Netspar
Network for Studies on Pensions, Aging and Retirement

Netspar THESES

Titus Galama

A Theory of Socioeconomic Disparities in Health

PhD Thesis 2011-029

A Theory of Socioeconomic Disparities in Health

A Theory of Socioeconomic Disparities in Health

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit van Tilburg op gezag van de rector magnificus prof. dr. Ph. Eijlander, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op maandag 27 juni 2011 om 16.15 uur door

Titus Johannes Galama,

geboren op 13 september 1970 te Zaandam.

Promotiecommissie

Promotores:	Prof. dr. ir. A. Kapteyn
	Prof. dr. A.H.O. van Soest
Overige leden:	Prof. dr. I. Ehrlich
	Prof. dr. M. Grossman
	Prof. dr. P. Kooreman
	Prof. dr. E.K.A. van Doorslaer

Acknowledgements

I am in debt to a number of people who, in various ways, have contributed to this thesis. First, and most importantly, I thank my beautiful and intelligent wife, Michelle, for her support of my various crazy endeavors, including this thesis. I had a successful career in Astrophysics, but got interested in business and management and pursued an MBA. Michelle came along with me to Singapore, then Paris, and found jobs to support a (substantial) student loan, traveling and living expenses (not to mention the opportunity costs at that age) as a result of my decision to change careers. She also gave me two beautiful, sweet and smart children (as a scientist I can say that I have established these facts objectively :-)). Amandine (four years of age) and Lucas (one year) have, whenever an opportunity occurred, very strategically and capably cooperated to distract me from my thesis research. Though I thank them for the joy they have provided (and still do), which no doubt has made me more productive, so that on net their contribution to my thesis has most definitely been positive.

After my MBA, and some years in strategy consulting, it became clear to me that at heart I am a scientist and I was fortunate enough to land a job at the RAND Corporation in Santa Monica. There I started to work with Arie Kapteyn and Jim Hosek.

No doubt, my thesis advisor Arie Kapteyn has been a most important individual supporting the development of this PhD thesis – and I thank him wholeheartedly. Without his support this thesis would not have been possible. Arie has a management style that I admire – which I believe is basically trusting that good people will do good work and then creating an environment and providing the means to support them. When I arrived at the RAND Corporation in early 2006 it wasn't immediately obvious how I could be best put to use in the RAND environment. Arie, however, decided early on that I might be worthy of some investment. He gave me an interesting task – to build an economic theory of health and retirement – and gave me plenty of support in terms of precious research time to familiarize myself with what was at that time (to me) a very new world of Economics research. This investment ultimately led to this PhD thesis. Arie, I am

immensely grateful for your trust and support and I hope to continue to benefit from future opportunities to conduct research with you.

I thank my other thesis advisor Arthur van Soest for the quality of the comments and feedback he provided on the papers that constitute my thesis, for selecting and inviting the members of the committee and for his support with the administrative side of the PhD thesis and the defense.

In addition to Arie, Jim Hosek has also been a very important mentor to me and I thank him for the numerous discussions we have had on topics of Economic and other interest. Arie and Jim are, in my opinion, great thinkers and I have been fortunate to receive significant amounts of their time to learn about Economic thought and concepts, despite their busy agendas. I have learned that many of the skills I developed in Physics could be applied to Economics. Some of what I needed to get used to was a different jargon.[1] But I also learned to appreciate that there are important differences between Economic thought and methods from those in Physics and Arie and Jim have been important mentors in this regard.

A first visit to Eddy van Doorslaer's research group at Erasmus University in the Netherlands can be credited with providing the motivation to begin developing a theory of socioeconomic status and health over the life cycle. This has led to a very fruitful collaboration with the Erasmus team. I want to thank Eddy in particular for his generous hospitality during my regular visits of Erasmus University and I am very pleased that the RAND and Erasmus collaboration, centered around empirical testing and continued development of the theoretical framework developed in this thesis, has been solidified by grant R01AG037398 from the National Institute of Aging. Eddy, I look forward to continue working with you in this collaboration.

In addition, it has been a pleasure to work with then graduate student, now Dr. Hans van Kippersluis, of the Erasmus team. Hans has been instrumental in developing the theoretical framework described in Chapter 5. He also deserves credit for helping me think through the theory developed in Chapter 4. Hans, I really enjoy working with you and I look forward to continuing our collaboration for a long time to come.

I want to thank Arie Kapteyn, Eddy van Doorslaer, Jim Smith, Erik Meijer, Hans van Kippersluis, Tom van Ourti, Owen O'Donnell, Mauricio Avendano and Megan Beckett for their excellent contributions to proposals and for the comments and feedback provided on the theoretical framework and papers presented in this thesis. I also want to thank

[1] Don't be fooled: the marginal cost of X just means they took the derivative of the cost of X and endogenous / exogenous means it is in- or outside of the model (at least roughly).

Erik Meijer for patiently answering numerous mathematical, economic and econometric questions of mine.

I am very grateful to Michael Grossman for extensive discussions during the NBER summer meeting of 2010 and for the subsequent frequent and extensive exchanges via email, in which Michael, Hans van Kippersluis and I debated the properties of Michael's seminal theory of health capital (Grossman 1972a, 1972b) that provides the foundation for the work presented in this thesis. I am honored to have Michael on the thesis committee.

Likewise, I am very grateful for exchanges with Isaac Ehrlich regarding his influential work on a theory of longevity (Ehrlich and Chuma, 1990), based on the Grossman model. I am also honored to have you on my thesis committee and I look forward to continuing to work with you on the organization of a conference at RAND this year.

I want to thank Rosalie Pacula for her enthusiasm about the work I do and for her detailed comments on Chapter 5. I thank Raquel Fonseca and Pierre-Carl Michaud for their contributions to Chapter 3. I thank Tania Gutsche for organizational and Christopher Dirks and Sloan Fader for administrative support. I thank seminar and meeting participants at the University of Lausanne, Switzerland (November, 2010); Harvard Center for Population and Development studies (July, 2010); University of Southern California (June, 2010); University of Tilburg (March 2009); Tinbergen seminar series at the Erasmus University in Rotterdam (March 2009); NETSPAR conference, Amsterdam (March 2009); and the American Economic Association (AEA) meeting (Jan 2009) for useful discussions and comments. In particular I would like to thank Peter Kooreman for his detailed comments on Chapter 3, provided at the NETSPAR conference.

I thank my mother, Marieke Kuipers, and father, Joep Galama, for instilling a love for science and education in me.

In addition to grant R01AG037398, this research was further made possible by National Institute on Aging grants R01AG030824, P30AG012815 and P01AG022481.

Titus Galama,
Santa Monica, March 2011

Contents

List of Tables

List of Figures

Chapter 1

Introduction

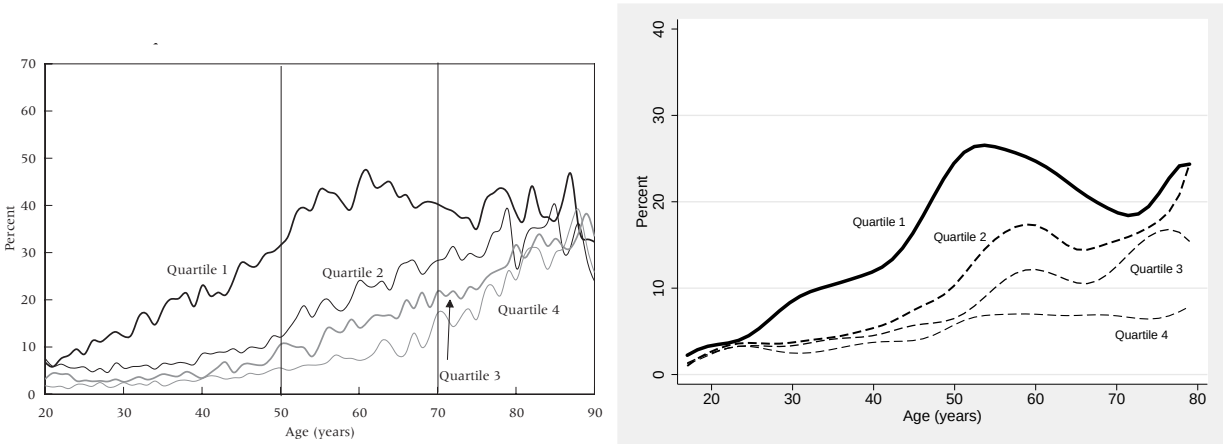
1.1 Background

One of the most remarkable findings in population health is the strong relationship between health and socio-economic status (SES). Figure 1.1 displays the principal features of the SES health gradient in the U.S. (left-hand side) and the Netherlands (right-hand side) by plotting at each age the fraction of people who self-report themselves in poor or fair health by age-specific household income quartiles (quartile 1 representing the lowest and quartile 4 the highest household incomes). At each age a downward movement in income is associated with poorer health.

The health differences by income quartile are large. For example, in the U.S. at around age 60 the fraction in poor or fair health in the top income quartile, at about 8 percent, is some 35 percentage points smaller than the fraction in the lowest income quartile, at about 44 percent (left-hand side of Figure 1.1). Similarly, Case and Deaton (2005) show how in the United States, a 20 year old low-income (bottom quartile of family income) male, on average, reports to be in similar health as a 60 year old high-income (top quartile) male. In Glasgow, U.K., life expectancy of men in the most deprived areas is 54 years, compared with 82 years in the most affluent (Hanlon et al. 2006).

Not only do low SES individuals start adulthood in worse health but their health also deteriorates faster with age than the health of their high SES peers. In cross sectional data the disparity in health between low and high SES groups appears to increase over the life cycle until ages 50-60, after which it narrows (see, e.g., Figure 1.1). Similar patterns hold for other measures of SES, such as education and wealth and other indicators of health, such as onset of chronic diseases, disability and mortality (e.g., Adler et al. 1994; Marmot, 1999; Smith, 1999).

Figure 1.1: Percent reporting fair or poor health by age-specific household income quartiles.



Notes: Percent reporting fair or poor health (bottom two categories of self-reported health) by age-specific household income quartiles. Left: U.S. National Health Interview Surveys, 1991-1996, taken from Smith (2004). Right: Dutch Statistics Netherlands (CBS) Health Interview Surveys, 1983-2000 (courtesy Hans van Kippersluis).

These patterns are remarkably similar between countries with relatively low levels of protection from loss of work and health risks, such as the U.S., and those with stronger welfare systems, such as the Netherlands (compare the left with the right-hand side of Figure 1.1; House et al. 1994; Kunst and Mackenbach, 1994; Preston and Elo, 1995; Smith 1999; 2004; 2007; Case and Deaton, 2005; van Kippersluis et al. 2010).

There is a widespread view that these large disparities in health between SES groups represent an infringement of social justice. The notion is that such inequities in health are avoidable and arise because of the circumstances in which people grow, live, work, and age, and the systems put in place to deal with illness (e.g., CSDH, 2008). With this viewpoint, the World Health Organization's (WHO) Commission on the Social Determinants of Health (CSDH) has called for global action on the social determinants of health with the aim of achieving health equity within a generation (CSDH, 2008).

This noble aim is, however, hampered by the fact that the causes of socioeconomic health disparities are not well understood. Studies across multiple disciplines (including epidemiology, sociology, demography, psychology, evolutionary biology and economics) reveal that the proposed explanations are multiple and diverse, that consensus on their importance is lacking and that it has been difficult to establish causality and even harder to firmly establish underlying mechanisms (e.g., Cutler et al. 2011). For example, education is found to have a causal protective effect on health (Lleras-Muney, 2005; Oreopoulos,

2006; van Kippersluis et al. 2011) but it is not known exactly how the more educated achieve their health advantage.

Some of the proposed mechanisms imply that SES influences ("causes") health, others imply the reverse path of causation, and some imply that SES and health are jointly determined, without direct causal link. Some mechanisms may fall in all three categories. Proposed explanations for the SES-health gradient include: access to medical care, health-enabling labor-force attachment, health behaviors (e.g., smoking, drinking, exercise), psychosocial and environmental risk factors, neighborhood social environment, social relationships and supports, sense of control, fetal and early childhood conditions, and physical, chemical, biological and psychosocial hazards and stressors at work. So-called "third factor" explanations posit that individual differences, e.g., in time preferences and the ability to delay gratification, affect SES and health in similar ways and thereby give rise to the SES-health gradient. Many of these explanations have been shown to explain each a piece of the puzzle (for a review see Galama and van Kippersluis, 2010 [Chapter 5]).

Advancement of understanding of the relative importance of the causal mechanisms responsible for the observed relationships is hampered by the lack of a sufficiently comprehensive theory. The significant social and economic patterning of disease suggests that social interventions have great potential for improving the health of, in particular, disadvantaged groups, and knowing qualitatively and quantitatively how these mechanisms operate informs the development of effective social interventions. Without knowledge of the mechanisms, it is difficult to design policies that are effective in reducing disparities (Deaton, 2002). Thus, integrating the roles of proposed mechanisms and their long-term effect into a comprehensive framework is a crucial first step towards designing and evaluating effective policy. It allows researchers across multiple disciplines to assess the relative importance of each proposed mechanism, the interaction between mechanisms, and to disentangle the differential patterns of causality. Case and Deaton (2005) argue that it is extremely difficult to understand the relationships between health, education, income and labor-force status without some guiding theoretical framework. It is therefore no surprise that several authors (e.g., Case and Deaton, 2005; Cutler et al. 2011) have pointed to the absence of a theory of SES and health over the life cycle and have emphasized the importance of developing one.

