress, and (v) a rich set of socioeconomic variables including income and wealth, education, parental resources and demographics (e.g., household composition, foreign born). We provide more detail on the sample selection and full list of variables in Appendix I. Since these mediators are observed in the registry data, we can estimate the Shapley-Owen contributions using the full population data, as shown in Panel A of Figure 8.

Panel B of Figure 8 compares the Shapley-Owen contribution of the registry-based measures to self-reported health behaviors as measured in the *Gezondheidsmonitor* national health survey, including smoking, drinking and physical activity, and BMI.[37] The health survey sample is relatively large, with around 400k individuals, but still substantially smaller than the Dutch population. Hence to calculate the Owen-Shapley values, we use a two-stage approach: in the first step we fit a model of CDI growth on a) age and gender alone, and b) all registry-based measures, and we then compute residuals for each. In the second step, we use 50% of the health survey as a training sample to fit a model of the first-stage residuals on health behaviors, then compute residuals on the holdout sample. This leaves us with four sets of residuals on the health survey sample, corresponding to each permutation of including or excluding health behavior or

[35] To elaborate on the example, suppose now parental health explains 10 percent of the variation in CDI by itself, and that all three factors (behavior, geography, parental health) jointly explain 50 percent. This would imply that the parental health variation is fully additional to behavior 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 25pp to geography, 15pp to health behaviors, 10pp to parental health.

[36] These spatial data are observed at the six-digit residential postcode level, which corresponds to around 40 residents per postcode.

[37] We control for BMI in the absence of direct information on nutrition, but this of course captures health more broadly, meaning we potentially over-state the role of health behaviors.

other registry based mediators. The R^2 is calculated from the holdout sample residuals. Out-of-sample statistics were used because sensitivity testing revealed that the standard regressor count adjustment $\frac{n-k-1}{n-1}$ was not sufficient to mitigate overfitting in the survey sample. As with Panel A, the registry and behaviour contributions to the R^2 are additive, but have been presented side-by-side to easily visualize the imprecision of the estimates in the survey sample.

Results A few striking patterns emerge from the analysis of registry-based data, as shown in Panel A of Figure 8.[38] Socioeconomic variables (education, demographics, income and wealth) play a dominant role throughout the life-cycle. Together they are responsible for over 80 percent of the explained variation between ages 20-29, driven predominantly by education. This declines to around 60-65 percent between in the forties and fifties, but is back over 80 percent between ages 60-69. For older ages, education is replaced by income as the key socioeconomic mediator, though this is partly driven by sparser education coverage among older cohorts. While a large role of social determinants has been conjectured before in the literature, this contribution is in addition to the variation explained by the other measured factors, where any commonly explained variation has been equally apportioned. Conversely, some of the other factors, while notable, are less important quantitatively. Employment factors, including status, sector and ranks, jointly account for 12 percent of the explained variation for 20-29 year olds, rising to 19 percent for those aged 50-59. Parents' measurable health also peaks at about 14 percent for those aged 40-49. The set of spatial information contributes another 7-10 percent of the overall explained variation.

The contributions of the registry-based measures are compared to the estimated contributions of health behaviors in Panel B of Figure 8. Overall, the precision is much lower, given that the survey sample is a small fraction of the population. This, by itself, highlights the importance of population-scale data to perform comprehensive mediator analysis. Indeed, we cannot detect any positive contribution of health behaviors to CDI growth at younger ages. However, it does increase to around a third of the explained variation for those aged 40-49, and just under half for 50-59 year olds. Only for those aged 60-69, for whom the coverage of education and employment is limited, does behavior dominate other mediators, accounting for around two-thirds of the explained variation.[39]

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

[39] While the primary object of interest is the flow of health, measured as the growth in CDI, this is not commonly

The estimated patterns are descriptive and uncovering causal pathways across types of mediating factors is challenging. Nevertheless, prior work has made assessments about the importance of specific factors separately, often highlighting the importance of individual health behaviors.[40] A few papers have devised a comprehensive account of the different factors jointly, by combining estimates from separate studies (e.g., McGovern 2014).[41] Combining results from Panels A and B of Figure 8 we can provide a similar assessment, based on a common methodology and context. Behaviour and BMI explain just below one third of lifetime variation in CDI incidence, if we take the the imprecise estimates for 20's and 30's as zero. This figure is consistent with prior estimates, but driven by the older ages. In comparison, socioeconomic factors explain just over half of the lifetime variation.[42]

6.3 Estimated Effects of Mediators on Health

The previous section quantified the share of the overall variation in aging that can be attributed to different mediators. In this section we describe how one can easily mis-estimate (or mis-interpret) these contributions when facing data constraints.

Figure 9 shows the estimates for different mediators in our baseline regression in (14), using CDI growth ($dCDI$) as the dependent variable and controlling for all mediators jointly, including socioeconomic factors.[43] In line with the analysis previously described, income and wealth have negative relationships, albeit relatively modest compared to the education gradient; those with graduate studies have 0.03pp lower CDI growth than those without a high school certificate. We

observed in survey data. We repeat the Shapley-Owen exercise for CDI levels, shown in Appendix Figure C.11. Strikingly, employment factors become a much more important mediator, but this is because the status of being on UI or disability benefits is a strong predictor of an elevated CDI *level*, more so than predicting subsequent CDI *growth* (see also Figure 9). The increase in the prominence of employment status is offset by lower contributions of education and demographic variables. The contribution of health behaviors, shown in Panel B, is much more precise than for growth, given that the variation in CDI levels is more predictable.

[40] 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."

[41] 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."

[42] Among the socioeconomic factors, education levels explain 22% of the lifetime variation, followed by income & wealth at 17%, and demographics at 13%. In addition to these socioeconomic factors (51%) and health behaviors (31%), we find that sector and spatial mediators explain 7% and 5% respectively, parental health 4%, while employment status and within-firm pay-rank explain less than 2%.

[43] All other admin-based mediators are used as controls for the left-hand panel, and both other admin-based mediators and health behaviour mediators are used as controls for the right-hand panel. A breakdown of these results by gender is shown in Appendix Figure I.1. Similarly, results for specific age groups are shown in Appendix Figure I.2.

find significant municipality effects; being in the worst decile of municipalities corresponds to CDI growth that is 0.02pp higher than when in the top decile of municipalities. The gradient is very similar for sector of employment. Health behaviors matter too, but perhaps less than expected. Being a smoker increases CDI growth by 0.02pp and being obese increases CDI growth by 0.04pp relative to a healthy weight. While the estimated differences in CDI growth control for all other factors observable in the registry data, including socioeconomic 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.[44]

Now the estimates change considerably for more 'naive' approaches that one may be limited to due to data constraints. First, Figure 9 shows that the strength 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 is more sensitive to the reverse causality of health on potential mediators. As was found with the Shapley-Owen analysis, we find that the estimated coefficients on employment factors (e.g., being on social assistance) and health behaviors increase in relative magnitude, but this is likely including some reverse causality from individuals' health on how they behave and what work they can do. 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.[45] The estimated relationship between health and health behaviors, including for smoking and BMI, again becomes substantially larger. However, this is now including the relationship between health and other correlated factors including socioeconomic differences.[46]

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

[44] 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. Furthermore, I.1 indicates that there is a gender component to this effect, where drinking is more protective among women than men.

[45] This is a similar exercise to Figure 6 of Finkelstein, Gentzkow, and Williams (2021), and Figure 8 of Chetty et al. (2016) that study correlates of geographic variation in mortality.

[46] See also Darden, Gilleskie, and Strumpf (2018) who come to a similar conclusion regarding the estimated mortality effect of smoking.

some conclusions may thus have been skewed due to data challenges.[47] The granularity of the data allows us to control for various factors jointly and this shows that while observed health behaviors are strongly correlated with chronic illness, other observable factors are at least as important for explaining the observed variation in health.

7 Conclusion

Mackenbach (2019, p. 178) articulated the long-standing knowledge gap in understanding health and inequality as follows: "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 seeks to make progress to fill this gap in our understanding. We exploit rich and comprehensive data on the entire population of the Netherlands to directly measure health, income and other relevant factors in one and the same setting. This allows for a comprehensive and transparent account of health inequalities and how they arise over the life-cycle.

We have shown that chronic diseases explain a substantial portion of the income gradient in mortality as well as in 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 is linear in income, 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 aims to make advances relative to the related work in epidemiology, either at the national or global level, but 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 (e.g., Wang et al. 2016; Murray et

[47] Appendix Figure C.13 illustrates this further by showing that health behaviors by themselves can explain about half of the gap in CDI levels between high and low income individuals, even at younger ages.

al. 2020). As mentioned, our paper can be seen as a recalibration of the potential importance of specific mechanisms and as such provides an ideal roadmap for further empirical work.

The chronic disease index that 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 socioeconomic differences in mortality, either due to differences in access to healthcare, differences in communicable or differences in acute disease. Both advantages make the CDI particularly valuable for further work.

References

Abdalla, Salma M, Samuel B Rosenberg, Nason Maani, Catalina Melendez Contreras, Shui Yu, and Sandro Galea. 2025. "Income, education, and the clustering of risk in cardiovascular disease in the US, 1999–2018: an observational study." *The Lancet Regional Health–Americas* 44.

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.

Avendano, Mauricio, M Maria Glymour, James Banks, and Johan P Mackenbach. 2009. "Health disadvantage in US adults aged 50 to 74 years: a comparison of the health of rich and poor Americans with that of Europeans." *American journal of public health* 99 (3): 540–548.

Banks, James, Alastair Muriel, and James P Smith. 2010. "Disease prevalence, disease incidence, and mortality in the United States and in England." *Demography* 47 (Suppl 1): S211–S231.

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, Neil Duzett, Adriana Lleras-Muney, Nolan Pope, and Joseph Price. 2024. "Intergenerational Transmission of Lifespan in the US." *NBER Working Paper* 31034 (March).

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.

Bleich, Sara N, Marian P Jarlenski, Caryn N Bell, and Thomas A LaVeist. 2012. "Health inequalities: trends, progress, and policy." *Annual review of public health* 33 (1): 7–40.

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).

Borella, Margherita, Francisco A Bullano, Mariacristina De Nardi, Benjamin Krueger, and Elena Manresa. 2024. *Health inequality and health types.* Technical report. National Bureau of Economic Research.

Campbell, Emma, Ellie Macey, Chris Shine, Vahé Nafilyan, Nathan Cadogan Clark, Piotr Pawelek, Isobel Ward, Andrew Hughes, Veena Raleigh, Amitava Banerjee, et al. 2023. "Sociodemographic and health-related differences in undiagnosed hypertension in the health survey for England 2015–2019: a cross-sectional cohort study." *EClinicalMedicine* 65.

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.

Case, Anne, Darren Lubotsky, and Christina Paxson. 2002. "Economic status and health in childhood: The origins of the gradient." *American economic review* 92 (5): 1308–1334.

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, Yiqun, Petra Persson, and Maria Polyakova. 2022. "The roots of health inequality and the value of intrafamily expertise." *American Economic Journal: Applied Economics* 14 (3): 185–223.

Chetty, Raj, Janet Currie, John N. Friedman, Ines Guix Sauquet, Nathaniel Hendren, Michael Stepner, Harvey Barnhard, Dhruv Gaur, Tyler Jacobson, Emma Lee, et al. 2025. *Growing Class Gaps and Shrinking Race Gaps in Life Expectancy, 2001-2019.* Work in Progress, presented at the National Bureau of Economic Research.

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.

Currie, Janet. 2009. "Healthy, Wealthy, and Wise: Socioeconomic Status, Poor Health in Childhood, and Human Capital Development." *Journal of Economic Literature* 47, no. 1 (March): 87–122. https://doi.org/10.1257/jel.47.1.87. https://www.aeaweb.org/articles?id=10.1257/jel.47.1.87.

Currie, Janet, and Mark Stabile. 2003. "Socioeconomic status and child health: why is the relationship stronger for older children?" *American Economic Review* 93 (5): 1813–1823.

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.

Dalgaard, Carl-Johan, and Holger Strulik. 2014. "Optimal aging and death: understanding the Preston curve." *Journal of the European Economic Association* 12 (3): 672–701.

Danesh, Kaveh, William Parker, Mieke Aarts, Jonathan Kolstad, and Johannes Spinnewijn. 2025. "Socioeconomic differences in cancer incidence, stage, and survival in the Netherlands, 2011-2017." Working Paper.

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 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.

Dharmayat, Kanika, Maria Woringer, Nikolaos Mastellos, Della Cole, Josip Car, Sumantra Ray, Kamlesh Khunti, Azeem Majeed, Kausik K Ray, and Sreenivasa Rao Kondapally Seshasai. 2020. "Investigation of Cardiovascular Health and Risk Factors Among the Diverse and Contemporary Population in London (the TOGETHER Study): Protocol for Linking Longitudinal Medical Records." *JMIR Res Protoc* 9, no. 10 (August): e17548. ISSN: 1929-0748. https://doi.org/10.2196/17548. http://www.ncbi.nlm.nih.gov/pubmed/33006568.

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.

Godøy, Anna, and Ingrid Huitfeldt. 2020. "Regional variation in health care utilization and mortality." *Journal of Health Economics* 71:102254.

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 Kopecky, and Kai Zhao. 2025. "How Important Is Health Inequality for Lifetime Earnings Inequality?" Rdaf030, *The Review of Economic Studies,* https://doi.org/10.1093/restud/rdaf030. https://doi.org/10.1093/restud/rdaf030.

Hosseini, Roozbeh, Karen A. Kopecky, and Kai Zhao. 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.

Khanolkar, Amal R, Nishi Chaturvedi, Valerie Kuan, Daniel Davis, Alun Hughes, Marcus Richards, David Bann, and Praveetha Patalay. 2021. "Socioeconomic inequalities in prevalence and development of multimorbidity across adulthood: a longitudinal analysis of the MRC 1946 national survey of health and development in the UK." *PLoS medicine* 18 (9): e1003775.

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.

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.

Lleras-Muney, Adriana, Hannes Schwandt, and Laura Wherry. 2024. *Poverty and Health.* Working Paper, Working Paper Series 32866. National Bureau of Economic Research, August. https://doi.org/10.3386/w32866. http://www.nber.org/papers/w32866.

Loucks, Eric B, Kristjan T Magnusson, Stephen Cook, David H Rehkopf, Earl S Ford, and Lisa F Berkman. 2007. "Socioeconomic position and the metabolic syndrome in early, middle, and late life: evidence from NHANES 1999–2002." *Annals of epidemiology* 17 (10): 782–790.

Lu, Yuan, Xin Xin, Chungsoo Kim, Jordan Asher, Harlan Krumholz, and John Brush. 2024. "Leveraging Electronic Health Records to Assess Neighborhood Advantages and Risk of Cardiovascular Outcomes Among Hypertensive Patients." *Hypertension* 81 (Suppl_1): AMP20–AMP20.

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.

Martinson, Melissa L. 2012. "Income inequality in health at all ages: a comparison of the United States and England." *American Journal of Public Health* 102 (11): 2049–2056.

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.

Nesson, Erik, and Joshua J Robinson. 2017. "The measurement of health and the connection between health inequality and income." *Available at SSRN 2994011.*

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.

Pais, Jeremy. 2014. "Cumulative structural disadvantage and racial health disparities: The pathways of childhood socioeconomic influence." *Demography* 51:1729–1753.

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.

Poulsen, Melissa N, Annemarie G Hirsch, Lorraine Dean, Jonathan Pollak, Joseph DeWalle, Katherine Moon, Meghann Reeder, Karen Bandeen-Roche, and Brian S Schwartz. 2024. "Community credit scores and community socioeconomic deprivation in association with type 2 diabetes across an urban to rural spectrum in Pennsylvania: a case–control study." *BMJ Public Health* 2 (1).

Russo, Nicolo, Rory McGee, Mariacristina De Nardi, Margherita Borella, and Ross Abram. 2024. *Health Inequality and Economic Disparities by Race, Ethnicity, and Gender.* Working Paper, Working Paper Series 32971. National Bureau of Economic Research, September. https://doi.org/10.3386/w32971. http://www.nber.org/papers/w32971.

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.

Seeman, Teresa, Sharon S Merkin, Eileen Crimmins, Brandon Koretz, Susan Charette, and Arun Karlamangla. 2008. "Education, income and ethnic differences in cumulative biological risk profiles in a national sample of US adults: NHANES III (1988–1994)." *Social science & medicine* 66 (1): 72–87.

Shui, Ailun, Gerard J van den Berg, Jochen O Mierau, and Laura Viluma. 2025. "Lifetime Trajectories and Drivers of Socioeconomic Health Disparities: Evidence from Longitudinal Biomarkers in the Netherlands."

Singh, Gopal K, Gem P Daus, Michelle Allender, Christine T Ramey, Elijah K Martin, Chrisp Perry, Andrew A De Los Reyes, and Ivy P Vedamuthu. 2017. "Social determinants of health in the United States: addressing major health inequality trends for the nation, 1935-2016." *International Journal of MCH and AIDS* 6 (2): 139.

Smith, James P. 2004. "Unraveling the SES: health connection." *Population and development review* 30:108–132.

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.

Timmermans, Erik J, Jeroen Lakerveld, Joline WJ Beulens, Dorret I Boomsma, Sophia E Kramer, Mirjam Oosterman, Gonneke Willemsen, Mariska Stam, Giel Nijpels, Carlo Schuengel, et al. 2018. "Cohort profile: the geoscience and health cohort consortium (GECCO) in the Netherlands." *BMJ open* 8 (6): e021597.

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.

World Health Organization. 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.

Zaninotto, Paola, George David Batty, Sari Stenholm, Ichiro Kawachi, Martin Hyde, Marcel Goldberg, Hugo Westerlund, Jussi Vahtera, and Jenny Head. 2020. "Socioeconomic inequalities in disability-free life expectancy in older people from England and the United States: a cross-national population-based study." *The journals of gerontology: Series A* 75 (5): 906–913.

Zhu, Yinjie, Louise H Dekker, and Jochen O Mierau. 2023. "Socio-economic gradients in diagnosed and undiagnosed type 2 diabetes and its related health complications." *Nutrition, Metabolism and Cardiovascular Diseases* 33 (1): 90–94.

A Tables

Table 1: 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 2: 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 in our dynamic decomposition. The effects are expressed as the change in the CDI gap between low and high income individuals for 10-year age bins, multiplied by 100 (i.e., percentage point changes in the CDI). More detail on the dynamic decomposition is provided in Appendix Section G. See Appendix Figure G.3 and Appendix Table G.1 for the ageing and sorting effects for different income quintiles and education groups.