The aim of this thesis is to make a contribution to a theory of socioeconomic disparities in health over the lifecycle. That limited progress so far has been made in constructing such a theory can probably be understood as a consequence of the following. Some of these mechanisms have direct short-term effects, but most operate over the longer term,

for example, through a relatively small but persistent effect on the health deterioration rate or asset accumulation. Health disparities, as well as SES differences (e.g., wealth) accumulate over the life course, and are considerably larger at old ages. In other words, in order to fully assess the contribution of each explanation it is essential that we take a life-course approach. A suitable framework in which multiple mechanisms and their cumulative long-term effects can be studied is a structural model of SES and health over the life cycle. Structural economic life-cycle models, in which individuals maximize their life-time utility over their decision options (such as consumption and saving) subject to budget and other constraints, have provided valuable insight into economic behavior such as consumption, saving, and labor-force participation. However, up to very recently, life-cycle models of health, medical care, and SES, suffered serious technical difficulties.

Chapters 2, 3 and 4 of this thesis are therefore aimed at addressing these technical issues. Chapter 5 then presents a theory of socioeconomic disparities in health over the lifecycle.

1.2 Overview of this thesis research

This thesis research began with a simple idea: to construct a theory of health and retirement. Economists have argued that an important part of the health differences by financial indicators of SES can be explained by the fact that bad health impinges on the ability to work, thereby reducing income (Smith, 1999, 2004, 2007). Retirement is thus an essential component of a theory of SES and health.

Our approach was to integrate the retirement decision into the formulation of the canonical model of the demand for health and health investment due to Grossman (1972a, 1972b). In Grossman's human capital framework individuals demand medical care for the consumption benefits (health provides utility) as well as production benefits (healthy individuals have greater earnings) that good health provides. Arguably the model has been one of the most important contributions of Economics to the study of health behavior. The model has become the standard (textbook) framework for the economics of the demand for health and medical care, and theoretical extensions and competing economic models are still relatively few.

Integrating the retirement decision into the health production literature (the literature spawned by Grossman's seminal 1972 papers) was, however, not as straightforward as one had hoped. An important artifact of the solution for health was a discontinuity near the age of retirement (see Galama et al., 2008 [Chapter 3]). The theory predicted that immediately following retirement health would fall (or in some cases increase) instantaneously

due to the substitution of health for leisure and the disappearance of the production benefit of health (during retirement health does not provide a production benefit as retirees do not earn wages). This cannot be correct. In the health production literature, health is a stock and in contrast to flows (such as health investment and consumption) it cannot be adjusted instantaneously. Health can only change gradually through health investment and biological aging.

This curious feature of the solution for health relates to an observation first made by Wolfe (1985). Wolfe noted that, in the health production literature, health is characterized by a so-called "bang-bang" solution. If, at any time, the health stock is not at its "optimal" level, individuals invest a large positive (or negative, depending on the direction of the adjustment) amount of medical care (or other forms of health investment) in a single period.[1] Wolfe (1985) further noted that there is no reason to expect the initial endowment of health to exactly equal the "optimal" level of health and that in fact humans may have been endowed with "excessive" health (see Wolfe, 1985, and Galama et al. 2008 [Chapter 3]). Individuals might then prefer to exchange health for consumption. But, because individuals cannot "sell" their health through negative health investment, the optimal decision is to initially not invest in health (this represents a corner solution). Health then deteriorates gradually as a result of the biological aging process. At a certain age health may reach the "optimal" health level and the individual begins to counter the aging process by investing in health. Wolfe interprets this onset of *"... a discontinuous mid-life increase in health investment ..."* with retirement.

Inspired by these findings, Chapter 2 (Galama and Kapteyn, 2009) explores a generalized solution to Grossman's model of health capital, relaxing the widely used assumption that individuals can adjust their health stock instantaneously to an "optimal" level without adjustment costs. The model then predicts the existence of a health threshold above which individuals do not demand medical care (a corner solution). We find that the generalized solution can account for a greater number of observations than can the traditional solution. Importantly it can deal with a significant criticism of health production models: that the predicted positive association between health and medical care is consistently rejected by the data (e.g., Wagstaff, 1986a; Zweifel and Breyer, 1997, p. 62). Chapter 2 also provides structural and reduced form equations to facilitate empirical tests of our generalized solution.

Chapter 3 (Galama et al. 2008) then formulates a stylized structural model of health, wealth accumulation and retirement decisions, utilizing the generalized solution developed

[1] In a continuous time formulation individuals would consume an infinitely large amount of medical care in an infinitesimally small period of time.

in Chapter 2. We derive analytic solutions for the time paths of consumption, health, health investment, savings and retirement. Exploring the properties of corner solutions we find that advances in population health decrease the retirement age, while at the same time individuals retire when their health has deteriorated. This potentially explains why retirees point to deteriorating health as an important reason for early retirement, while retirement ages have continued to fall in the developed world, despite continued improvements in population health and mortality. The model further predicts that workers with higher human capital invest more in health and because they stay healthier retire later than those with lower human capital whose health deteriorates faster.

While the corner solutions employed in Chapters 2 and 3 initially appeared promising, issues with the characteristics of the solutions for health and health investment remained. For example, the model's predictions seem caricatures of real life: in the corner solution healthy individuals do not invest in health at all for periods of time, while in reality most people see the doctor at least once per year. Further, while the "bang-bang" issue appears to have been addressed for individuals whose initial health is above the health threshold (but see Galama, 2011 [Chapter 4]) this is not the case for solutions where initial health is below the threshold. A review of the literature highlighted at least five main limitations of health production models. Briefly these are: a) the indeterminacy problem ("bang-bang" solution) for investment in health (Ehrlich and Chuma, 1990), b) the inability of the model to predict the observed negative relation between health and the demand for medical care (e.g., Wagstaff, 1986a; Zweifel and Breyer, 1997), c) the inability to explain differences in the health deterioration rate (not just the level) between socioeconomic groups (e.g., Case and Deaton, 2005), d) the lack of "memory" in the model solutions (e.g., Usher, 1975) and e) the need to assume that the biological aging rate is increasing with age to ensure that life is finite and health falls with age and to reproduce the observed rapid increase in medical care near the end of life (e.g., Case and Deaton, 2005).

Ehrlich and Chuma (1990) point out that under the constant returns to scale (CRTS) health production process assumed in the health production literature, the marginal cost of investment is constant, and no interior equilibrium for health investment exists. The authors argue that this is a serious limitation of health production models. Their finding suggests that introducing diminishing returns to scale (DRTS) in the health production process might be an avenue worth pursuing in order to address the alleged technical issues associated with health production models.

At the time, however, it was far from obvious that introducing DRTS in the health production process would bear fruit. First, Ehrlich and Chuma's claim was disputed (e.g., Reid, 1998; Grossman, 2000). Reid (1998) argued that *"... the authors [Ehrlich and*

Chuma] fail to substantiate either claim [bang-bang and indeterminacy] ...". This may have been because Ehrlich and Chuma's argument is brief and technical. Further, Ehrlich and Chuma's finding that health investment is undetermined (under the usual assumption of a CRTS health production process) was incidental to their main contribution of modeling the demand for longevity and the authors did not explore the full implications of a DRTS health production process. Second, a DRTS health production process was believed to increase the complexity of the problem substantially, rendering theoretical and econometric analysis very difficult (e.g., Grossman, 2000, p. 364). This notion may have been reinforced by the fact that Ehrlich and Chuma (1990) had to resort to comparative dynamics to illustrate the properties of the model. This technique (Oniki, 1973) is essentially a sensitivity analysis in which the directional effect of a parameter change can be investigated. Ehrlich and Chuma's (1990) insightful work is therefore limited to generating directional predictions. Third, it was not apparent that the introduction of DRTS in the health production process would substantially change the nature of the model. For example, there was the notion that introducing DRTS would result in individuals reaching the desired stock gradually rather than instantaneously (e.g., Grossman, 2000, p. 364) – perhaps not a sufficiently important improvement to warrant the increased level of complexity. Last, plausibly as a result of the above factors the health production literature never adopted a DRTS health production process,[2] i.e. developing a health production model with a DRTS health production process was relatively uncharted territory.

Chapter 4 (Galama, 2011) presents a theory of the demand for health, health investment and longevity based on Grossman (1972a, 1972b) and the extended version of this model by Ehrlich and Chuma (1990). In this chapter I make several contributions to the literature. First, I argue for a different interpretation of the health stock equilibrium condition, one of the most central relations in the health production literature: this relation determines the optimal level of health investment (and not the health stock as is assumed in the health production literature). Second, I show that this alternative interpretation necessitates the assumption of DRTS in the health production process, or no solution to the optimization problem exists (Ehrlich and Chuma, 1990). Third, I provide a detailed assessment of the implications of the alternative interpretation of the first-order condition for health investment and of the assumption of DRTS in the health production process, and show that this can address the five technical difficulties discussed above. In contrast to the health production literature I predict a negative correlation between health investment and health, that the health of wealthy and educated individuals declines more

[2]To the best of my knowledge the only exception is an unpublished working paper by Dustmann and Windmeijer (2000) who take the model by Ehrlich and Chuma (1990) as their point of departure.

slowly and that they live longer, that current health status is a function of the initial level of health and the histories of prior health investments made, that health investment rapidly increases near the end of life and that length of life is finite as a result of limited life-time resources (the budget constraint). Fourth, I derive structural relations between health and health investment (e.g., medical care) that are suitable for empirical testing. These structural relations contain the CRTS health production process as a special case, thereby allowing empirical tests to verify or reject this common assumption in the health production literature. Last, I find that the theory does not support the common notion that individuals aspire to a certain "optimal" level of the health stock. Rather, given any level of their health stock individuals decide about the optimal level of health investment.

With these essential issues addressed our formulation can account for a greater number of observed empirical patterns and suggests that the Grossman model provides a suitable foundation for the development of a life-cycle model of the SES-health gradient. Chapter 5 (Galama and van Kippersluis, 2010) completes this thesis research and presents a life-cycle model that incorporates multiple mechanisms explaining (jointly) a large part of the observed disparities in health by SES. The framework includes simplified representations of major mechanisms, which allows us to improve our understanding of their operational roles in explaining the SES health gradient and make predictions. Our starting point is the health production literature spawned by Grossman (Grossman, 1972a; 1972b) and the extensions presented by Ehrlich and Chuma (1990) and Case and Deaton (2005). Our contribution is as follows.

First, we employ the alternative interpretation of the equilibrium condition for health as determining the optimal level of health investment (as in Galama, 2011 [Chapter 4]). This interpretation addresses the five before mentioned limitations of health production models.

Yet, utilization of medical services and access to care explain only part of the association between SES and health (e.g., Adler et al. 1993). Our second contribution is therefore to incorporate many potential mechanisms in the model that could explain disparities in health by SES and to include a multitude of potential bi-directional pathways between health and dimensions of SES. One important concept in our work is "job-related health stress", which can be interpreted broadly and can range from physical working conditions (e.g., hard labor) to the psychosocial aspects of work (e.g., low status, limited control, repetitive work, etc). The notion here is that job-related health stress can include any aspect of work that is detrimental to health and as such is associated with a wage premium (a compensating wage differential). Other important features of the model are lifestyle

factors (preventive care, healthy and unhealthy consumption), curative (medical) care, labor force withdrawal (retirement) and mortality.

We find that greater initial wealth, permanently higher earnings (over the life cycle) and a higher level of education induce individuals to invest more in curative and in preventive care, shift consumption toward healthy consumption, and enable individuals to afford healthier working environments (associated with lower levels of physical and psychosocial health stresses) and living environments. The mechanism through which initial wealth, permanent income and education operates is by increasing the demand for curative care and raising the marginal cost of curative care. A higher marginal cost of curative care, in turn, increases the health benefit of (and hence demand for) preventive care and healthy consumption, and the health cost of (and hence reduced demand for) unhealthy working and living environments, and unhealthy consumption. Jointly these behavioral choices gradually lead to growing health advantage with age. Further, the model predicts an initial widening and potentially a subsequent narrowing of the SES-health gradient, as low SES individuals increase their health investment and improve their health-related behavior faster as a result of their worse health. Results from earlier studies (Ehrlich and Chuma, 1990; Ehrlich, 2000; Galama et al. 2008 [Chapter 3]) suggest that the more rapidly worsening health of low SES individuals could lead to early withdrawal from the labor force, potentially widening the gradient in early and mid age, and shorter life spans, potentially narrowing the gradient in late age. Our model thus holds promise in explaining empirical health patterns. Such a model has not been available before and economists have highlighted the significance of its development (e.g., Cutler et al. 2011; Case and Deaton, 2005).

Chapter 2

Grossman's Missing Health Threshold

We present a generalized solution to Grossman's model of health capital (1972a, 1972b), relaxing the widely used assumption that individuals can adjust their health stock instantaneously to an "optimal" level without adjustment costs. The Grossman model then predicts the existence of a health threshold above which individuals do not demand medical care. Our generalized solution addresses a significant criticism: the model's prediction that health and medical care are positively related is consistently rejected by the data. We suggest structural and reduced form equations to test our generalized solution and contrast the predictions of the model with the empirical literature.

This chapter is based upon:

Galama, T.J. and A. Kapteyn (2009), "Grossman's Missing Health Threshold", *RAND Working Paper*, WR-684.

2.1 Introduction

Grossman's model of health capital (1972a, 1972b, 2000) is considered a breakthrough in the economics of the derived demand for medical care. In Grossman's human capital framework individuals demand medical care (e.g., invest time and consume medical goods and services) for the consumption benefits (health provides utility) as well as production benefits (healthy individuals have greater earnings) that good health provides. The model has been employed widely to explore a variety of phenomena related to health, medical care, inequality in health, the relationship between health and socioeconomic status, occupational choice, etc (e.g., Muurinen and Le Grand, 1985; Case and Deaton, 2005; Cropper, 1977).

Yet the Grossman model has also received significant criticism. For example, the model has been criticized for its simplistic deterministic nature (e.g., Cropper, 1977, Dardanoni and Wagstaff, 1987), for not determining length of life (e.g., Ehrlich and Chuma, 1990), for allowing complete health repair (Case and Deaton, 2005), and for its formulation in which medical investment in health has constant returns which is argued to lead to an unrealistic "bang-bang" solution (e.g., Ehrlich and Chuma, 1990). The criticism has led to theoretical and empirical extensions of the model (often by the same authors who provided the criticism), which to a large extent address the issues identified.[1] For an extensive review see Grossman (2000) and the work referenced therein.

However, there is one most significant criticism that thus far has not satisfactorily been addressed. Zweifel and Breyer (1997; p. 62) reject the Grossman model's central proposition that the demand for medical care is derived from the demand for good health: *" ... the notion that expenditure on medical care constitutes a demand derived from an underlying demand for health cannot be upheld because health status and demand for medical care are negatively rather than positively related ..."* In a review of the empirical literature Zweifel and Breyer conclude that the model's prediction that health and medical care should be positively related (healthy individuals consume more medical goods and services) is consistently rejected by the data. For example, Cochrane et al. (1978) find in a study of various determinants of mortality across various countries that indicators of medical care usage are positively related to mortality. And more specifically, Wagstaff (1986a) and Leu and Gerfin (1992), in estimating structural and reduced form equations

[1] With the exception perhaps of the "bang-bang" solution and for allowing complete health repair, which we will discuss briefly in this work.

of the Grossman model, find that measures of medical care are negatively correlated with measures of health and that the relationships are highly significant.[2]

It is of importance that this criticism be addressed. Dismissal of the central proposition of the Grossman model essentially amounts to rejecting the model itself. And a model of health and medical care should at a minimum predict the correct sign of the relationship between the two.

Several authors have sought to explain the consistently negative relation between health and medical care in empirical studies. For example, Grossman argues that the observed negative relation could be attributed to biases that arise if the conditional demand function is estimated with health treated as exogenous (Grossman 2000; p. 386). Further, Grossman (2000; pp. 369-370) shows that the model does not always produce the incorrect sign for the relationship between health and investment in medical care. For the pure investment model and assuming that the "natural" deterioration rate increases with age (a necessary assumption for the health stock to decline with age in Grossman's formulation), Grossman finds that investment in medical care increases with age while the health stock falls with age if the elasticity of the marginal production benefit of health with respect to health is less than one (Grossman refers to this as the MEC schedule). Thus it is the relation between earnings and health (the marginal production benefit of health or MEC schedule) that is responsible for the observed negative relation.

Muurinen and Le Grand (1985), in attempting to explain the positive relation between mortality and medical care usage found by Cochrane et al. (1978), suggest that the negative relation between indicators of health and of medical care (apart from suggesting that medical care is actually harmful) could be explained by differences in socioeconomic status. Individuals with fewer resources derive relatively higher production benefits from their health stock. They thus would have relatively greater usage of the stock (i.e., higher rates of health deterioration) which would require higher medical care to compensate for health losses. But if health cannot be completely repaired due to the increased use-intensity they would have inferior health states. High mortality would then be positively correlated with use of health services.

Wagstaff (1986a) provides a detailed discussion of potential reasons why estimates of the Grossman model may lead to a negative relation between measures of medical care usage and measures of health. On the one hand, one might argue that the coeffi-

[2]Numerous other studies do not specifically test Grossman's structural and reduced form equations, but broadly test similar relations between measures of health and measures for the demand for medical goods and services, controlling for relevant demographic and other characteristics. These studies find similar results. See section 2.4 for a discussion.

cients determined in Wagstaff (1986a) and similar analyses are not reliable estimates of the model's parameters. For example, Wagstaff suggests that in moving from the theoretical to the empirical model inappropriate assumptions may have been introduced (see Wagstaff, 1986a, for details). Or the identification of medical care with market inputs may insufficiently characterize health inputs if non-medical inputs are important in the production of health. On the other hand, one may take the estimates at face value and seek explanations in terms of the underlying model. Interestingly, Wagstaff (1986a) suggests that, contrary to what is assumed in Grossman's theoretical work, the negative relationship may reflect a non-instantaneous adjustment of health capital to its "optimal" value.[3] This, Wagstaff argues, may be the result of a constraint on medical care or be due to the existence of adjustment costs. Wagstaff finds in subsequent analysis (Wagstaff, 1993) that a reformulation of Grossman's empirical model with non-instantaneous adjustment is not only more consistent with Grossman's theoretical model but also with the data.

Indeed, in earlier theoretical work building on a simplified version of the Grossman model (Galama et al. 2008; see Chapter 3) we concluded that the widely employed assumption in the Grossman literature that any health "excess" or "deficit" can be adjusted instantaneously and at no adjustment cost may be too restrictive. Any "excess" in health capital cannot rapidly dissipate as individuals with "excessive" health can at best decide not to consume medical care.[4] As a consequence their health deteriorates at the natural deterioration rate $d(t)$ (i.e., non instantaneous) until health reaches Grossman's "optimal" level. Thus an individual's health is not always at the predicted "optimal" level. While the widely employed assumption that an individual's health follows Grossman's solution for the "optimal" path allows one to derive simple model predictions for empirical validation (and indeed this may be the primary reason for its use), it is otherwise unnecessary and is not demanded by theory. Importantly, Wagstaff's (1993) work suggests that individuals do not adjust their health stocks instantaneously. In other words, not only is there no theoretical basis for the assumption, empirical evidence suggests the assumption is not valid.

In this paper we relax the widely used assumption that individuals can adjust their health stock to Grossman's "optimal" level instantaneously. We do not restrict an in-

[3]Throughout this paper we will refer to Grossman's solution for the optimal health level as "optimal" health (using quotation marks) to reflect the fact that the Grossman solution is not always the optimal solution. Grossman's solution is optimal only in the absence of corner solutions. In this work we explore corner solutions in which individuals do not consume medical care for periods of time. The Grossman solution is then strictly speaking not the optimal solution.

[4]In other words medical care is restricted to be non-negative and the situation where individuals do consume medical care represents a corner solution.

dividual's health path to Grossman's "optimal" solution but allow for corner solutions where the optimal response for healthy individuals is to not consume medical goods and services for some period of time. We then find that the Grossman model predicts a substantially different pattern of medical care over the life-time than previously was assumed. Healthy individuals initially do not demand medical care till their health has deteriorated to a certain threshold level given by Grossman's "optimal" health. Subsequently their health evolves as the Grossman solution for the "optimal" path as individuals begin to demand medical care. In other words, Grossman's "optimal" health level is in fact a "health threshold" rather than an "optimal" trajectory. This simple pattern potentially addresses the most damning criticism: we find that the Grossman model predicts that healthy individuals (those above the threshold) do not consume medical care, but the unhealthy (at the threshold) do. Grossman's model thus predicts that healthy individuals demand less medical care, not the opposite, in agreement with the empirical literature.

Our working hypothesis is that a significant share of the population is healthy for much of their life. In our definition the healthy do not demand medical care. This would help explain the observed negative relation between measures of health and measures of medical care. Further, as we will see, this hypothesis can explain a number of other empirical facts.

A consequence of the assumption that a significant share of the population is healthy for much of their life, combined with the threshold nature of the demand for medical care, is that health investment in the Grossman model is to be strictly interpreted as medical care. It is the type of health investment (own time inputs and purchases of goods and services in the market) that individuals engage in when they are unhealthy and seek to "repair" their health. The Grossman literature sometimes views health investment as including a wide range of other types of investments, such as: preventive care (e.g., medical check ups), healthy dieting, and sports / exercise. Strictly speaking, the Grossman model does not contain the concept of healthy or unhealthy consumption nor of preventive care. In contrast to medical care, individuals engage in such activities when they are healthy as well as when they are unhealthy. In other words, these types of health investment are not part of the current formulation of the Grossman model where health investments take place only when individuals are unhealthy. The Grossman model, however, does offer an alternative way to include such health investments, by slowing the deterioration rate. For example, Case and Deaton (2005) model the effect of healthy consumption (e.g., healthy dieting, sports / exercise) as slowing and unhealthy consumption (e.g., smoking, excessive alcohol consumption) as accelerating the rate of deterioration. Preventive care

may operate in a similar manner. Here we consider these extensions as beyond the scope of the current paper.

As mentioned before, we are motivated by the lack of a theoretical justification in the Grossman literature for employing the assumption that health is always at Grossman's "optimal" level (see Galama et al. 2008 [Chapter 3]) and by Wagstaff's (1993) empirical analysis that suggests the assumption is not valid. A further motivation comes from the observation that the above attempts to explain the observed negative relationship between measures of health and measures of medical care do not pass the principle of Occam's razor when compared to the simple explanation put forward here that individuals cannot adjust their health stocks instantaneously (Wagstaff 1986a, 1993; Galama et al. 2008 [Chapter 3]). Our proposed explanation is the simplest in that we adopt the Grossman model as is and make one fewer assumption than is commonly made in the Grossman literature.

The aim of this paper is to investigate the solutions and predictions of the Grossman model without restricting the solutions to Grossman's so-called "optimal" solution by allowing for corner solutions. We proceed as follows. In section 2.2, we reformulate the Grossman model in continuous time allowing for corner solutions, solve the optimal control problem and derive first-order conditions for consumption and health. In section 2.3 we present structural form and reduced form solutions for health, medical care and consumption to enable empirical testing of our reformulation of the Grossman model. In section 2.4 we contrast the predictions of our generalized solution of the Grossman model with the traditional solution and with the empirical literature. We conclude in section 2.5 and provide detailed derivations in the Appendix.

2.2 General framework: the full Grossman model

We present the original human-capital model of the derived demand for health by Grossman (Grossman, 1972a, 1972b, 2000) in continuous time (see also Wagstaff, 1986a; Wolfe, 1985; Zweifel and Breyer, 1997; Ehrlich and Chuma, 1990). Health is treated as a form of human capital (health capital) and individuals derive both consumption (health provides utility) and production benefits (health increases earnings) from it. The demand for medical care is a derived demand: individuals demand "good health", not the consumption of medical care. In the original formulation of the Grossman model (Grossman, 1972a, 1972b, 2000) health yields an output of healthy time and consumption and medical care constitute both own-time inputs and goods or services purchased in the market. Simplified versions of the Grossman model have been presented by Case and Deaton (2005)

who assume consumption and production benefits are functions of health rather than healthy time, Wolfe (1985) who assumes health does not provide utility, and Case and Deaton (2005) and Wagstaff (1986a) who do not include time inputs into the production of consumption nor in the production of medical care. For an excellent review of the basic concepts of the Grossman model see Muurinen and Le Grand (1985).

Individuals maximize the life-time utility function

$$\int_0^T U\{C(t), s[H(t)]\}e^{-\beta t}dt, \tag{2.1}$$

where T denotes total life time, β is a subjective discount factor and individuals derive utility $U\{C(t), s[H(t)]\}$ from consumption $C(t)$ and from reduced sick time $s[H(t)]$. Sick time is assumed to be a function of health $H(t)$. Time t is measured from the time individuals begin employment. Utility decreases with sick time $\partial U(t)/\partial s(t) \leq 0$ and increases with consumption $\partial U(t)/\partial C(t) \geq 0$. Sick time decreases with health $\partial s(t)/\partial H(t) \leq 0$. Further we assume diminishing marginal benefits: $\partial^2 U(t)/\partial^2 s(t) \geq 0$ and $\partial^2 U(t)/\partial^2 C(t) \leq 0$.

The objective function (2.1) is maximized subject to the following constraints:

$$\dot{H}(t) = I(t) - d(t)H(t), \tag{2.2}$$

$$\dot{A}(t) = \delta A(t) + Y\{s[H(t)]\} - p_X(t)X(t) - p_m(t)m(t), \tag{2.3}$$

and we have initial and end conditions: $H(0)$, $A(0)$ and $A(T)$ are given.

$\dot{H}(t)$ and $\dot{A}(t)$ in equations (2.2) and (2.3) denote time derivatives of health $H(t)$ and assets $A(t)$. Health (equation 2.2) can be improved through medical health investment $I(t)$ (medical care) and deteriorates at the "natural" health deterioration rate $d(t)$. Using equation (2.2) we can write $H(t)$ as a function of medical care $I(t)$ and initial health $H(0)$,

$$H(t) = H(0)e^{-\int_0^t d(s)ds} + \int_0^t I(x)e^{-\int_x^t d(s)ds}dx. \tag{2.4}$$

Assets $A(t)$ (equation 2.3) provide a return δ (the interest rate), increase with income $Y\{s[H(t)]\}$ and decrease with purchases in the market of goods $X(t)$ and medical goods and services $m(t)$ at prices $p_X(t)$ and $p_m(t)$, respectively. Income $Y\{s[H(t)]\}$ is assumed to be a decreasing function of sick time $s[H(t)]$.