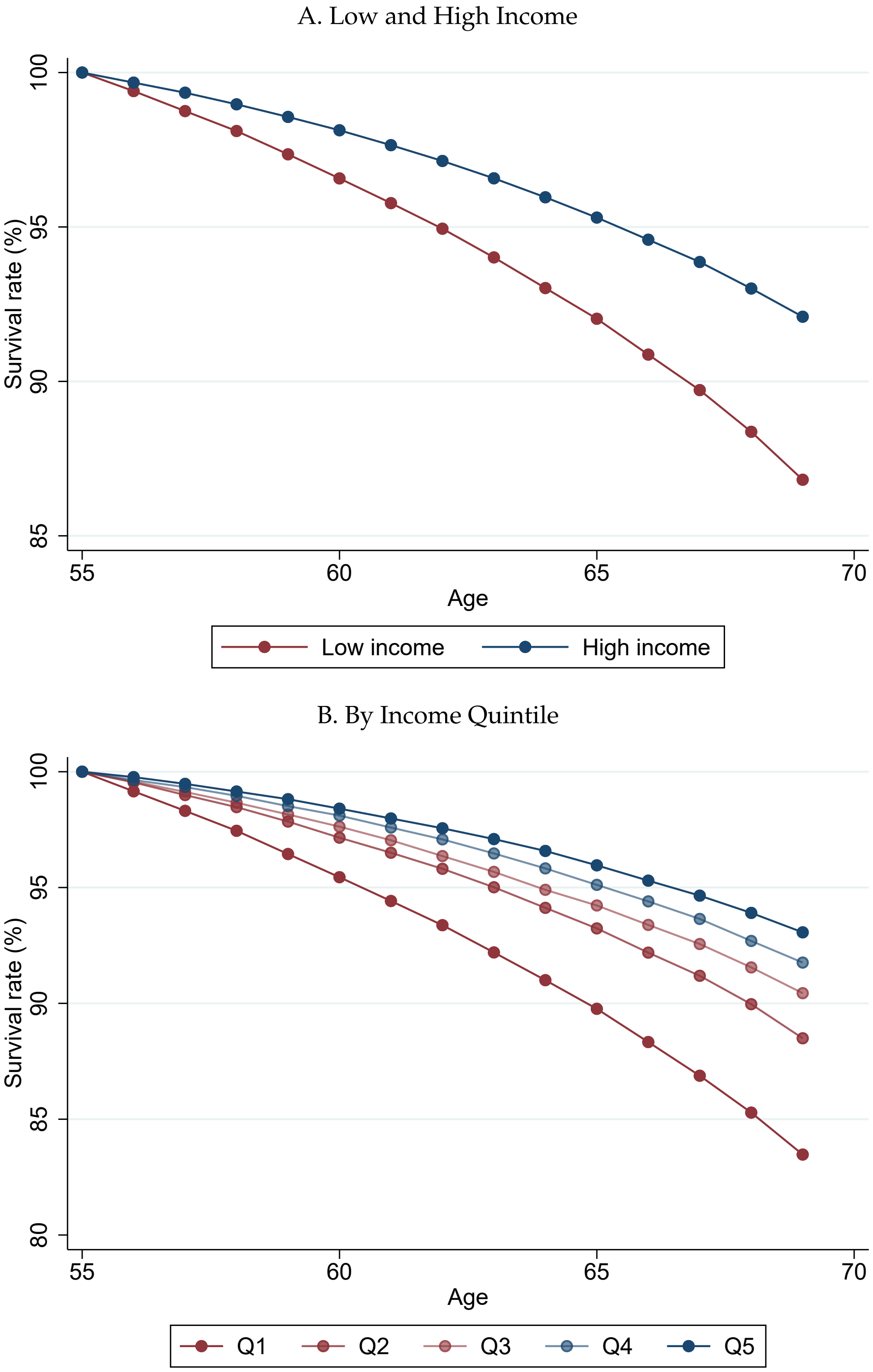
Table 3: Counterfactual Analysis

	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: The table reports simulated biological ages, life expectancy, and lifetime cost figures. In the biological age calculations, the CDI is simulated over the life-cycle 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



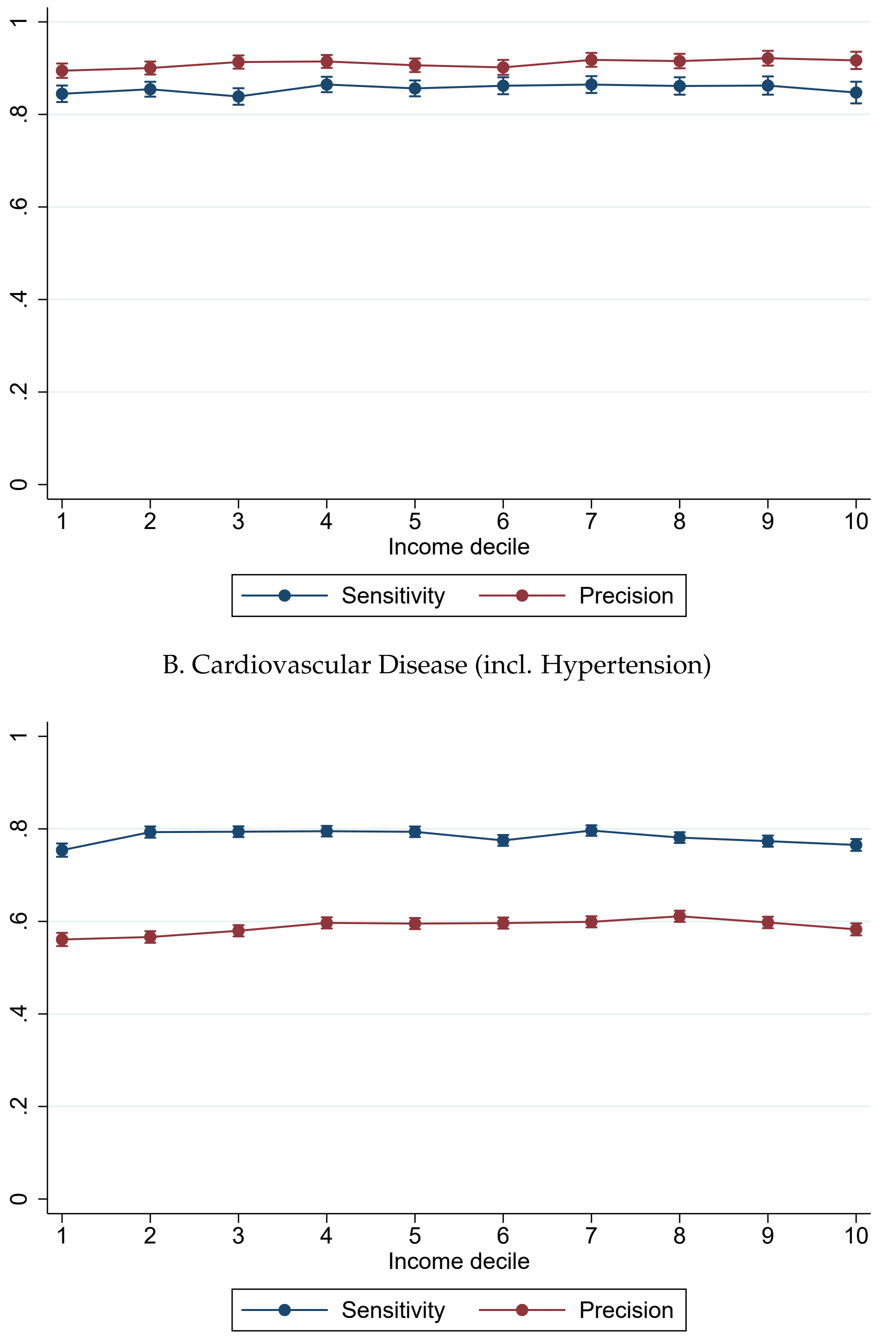
A. Low and High Income

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

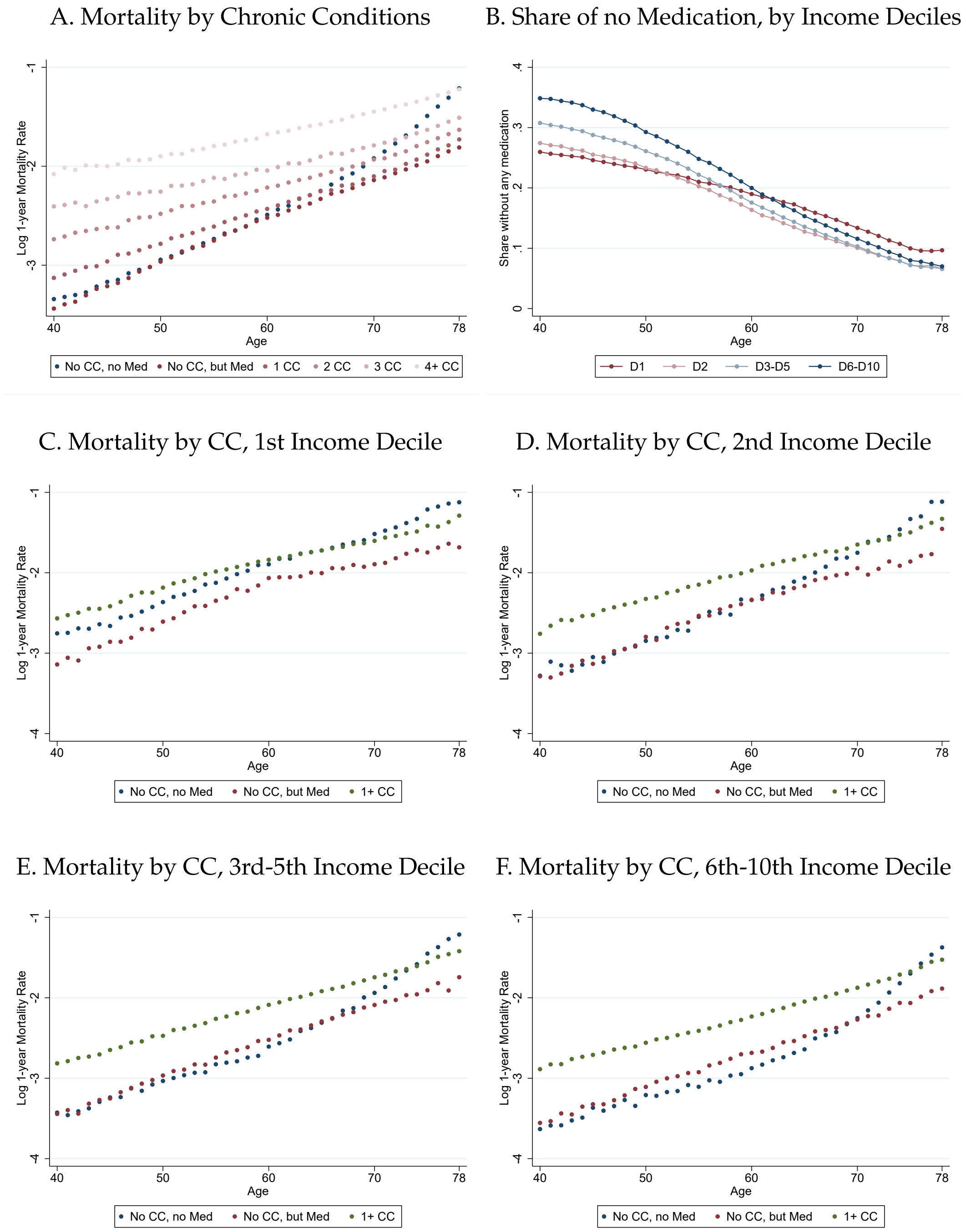
A. Diabetes



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).

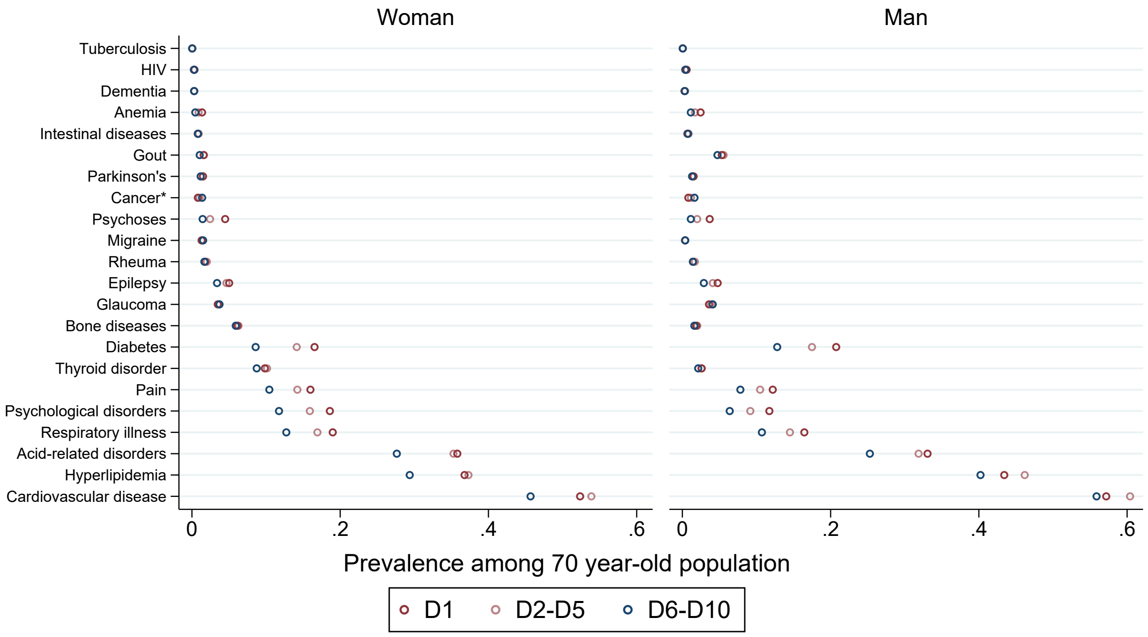
Figure 3: Assessing Under-Diagnosis by Income Deciles



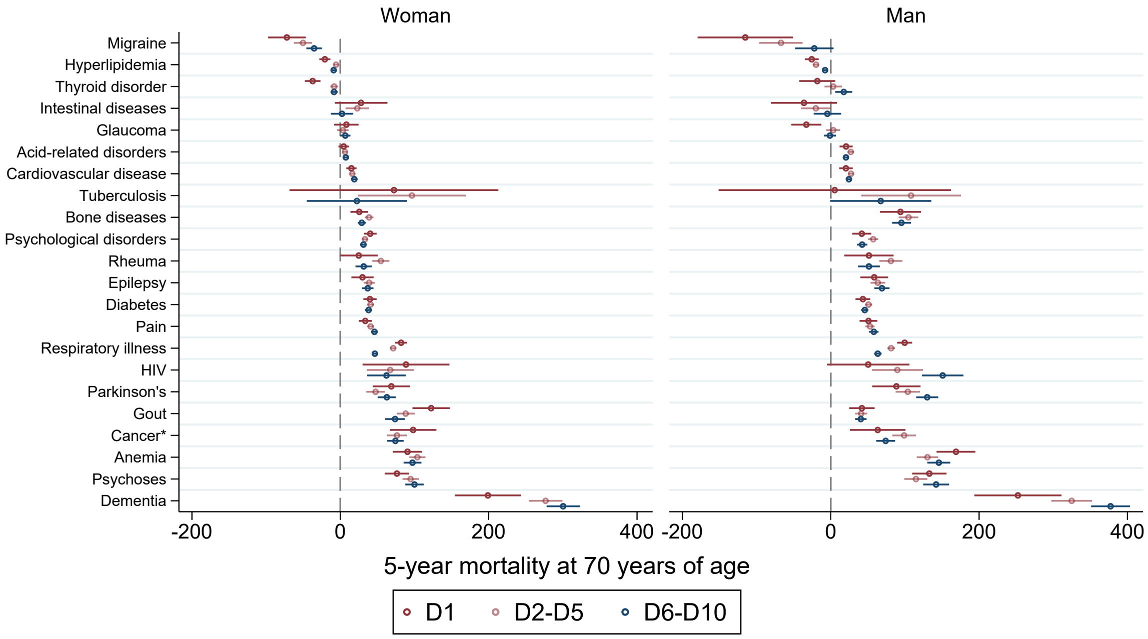
Note: Panel A reports one-year mortality by number of chronic conditions on a logarithmic scale. The category without chronic conditions is divided into a group taking some prescription medication for non-chronic illnesses and a group with individuals taking no prescription medication at all. This panel 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 prescription 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

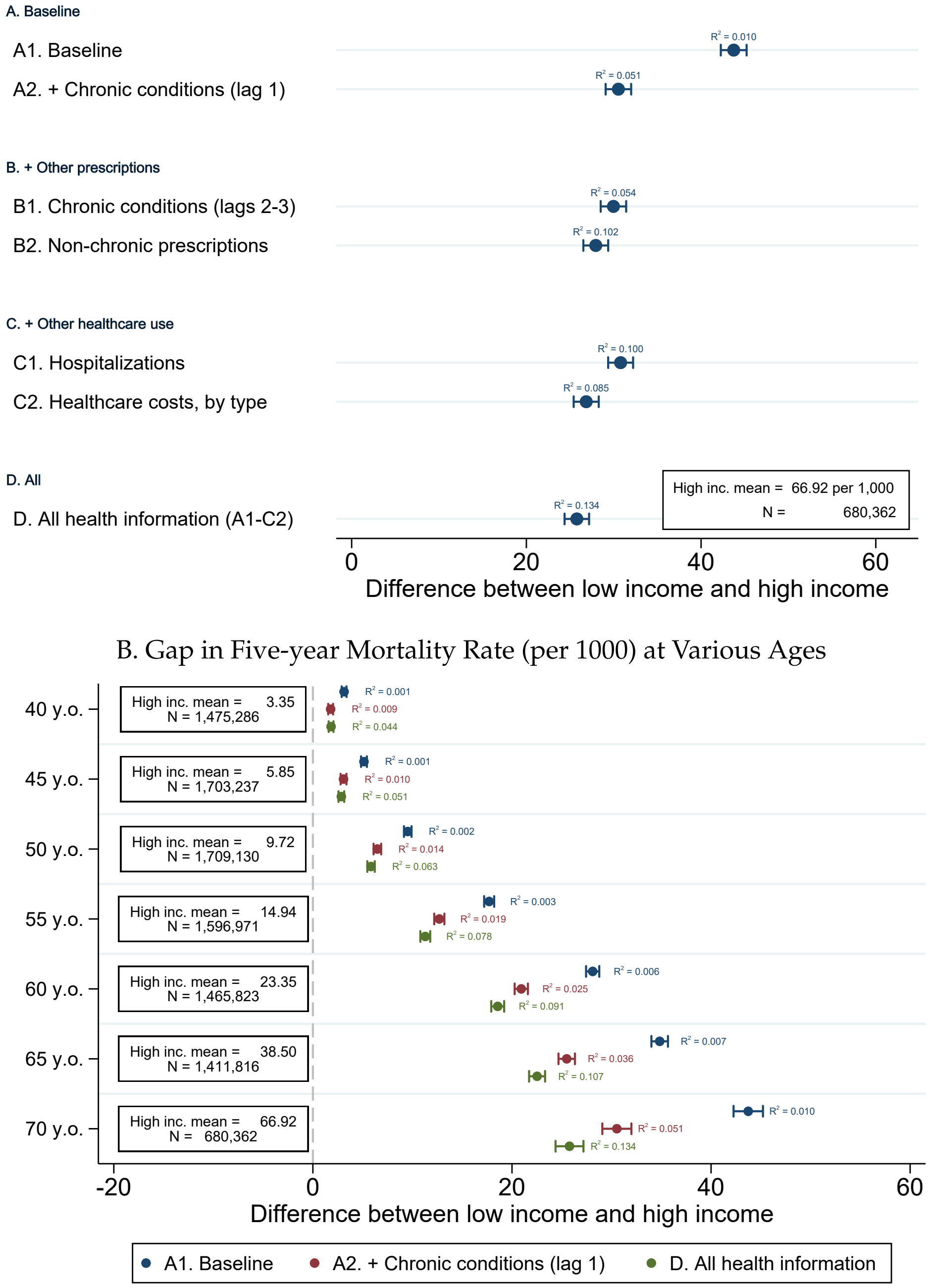


B. Effect of Chronic Conditions on Five-Year Mortality at Age 70 (per 1000)



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). We pool all observations for years between 2013 and 2021. 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

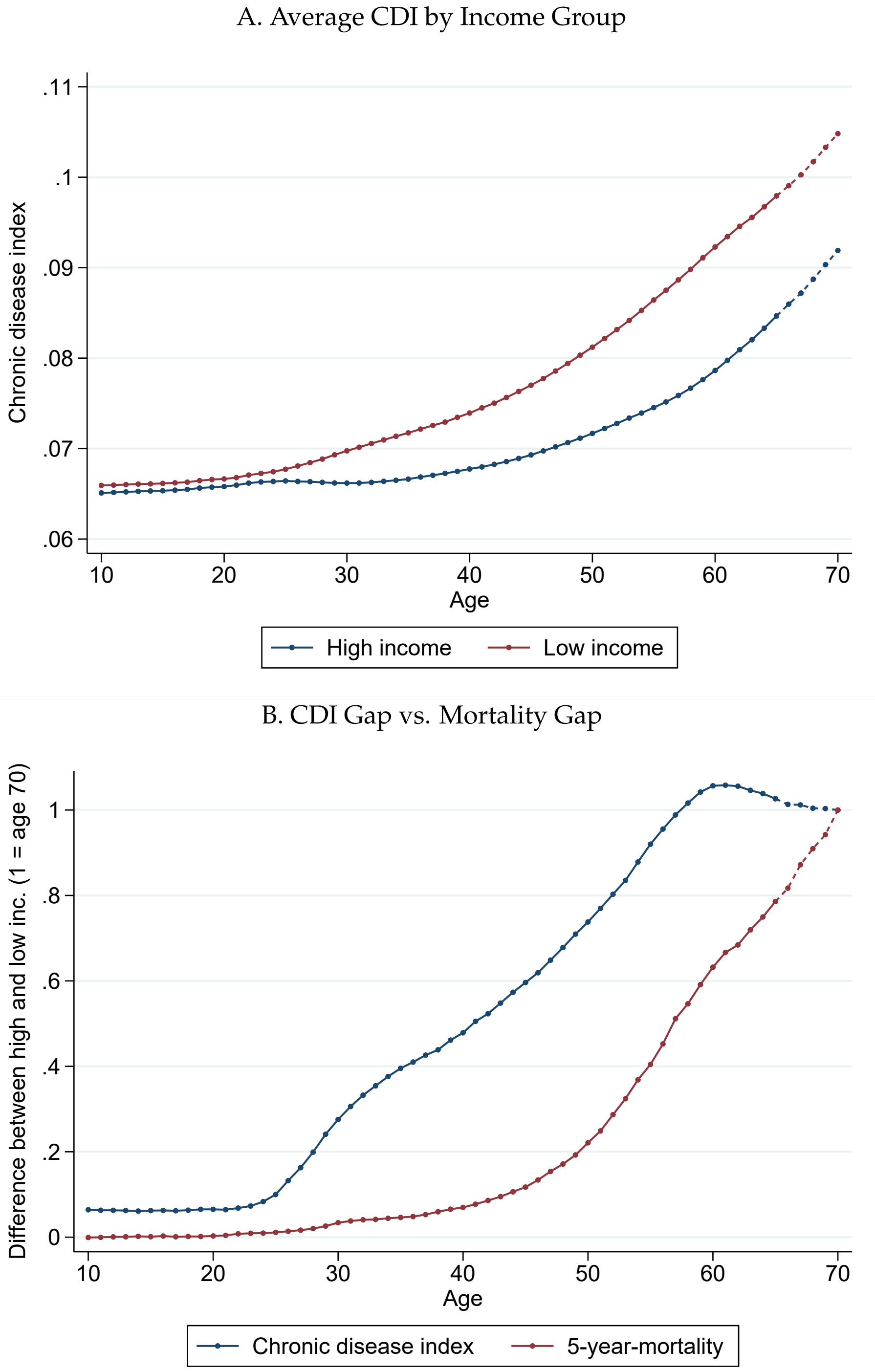


A. Gap in Five-year Mortality Rate (per 1000) at 70 Years of Age

B. Gap in Five-year Mortality Rate (per 1000) at Various Ages

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. We pool the observations for years between 2013 and 2016. For more information, refer to Appendix Section E.

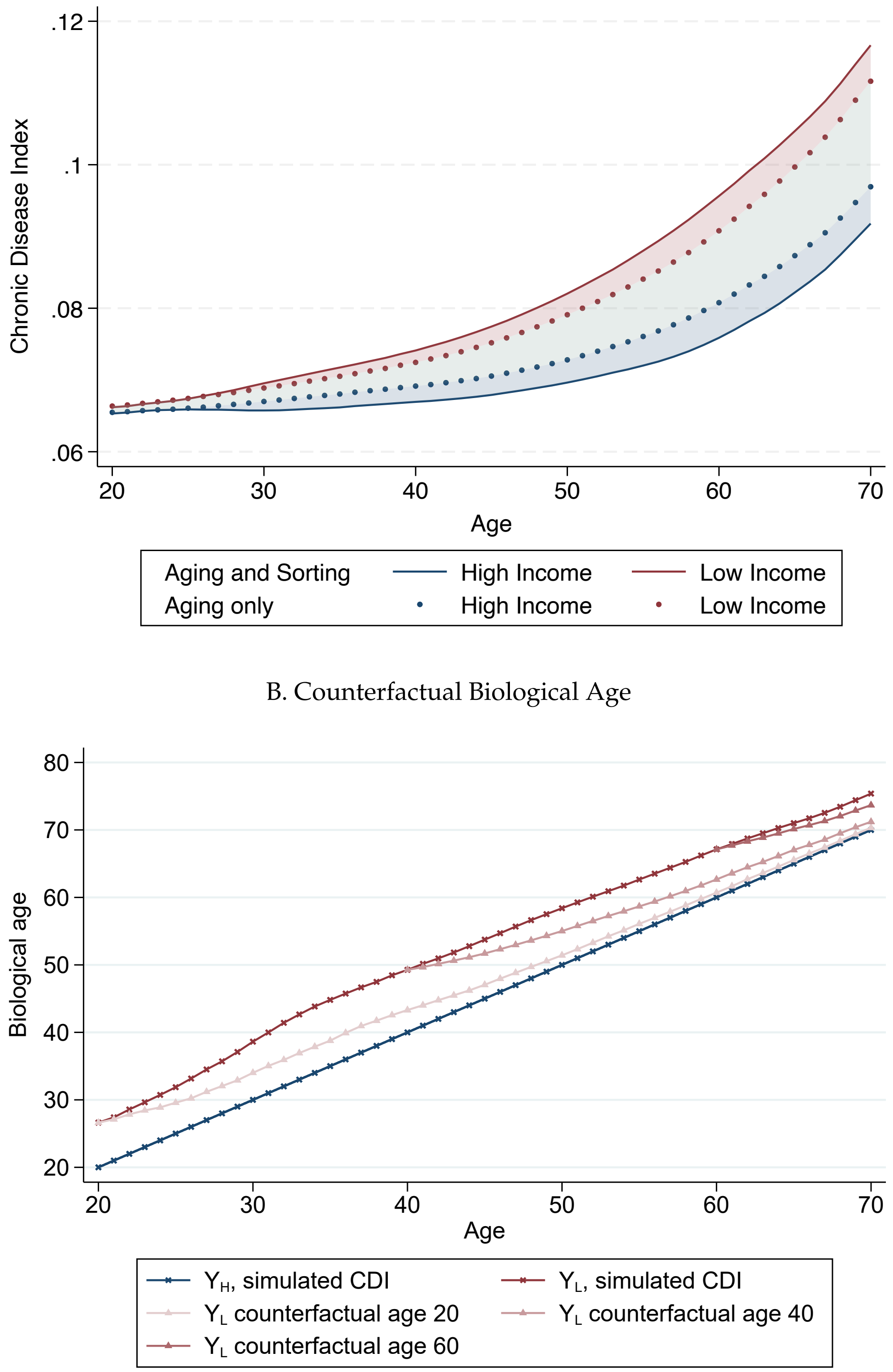
Figure 6: Health Gap over the Lifecycle


A. Average CDI by Income Group

B. CDI Gap vs. Mortality Gap

Note: Panel A plots the average chronic disease index by income group over the life-cycle. 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

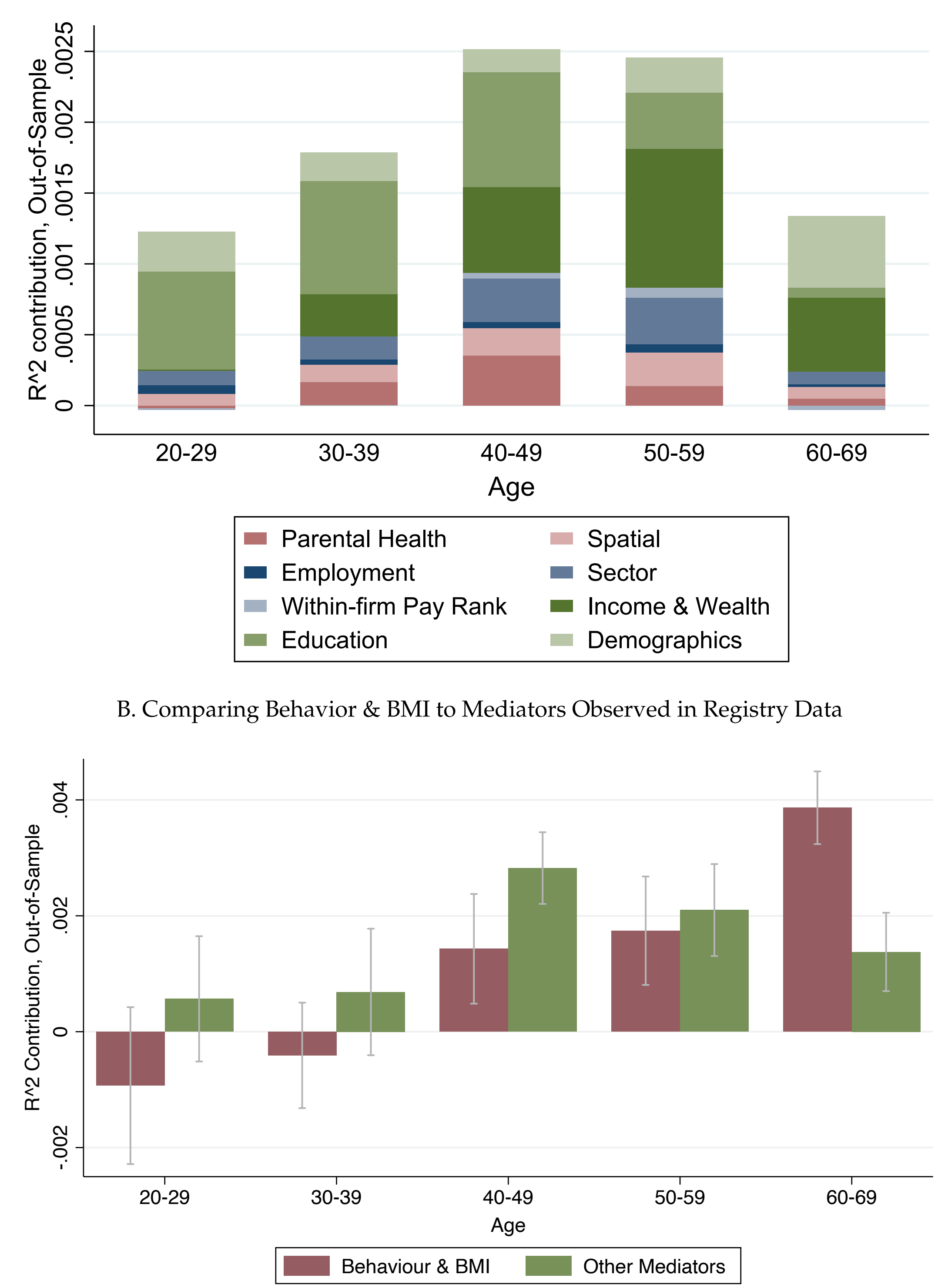


A. Differential Aging and Health-based Sorting

B. Counterfactual Biological Age

Note: Panel A shows the evolution of the CDI when it is simulated using either a) only aging effects, or b) aging and health-based sorting effects. The simulated CDI's start from the observed CDI at age 10, and use the components defined in equation (12), to simulate the CDI at all later ages. The teal shaded area represents the health gap due to differential aging. The blue and red areas are the gaps due to positive and negative sorting, for high and low incomes, respectively. Panel B shows biological ages for different scenarios. In the baseline scenario, the high and low income CDI are simulated based on their respective aging effects. In the counterfactual scenarios, the high income aging effect is used to simulate the Low Income CDI from different ages onwards. Table 3 shows the impact of these counterfactuals on life expectancy and lifetime health expenditures.

Figure 8: Shapley-Owen decomposition of CDI growth

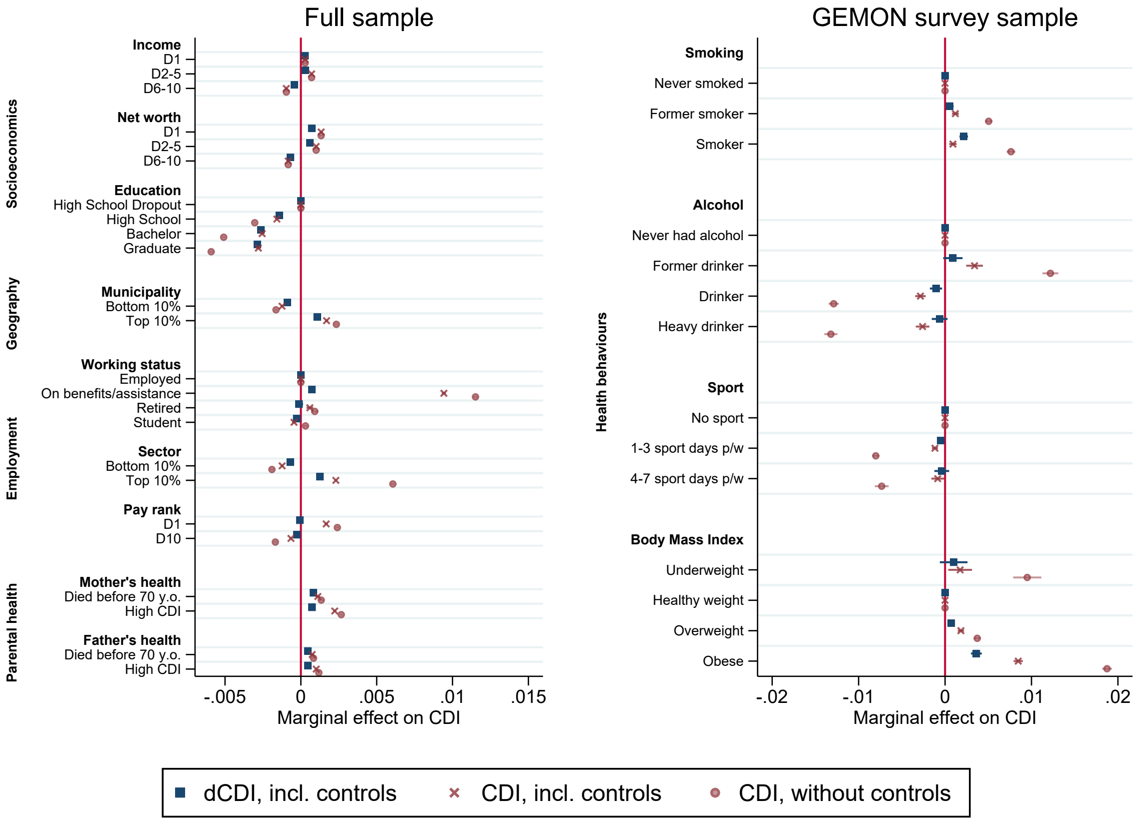


A. Decomposing Mediators Observed in Registry Data

B. Comparing Behavior & BMI to Mediators Observed in Registry Data

Note: The figure shows the results of a Shapley-Owen decomposition of five-year log CDI growth on age, gender, and the sets of mediators reported in the legend, based on equation (14). Separate decompositions are carried out for each age bin. The stacked bars in Panel A represent the contribution of each set of mediators to the overall R^2. Detailed information on the mediators in each group is provided in section 6.2. Panel B presents results from a two-stage Shapley-Owen approach, where the behavior & BMI contributions are estimated on the health survey subsample, and the contribution of all other mediators are estimated on the full population sample. Information on the estimation approach and 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 "dCDI, incl. controls" regresses five-year CDI growth ($dCDI$) on the comprehensive set of controls used in the Shapley-Owen Decomposition, as reported in the figure. Specification "CDI, incl. controls" uses the CDI level (CDI) rather than CDI growth ($dCDI$) as dependent variable and uses the same comprehensive set of controls. Finally, specification "CDI, without controls" shows the coefficients from separate regressions of the CDI level on each mediator separately. All regressions control for age and gender fixed effects.

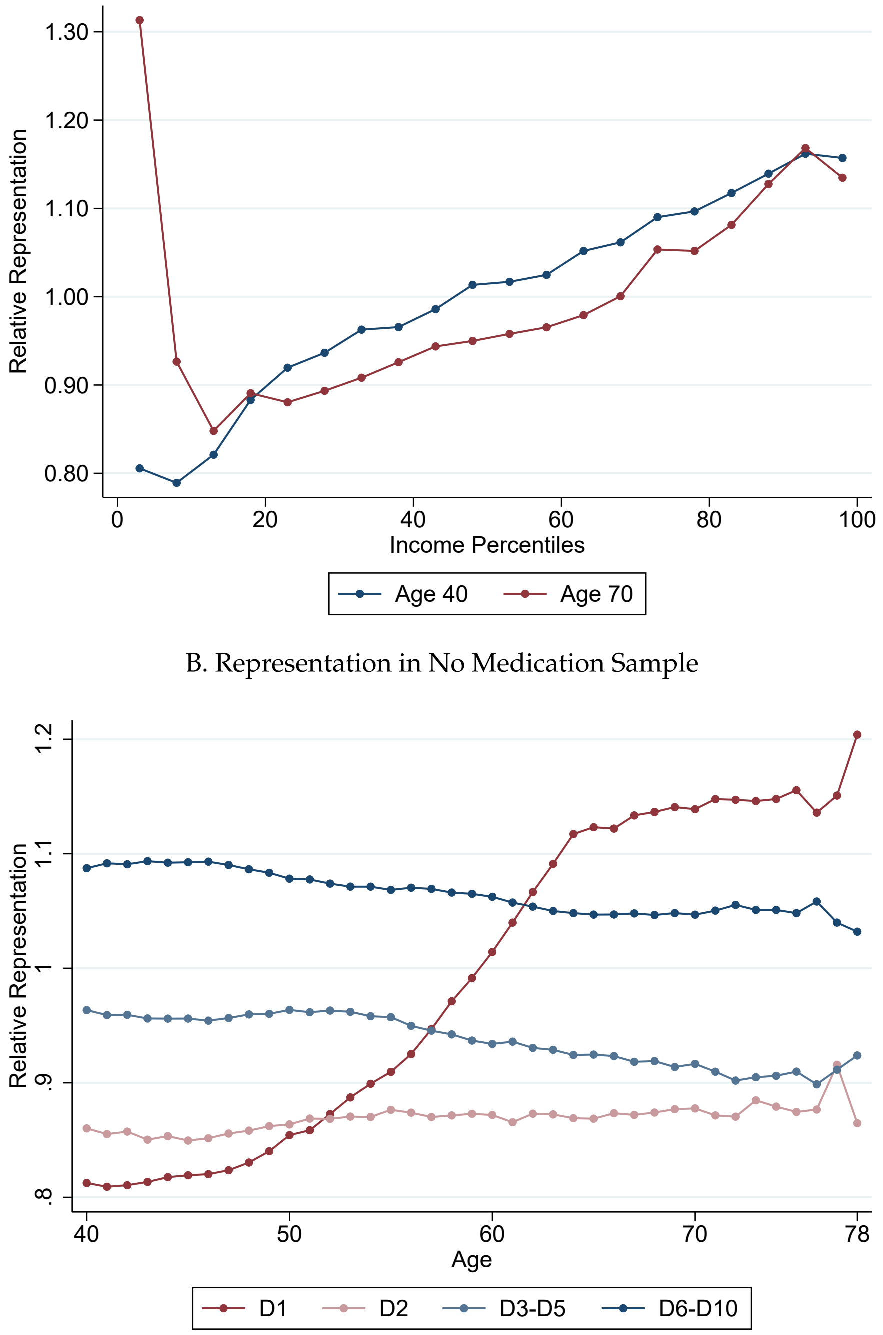
C Appendix: Additional Tables and Figures

Table C.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 our analysis sample, at selected ages and in selected calendar years.