Integrating equation (2.3) over the life time we obtain the life-time budget constraint

$$\begin{aligned} &\int_0^T p_X(t)X(t)e^{-\delta t}dt + \int_0^T p_m(t)m(t)e^{-\delta t}dt = \\ &A(0) - A(T)e^{-\delta T} + \int_0^T Y\{s[H(t)]\}e^{-\delta t}dt. \end{aligned} \tag{2.5}$$

The left-hand side of (2.5) represents life-time consumption of market goods and life-time consumption of medical goods and services, and the right-hand side represents life-time financial resources in terms of life-time assets and life-time earnings.

Goods $X(t)$ purchased in the market and own time inputs $\tau_C(t)$ are used in the production of consumption $C(t)$. Similarly medical goods and services $m(t)$ and own time inputs $\tau_I(t)$ are used in the production of medical care $I(t)$. The efficiencies of production are assumed to be a function of the consumer's stock of knowledge E (an individual's human capital exclusive of health capital [e.g., education]) as it is generally believed that the more educated are more efficient consumers of medical care (see, e.g., Grossman 2000),

$$I(t) = I[m(t), \tau_I(t); E], \tag{2.6}$$

$$C(t) = C[X(t), \tau_C(t); E]. \tag{2.7}$$

The total time available in any period $\Omega(t)$ is the sum of all possible uses $\tau_w(t)$ (work), $\tau_I(t)$ (medical care), $\tau_C(t)$ (consumption) and $s[H(t)]$ (sick time),

$$\Omega(t) = \tau_w(t) + \tau_I(t) + \tau_C(t) + s[H(t)]. \tag{2.8}$$

In this formulation one can interpret $\tau_C(t)$, the own-time input into consumption $C(t)$ as representing leisure.

Income $Y\{H[s(t)]\}$ is taken to be a function of the wage rate $w(t)$ times the amount of time spent working $\tau_w(t)$,

$$Y\{H[s(t)]\} = w(t)\left\{\Omega(t) - \tau_I(t) - \tau_C(t) - s[H(t)]\right\}. \tag{2.9}$$

So far we have simply followed Grossman's formulation in continuous time. See Wagstaff (1986a), Wolfe (1985), Zweifel and Breyer (1997), and Ehrlich and Chuma (1990) for similar formulations. Our formulation differs however in one crucial respect from prior work: we explicitly impose the constraint that medical care is non-negative for all ages and allow for corner solutions in which individuals do not demand medical care ($I(t) = 0$).

2.2.1 Periods where individuals do not demand medical care: $I(t) = 0$

It is commonly assumed that any initial "excess" in health capital can be shed and any "deficit" can be repaired over a small period of time and at negligible cost. In other words, individuals are capable of ensuring that their health is at a certain desirable or "optimal" level (e.g., Grossman, 1972a, 1972b, 2000; Case and Deaton, 2005; Muurinen, 1982; Wagstaff, 1986a; Zweifel and Breyer, 1997; Ehrlich and Chuma, 1990; Ried,

1998).[5] This assumption is not necessarily always stated explicitly. The literature generally assumes that there are no corner solutions. In making this assumption the literature restricts the solution to Grossman's "optimal" solution. While this allows one to derive simple model predictions for empirical validation, it is unnecessary.

It is useful to view medical health investment $I(t)$ as encompassing activities related to health repair (e.g., purchases of medical goods and services and own-time inputs) and to view health-damaging environments (e.g., work and living environments, etc) as affecting the rate $d(t)$ at which health capital deteriorates (see, e.g., Wagstaf, 1986a; Case and Deaton, 2005). Similar to Grossman (1972a, 1972b, 2000) we treat the health deterioration rate $d(t)$ as strictly exogenous.

Healthy individuals, those with health levels above the "optimal" level, may desire to substitute health capital for more liquid capital. In other words, individuals may wish to "sell" their health. But, as equation (2.4) shows individuals cannot "choose" health optimally. Instead they can consume medical care (medical health investment) $I(t)$ optimally. But medical care $I(t)$, viewed as health-promoting cannot be traded (individuals cannot "sell" health through negative medical health investment) and is therefore positive for all ages $I(t) \geq 0$. As a result health cannot deteriorate faster than the health deterioration rate $d(t)$. This corresponds to the corner solution $I(t) = 0$.

Thus, we have the following optimal control problem: the objective function (2.1) is maximized with respect to the control functions $C(t)$ and $I(t)$ and subject to the constraints (2.2 and 2.3). The Lagrangean or generalized Hamiltonian (see, e.g., Seierstad and Sydsaeter 1987) of this problem is:

$$\begin{aligned} \Im &= U\{C(t), s[H(t)]\}e^{-\beta t} + q_H(t)\{I(t) - d(t)H(t)\} \\ &+ q_A(t)\{\delta A(t) + Y\{s[H(t)]\} - p_X(t)X(t) - p_m(t)m(t)\} + q_I(t)I(t), \end{aligned} \tag{2.10}$$

where $q_H(t)$ is the adjoint variable associated with the differential equation (2.2) for health $H(t)$, $q_A(t)$ is the adjoint variable associated with the differential equation (2.3) for assets $A(t)$, and $q_I(t)$ is a multiplier associated with the condition that health investment is non negative, $I(t) \geq 0$.

[5]While many authors realize that medical health investments cannot be negative (i.e. that corner solutions exist), the literature has not fully explored the implications of this constraint.

2.2.2 First-order conditions

The first-order condition for maximization of (2.1) with respect to consumption, subject to the conditions (2.2) and (2.3) is (see the Appendix for details)

$$\partial U(t)/\partial C(t) = q_A(0)\pi_C(t)e^{(\beta-\delta)t}, \tag{2.11}$$

where the Lagrange multiplier $q_A(0)$ is the shadow price of wealth (see, e.g., Case and Deaton 2005) and $\pi_C(t)$ is the marginal cost of consumption $C(t)$

$$\pi_C(t) \equiv \frac{p_X(t)}{\partial C(t)/\partial X(t)} = \frac{w(t)}{\partial C(t)/\partial \tau_C(t)}. \tag{2.12}$$

The first-order condition for maximization of (2.1) with respect to health, subject to the conditions (2.2) and (2.3) is (see the Appendix for details)

$$\frac{\partial U(t)}{\partial s(t)}\frac{\partial s(t)}{\partial H(t)} \equiv q_A(0)\left[\pi_H(t) - \varphi_H(t)\right]e^{(\beta-\delta)t} + \left[\dot{q}_I(t) - q_I(t)d(t)\right]e^{\beta t}, \tag{2.13}$$

where $\pi_H(t)$ is the user cost of health capital at the margin,

$$\pi_H(t) \equiv \pi_I(t)\left[d(t) + \delta - \tilde{\pi}_I(t)\right], \tag{2.14}$$

$\pi_I(t)$ is the marginal cost of medical health investment $I(t)$ (see equation 10 in Grossman, 2000)

$$\pi_I(t) \equiv \frac{p_m(t)}{\partial I(t)/\partial m(t)} = \frac{w(t)}{\partial I(t)/\partial \tau_I(t)}, \tag{2.15}$$

$\tilde{\pi}_I(t) \equiv \dot{\pi}(t)/\pi(t)$, and $\varphi_H(t)$ is the marginal production benefit of health

$$\varphi_H(t) \equiv \frac{\partial Y(t)}{\partial s(t)}\frac{\partial s(t)}{\partial H(t)}. \tag{2.16}$$

Note that we have to impose that the user cost of health capital at the margin exceeds the marginal production benefits of health $\pi_H(t) > \varphi_H(t)$. Without this condition, the consumption of medical care would finance itself by increasing wages by more than the user cost of health. As a result of this, consumers would choose infinite medical care paid for by infinite earnings increases to reach infinite health.

Equations (2.11) and (2.13) describe the first-order conditions for the constrained optimization problem. Equation (2.11) is similar to equation 4a by Wagstaff (1986a) and equation 6 by Case and Deaton (2005). Equation (2.13) is similar to equations 13, 1-13 and 11 of Grossman (1972a), (1972b) and (2000), respectively, equation 4b by Wagstaff (1986a), equation 3.5 of Zweifel and Breyer (1997), and equation 6 by Case and Deaton (2005), for $q_I(t) = 0$ (i.e., $I(t) > 0$).[6] The essential difference between our results and those of fore mentioned authors is in the term $q_I(t)$ which is non-vanishing for $I(t) = 0$.

[6] Various other authors have presented first-order conditions for the Grossman model. The list provided here is not exhaustive.

2.2.3 Grossman's solutions for consumption and health

The first-order condition (2.13) contains an expression in the multiplier $q_I(t)$ which is non-vanishing ($q_I(t) \neq 0$) for corner solutions in which individuals do not demand medical care ($I(t) = 0$). Let's first focus on the solution where $q_I(t) = 0$. This special case corresponds to the solutions found by Grossman (1972a, 1972b, 2000). The first-order condition (2.13) determines the "optimal" level of health for the "traditional" Grossman solution.

Denoting Grossman's "optimal" solutions for consumption, consumption goods, medical care, medical goods and services, own time input into the production of consumption, own time input into the production of medical care, sick time and health by $C_*(t)$, $X_*(t)$, $I_*(t)$, $m_*(t)$, τ_{C_*}, τ_{I_*}, $s_*(t)$, and $H_*(t)$, we have:

$$\partial U(t)/\partial C_*(t) = q_{A_*}(0)\pi_{C_*}(t)e^{(\beta-\delta)t}, \tag{2.17}$$

and,

$$\begin{aligned}\frac{\partial U(t)}{\partial s_*(t)}\frac{\partial s_*(t)}{\partial H_*(t)} &= q_{A_*}(0)\left\{\pi_{I_*}(t)\left[d(t)+\delta-\tilde{\pi_{I_*}}(t)\right]-\frac{\partial Y(t)}{\partial s_*(t)}\frac{\partial s_*(t)}{\partial H_*(t)}\right\}e^{(\beta-\delta)t} \\ &\equiv q_{A_*}(0)\left[\pi_{H_*}(t)-\varphi_{H_*}(t)\right]e^{(\beta-\delta)t}. \end{aligned}\tag{2.18}$$

The first-order condition (2.17) determines the level of consumption. It requires the marginal benefit of consumption to equal the product of the shadow price of wealth $q_{A_*}(0)$, the marginal cost of consumption $\pi_{C_*}(t)$, and a time varying exponent that either grows or decays with time, depending on the difference $\beta-\delta$ between the time preference rate β and the interest rate δ. Increasing lifetime resources will lower $q_{A_*}(0)$[7] and hence increase consumption. The marginal cost of consumption $\pi_{C_*}(t)$ increases with the price $p_{X_*}(t)$ of consumption goods $X_*(t)$ and with wages $w(t)$, and decreases with the efficiency of consumption goods in producing consumption, $\partial C_*(t)/\partial X_*(t)$ and with the efficiency of time inputs $\tau_{C_*}(t)$ in producing consumption, $\partial C_*(t)/\partial \tau_{C_*}(t)$ (see equation 2.12). Since the marginal benefit of consumption $\partial U(t)/\partial C_*(t)$ is a decreasing function of consumption $C_*(t)$, higher prices of consumption goods $p_{X_*}(t)$, higher wages $w(t)$ and lower efficiencies $\partial C_*(t)/\partial X_*(t)$ and $\partial C_*(t)/\partial \tau_{C_*}(t)$[8] lower the equilibrium level of consumption $C_*(t)$.

The marginal benefit of health (equation 2.18) equals the product of the shadow price of wealth $q_{A_*}(0)$, the user cost of health capital at the margin $\pi_{H_*}(t)$ minus the marginal production benefits of health $\varphi_{H_*}(t)$, and a time varying term with exponent $-(\beta-\delta)t$.

[7] This result can be obtained by substituting the solutions for consumption, health, and medical care in the budget constraint (equation 2.5) and solving for $q_A(0)$. See, for example, Galama et al. (2008) [Chapter 3].

[8] I.e., where large increases in $X_*(t)$ and/or $\tau_{C_*}(t)$ result in an insignificant increase in $C_*(t)$.

Since the marginal benefit of health $[\partial U(t)/\partial s_*(t)][\partial s_*(t)/\partial H_*(t)]$ is a decreasing function in health $H_*(t)$, lower lifetime resources (higher $q_{A_*}(0)$), higher user cost of health capital $\pi_{H_*}(t)$ and lower production benefits of health $\varphi_{H_*}(t)$ will lower the level of health $H_*(t)$. The user cost of health capital (see equations 2.15 and 2.14) increases with the price $p_{m_*}(t)$ of medical goods/services, with wages $w(t)$, the health deterioration rate $d(t)$ and the rate of return on assets δ (reflecting an opportunity cost). The user cost of health capital decreases with the efficiency of medical goods/services in producing medical care, $\partial I(t)/\partial m_*(t)$, the efficiency of time input $\tau_{I_*}(t)$ in producing medical care, $\partial I_*(t)/\partial \tau_{I_*}(t)$, and with $\tilde{\pi}_{I_*}(t)$, the rate of relative change in the marginal cost of medical care π_{I_*}. The marginal production benefit of health $\varphi_{H_*}(t)$ (equation 2.16) increases with the extent to which health increases earnings $[\partial Y(t)/\partial s_*(t)][\partial s_*(t)/\partial H_*(t)]$.