Figure C.1: Under-Diagnosis in First Income Decile

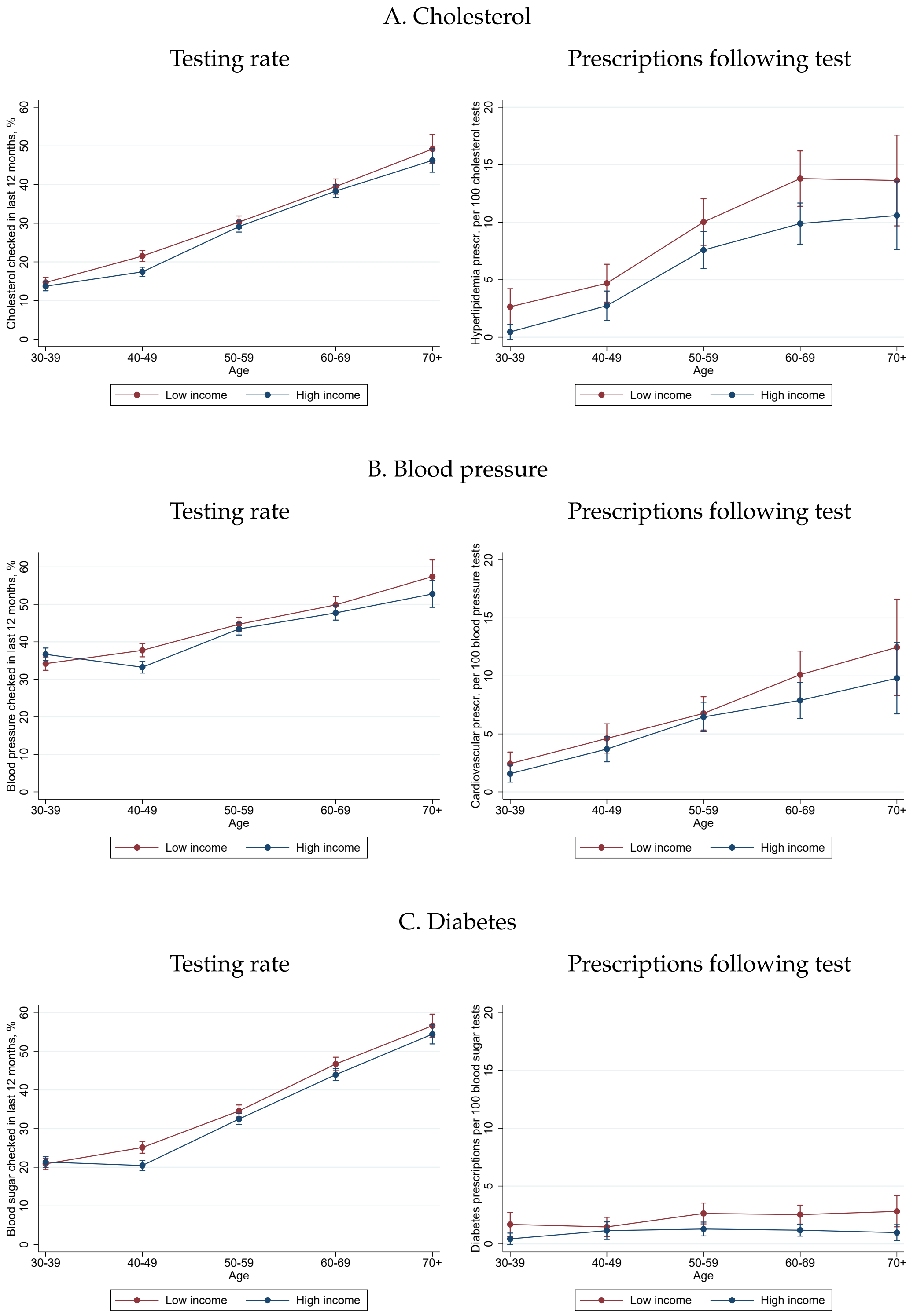


A. Relative Representation in No Medication Group

B. Representation in No Medication Sample

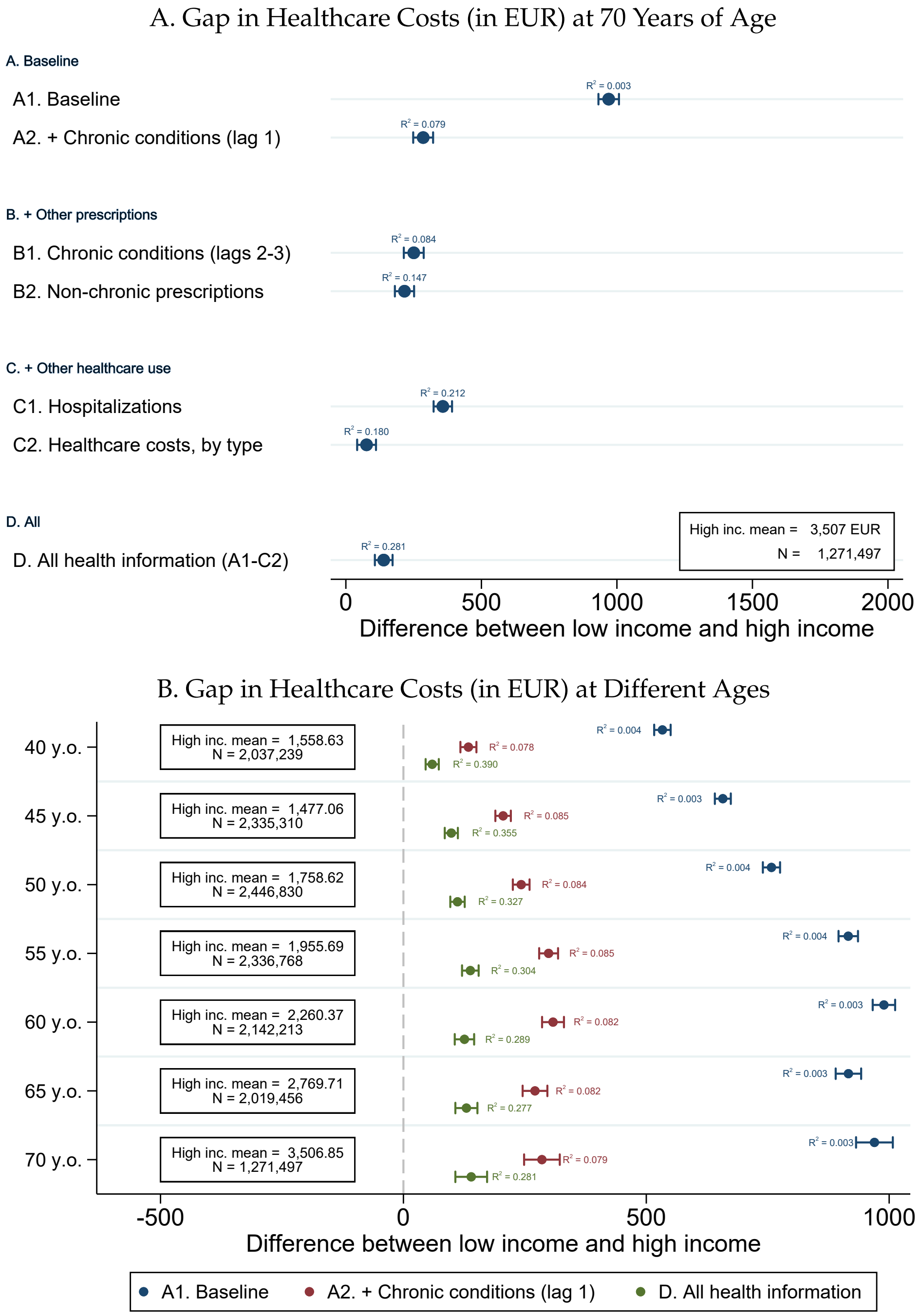
Note: This figure presents evidence for under-diagnosis among very low incomes. Panel A reports relative representation in the sample of people without any prescription medication by income ventile at age 40 and age 70. Relative representation is defined as the share of people within the no prescription 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 prescription 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 report responses to the question "*Have you had a* [Cholesterol/Blood pressure/Blood sugar] *test in the past 12 months?*". The right-hand panels then plot the share of those who report being tested in the past 12 months 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

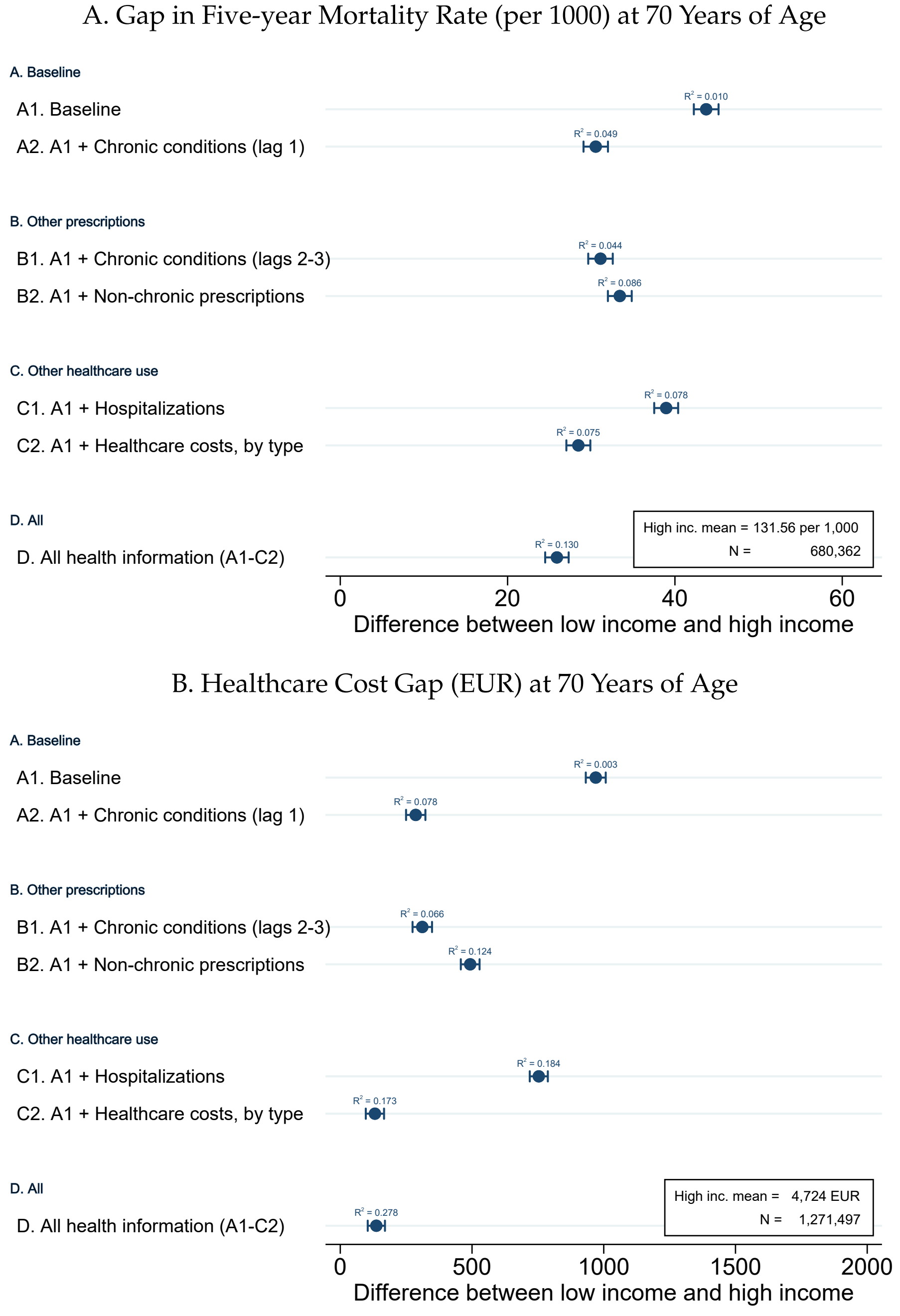


A. Gap in Healthcare Costs (in EUR) at 70 Years of Age

B. Gap in Healthcare Costs (in EUR) at Different Ages

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.

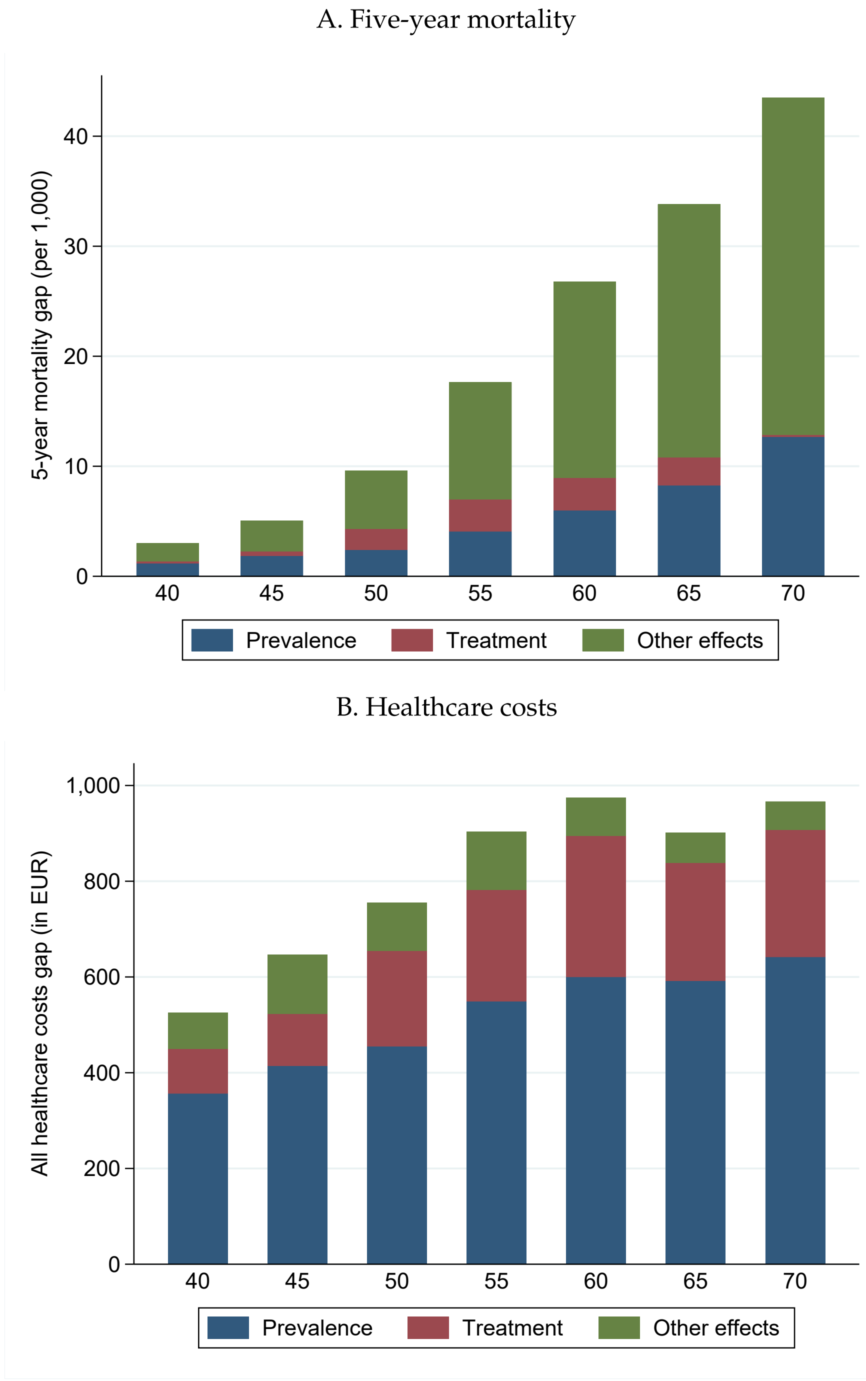
Figure C.4: Mortality and Healthcare Cost Gap, Separate Controls


A. Gap in Five-year Mortality Rate (per 1000) 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 coefficients 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

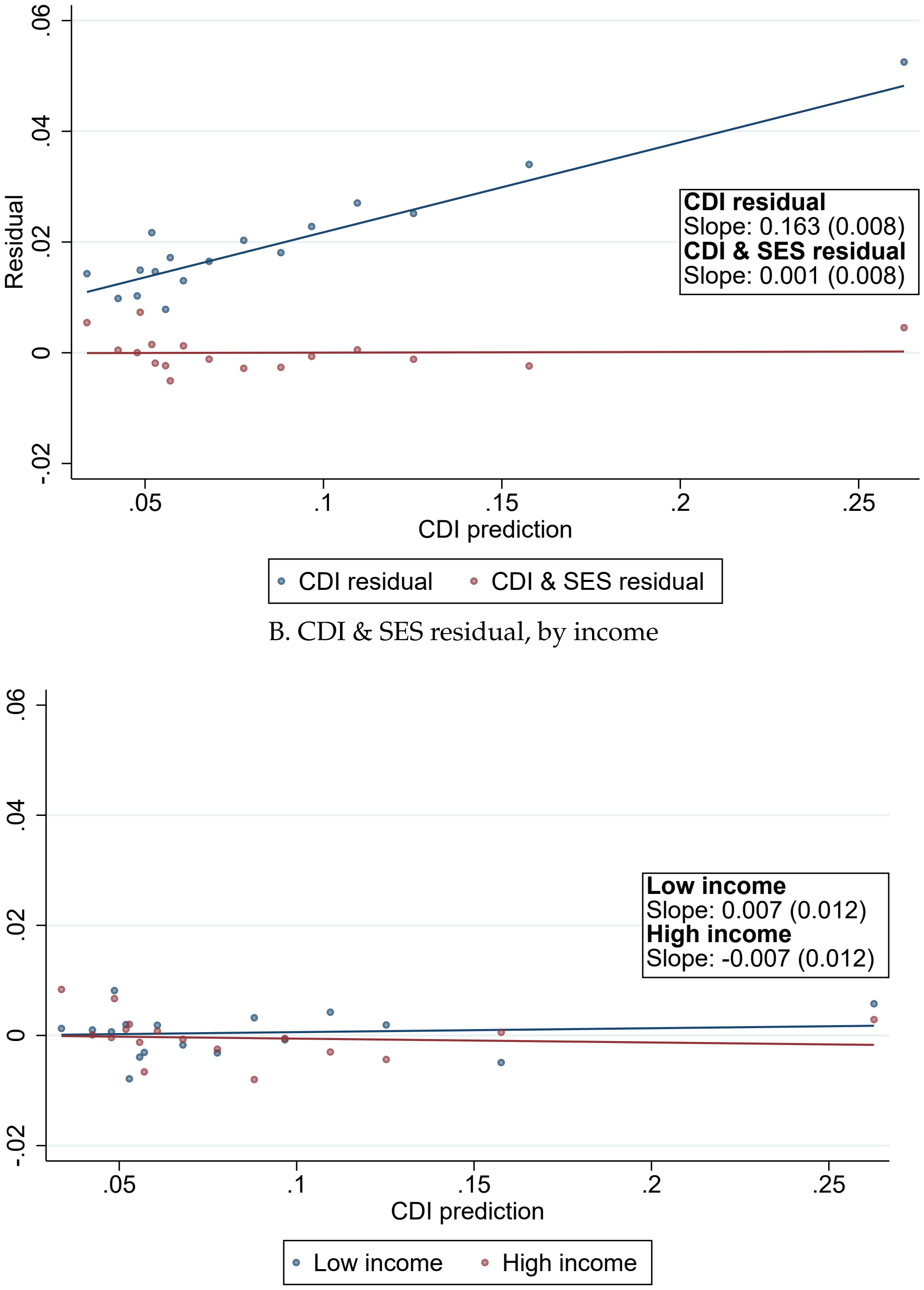

A. Five-year mortality

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 "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

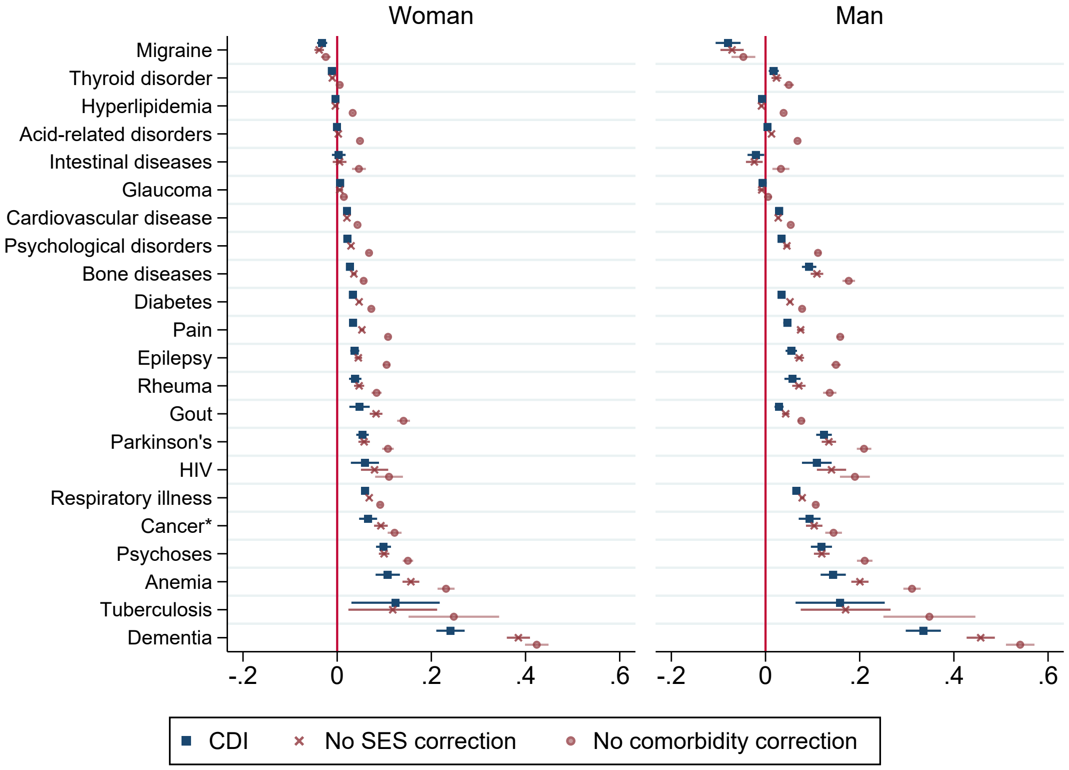
A. Residual mortality risk, controlling for CDI alone or CDI & SES



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 (9): $E\left[\hat{\zeta}_i \mid CDI_i\right]$. 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

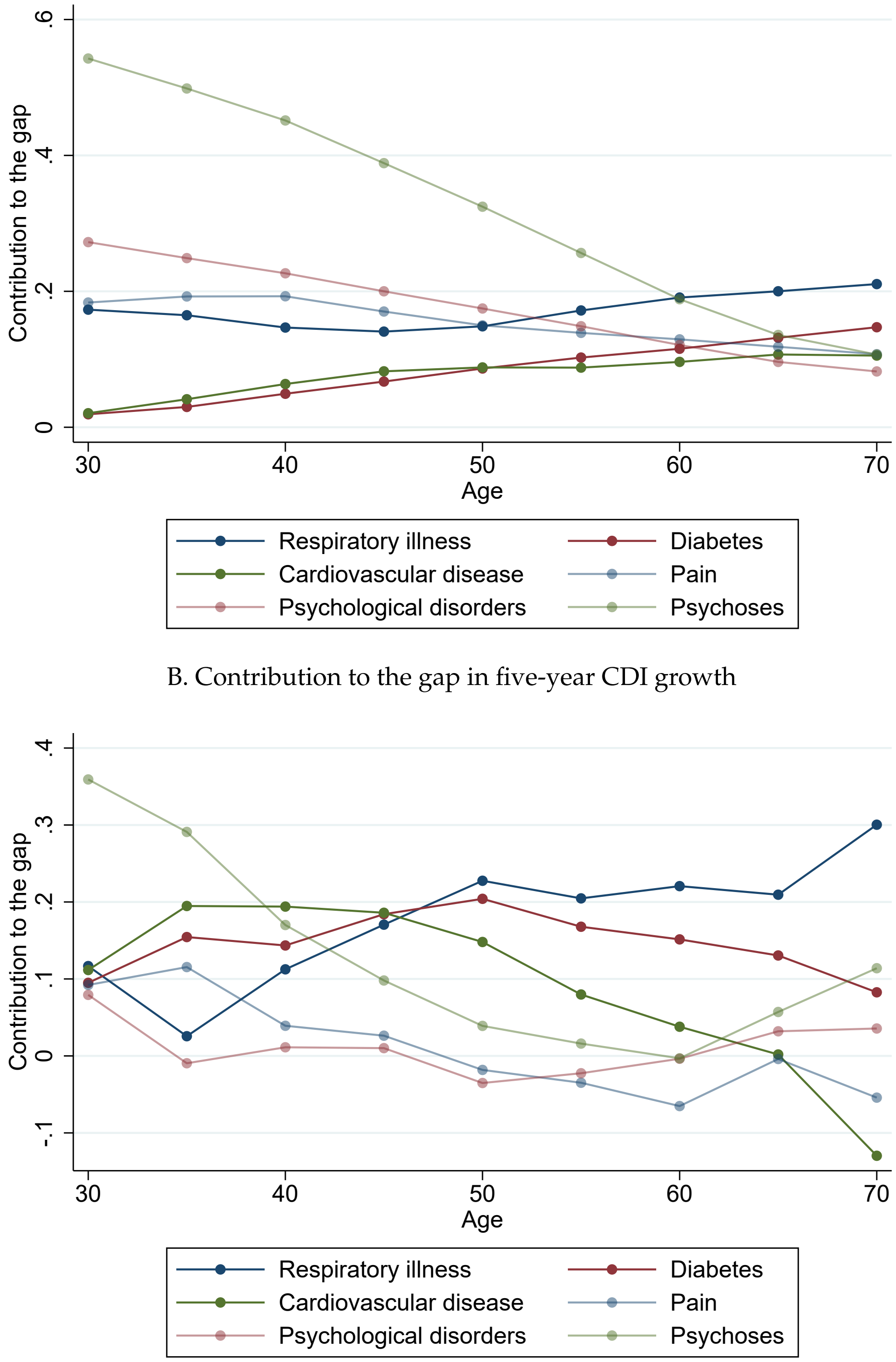


Note: This figure presents the marginal effects of each chronic condition on predicted CDI, by gender. The marginal effects are defined for each gender as follows:

$\beta^j = E[CDI_{70} \mid c^j_{69} = c^j_{68} = c^j_{67} = 1, c^{-j} = \overline{c^{-j}}] - E[CDI_{70} \mid c^j_{69} = c^j_{68} = c^j_{67} = 0, c^{-j} = \overline{c^{-j}}] \quad \forall j = 1, ..., 22.$

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

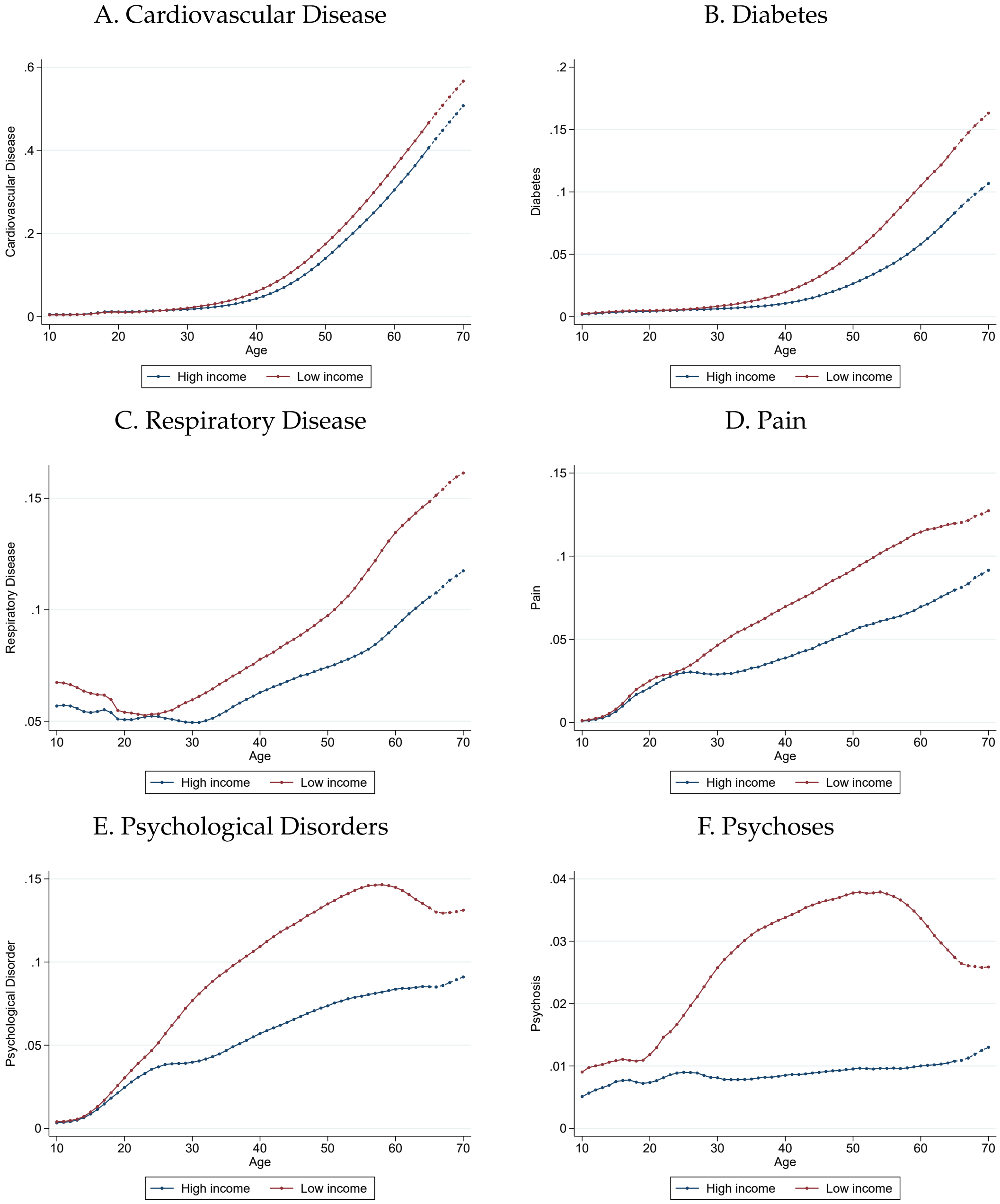


A. Contribution to the gap in CDI levels

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



A. Cardiovascular Disease

B. Diabetes

C. Respiratory Disease

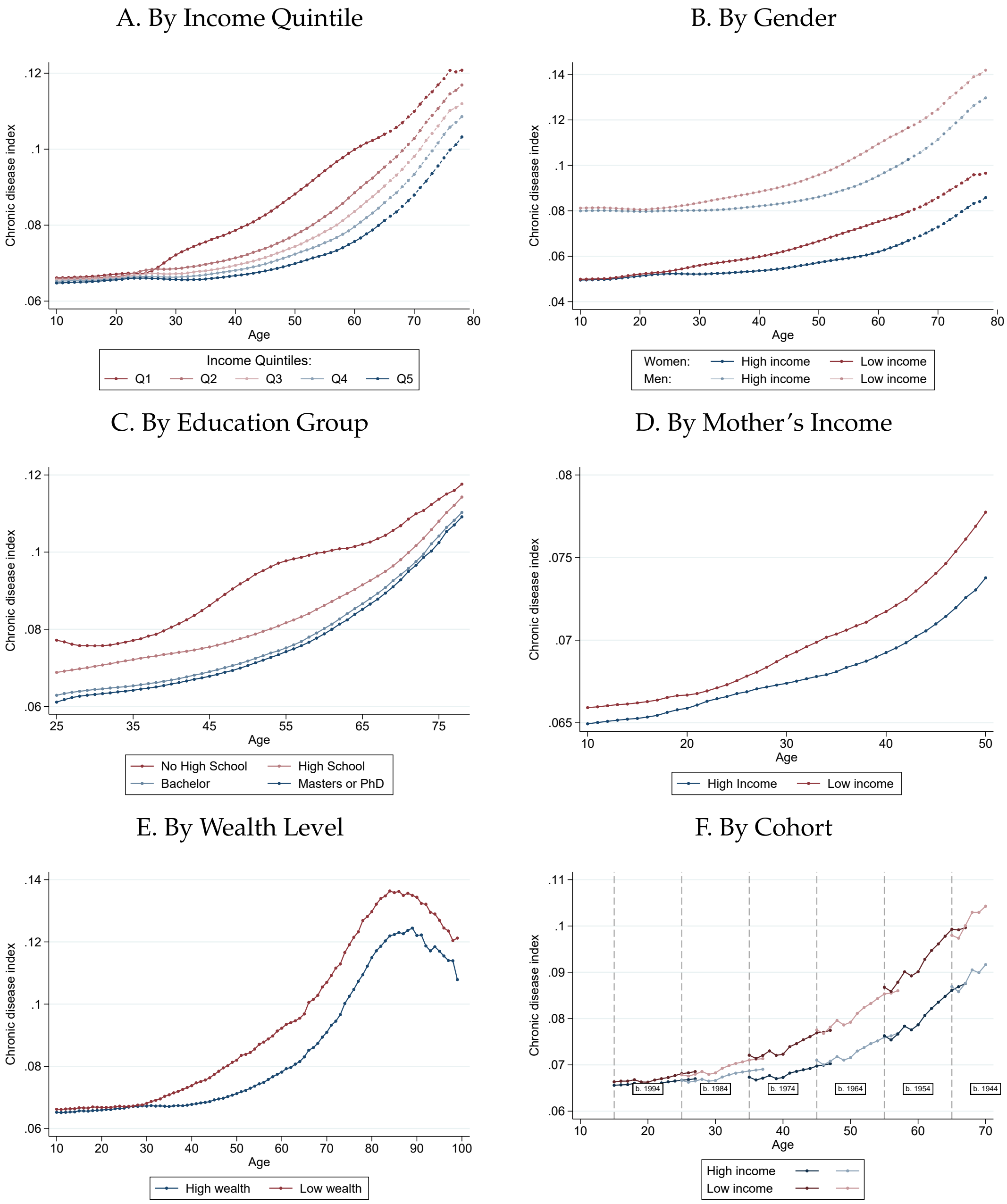
D. Pain

E. Psychological Disorders

F. Psychoses

Note: All panels show the life-cycle 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


A. By Income Quintile

B. By Gender

C. By Education Group

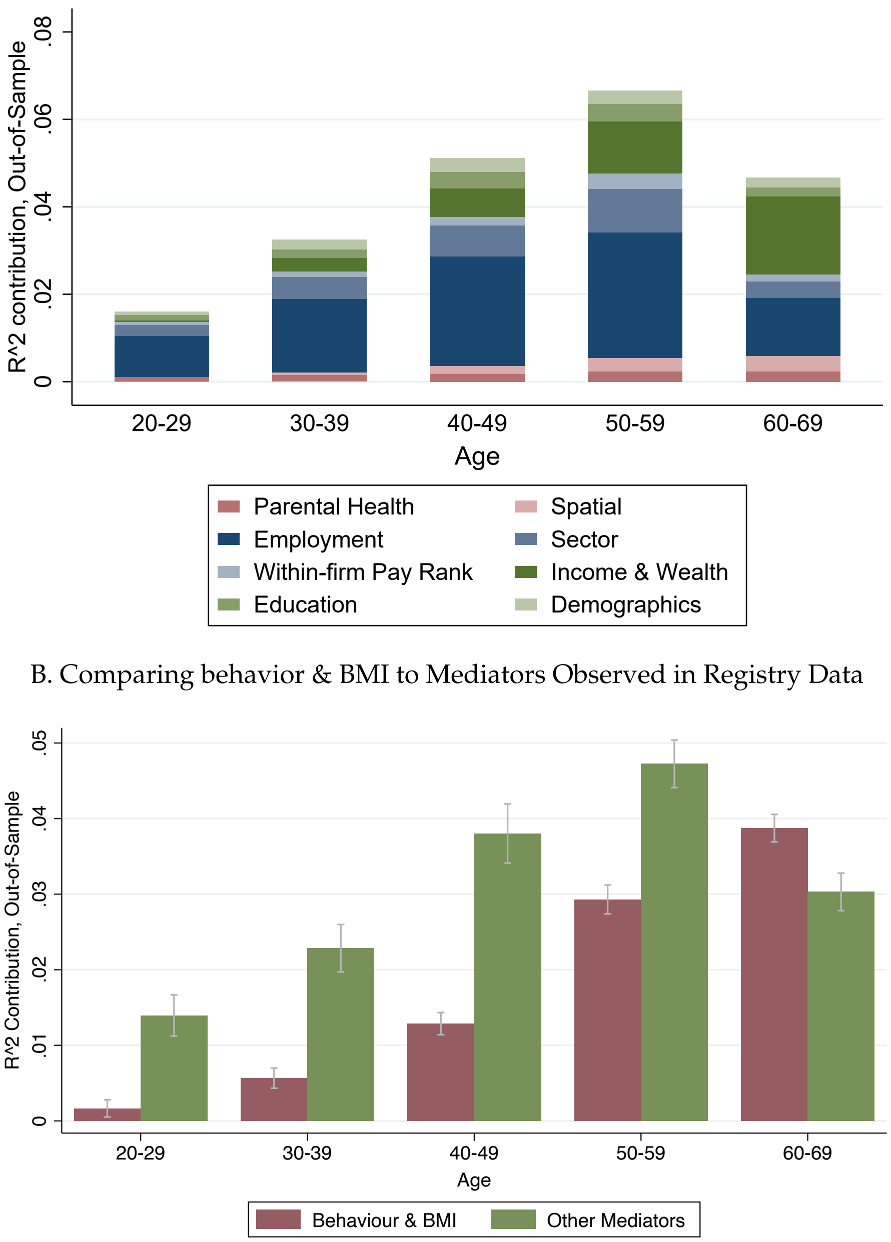
D. By Mother's Income

E. By Wealth Level

F. By Cohort

Note: This figure shows the evolution of the CDI across different subgroups and socioeconomic 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.

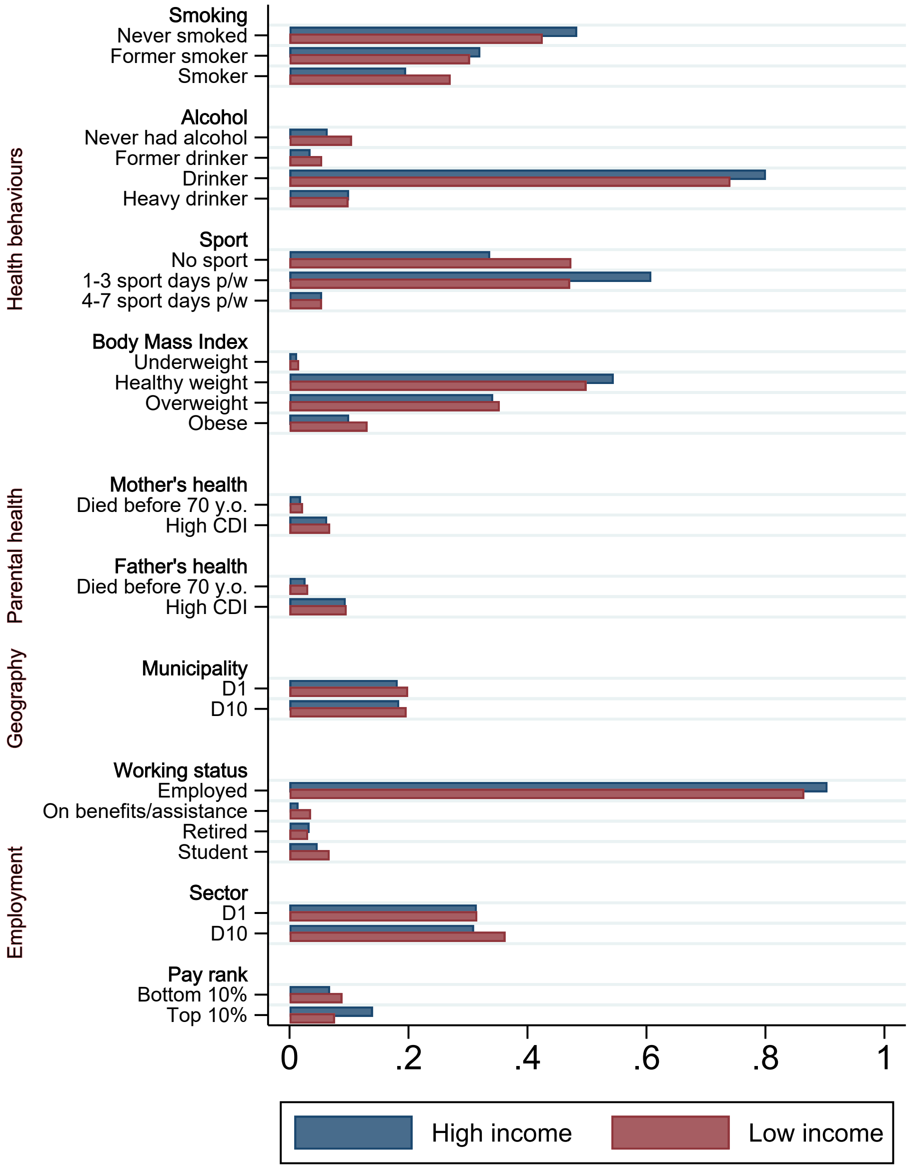
Figure C.11: Shapley-Owen Decomposition Using CDI Levels


A. Decomposing Mediators Observed in Registry Data

B. Comparing behavior & BMI to Mediators Observed in Registry Data

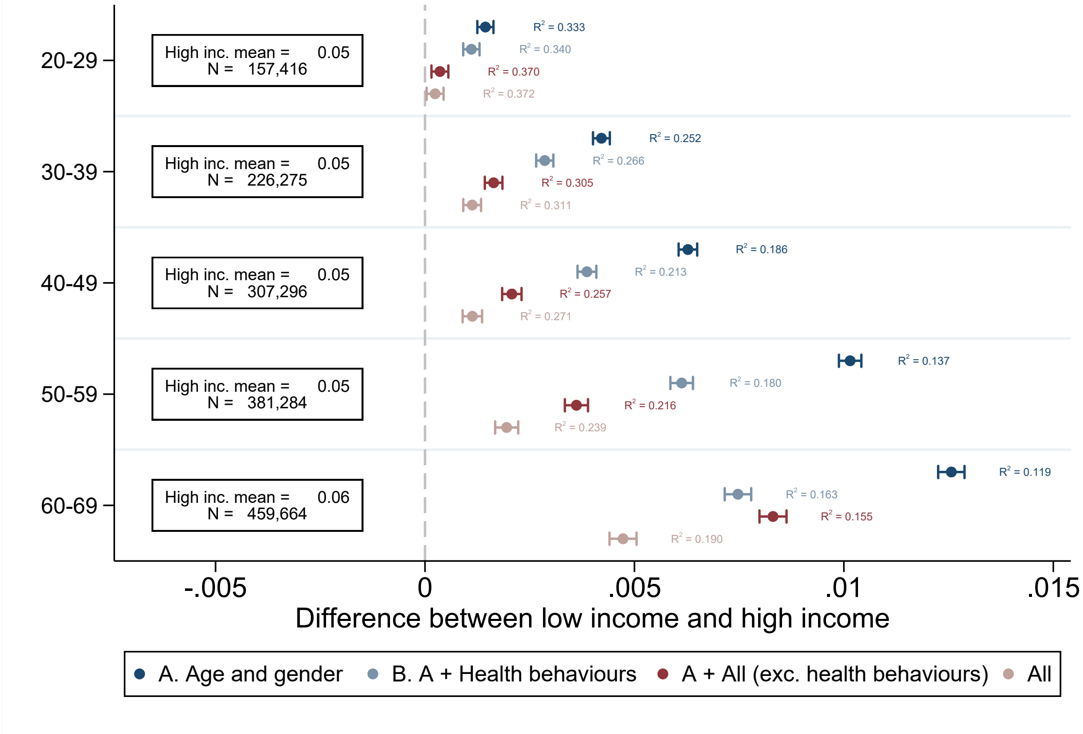
Note: This shows the results of a Shapley-Owen decomposition of the CDI levels at time t on age, gender, and the set of mediators reported in the legend. For both panels, separate decompositions are carried out for each age bin. The stacked bars in Panel A represent the contribution of each set of mediators to the overall R^2. Mediators are treated as indicators. Detailed information on the mediators in each group is provided in section 6.2:. Panel B presents results from a two-stage Shapley-Owen approach, where the behavior & BMI contributions are estimated on the *Gezondheidsmonitor* survey subsample, and all other mediator contributions are estimated on the full population sample. Information on the estimation approach and 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 "dCDI, incl. controls" 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 range shown in the vertical 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 indicator 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 behaviors 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 behaviors. The last set of regressions controls for all the predictors used in the Shapley-Owen decompositions.