A lower price of medical goods/services thus increases health. This is pertinent in a cross-country comparison, but also when comparing across the life-cycle, for instance if health care is subsidized for certain age groups (like Medicare in the U.S.) Also, more efficient medical care will lead to greater health. Efficiency can explain variations within a country (if for instance individuals with a higher education level are more efficient consumers of medical care, Goldman and Smith, 2002) or across countries (if health care is more efficient in one country than in another).

2.2.4 Corner solutions

We allow for corner solutions in which individuals do not demand medical care $I(t) = 0$. This situation occurs when individuals have initial health endowments $H(0)$ that are greater than Grossman's "optimal" level of health $H_*(0)$.

We follow a simple intuitive approach. The corner solution is associated with a non-vanishing Lagrange multiplier $q_I(t)$. The solution for consumption is still provided by the first-order condition (2.11) as this condition is independent of the Lagrange multiplier $q_I(t)$. The solution for medical care is simply

$$I(t) = 0. \tag{2.19}$$

We do not need to use the first-order condition (2.13) to obtain the solution for health. Using equation (2.4) and $I(x) = 0$ we have

$$H(t) = H(0)e^{-\int_0^t d(s)ds}. \tag{2.20}$$

In other words, in the absence of medical care health deteriorates at the natural deterioration rate $d(t)$. The corner solution is fully determined by equations (2.11), (2.19) and (2.20).

2.3 Empirical model

The Grossman literature assumes that an individual's health follows Grossman's "optimal" health path, $H_*(t)$ (e.g., Grossman, 1972a, 1972b, 2000; Case and Deaton, 2005; Muurinen, 1982; Wagstaff, 1986a; Zweifel and Breyer, 1997; Ehrlich and Chuma, 1990; Ried, 1998). In other words, the literature assumes that either the initial health endowment $H(0)$ is at or very close to Grossman's "optimal" health stock $H_*(0)$ or that individuals find this health level desirable and are capable of rapidly dissipating or repairing any "excess" or "deficit" in health.

Corner solutions, where individuals do not demand medical care ($I(t) = 0$), occur when individuals are healthy, i.e. $H(t) > H_*(t)$. Health then deteriorates at the natural deterioration rate $d(t)$ (see equation 2.20) until it reaches Grossman's level $H(t) = H_*(t)$. Individuals then begin to demand medical care $I(t) > 0$. In other words, the Grossman solution for the "optimal" health stock represents a health "threshold" instead. In our generalized solution of the Grossman model, $H_*(t)$ is the minimum health level individuals "demand" to be economically productive (production benefits of health) or satisfied (consumption benefits of health). Individuals only consume medical care when they are "unhealthy" (health levels at the threshold) and not when they are "healthy" (health levels above the threshold).

Wolfe (1985) assumes an initial surplus of health and is, to the best of our knowledge, the only researcher who has attempted to explore the consequences of corner solutions in Grossman's model in some detail. Wolfe employs a simplified Grossman model where health (or, alternatively, reduced sick time as in Grossman's original formulation) does not provide utility. Wolfe interprets the onset of *" . . . a discontinuous mid-life increase in health investment . . . "* with retirement. We however do not associate the discontinuous increase in medical health investment with retirement but with becoming unhealthy (health levels at the health threshold leading to consumption of medical care to improve health). We allow the onset of medical health investment to take place anytime during the life of individuals, including allowing for the possibility that the onset never occurs. While Wolfe (1985) provides a convincing argument that high initial health endowments

are plausible[9], we simply assume that initial health $H(0)$ can take any positive value (including values below the health threshold).

We distinguish three scenarios as shown in Figure 2.1. We show the simplest case in which the health threshold $H_*(t)$ is constant across age (e.g., for constant user cost of health capital $\pi_{H_*}(t) = \pi_{H_*}(0)$, constant production benefits of health $\varphi_{H_*}(t) = \varphi_{H_*}(0)$ and for $\beta = \delta$; see equations 2.17 and 2.18) but the scenarios are valid for more general cases. Scenarios A and B begin with initial health $H(0)$ greater than the initial health threshold $H_*(0)$ and scenario C begins with initial health $H(0)$ below the initial health threshold $H_*(0)$. In scenario A health $H(t)$ reaches the health threshold $H_*(t)$ during life (before the age of death T) at age t_1. In scenario B health $H(t)$ never reaches the health threshold $H_*(t)$ during the life of the individual. In scenario C individuals begin working life with health levels $H(0)$ below the initial health threshold $H_*(0)$.

In scenarios A and B the solution for health is determined by the corner solution presented in section 2.2.4 for young ages (scenario A) or all ages (scenario B). In scenario A, after health reaches the threshold level the solutions are determined by the "traditional" Grossman solution. In scenarios A and B we do not have to assume that individuals adjust their health to reach the health threshold.

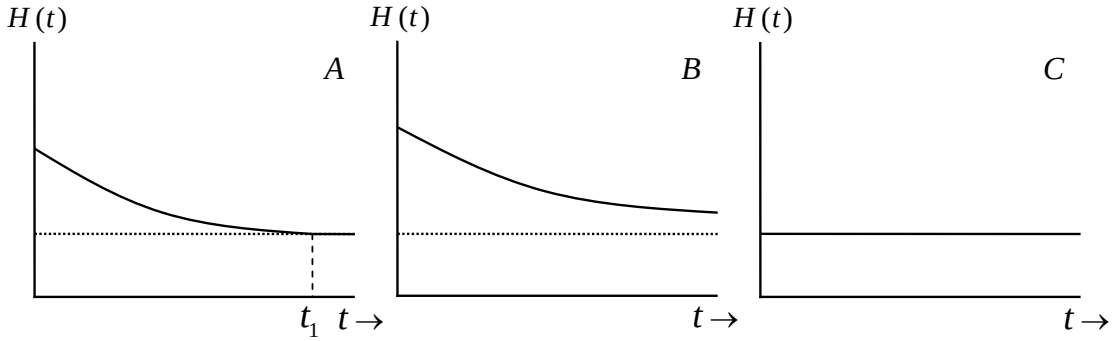
In contrast, in scenario C we follow the traditional Grossman model and assume that an individual is able to adjust his/her health level to reach the health threshold ("optimal" health). Individuals will invest initial assets $A(0)$ to improve initial health $H(0)$ such that initial health equals the initial health threshold $H(0) = H_*(0)$. These solutions have been criticized by Ehrlich and Chuma (1990) as being unrealistic "bang-bang" solutions; the adjustment takes place instantaneously. It is, however, not necessary to assume that the adjustment is instantaneous as individuals will have had ample time to consume medical care before they enter the labor force. There is also naturally an adjustment cost associated with these medical investments in the sense that such individuals begin their work life with fewer assets as a result of the purchase of medical care in the market before they entered the labor force. In other words, by the time individuals enter the labor force their health has gradually reached the health threshold and the adjustment

[9]On the grounds that *"... the human species, with its goal of self-preservation, confronts a different problem than the individual who seeks to maximize utility. The evolutionary solution to the former may entail an excessive health endowment in the sense that an individual might prefer to have less health and to be compensated with wealth in a more liquid form ..."* In other words, humans may have been endowed with "excessive" health as a result of our evolutionary history which required good physical condition to hunt and gather food, defend ourselves, survive periods of hunger etc. Today's demands on human's physical condition are essentially based on the utility of good health and on economic productivity, which in an increasingly knowledge-intensive environment may be significantly smaller than in pre-historic times.

cost is reflected in reduced assets. The health of such individuals will then continue to evolve along the health threshold (the "optimal" health path).

Further, as mentioned before, our working hypothesis is that most individuals are healthy for most of their life (health levels above the health threshold). A consequence of this is that scenario C, where initial health is below the initial health threshold, is less relevant for our discussion. That is, we do not disagree with Ehrlich and Chumas criticism of the Grossman model. The formulation could benefit from a more realistic incorporation of medical technology (allowed to instantaneously take effect in the Grossman model) or from diminishing returns to medical care so that a consumer doesn't demand such investment all at once (the solution Ehrlich and Chuma offer; see also Case and Deaton, 2005). For the purpose of the current research such extensions would complicate the model and provide relatively little benefit.

Figure 2.1: Three scenarios for the evolution of health.



Notes: t_1 in scenario A denotes the age at which health (solid line) has evolved towards the threshold health level (dotted line).

Following Grossman (1972a, 1972b, 2000) and Wagstaff (1986a) we derive structural and reduced form equations for empirical testing. Empirical tests of Grossman's model in the empirical literature have been based on estimating two sub-models (1) the "pure investment" model in which the restriction $\partial U(t)/\partial H(t) = 0$ is imposed and (2) the "pure consumption" model in which the restriction $\partial Y(t)/\partial H(t) = 0$ is imposed. To allow comparison with previous research we adopt the same restrictions and explore the same two sub-models. As Wagstaff (1986a) notes equation (2.18) can be transformed into a linear estimating equation with the restriction $\partial U(t)/\partial H(t) = 0$ or $\partial Y(t)/\partial H(t) = 0$, but this is not the case for the more general model. In addition, without imposing these restrictions analytical solutions for health, medical care and consumption cannot be obtained without making further assumptions. Lastly, the two sub models represent two

essential characteristics of health: health as a means to produce (investment) and health as a means to provide utility (consumption). We now discuss each sub-model in turn.

2.3.1 Pure investment model

In the following we follow Grossman (1972a, 1972b, 2000). We impose

$$[\partial U(t)/\partial s(t)][\partial s(t)/\partial H(t)] = 0, \tag{2.21}$$

assume that sick time is a power law in health

$$s(t) = \beta_0 + \beta_1 H(t)^{-\beta_2}, \tag{2.22}$$

where β_1 and β_2 are positive constants (e.g., Wagstaff, 1986a).[10] We thus have

$$[\partial Y(t)/\partial s(t)][\partial s(t)/\partial H(t)] = \beta_1\beta_2 w(t) H(t)^{-(\beta_2+1)}. \tag{2.23}$$

We further assume that medical health investment (medical care) is produced by combining own time and medical goods/services according to a Cobb-Douglass constant returns to scale production function

$$I(t) = \mu_I(t) m(t)^{1-k_I} \tau_I(t)^{k_I} e^{\rho_I E}, \tag{2.24}$$

where $\mu_I(t)$ is an efficiency factor, $1-k_I$ is the elasticity of medical care $I(t)$ with respect to medical goods/services $m(t)$, k_I is the elasticity of medical care $I(t)$ with respect to health time input $\tau_I(t)$, and ρ_I determines the extent to which education E improves the efficiency of medical care $I(t)$. Further, the ratio of the marginal product of medical care with respect to medical goods/services $\partial I(t)/\partial m(t)$ and the marginal product of medical care with respect to own time investment $\partial I(t)/\partial \tau_I(t)$ equals the ratio of the price of medical goods/services $p_m(t)$ to the wage rate $w(t)$ (representing the opportunity cost of time; see equation 2.15)

$$\frac{\partial I(t)/\partial m(t)}{\partial I(t)/\partial \tau_I(t)} = \frac{p_m(t)}{w(t)} = \frac{1-k_I}{k_I}\frac{\tau_I(t)}{m(t)}. \tag{2.25}$$

Lastly, we follow Wagstaff (1986a) and Cropper (1981) and assume the health deterioration rate $d(t)$ to be of the form

$$d(t) = d_\bullet e^{\beta_3 t + \boldsymbol{\beta_4}\mathbf{X}(t)}, \tag{2.26}$$

where $d_\bullet \equiv d(0)e^{-\boldsymbol{\beta_4}\mathbf{X}(0)}$ and $\mathbf{X}(t)$ is a vector of environmental variables (e.g., working and living conditions, hazardous environment, etc) that affect the deterioration rate. The vector $\mathbf{X}(t)$ may include other exogenous variables that affect the deterioration rate, such as education (Muurinen, 1982).

[10] But note that negative values can be allowed as long as $\beta_1\beta_2 > 0$

Health threshold

Structural form equations

The structural form equation for the health "threshold" (Grossman's solution for "optimal" health) is as follows (see the Appendix for details)

$$\begin{aligned} \ln H(t) &= \beta_5 + \epsilon(1-k_I)\ln w(t) - \epsilon(1-k_I)\ln p_m(t) + \epsilon\rho_I E - \epsilon(\beta_3+\beta_6)t - \epsilon\boldsymbol{\beta_4}\mathbf{X}(t) \\ &- \epsilon\ln d_\bullet - \epsilon\ln\{1 + d_\bullet^{-1}e^{-\beta_3 t - \boldsymbol{\beta_4}\mathbf{X}(t)}[\delta - k_I\tilde{w}(t) - (1-k_I)\tilde{p_m}(t) - \beta_6]\}, \quad (2.27) \end{aligned}$$

where $\epsilon \equiv (\beta_2+1)^{-1}$, the constant $\beta_5 \equiv \epsilon\ln(\beta_1\beta_2) + \epsilon\ln[k_I^{k_I}(1-k_I)^{(1-k_I)}] + \epsilon\ln\mu_I(0)$, and we allow medical technology $\mu_I(t) = \mu_I(0)e^{-\beta_6 t}$ to depend on age (e.g., the efficiency of medical goods/services $m(t)$ and own time inputs $\tau_I(t)$ in improving health could diminish with age).[11] It is customary to assume that the term $\ln d_\bullet$ in equation (2.27) is an error term with zero mean and constant variance $\xi_1(t) \equiv -\ln d_\bullet$ (as in Wagstaff, 1986a, and Grossman, 1972a, 1972b, 2000) and that the term $\ln[1+\delta/d(t)-\tilde{\pi_I}(t)/d(t)]$ (the last term in equation 2.27) is small or constant (see, e.g., Grossman, 1972a, 2000),[12] or that it is time dependent $\ln[1+\delta/d(t)-\tilde{\pi_I}(t)/d(t)] \propto t$ (e.g, Wagstaff, 1986a). We do not have to make these assumptions as in our generalized solution of the Grossman model the rate of deterioration $d(t)$ is observable for those times that individuals do not demand medical care (i.e., for corner solutions). While we assume that the last term in equation (2.27) is small, our formulation allows us to estimate and test this common assumption.