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.

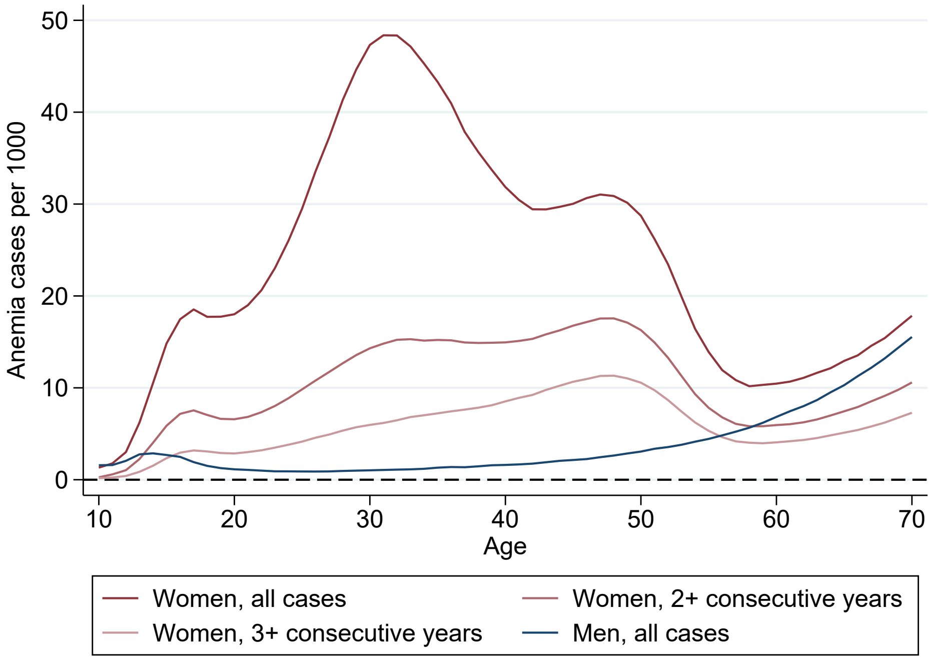
- **HIV**: Huber et al. (2013) use J05AE (protease inhibitors), J05AG (Non-nucleoside reverse transcriptase inhibitors) and J05AR (Antivirals for treatment of HIV infections, combinations). We use J05A (direct acting antivirals). To reduce the number of false positives due to antivirals for acute conditions, 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.

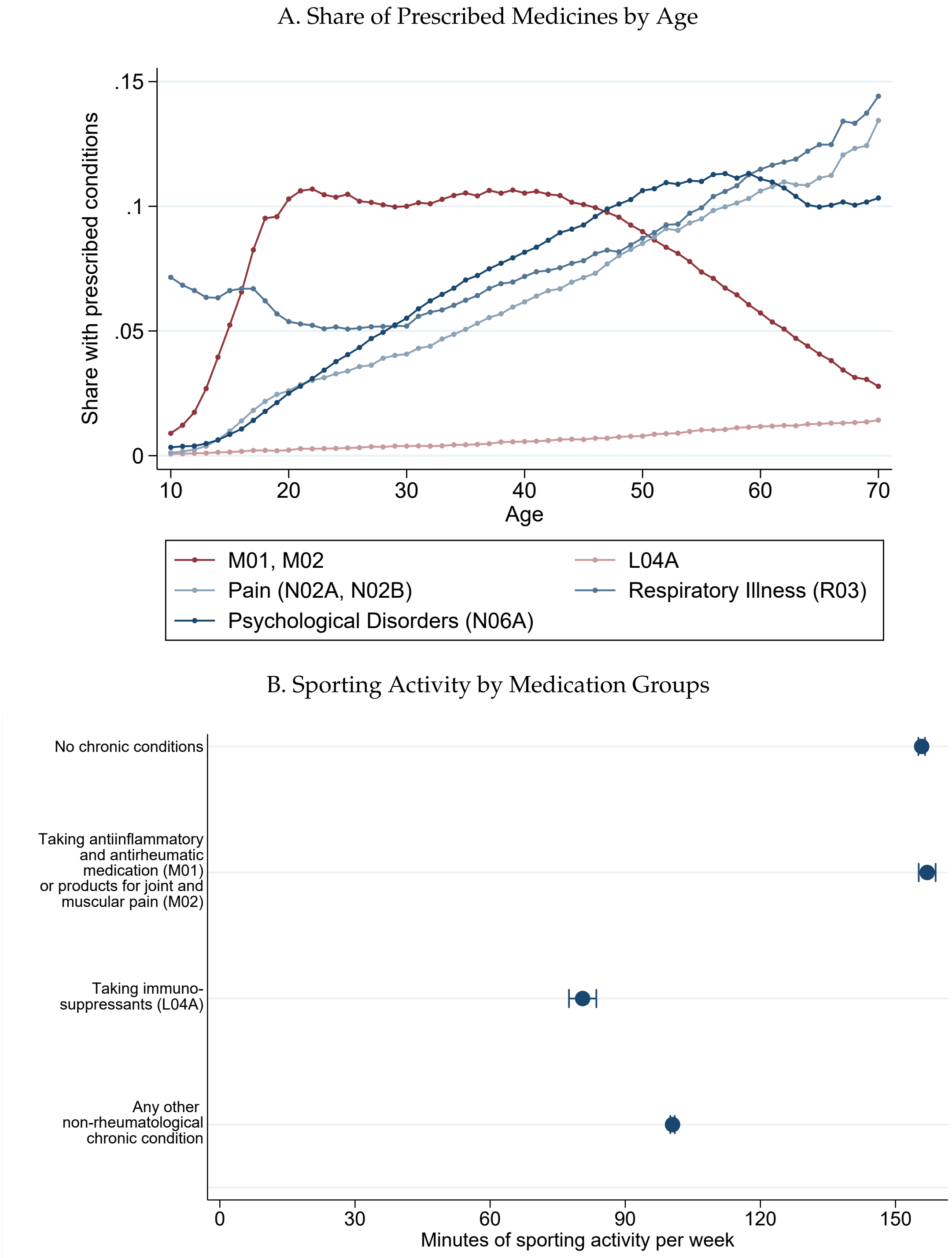
- **Rheumatic conditions**: Huber et al. (2013) use M01 (Antiflammatory and antirheumatic products), M02 (topical products for joint and muscle pain), L04AA (selective immunosuppressants) and L04AB (tumor necrosis factor alpha inhibitors), among others. We use the upper group L04A (immunosuppressants), 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: Identifying 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" indicator from regressions of the relevant outcome (mortality or costs) on that indicator 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 indicator 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. Observations are pooled from 2013-2016.[48]Specification *A1 Baseline* regresses the relevant outcome on the low income indicator, age indicators, 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 Non-chronic prescriptions*, 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 Lasso's 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 Lasso's are reported in Appendix Figure E.1.

Specification *C1 Hospitalizations* adds to A2 information on hospitalizations: primary diagnosis ICD codes from the previous year; 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 indicator controlling for the union of all the independent variables employed in all the previous specifications.

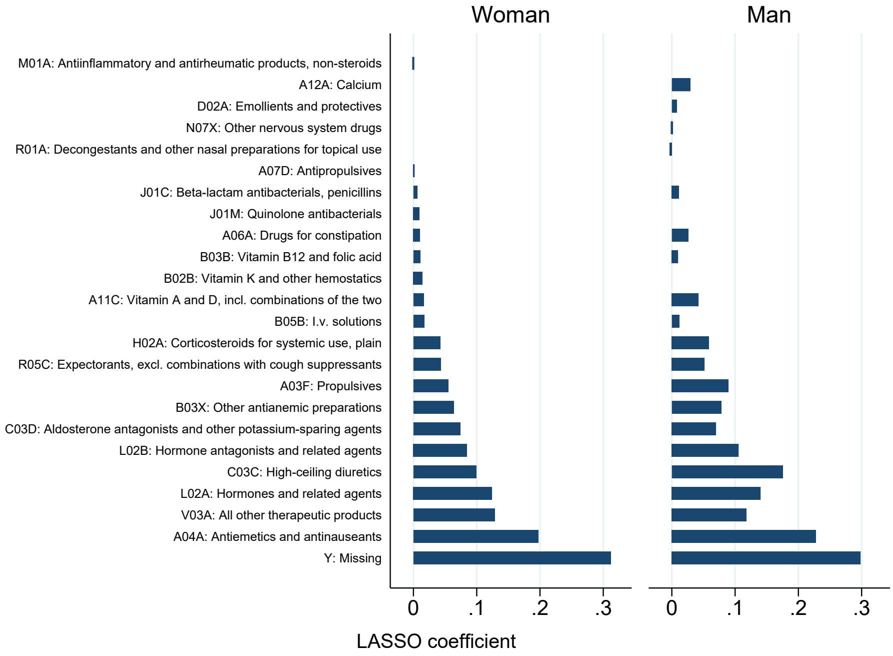
[48] For individuals aged 65 and older, pre-retirement income is defined as the average income from 60 to 62 years of age. Since data on income is only available starting in 2003, the oldest cohort whose pre-retirement income is available turns 70 year old in 2013. Since we use data on mortality until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

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.

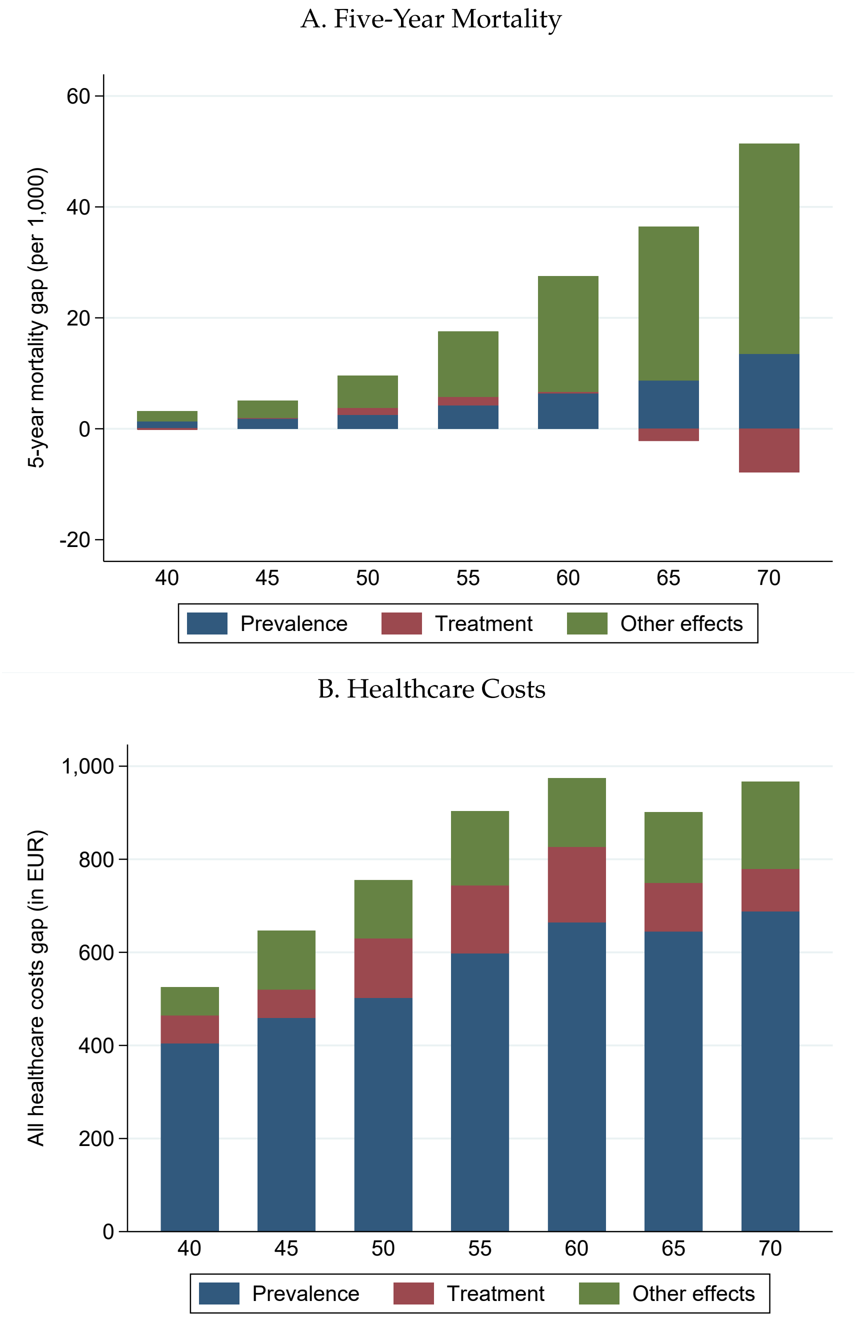
Appendix Figure E.2 uses chronic condition indicators from periods $t-3$, $t-2$, and $t-1$. Moreover, based on the distribution of chronic conditions in $t-1$, the most frequent twoway interactions between chronic conditions are retrieved and added to the Oaxaca-Blinder regression for all three lags 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.

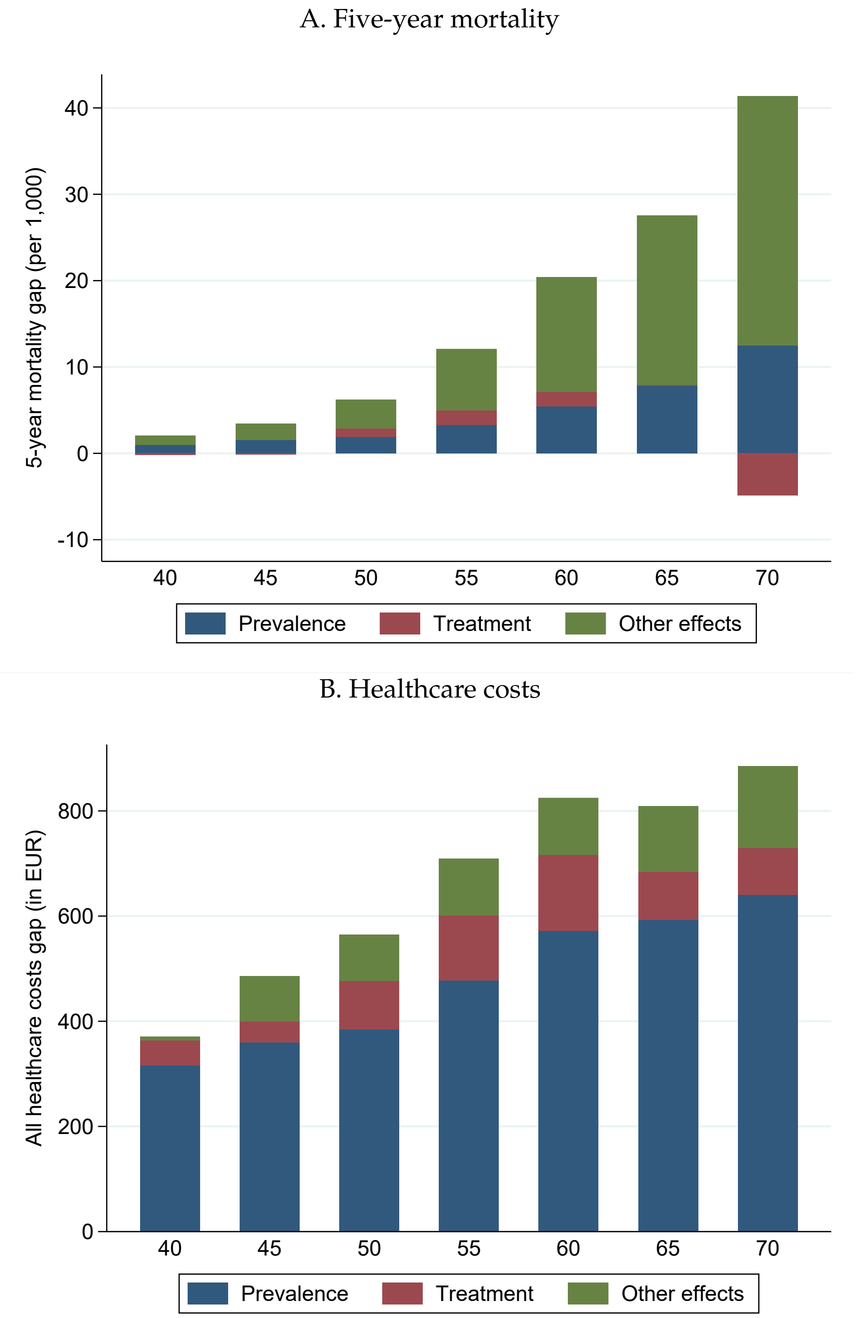
Figure E.2: Oaxaca-Blinder Decomposition, Using More Lagged and Interacted Chronic Conditions


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 "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 intergroup difference in chronic condition endowments; the "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 **.

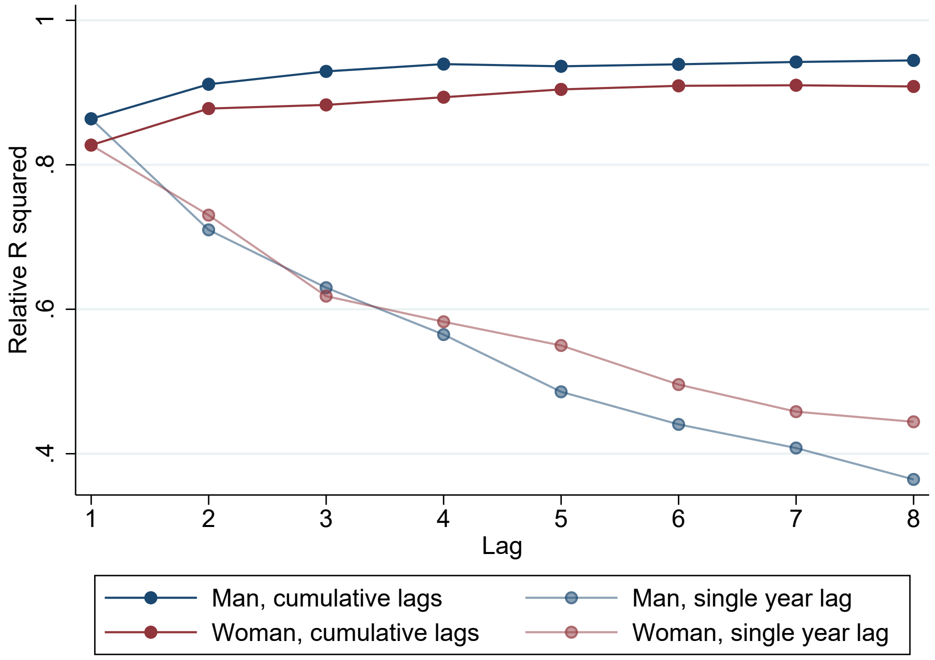
F.1 CDI model robustness

This section describes a series of considerations to understand the robustness 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 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 range and have been used elsewhere in the literature. We have tested these two alternatives, but they do not yield any increases in predictive power, and given the the model is sufficiently regularised with the variable selection process, there are