The demand for health (equation 2.27) thus increases with wages $w(t)$ and with education E and decreases with prices $p_m(t)$ and the health deterioration rate (terms $d_\bullet$, β_3 and $\boldsymbol{\beta_4}\mathbf{X}(t)$). The relation with age t is ambiguous. To ensure that health declines with age, it is commonly assumed that health deterioration increases with age, $\dot{d}(t) > 0$ (i.e. that $\beta_3 > 0$).[13] But since wages $w(t)$ generally increase with years of experience (e.g., Mincer 1974) it is possible that the health threshold initially increases with age t.

[11] For example, elderly and frail patients may not be able to cope with certain aggressive chemotherapy regiments. Note also that advances in medical technology could be modeled by an increasing $\mu_I(0)$ with time (e.g., $\mu_I(0)$ increases with subsequent cohorts).

[12] This would require that the real interest rate δ and changes in the ratio of the price of medical goods/services and the efficiency of medical goods/services in producing medical care $\pi_I(t) = p_m(t)/[\partial I(t)/\partial m(t)]$ are much smaller than the health deterioration rate $d(t)$ or that changes in the interest rate and in $\tilde{\pi_I}(t)$ follow the same pattern as changes in $d(t)$ (so that the term is approximately constant).

[13] Assuming that the efficiency of medical care decreases with age $\beta_6 > 0$ provides an alternative means to achieve the same result.

The structural equation for the "optimal" consumption of medical goods/services is as follows

$$\begin{aligned} \ln m(t) &= \beta_7 + \ln H(t) + k_I \ln w(t) - k_I \ln p_m(t) - \rho_I E \\ &+ (\beta_3 + \beta_6)t + \boldsymbol{\beta_4}\mathbf{X}(t) + \ln d_\bullet + \ln[1 + \tilde{H}(t) d_\bullet^{-1} e^{-\beta_3 t - \boldsymbol{\beta_4}\mathbf{X}(t)}], \end{aligned} \tag{2.28}$$

where $\beta_7 \equiv -\ln \mu_I(0) - k_I \ln [k_I/(1-k_I)]$. It is customary to assume that the last term in equation (2.28), $\ln[1+\tilde{H}(t)/d(t)] = \ln[1+\tilde{H}(t) d_\bullet^{-1} e^{-\beta_3 t - \boldsymbol{\beta_4}\mathbf{X}(t)}]$, is small and can be ignored (Grossman, 1972b) or treated as an error term (Wagstaff, 1986a). This would require that the effective rate of change in health $\dot{H}(t)$ is smaller than $d(t)H(t)$. This assumption is perhaps not unreasonable if medical care is efficient and slows down the effective health decline $\dot{H}(t)$. Note, once more that in our generalized solution of the Grossman model $d(t)$ can be observed during times when corner solutions hold. The last term in equation (2.28) can thus be estimated. For small $\tilde{H}(t)/d(t)$, we have $\ln[1+\tilde{H}(t)/d(t)] \sim \tilde{H}(t)/d(t)$.

Equation (2.28) predicts that Grossman's "optimal" demand for medical goods/services and Grossman's "optimal" demand for health are positively related. This is the crucial prediction which empirical studies consistently reject. Further, the demand for medical goods/services increases with wages $w(t)$ and the health deterioration rate (terms $d_\bullet$, β_3 and $\boldsymbol{\beta_4}\mathbf{X}(t)$), and decreases with education E and prices $p_m(t)$.

The literature usually focuses on the equations for health (2.27) and medical care (2.28), but note that equation (2.11) provides a condition for consumption $C(t)$ as well, which, after making some reasonable assumptions, can be utilized to obtain expressions for consumption goods $X(t)$ (see the Appendix for details). The budget constraint (equation 2.5) then provides the solution for assets $A(t)$.

Reduced form equations

Wagstaff (1986a) notes that one way of overcoming the unobservability of health capital is to estimate reduced-from demand functions for health and medical goods/services. Combining (2.27) and (2.28) and eliminating any expression in health $H(t)$ we find (see the Appendix for details):

$$\begin{aligned} \ln m(t) &= \beta_8 + [k_I + \epsilon(1-k_I)] \ln w(t) - [k_I + \epsilon(1-k_I)] \ln p_m(t) \\ &- (1-\epsilon)\rho_I E - \epsilon[\beta_3 - (1-\epsilon)\beta_6]t - \epsilon \boldsymbol{\beta_4}\mathbf{X}(t) - \epsilon \ln d_\bullet \\ &- \epsilon \ln\{1 + d_\bullet^{-1} e^{-\beta_3 t - \boldsymbol{\beta_4}\mathbf{X}(t)} [\delta - k_I \tilde{w}(t) - (1-k_I)\tilde{p_m}(t) - \beta_6]\} \\ &+ \ln\{\epsilon(1-k_I)[\tilde{w}(t) - \tilde{p_m}(t)] - \epsilon(\beta_3 + \beta_6) - \epsilon \boldsymbol{\beta_4} \partial \mathbf{X}(t)/\partial t \\ &+ d_\bullet e^{\beta_3 t + \boldsymbol{\beta_4}\mathbf{X}(t)} + \epsilon \mathcal{O}(t)\}, \end{aligned} \tag{2.29}$$

where $\beta_8 \equiv \beta_5 + \beta_7$ and

$$\begin{aligned}\mathcal{O}(t) &= \frac{\tilde{d}(t)[\delta - k_I\tilde{w}(t) - (1-k_I)\tilde{p_m}(t) - \beta_6]}{[d(t) + \delta - k_I\tilde{w}(t) - (1-k_I)\tilde{p_m}(t) - \beta_6]} \\ &+ \frac{k_I[\frac{\ddot{w}(t)}{w(t)} - \tilde{w}(t)^2] + (1-k_I)[\frac{\ddot{p_m}(t)}{p_m(t)} - \tilde{p_m}(t)^2]}{[d(t) + \delta - k_I\tilde{w}(t) - (1-k_I)\tilde{p_m}(t) - \beta_6]}\end{aligned} \tag{2.30}$$

which we assume to be small (of the order $\tilde{d}(t) \times \delta, \tilde{d}(t) \times \tilde{w}(t)$, etc).

The demand for medical goods/services (equation 2.29) increases with wages $w(t)$ and the efficiency of medical care (term β_6), and decreases with prices $p_m(t)$, education E, and the health deterioration rate (terms $d_\bullet$, β_3 and $\boldsymbol{\beta_4}\mathbf{X}(t)$).[14]

Corner solution

We have (using equations 2.20 and 2.26)

$$\ln H(t) = \ln H(0) - d_\bullet \int_0^t e^{\beta_3 s + \boldsymbol{\beta_4}\mathbf{X}(s)} ds, \tag{2.31}$$

and

$$m(t) = 0. \tag{2.32}$$

Note that during periods in which the corner solutions hold it is in principle possible to determine the rate of deterioration $d_\bullet$ empirically. Hence we do not have to assume that the term $\ln d_\bullet$ in equations (2.27) and (2.28) is an error term.

Regime switching

The time t_1 when health has deteriorated to the "threshold" level must satisfy the following condition (given by equating 2.27 with 2.31):

$$\begin{aligned}\ln H(t_1) &= \beta_5 + \epsilon(1-k_I)\ln w(t_1) - \epsilon(1-k_I)\ln p_m(t_1) + \epsilon\rho_I E - \epsilon(\beta_3+\beta_6)t_1 - \epsilon\boldsymbol{\beta_4}\mathbf{X}(t_1) \\ &- \epsilon\ln d_\bullet - \epsilon\ln\{1 + d_\bullet^{-1}e^{-\beta_3 t_1 - \boldsymbol{\beta_4}\mathbf{X}(t_1)}[\delta - k_I\tilde{w}(t_1) - (1-k_I)\tilde{p_m}(t_1) - \beta_6]\} \\ &= \ln H(0) - d_\bullet \int_0^{t_1} e^{\beta_3 s + \boldsymbol{\beta_4}\mathbf{X}(s)} ds\end{aligned} \tag{2.33}$$

The model thus implies a switch of regimes at time t_1. Before t_1 the evolution of health is given by equation (2.31), whereas after t_1 it is given by (2.27). Empirically, this would generate a switching regression model with endogenous switching. Once health hits the "optimal" path, the process governing health switches from (2.31) to (2.27). Similarly, before t_1 the demand for medical goods/services is given by equation (2.32), whereas after t_1 it is given by (2.28) or, alternatively, by (2.29).

[14] For $0 < \epsilon < 1$.

2.3.2 Pure consumption model

In the following we follow Wagstaff (1986a). We impose

$$[\partial Y(t)/\partial s(t)][\partial[s(t)/\partial H(t)] = 0. \tag{2.34}$$

To convert (2.18) into estimable equations we have to specify a functional form for the utility function.

Utility specification

Grossman (1972a, 1972b, 2000) formulates his model in terms of sick time[15] and assumes that sick time $s(t)$ is a function of health $H(t)$; $s(t) = s[H(t)]$. An alternative formulation is provided by Case and Deaton (2005). Case and Deaton formulate a simplified Grossman model in which utility and income are functions of health $H(t)$ directly, rather than indirectly through sick-time $s(t)$ which in turn is assumed to be a function of health $s(t) = s[H(t)]$ (as in Grossman, 1972a, 1972b, 2000). Following Case and Deaton we write utility $U\{C(t), s[H(t)]\} = U[C(t), H(t)]$ and income $Y\{s[H(t)]\} = Y[H(t)]$ as functions of health $H(t)$ instead of sick time $s(t)$. Essentially both formulations are equivalent except that Case and Deaton's formulation is more general, allowing for example for earnings to be influenced not only by reductions in sick time but also increased worker efficiency resulting from good health. And, at any time we can revert back to the original specification in terms of sick time if deemed desirable.

We begin by noting that (see the first-order conditions 2.11 and 2.13)

$$\frac{\partial U(t)}{\partial H(t)} = \pi_C(t)^{-1}\left[\pi_H(t) - \varphi_H(t)\right]\frac{\partial U(t)}{\partial C(t)} + \left[\dot{q}_I(t) - q_I(t)d(t)\right]e^{\beta t}. \tag{2.35}$$

In other words, the marginal benefit of health $\partial U(t)/\partial H(t)$ is given by the function $\pi_C(t)^{-1}[\pi_H(t) - \varphi_H(t)]$ times the marginal benefit of consumption $\partial U(t)/\partial C(t)$ and an additional expression in $q_I(t)$. For Grossman's solutions we have $q_I(t) = 0$ and the additional term vanishes.

Equation (2.35) suggests that the marginal utility of health $\partial U(t)/\partial H(t)$ and the marginal utility of consumption $\partial U(t)/\partial C(t)$ are functions of both health $H(t)$ and consumption $C(t)$. To allow for this we specify the following constant relative risk aversion (CRRA) utility function:

$$U[C(t), H(t)] = \frac{1}{1-\rho}\left[C(t)^{\zeta}H(t)^{1-\zeta}\right]^{1-\rho}, \tag{2.36}$$

[15] One possible reason for this formulation is that the NORC data set the author employed in empirical testing of the model contained information on sick days

where ζ ($0 \leq \zeta \leq 1$) is the relative "share" of consumption versus health and ρ ($\rho > 0$) the coefficient of relative risk aversion.

The functional form for the utility function can account for the observation that the marginal utility of consumption declines as health deteriorates (e.g., Finkelstein, Luttmer and Notowidigdo, 2008). The authors find that a one-standard deviation increase in the number of chronic diseases is associated with an 11 percent decline in the marginal utility of consumption relative to this marginal utility when the individual has no chronic diseases (the 95 percent confidence interval ranges between 2 percent and 17 percent). This would rule out the strongly separable functional form for the utility function employed by Wagstaff (1986a), where the marginal utility of consumption is independent of health. While we follow Wagstaff (1986a) in most of the derivations we do not adopt his utility specification.

Health threshold

Structural form equations

The structural equation for the health "threshold" (Grossman's solution for "optimal" health) is as follows (see the Appendix for details)

$$\begin{aligned} \ln H(t) &= \beta_9 + \ln X(t) + \ln p_X(t) - k_I \ln w(t) - (1-k_I)\ln p_m(t) \\ &+ \rho_I E - (\beta_3 + \beta_6)t - \boldsymbol{\beta_4}\mathbf{X}(t) - \ln d_\bullet \\ &- \ln\{1 + d_\bullet^{-1} e^{-\beta_3 t - \boldsymbol{\beta_4}\mathbf{X}(t)}[\delta - k_I \tilde{w}(t) - (1-k_I)\tilde{p_m}(t) - \beta_6]\}, \end{aligned} \tag{2.37}$$

where $\beta_9 \equiv \ln \mu_I(0) - \ln(1-k_C) + \ln[k_I^{k_I}(1-k_I)^{(1-k_I)}] + \ln[(1-\zeta)/\zeta]$. The health threshold thus increases with consumption goods $X(t)$, prices for consumption goods $p_X(t)$, and education E and decreases with wages $w(t)$, prices of medical goods/services $p_m(t)$, and the health deterioration rate (terms $d_\bullet$, β_3 and $\boldsymbol{\beta_4}\mathbf{X}(t)$). The last term is generally assumed to be small and can be estimated in our formulation.