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



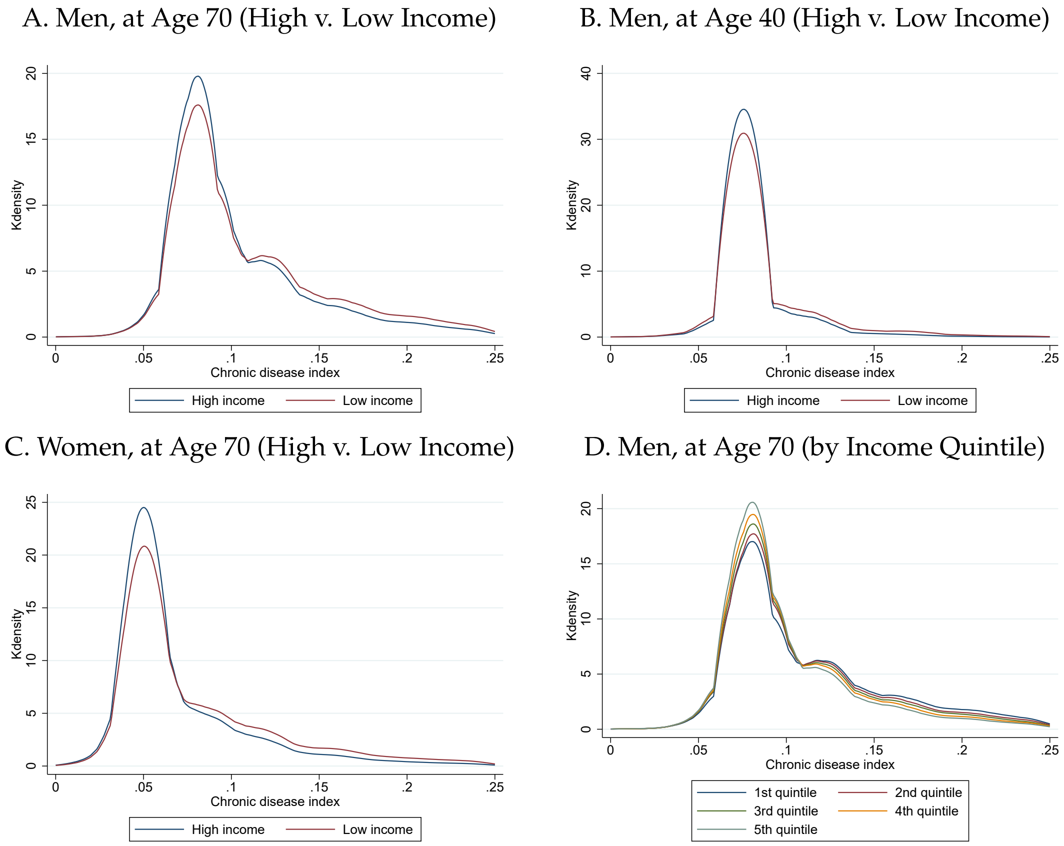
Note: This figure plots the predictive power of varying lags of chronic conditions on five-year mortality. Predictive power is measured as a relative R^2 statistic: $R^2_{s,t}/R^2_{(1,3),Int}$. The numerator $R^2_{s,t}$ is computed from a regression of five year mortality on the set of chronic condition 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)

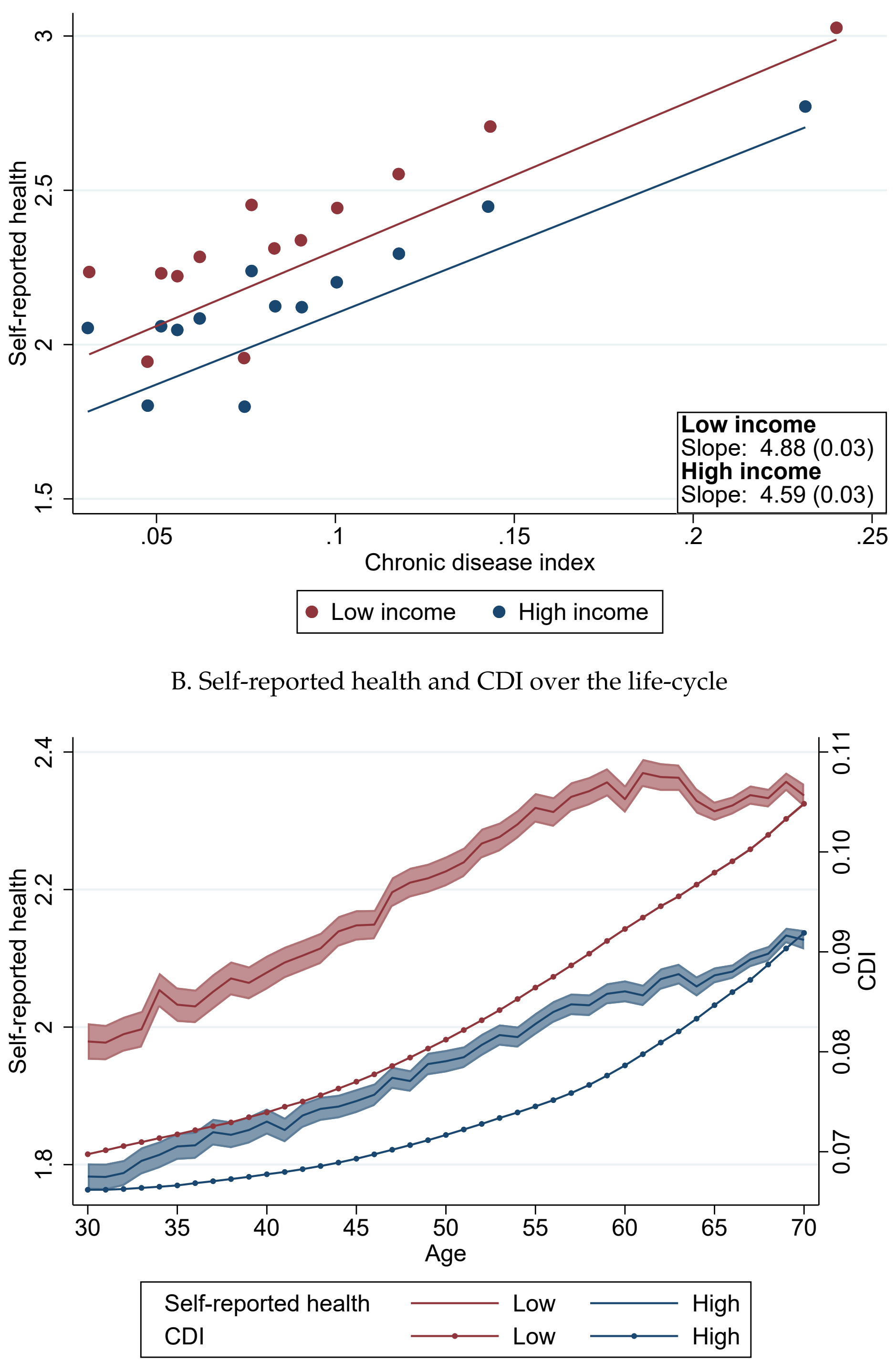
B. Men, at Age 40 (High v. Low Income)

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 life-cycle

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 vertical axis. A greater value denotes worse health. Panel B plots the evolution of self-reported health and the CDI over the life-cycle, 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.

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 (16) below. The CDI outperforms this benchmark in terms of R^2 (+10%), while the AUC statistic is a marginal improvement.

$$M_i^+ = CC_i^{*\prime}\beta_{CC}^+ + X_i^{*\prime}\beta_X^+ + \zeta_i^+ \tag{15}$$

$$\widehat{M}_i^+ = CC_i^{*\prime}\widehat{\beta}_{CC}^+ + \overline{X}_i^{*\prime}\widehat{\beta}_X^+ \tag{16}$$

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: 70 year old's are not materially positively selected in terms of survivorship bias, the under-diagnosis issue is not 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 also do 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 Dynamic Decomposition

In Figure 6, the average CDI for both income groups is shown over the life-cycle. 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 $E_{a+1}[CDI_{i,a+1}|Y_{a+1}] - E_a[CDI_{i,a}|Y_a]$, 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:

$$\begin{aligned} E_{a+1}[CDI_{i,a+1}|Y_{a+1}] - E_a[CDI_{i,a}|Y_a] &= [E_{a+1}(CDI_{i,a+1}|Y_{a+1}) - E_{a+1,a}(CDI_{i,a+1}|Y_{a+1})] \\ &+ [E_{a+1,a}(CDI_{i,a+1}|Y_{a+1}) - E_{a+1,a}(CDI_{i,a+1}|Y_a)] \\ &+ [E_{a+1,a}(CDI_{i,a+1} - CDI_{i,a}|Y_a)] \\ &+ [E_{a+1,a}(CDI_{i,a}|Y_a) - E_a(CDI_{i,a}|Y_a, S_{a+1})] \\ &+ [E_a(CDI_{i,a}|Y_a, S_a + 1) - E_a(CDI_{i,a}|Y_a)] \end{aligned} \tag{17}$$

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:

   $$Aging = E_{a+1,a}(CDI_{i,a+1} - CDI_{i,a}|Y_a) \tag{18}$$

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

2. **Health-based Sorting**: over the life-cycle, 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 $a+1$ between the members of Y_{a+1} and the members of Y_a:

   $$Sorting = E_{a+1,a}(CDI_{i,a+1}|Y_{a+1}) - E_{a+1,a}(CDI_{i,a+1}|Y_a) \tag{19}$$

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 CDI at age a between those individuals in Y_a who survived until age $a + 1$ and the mean CDI of all observed in income group Y_a.

$$Attrition = E_a(CDI_{i,a}|Y_a, S_{a+1}) - E_a(CDI_{i,a}|Y_a) \tag{20}$$

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

   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).

$$Exit = E_{a+1,a}(CDI_{i,a}|Y_a) - E_a(CDI_{i,a}|Y_a, S_{a+1}) \tag{21}$$

   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 $a - 1$).

$$Entry = E_{a+1}(CDI_{i,a+1}|Y_{a+1}) - E_{a+1,a}(CDI_{i,a+1}|Y_{a+1}) \tag{22}$$

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

$$Cohort = Exit + Entry \tag{23}$$

In our main life-cycle 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 life-cycle 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 2 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 simulates the CDI for both income groups with a) the aging component only, and b) the aging plus health-based sorting components. Appendix Figure G.2 repeats this exercise, also accounting 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 life-cycle decomposition framework, it is also possible to consider more groups based on income or other observable characteristics. Appendix Figure G.3 shows aging and sorting effects when the decomposition is performed by income quintile. Aging effects are strongest for the lowest quintile 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. The negative sorting effect on the bottom quintile peaks between 25-30, when young individuals are moving out of home and starting their careers. Towards retirement, the sorting effect becomes less apparent as income trajectories plateau. From 65-70, sorting is zero by construction since we use pre-retirement incomes. 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:

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

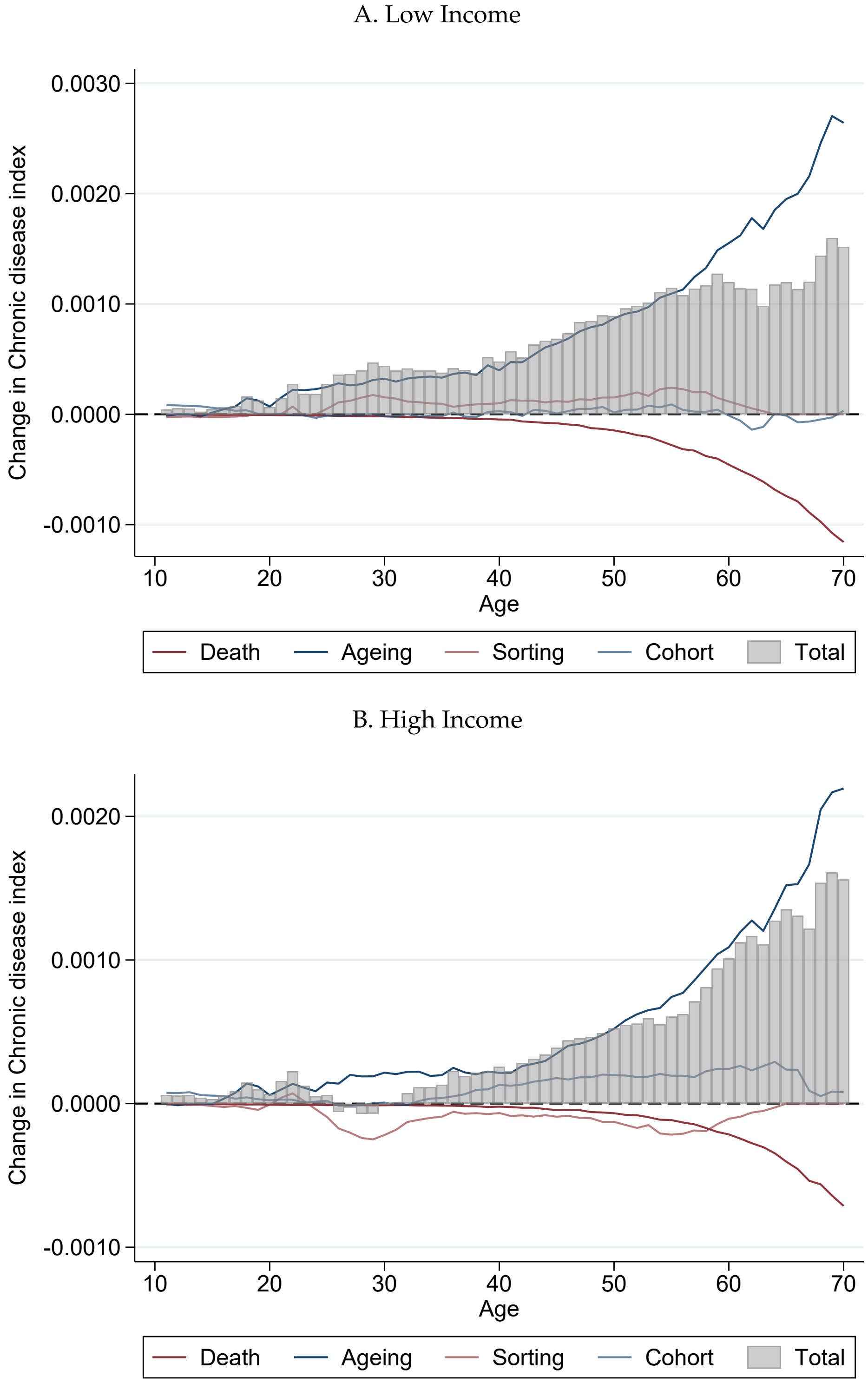
And sorting is:

$$Sorting_{MR} = E_{a+1,a,a-1}(CDI_{i,a+1}|Y_{a+1}) - E_{a+1,a,a-1}(CDI_{i,a+1}|Y_a, Y_{a-1}) \tag{25}$$

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.

Finally, we perform a robustness check on our estimated sorting effects. Sorting effects could be driven by individuals who change household composition, affecting the standardized household income in the process. To do so, we estimate the sorting effects separately for household that composition and households which have the same composition. Table G.2 reports the results of this robustness check, 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 2. This provides evidence that the sorting effect is not driven by individuals who move into a new household.

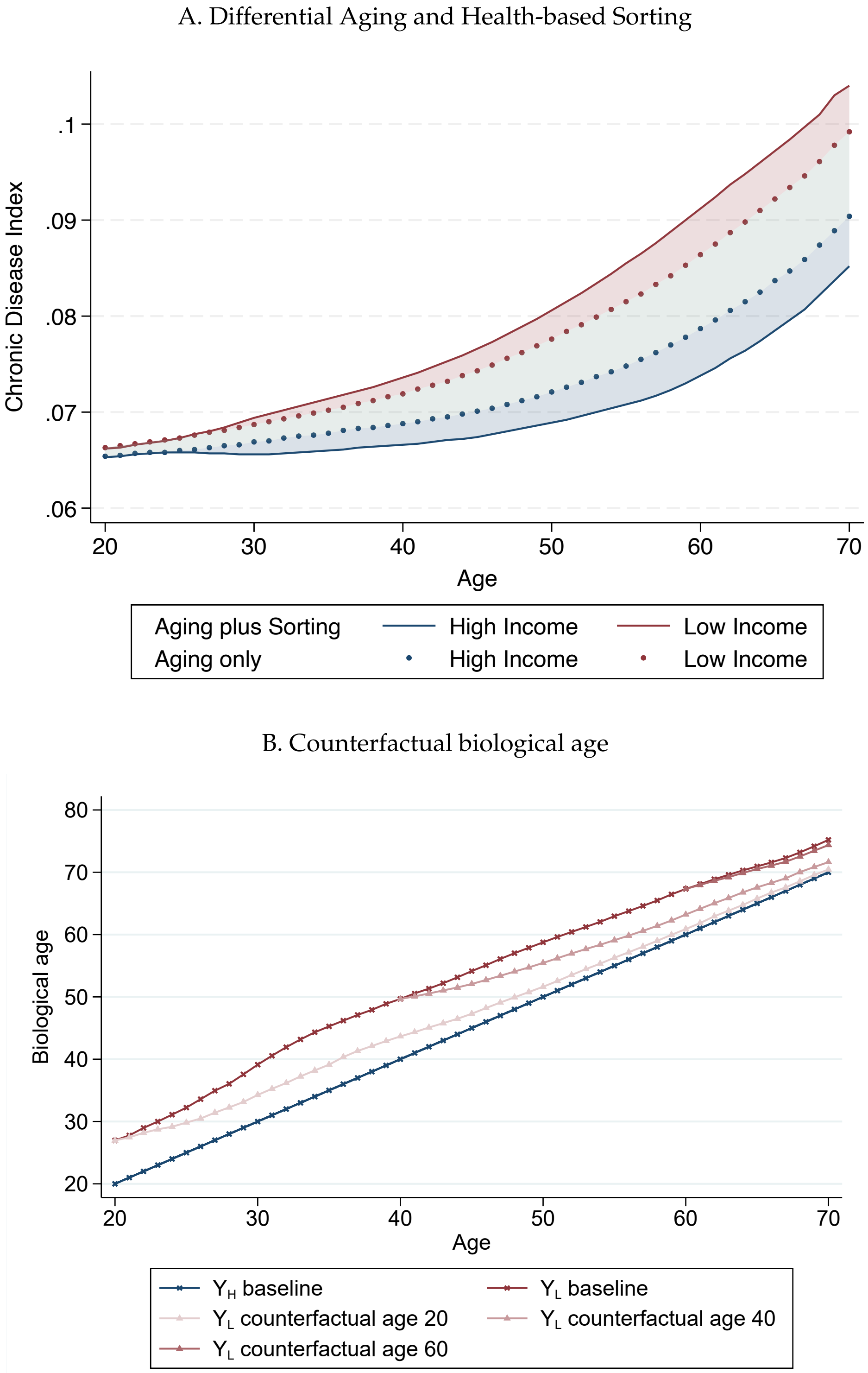
Figure G.1: Decomposition by Income group


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 difference between the low and high income group is shown for each of the effects in Table 2.

Figure G.2: Biological Aging, Accounting for Attrition due to Death



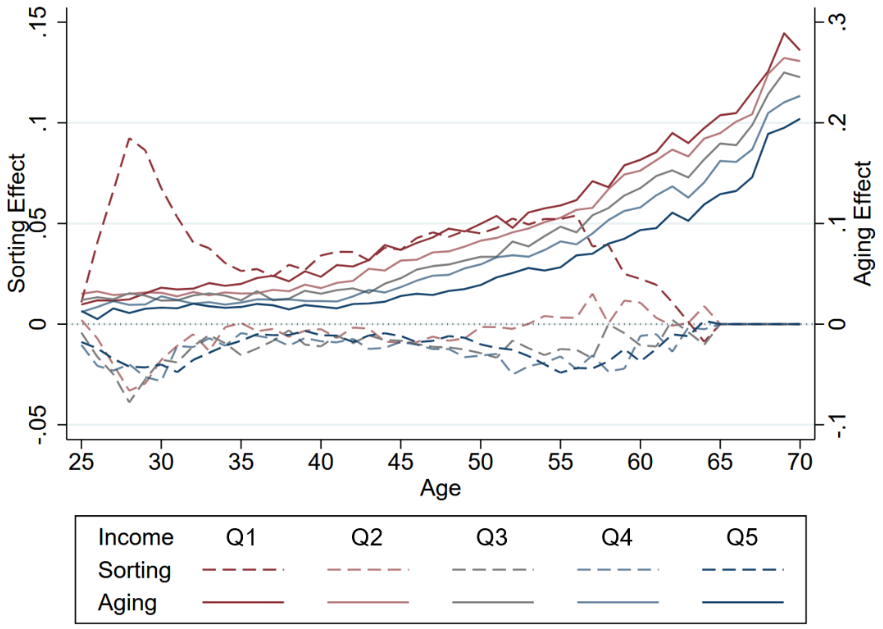
A. Differential Aging and Health-based Sorting

B. Counterfactual biological age

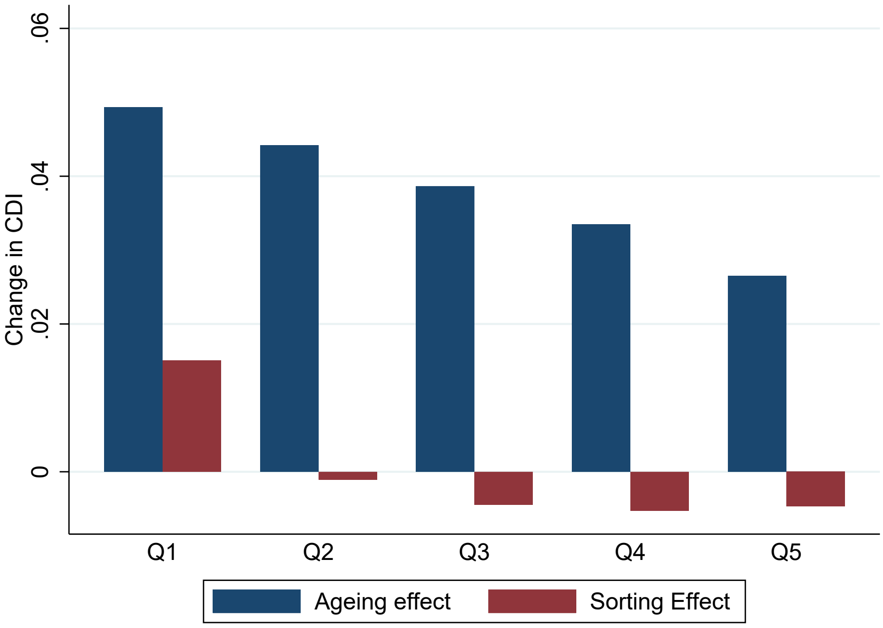
Note: Both panels are equivalent to Figure 7, but include attrition due to death effects. Panel A shows the simulated evolution of the CDI by income group with either a) only aging and attrition due to death effects, or b) aging, attrition due to death effects, and health based sorting effects. The teal shaded area represents the health gap due to differential aging. The blue and red areas are the gaps due to positive and negative sorting, for high and low incomes, respectively. 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: NON-LINEARITY IN SORTING VS. AGING EFFECTS

A. BY AGE

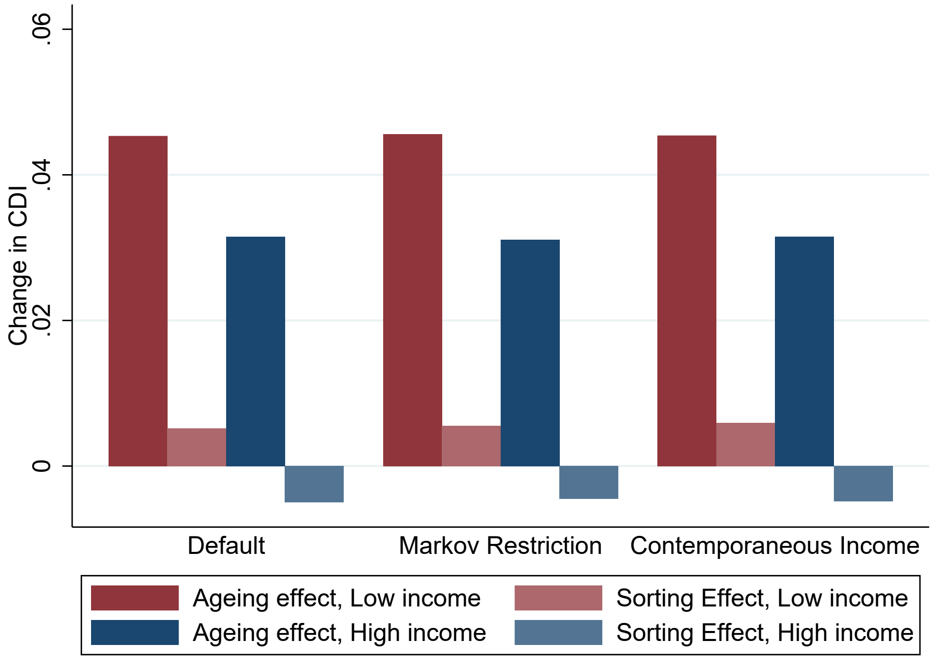


B. AGGREGATED FROM AGES 20-70



Note: Both panels report the results of the life-cycle decomposition by income quintile. Panel A shows this decomposition by age, while Panel B aggregates the effects from ages 20-70. All decompositions pool all observations in the period 2009-2021.

Figure G.4: Robustness Of Sorting vs. Aging Effects



Note: This shows cumulative aging and sorting effects for the default decomposition, and two alternatives. The 'Markov restriction' robustness check only considers individuals at age a who were in the same income group at a and $a - 1$ (based on their respective lagged incomes) and repeats the decomposition for this selected sample of individuals. The 'contemporaneous income' alternative reduces the income lag by one year, such that the average of Y_{a-3}, Y_{a-2} and Y_{a-1} is used to rank individuals' incomes. All decompositions are cumulative from ages 20-70 and 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 life-cycle 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							
Aging	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							
Aging	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							
Aging	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 life-cycle 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:

$$M_{i,a,Y} = \alpha_{a,Y} + \beta_a CDI_{i,a,Y} + \varepsilon_{i,a,Y} \tag{26}$$

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 3, we simulate alternative evolutions of the CDI based on the aging and attrition effects estimated in the life-cycle 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 (26) 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 (26) 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:

$$\mathbb{E}[A|A \geq 40] \approx \sum_{a=40}^{110} Pr(A = a|A \geq 40) \cdot a \tag{27}$$

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 3. 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 (26) :

$$k_{i,a,Y} = \gamma_{a,Y} + \delta_a CDI_{i,a,Y} + u_{i,a,Y} \tag{28}$$

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 weights 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, lifetime expected cost at age 40 is computed as:

$$\mathbb{E}\left[\sum_{a=40}^{\infty} K_a \mid A \geq 40\right] \approx \sum_{a=40}^{110} S_a \cdot K_a \tag{29}$$

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

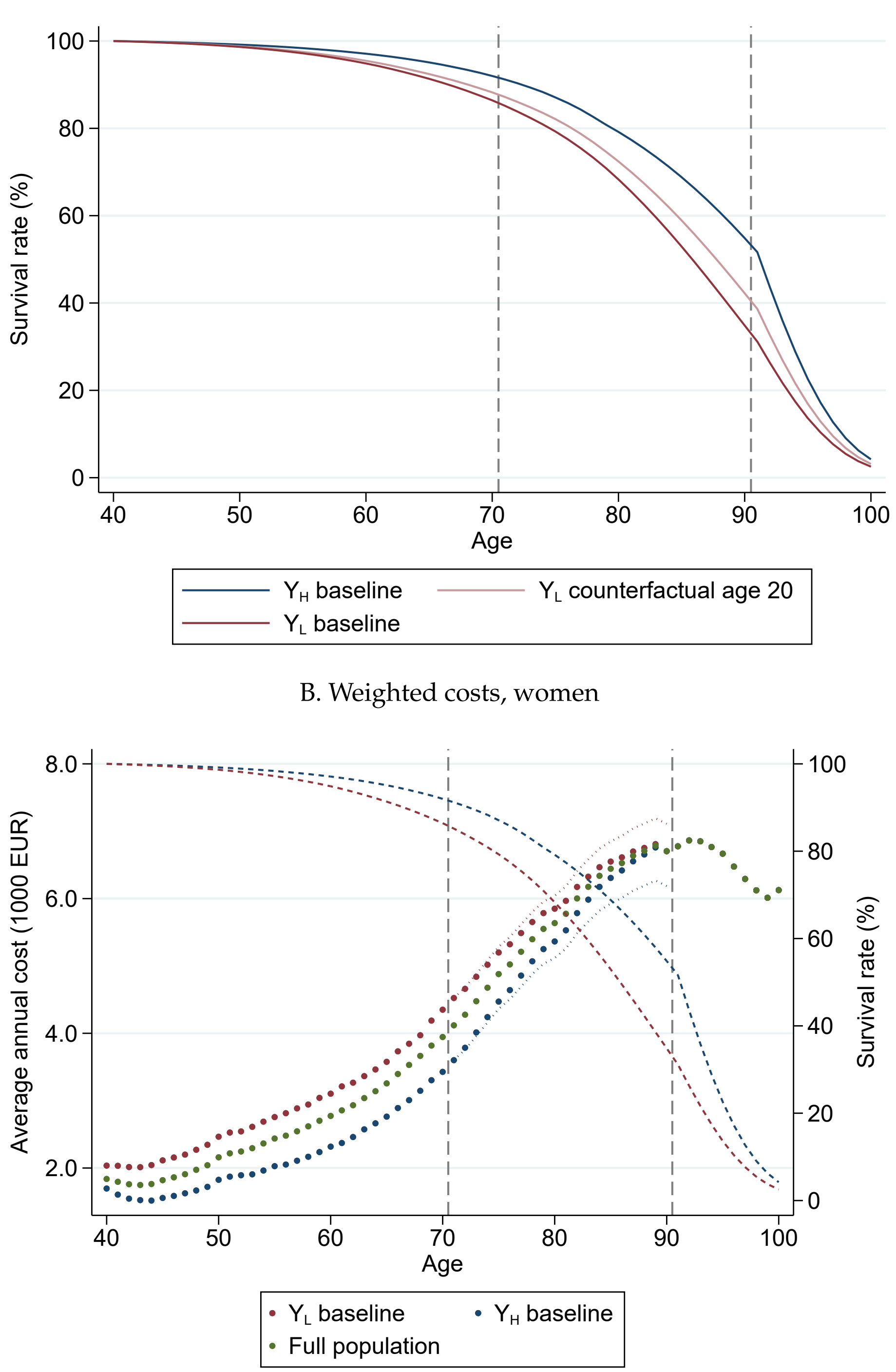
We use two alternative approaches with respect to the survival probabilities in our counterfactuals. In the first approach, we compute counterfactual healthcare costs when intervening at age 20, 40 or 60, but use the baseline low income survival probabilities computed in the life expectancy calculations. 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 3.

In the second approach, survival probabilities are adjusted for each counterfactual. More specifically, they are taken from the corresponding alternative estimated in the life expectancy calculations, using equation (27). 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 3.

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 life-cycle. 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 survival rates at higher ages.

I Mediators Analysis

Figure 8 reports the Shapley-Owen values for regression equation (14), 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. Panel A presents a Shapley-Owen decomposition for each of the eight mediator groups that are observed in the full registry data:

- **Parental health:** for each parent, we include a binary if they have died before 70, and if they are alive we include a binary if their CDI is above 0.15, approximately twice the population average.
- **Spatial data:** this comprises pollution exposure, green-space, food retail quality, healthcare proximity, population density, and mean residential property value. These spatial data are observed at the six-digit residential postcode level, which corresponds to around 40 residents per postcode. Each variable is ranked and partitioned into population-weighted deciles.
- **Employment status:** a categorical variable for whether employed or self-employed, on benefits/assistance, retired, or studying.
- **Employer industry sector:** For those that are employed, we observe the industry sector of they employer, broken into 70 categories.
- **Pay-rank:** for those that are employed in a firm with 50 or more employees, the within-firm FTE pay rank is computed and split into deciles.
- **Income & wealth:** standardized disposable household income and household net wealth, both constructed by CBS, are ranked within gender and age, and split into percentiles. Similarly for income and net wealth of parents, where observed.
- **Education level:** indicators for highest education level attained, split by below high-school, high-school level, bachelor and graduate studies.
- **Demographics:** indicators for household composition, whether foreign born, whether parents foreign born.

To allow for enough flexibility, all predictors are treated as binary indicators. In addition to the predictor groups already listed, each specification controls for age and gender indicators.

Panel B presents results from a two-stage Shapley-Owen approach, where the behavior & BMI contributions are estimated on the *Gezondheidsmonitor* survey subsample, and the contribution of all other mediators are estimated on the full population sample.

Some of the variables used in the decomposition have poor coverage. This 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.

The Appendix Figure C.11 repeats the decompositions from Figure 8, but instead the dependent variable is the log of the CDI in *levels* rather than growth. Again, in both cases, the base year used is 2013.

Figure 9 reports selected coefficients from a number of different linear regressions. In specification "dCDI, incl. controls", 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 born and for having foreign parents. Given the health behaviour variables are not observed in the full sample, these are not included as controls on the left-hand panel. In specification "CDI, incl. controls", the dependent variable is the same-year CDI level, and the same independent variables are used. Finally, specification "dCDI, without controls" aggregates the results of several regressions, whose dependent variable is the same-year CDI. Each regression has as an independent variable one of the factors shown in the 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 "dCDI, incl. controls" for specific subgroups of the population.

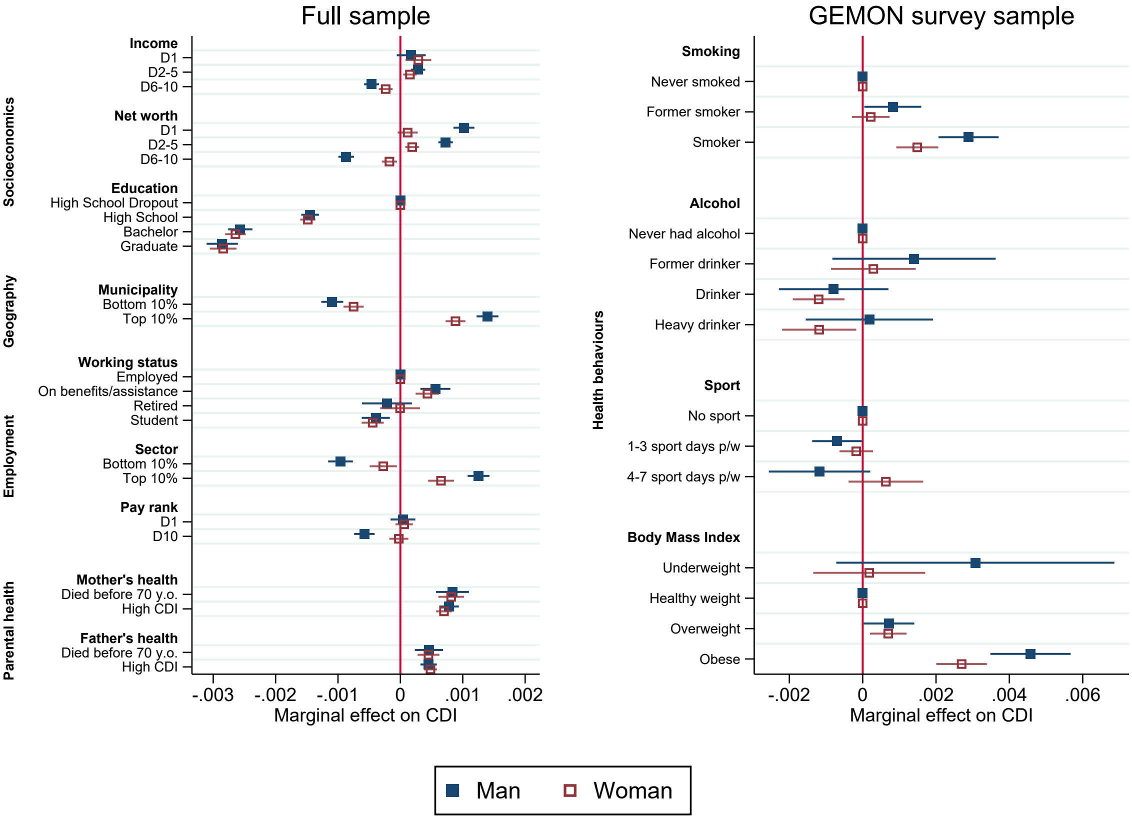
The figures report coefficients for sector and municipality aggregated into deciles, the top and bottom of which are shown. This results from an ex-post categorization, conducted as follows. The regressions use separate indicators for each municipality and sector, which are then aggregated into deciles 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

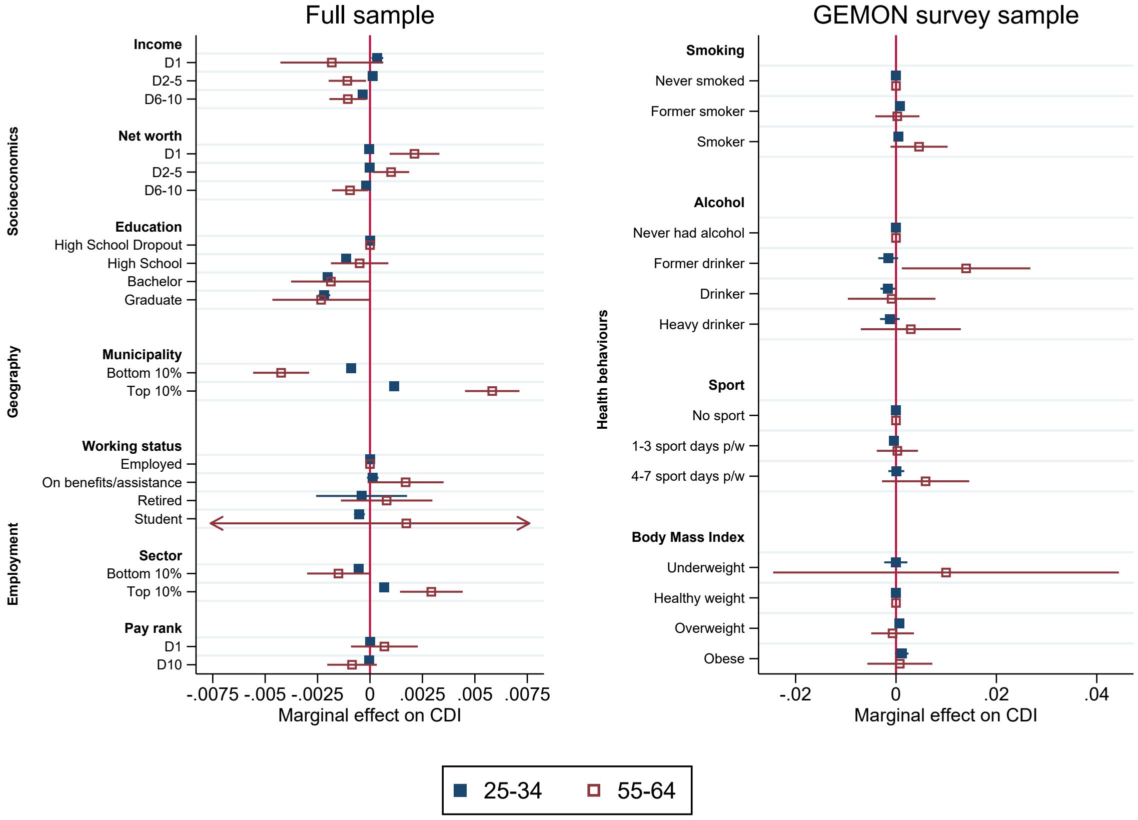
Note: Row "Observations" report the number of observations of the dependent variable (the within-individual five-year CDI growth) in 2013 for the age group in column. Row "Sample used" reports the number of observations actually used in the regressions. Row "Used, no filling in" reports the number of 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 "dCDI, incl. controls" from Figure 9.

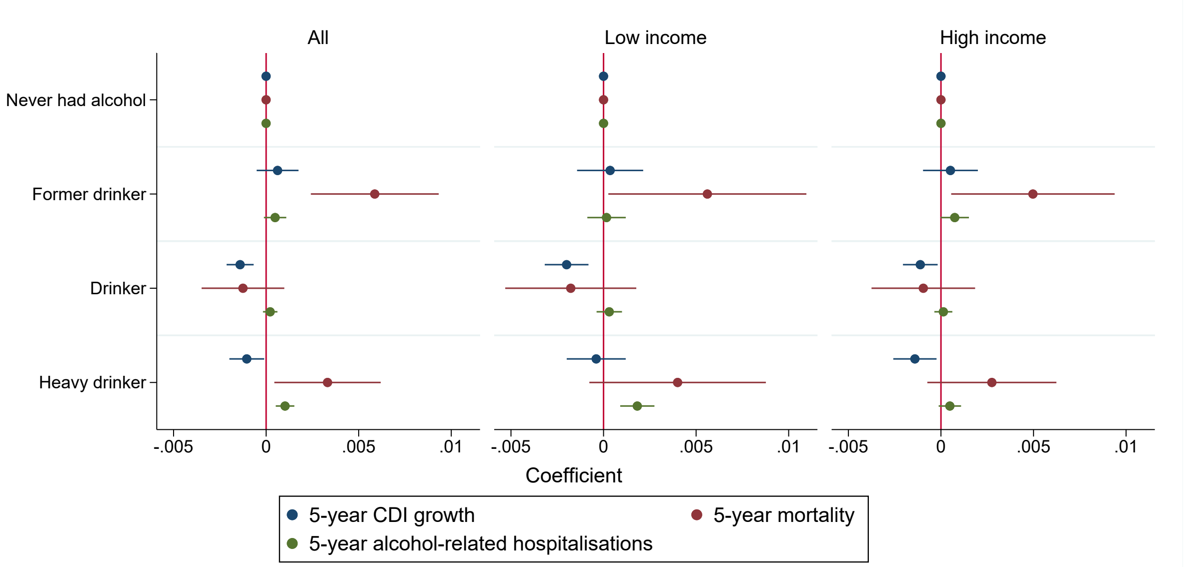
Figure I.2: MEDIATORS OF THE CDI, BY AGE



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 "dCDI, incl. controls" from Figure 9.

The confidence interval for student working status for 55-64 was truncated for reporting purposes.

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 "dCDI, incl. controls" from Figure 9.