The structural form equation for medical goods/services is the same as for the pure investment model (equation 2.28) and is repeated for convenience:

$$\begin{aligned} \ln m(t) &= \beta_7 + \ln H(t) + k_I \ln w(t) - k_I \ln p_m(t) - \rho_I E \\ &+ (\beta_3 + \beta_6)t + \boldsymbol{\beta_4}\mathbf{X}(t) + \ln d_\bullet + \ln[1 + \tilde{H}(t) d_\bullet^{-1} e^{-\beta_3 t - \boldsymbol{\beta_4}\mathbf{X}(t)}], \end{aligned} \tag{2.38}$$

where $\beta_7 \equiv -\ln \mu_I(0) - k_I \ln [k_I/(1-k_I)]$.

Combining equation (2.37) with (2.38) and eliminating any expression in health $H(t)$ we find (see the Appendix for details):

$$\begin{aligned} \ln m(t) &= \beta_{12} + \ln X(t) + \ln p_X(t) - \ln p_m(t) - \ln d_\bullet - \beta_3 t - \boldsymbol{\beta_4}\mathbf{X}(t) \\ &- \ln\{1 + d_\bullet^{-1} e^{-\beta_3 t - \boldsymbol{\beta_4}\mathbf{X}(t)}[\delta - k_I\tilde{w}(t) - (1-k_I)\tilde{p_m}(t) - \beta_6]\} \\ &+ \ln\{d_\bullet e^{\beta_3 t + \boldsymbol{\beta_4}\mathbf{X}(t)} - (1-k_I)\tilde{p_m}(t) - k_I\tilde{w}(t) - (\beta_3 + \beta_6) - \boldsymbol{\beta_4}\partial\mathbf{X}(t)/\partial t \\ &+ \tilde{X}(t) + \tilde{p_X}(t) + \mathcal{O}(t)\}, \end{aligned} \tag{2.39}$$

where $\beta_{12} \equiv \beta_7 + \beta_9$, and the expression for $\mathcal{O}(t)$ is provided by equation (2.30).

Reduced form equations

Note that the health threshold (equation 2.37) is expressed directly as a function of consumption goods $X(t)$. This relation is different from the one found by Wagstaff (1986a; his equation 12), which is the result of our choice for the functional form of the utility function (equation 2.36). Wagstaff (1986a) finds that health $H(t)$ is a function of the shadow price of wealth $q_A(0)$. We can obtain a similar reduced form expression to the one found by Wagstaff (1986a) by using the first-order condition (2.11) and making some reasonable assumptions to obtain an expression for consumption good $X(t)$. We then find (see the Appendix for details):

$$\begin{aligned} \ln H(t) &= \beta_{10} - \chi(1/\rho\chi - 1)(1-k_C)\ln p_X(t) - \chi(1-k_I)\ln p_m(t) \\ &- \chi[k_I + (1/\rho\chi - 1)k_C]\ln w(t) + \chi[\rho_I + (1/\rho\chi - 1)\rho_C]E \\ &- \chi[(\beta_3 + \beta_6) + (1/\rho\chi - 1)\beta_{11} + (\beta - \delta)/\rho\chi]t - \chi\boldsymbol{\beta_4}\mathbf{X}(t) \\ &- \chi\ln d_\bullet + \ln q_A(0)^{-1/\rho} \\ &- \chi\ln\{1 + d_\bullet^{-1} e^{-\beta_3 t - \boldsymbol{\beta_4}\mathbf{X}(t)}[\delta - k_I\tilde{w}(t) - (1-k_I)\tilde{p_m}(t) - \beta_6]\}, \end{aligned} \tag{2.40}$$

where

$$\begin{aligned} \beta_{10} &\equiv \chi\ln\mu_I(0) + \chi(1/\rho\chi - 1)\ln\mu_C(0) + \chi\ln[k_I^{k_I}(1-k_I)^{(1-k_I)}] \\ &+ \chi(1/\rho\chi - 1)\ln[k_C^{k_C}(1-k_C)^{(1-k_C)}] + \chi\ln[(1-\zeta)/\zeta] + \ln\zeta^{1/\rho}, \end{aligned}$$

and

$$\chi \equiv \frac{1 + \rho\zeta - \zeta}{\rho}, \tag{2.41}$$

and we allow the efficiency of consumption to depend on age $\mu_C(t) = \mu_C(0)e^{-\beta_{11}t}$.

An expression for the shadow price of wealth $q_A(0)$ in equation (2.40) can be obtained by using the life-time budget constraint (equation 2.5), substituting the solutions for consumption, health, and medical care and solving for $q_A(0)$ (see, for example, Galama

et al. 2008 [Chapter 3]). The shadow price of wealth $q_A(0)$ is found to be a complicated function of wealth (assets, life-time income), wages $w(t)$, prices $p_m(t)$, $p_X(t)$, education E and the health deterioration rate (terms $d_\bullet$, β_3 and $\boldsymbol{\beta_4}\mathbf{X}(t)$). Wagstaff (1986a) provides a simple approximation for the shadow price of wealth $q_A(0)$ (his equations 15 and 16) which may be easier to use in empirical testing of the model.

Assuming that both medical goods / services $m(t)$ and time input $\tau_I(t)$ increase medical care suggests $0 \leq k_I \leq 1$, and if education E increases the efficiency of medical care then $\rho_I > 0$ (see equation 2.24). Similarly we have $0 \leq k_C \leq 1$ and $\rho_C > 0$ (see equation 2.63). Finkelstein, Luttmer and Notowidigdo (2008) provide evidence that the marginal utility of consumption declines as health deteriorates. Assuming further diminishing marginal benefits of health $\partial^2 U(t)/\partial^2 H(t) < 0$ we find $1 < \chi < 1 + 1/\rho$ (and hence $0 < \rho < 1$ and $1/\rho\chi > 1$).

For these parameter values we find that the health threshold (equation 2.40) increases with education E, wealth $q_A(0)^{-1/\rho}$, and decreases with the price of consumption goods $p_X(t)$, the price of medical care $p_m(t)$, wages $w(t)$, and the health deterioration rate (terms $d_\bullet$, β_3 and $\boldsymbol{\beta_4}\mathbf{X}(t)$). The health threshold could increase or decrease with age depending on the sign of $\chi(\beta_3 + \beta_6) + \chi(1/\rho\chi - 1)\beta_{11} + [(\beta - \delta)/\rho]$ and on the evolution of wages $w(t)$ with years of experience (e.g., Mincer, 1974).

Combining equation (2.38) with (2.40) we find:

$$\begin{aligned}
\ln m(t) &= \beta_{13} - \chi(1/\rho\chi - 1)(1 - k_C)\ln p_X(t) - [k_I + \chi(1 - k_I)]\ln p_m(t) \\
&- \chi[(1 - 1/\chi)k_I + (1/\rho\chi - 1)k_C]\ln w(t) + \chi[(1 - 1/\chi)\rho_I + (1/\rho\chi - 1)\rho_C]E \\
&- \chi[(1 - 1/\chi)(\beta_3 + \beta_6) + (1/\rho\chi - 1)\beta_{11} + (\beta - \delta)/\rho\chi]t \\
&- (\chi - 1)\boldsymbol{\beta_4}\mathbf{X}(t) - (\chi - 1)\ln d_\bullet + \ln q_A(0)^{-1/\rho} \\
&- \chi\ln\{1 + d_\bullet^{-1}e^{-\beta_3 t - \boldsymbol{\beta_4}\mathbf{X}(t)}[\delta - k_I\tilde{w}(t) - (1 - k_I)\tilde{p_m}(t) - \beta_6]\}, \qquad (2.42)
\end{aligned}$$

where $\beta_{13} \equiv \beta_7 + \beta_{10}$. The demand for medical goods/services (equation 2.42) increases with education E, wealth $q_A(0)^{-1/\rho}$, and decreases with the price of consumption goods $p_X(t)$, the price of medical goods/services $p_m(t)$, wages $w(t)$, and the health deterioration rate (terms $d_\bullet$, β_3 and $\boldsymbol{\beta_4}\mathbf{X}(t)$). The health threshold could increase or decrease with age depending on the sign of $\chi(1 - 1/\chi)(\beta_3 + \beta_6) + \chi(1/\rho\chi - 1)\beta_{11} + [(\beta - \delta)/\rho]$ and on the evolution of wages $w(t)$ with years of experience (e.g., Mincer, 1974).

Corner solution

The solutions are given by the corner solutions (2.31) and (2.32) derived in section 2.2.4.

Regime switching

The time t_1 when health has deteriorated to the "threshold" level must satisfy the following condition (given by equating 2.37 or 2.40 with 2.31):

$$\begin{aligned}
\ln H(t_1) &= \beta_9 + \ln X(t_1) + \ln p_X(t_1) - k_I \ln w(t_1) - (1-k_I)\ln p_m(t_1) \\
&+ \rho_I E - (\beta_3+\beta_6)t_1 - \boldsymbol{\beta_4}\mathbf{X}(t_1) - \ln d_\bullet \\
&- \ln\{1 + d_\bullet^{-1} e^{-\beta_3 t_1 - \boldsymbol{\beta_4}\mathbf{X}(t_1)}[\delta - k_I\tilde{w}(t_1) - (1-k_I)\tilde{p_m}(t_1) - \beta_6]\} \\
&= \beta_{10} - \chi(1/\rho\chi - 1)(1-k_C)\ln p_X(t_1) - \chi(1-k_I)\ln p_m(t_1) \\
&- \chi[k_I + (1/\rho\chi - 1)k_C]\ln w(t_1) + \chi[\rho_I + (1/\rho\chi - 1)\rho_C]E \\
&- \chi[(\beta_3+\beta_6) + (1/\rho\chi - 1)\beta_{11} + (\beta-\delta)/\rho\chi]t_1 - \chi\boldsymbol{\beta_4}\mathbf{X}(t_1) \\
&- \chi\ln d_\bullet + \ln q_A(0)^{-1/\rho} \\
&- \chi\ln\{1 + d_\bullet^{-1} e^{-\beta_3 t_1 - \boldsymbol{\beta_4}\mathbf{X}(t_1)}[\delta - k_I\tilde{w}(t_1) - (1-k_I)\tilde{p_m}(t_1) - \beta_6]\} \\
&= \ln H(0) - d_\bullet \int_0^{t_1} e^{\beta_3 s + \boldsymbol{\beta_4}\mathbf{X}(s)} ds. \qquad (2.43)
\end{aligned}$$

Similar to the previous discussion for the pure investment model, the model thus implies a switch of regimes at time t_1. Before t_1 the evolution of health is given by equation (2.31), whereas after t_1 it is given by (2.37) or by (2.40). Empirically, this would generate a switching regression model with endogenous switching. Once health hits the optimal path, the process governing health switches from (2.31) to (2.37), or alternatively to (2.40). Similarly, before t_1 medical care is given by equation (2.32), whereas after t_1 it is given by (2.38) or alternatively (2.39) or (2.42).

2.4 Model Predictions

The Grossman model has been tested in a number of empirical studies on a variety of datasets from different countries (Grossman, 1972a; Wagstaff 1986a, 1993; Leu and Doppman, 1986; Leu and Gerfin, 1992; van Doorslaer, 1987; Van de Ven and van der Gaag, 1982; Erbsland, Ried and Ulrich, 2002; Gerdtham et al. 1999; Gerdtham and Johannesson, 1999).[16] Despite the large variety in methodologies and the diversity in cultural and institutional environments these datasets represent, the studies are broadly in agreement

[16]Grossman (1972a) employs the 1963 health interview survey conducted by the National Opinion Research Center (NORC) of the U.S. civilian noninstitutionalized population. Grossman employs measures of sick time and self-reported health and restricts the dataset to individuals with positive sick time. Wagstaff (1986a) employs the 1976 Danish Welfare Survey (DWS) and uses principal components analysis (PCA) to derive a smaller number of health components from a long list of health indicators. Wagstaff also uses the wealth of DWS measures of work environment and use-related health depreciation. Measures

with one another and confirm the predictions of the Grossman model for the demand for *health*. Health is found to increase with income (wages, life-time earnings), and education, and decreases with age, the price of medical goods/services, being single, and with environmental factors, such as, physically and mentally demanding work environments, manual labor, psychological stress factors.[17]

While reduced form estimates of the demand for *medical care* are generally in agreement with the predictions of the Grossman model, this is not true for structural estimates (see Wagstaff, 1986a). Structural estimates allow for direct testing of the relationship between health (most often a latent health variable is employed) and medical care. The most noticeable feature of such structural estimates is the consistently negative relationship between health and medical care (healthy individuals do not go to the doctor). But this relationship is predicted to be positive in the traditional solution of the Grossman model (see equation 2.28; those who consume more medical care are healthier). Further, the negative relationship between health and medical care is found to be the most sta-

of medical care employed are general practitioner visits, weeks in hospital and number of complaints for which medicine are taken. Wagstaff (1993) employs the Danish Health Survey (DHS) and uses a latent variable health model (multiple indicators multiple causes; MIMIC). Leu and Doppman (1986) employ a latent health variable, latent earnings and latent transfer income model based on Socio-medical indicators for the population of Switzerland (SOMIPOPS) data combined with the Swiss income and wealth study (SEVS). General practitioner consultations, hospital days and sick days are used as measures of medical care. Leu and Gerfin (1992) employ the same datasets as Leu and Doppman (1986) but follow a different methodology (health is a latent variable but no other latent variables are employed). Van Doorslaer (1987) estimates a latent health and latent medical knowledge variable model to the Health Interview Survey of the Belgian National Health Research Project on Primary Health Care conducted in 1976 among the Dutch-speaking (Flemish) population. Van de Ven and van der Gaag (1982) employ a MIMIC model with latent health and data from a health-care survey among 8000 households in the Netherlands. Erbsland, Ried and Ulrich (2002) use data from the German Socio Economic Panel (SOEP). They use a model with a latent health and a latent environment variable and restrict the analysis to the working population and to those with a positive demand for medical services. Gerdtham et al. (1999) use a rating scale and a time-trade off method to obtain measures of health as well as a self-reported measure of health from data collected in Uppsala County in Sweden. Gerdtham and Johannesson (1999) use a self-reported health measure from 1991 data of the Level of Living Survey (LNU), a random sample of the Swedish population. Both Gerdtham et al. (1999) and Gerdtham and Johannesson (1999) provide estimates of the demand for health but no structural estimates for the demand of medical care.

[17] In addition, these studies find that health increases with healthy behavior (sports, healthy eating and sleeping habits) and decreases with being overweight and with smoking. Females are found to be in lower health. And, moderate alcohol consumption is found to have a positive or negligible impact on health (e.g., Gerdtham et al. 1999, Leu and Doppman, 1986). Since the effect of consumption (healthy and unhealthy forms) on health as well as health behaviors (exercise, sleeping habits) and gender differences are not part of the Grossman model we do not discuss these here.

tistically significant of any relationship between medical care and any of the independent variables (see, e.g., Grossman, 1972a; Wagstaff, 1986a, 1993; Leu and Doppman, 1986; Leu and Gerfin, 1992; van Doorslaer, 1987; Van de Ven and van der Gaag, 1982; Erbsland, Ried and Ulrich, 2002).

We assume that each of the scenarios A, B and C occur in reality (see Figure 2.1). In other words, that there exist *healthy* individuals who consume medical care during some part of their life (scenario A; initial health above the initial health threshold and the threshold reached during life), *very healthy* individuals who never consume medical care (scenario B; initial health well above the initial health threshold and the threshold never reached), and *ill* individuals who consume medical care their entire life (scenario C; initial health at the health threshold). We do not a-priori know the distribution of healthy, very healthy and ill individuals in the population but if a statistically significant share of individuals have initial health endowments $H(0)$ above the initial health threshold $H_*(0)$ (scenarios A and B) then empirical tests should be able to distinguish between the interpretation of the Grossman model advocated here (represented by the joint occurrences of scenarios A, B and C) and the interpretation adopted in the literature (represented by scenario C only).

In the following we will contrast the predictions of our interpretation of the Grossman model with the more generally held interpretation and with empirical observations from the literature.

2.4.1 Similarities

The predictions for the demand for health and for medical care for unhealthy individuals (those individuals whose health is at the threshold) in our generalized solution of the Grossman model are, with the exception of some minor differences in formulation, the same as for the original solution of the Grossman model. Those predictions have largely been verified in the empirical literature, with the exception of the relation between the demand for health and the demand for medical care (see for details the earlier discussion and references therein). We summarize our predictions in Table 2.1.

Our generalized solution of the Grossman model broadly replicates the predictions of the traditional solution of the Grossman model. This can be seen as follows. Since the empirical literature has not distinguished between healthy and unhealthy individuals (a concept introduced in this work) a mixture of healthy and unhealthy individuals will have been included in the samples investigated. If at any time the proportion of unhealthy individuals (those whose health is at the health threshold and who behave according to

the traditional Grossman solution) is significant this could produce the observed relationships, with the exception of the relation between health and medical care. The reason that the relationship between health and medical care is different stems from the significantly different behavior between healthy and unhealthy individuals. The healthy do not consume medical care while the unhealthy do. If both healthy and unhealthy individuals are included in a sample this would produce the observed strong negative relationship between measures of health and measures of medical care. At the same time, if we can restrict the sample to the unhealthy, we should observe the positive relationship between health and medical care as predicted by Grossman.[18]

As Table 2.1 shows we expect health to decrease with the price of medical goods/services $p_m(t)$, unhealthy environmental factors ($\mathbf{X}(t)$, $d_\bullet$, β_3), and increase with education E[19] and with the efficiency of medical care $-\beta_6$. The relation with age t is ambiguous as wages $w(t)$ increase with working experience (e.g., Mincer, 1974) potentially countering the "aging" variables β_3, β_6, $\beta - \delta$. The effect of wages $w(t)$ is unclear, with a positive effect on health in the pure investment (PI) and a negative effect on health in the pure consumption (PC) model. Do note however that the predictions for the PC model have less predictive power than for the PI model. The structural form equation (2.37) includes consumption good $X(t)$, an endogenous variable, which in turn is a function of exogenous variables, such as wages $w(t)$, the price of medical goods/services $p_m(t)$, education E, etc. The inclusion of consumption good $X(t)$ in the structural form equation may distort the relationships between health and the exogenous variables. While the structural form equation (2.40) does not suffer from this problem, the predictions shown in the table depend on assumptions about model parameters (see table note *b* in Table 2.1). In addition, the shadow price of wealth $q_A(0)$ is a complicated function of various exogenous variables over the life cycle. Equation (2.40) thus suffers from a similar lack of transparency.

With regard to the demand for medical care, Table 2.1 shows that we expect the demand for medical goods/services to decrease with the price of medical goods/services $p_m(t)$, education E and the efficiency of medical care $-\beta_6$, and to increase with health

[18]Note that in retirement there is no production benefit from health as income (a pension / savings) is independent of the health status of the individual. Whether individuals demand less health as a result is unclear. The increased availability of leisure could reduce or increase the demand for health depending on whether leisure is a substitute or compliment of health (see for a discussion Galama et al. 2008 [Chapter 3]). Given potential differences in the demand for health between workers and retirees it may be necessary to distinguish between workers and retirees to potentially establish the positive relationship between health and medical care.

[19]Note that education could possible enter through lowering the rate of health deterioration $d(t)$ in addition, or as an alternative, to increasing the efficiency of medical care; see, e.g., Muurinen (1982)

Table 2.1: Relationships between the health threshold, the demand for medical care and various model variables, for the pure investment and pure consumption models.

	Health			Medical care
	PI[a]	PC	PC (Full)[b]	PI/PC
	Eq. 2.27	Eq. 2.37	Eq. 2.40	Eqs. 2.28, 2.38
Health $H(t)$	n/a	n/a	n/a	+
Wages $w(t)$	+	-	-	+
Price of medical goods/services $p_m(t)$	-	-	-	-
Education E	+	+	+	-
Age t	?	?	?	?
(Un)healthy environment $\mathbf{X}(t)$, $d_\bullet$, β_3	(-)+	(-)+	(-)+	(+)-
Consumption good $X(t)$	n/a	+	n/a	n/a
Price of consumption good $p_X(t)$	n/a	+	-	n/a
Life-time wealth $q_A(0)^{-1/\rho}$	n/a	n/a	+	n/a
Efficiency of medical care $-\beta_6$	+	+	+	-

Notes: Health threshold denoted by "Health"; demand for medical care denoted by "Medical care"; pure investment model denoted by "PI" and pure consumption model by "PC". Equation numbers (Eq.) refer to the structural form equations in section 2.3.

[a] Relations are valid for $\epsilon = 1/(1+\beta_2) > 0$.

[b] For plausible parameter choices. Precise relationships and conditions under which relations are valid are provided in section

$H(t)$, wages $w(t)$ and unhealthy environmental factors ($d_\bullet$, β_3, $\mathbf{X}(t)$). The predictions for the PI and PC models are the same. As discussed earlier the positive relationship between health and medical care is expected to be observable only if the sample can be restricted to unhealthy individuals.

2.4.2 Differences

In addition to the above predictions of our generalized solution of the Grossman model that are the same as in the traditional solution of the model, there are a number of distinctly different predictions. Those are discussed in detail below. We denote the predictions of the more generally held interpretation of the Grossman model by *"Optimal" stock*, our interpretation of the Grossman model by *Health threshold*, and the empirical observations from the literature by *Empirical literature*.

1. **Medical care and health are negatively correlated if measured across healthy and unhealthy individuals**

   "Optimal" stock: Health and medical care are positively correlated (see equations 2.28 and 2.38), i.e. individuals who consume more medical care are healthier.

   Health threshold: Healthy individuals ($H(t) > H_*(t)$) do not consume medical care, while unhealthy individuals ($H(t) = H_*(t)$) do. I.e. healthy individuals do not go to the doctor much, do not take much medicine, are not found to stay often in hospitals. Measured across a sample of healthy and unhealthy individuals we expect unhealthy individuals to consume more medical care than healthy individuals.

   Empirical literature: As discussed earlier the most striking feature of structural form estimates of the demand for *medical care* (see, e.g., Grossman, 1972a; Wagstaff, 1986a, 1993; Leu and Doppman, 1986; Leu and Gerfin, 1992; van Doorslaer, 1987; Van de Ven and van der Gaag, 1982; Erbsland, Ried and Ulrich, 2002) is the persistent and highly statistically significant negative relation found between measures of health and measures of medical care. The studies employ a variety of methodologies and a variety of datasets representing different cultural and institutional settings in a number of different countries (Europe and U.S.), yet their findings are largely in agreement with one another. None of these studies separate a healthy from an unhealthy population and hence we expect to observe a strong negative correlation between health and medical care if the population consists of both healthy and unhealthy individuals.[20]

2. **Healthy people do not consume medical care**

   "Optimal" stock: In the standard solution of the Grossman model individuals consume medical care at all ages.

   Health threshold: In our generalized solution healthy individuals (individuals whose health $H(t)$ is above the threshold $H_*(t)$) do not consume medical goods/services,

[20] Grossman (1972a) however selected a sub sample of the NORC dataset by restricting the data to those individuals that reported positive sick time and Erbsland, Ried and Ulrich (2002) restricted the sample to individuals reporting positive demand for health services. Interestingly Grossman (1972a) shows the least statistically significant negative relation between health and medical outlays of all the studies (t-stat of -5.84 [see Table 7 OLS estimates]). Erbsland, Ried and Ulrich (2002) report t-values of around -10 for three measures of medical care usage. Other studies, on the other hand, report values of at least -10 and up to -90. Perhaps the restriction of the samples to individuals that report positive sick time or positive medical care partially limited the sample to unhealthy respondents.

i.e. we would expect some fraction of the population at any given time to not consume medical goods/services.

Empirical literature: We would expect that healthy people pay few visits to the doctor (perhaps only to prevent illness, such as for a "health check up") and that they do not require much medical care (hospital stays, use medicine, etc). For example, Wagstaff (1986a) observes that 48% of the 1976 Danish Welfare Survey (DWA) sample he employed recorded zero general practitioner visits and 46.5% recorded zero weeks in hospital.

3. **Effective health deterioration slows when individuals reach the health threshold**

   "Optimal" stock: In the standard solution of the Grossman model health evolves as Grossman's "optimal" health stock, i.e. we do not expect to see discontinuous changes in the evolution of health.

   Health threshold: Healthy people ($H(t) > H_*(t)$) do not consume medical goods/services and their health deteriorates at the "natural" deterioration rate $\dot{H}(t) = -d(t)H(t)$. When, as a result of health deterioration their health reaches the health threshold $H(t) = H_*(t)$ (i.e., they have become unhealthy by our definition) they begin to consume medical goods / services and their health deteriorates at a lower effective rate $\dot{H}(t) = I(t) - d(t)H(t)$. If medical care improves one's health (e.g., medical care is effective), we expect to observe slower effective health deterioration $\dot{H}(t)$ or even health improvement when individuals reach the health threshold and begin to consume medical goods/services).[21]

   Empirical literature: Van Kippersluis et al. (2008) examine inequality in self-reported health (SRH) as a function of income in 11 European countries. The authors transform the ordinal SRH information onto a cardinal scale using utility scores for the SRH categories taken from the 2001 Canadian Community Household Survey (CCHS). The authors find a remarkable consistency in the pattern of health with age. In most countries health deteriorates gradually from early adulthood until around age 50 after which it generally levels off before accelerating rapidly after age 70. The authors find this middle-age plateau (ages 50-70) rather puzzling, but it would be consistent with a slowing of the decline in health resulting from increased medical care as the average individual reaches a health threshold. After age 70, as terminal illnesses set in, health again declines rapidly.

[21] Note the distinction between the effective health deterioration rate $\dot{H}(t)$ and the "natural" health deterioration rate $d(t)$.

Smith (2004, 2007) uses self-reported health (SRH) status from the National Health Interview Survey (NHIS) and PSID to show how disparity in health between low- and high-income individuals (the so-called socio-economic status [SES]-health gradient) increases with age till about age 60 after which the disparity narrows (see Van Doorslaer et al. 2008 for an excellent review of the literature on the SES-health gradient over the life cycle). The percentage of individuals reporting excellent or very good health status declines rapidly till age 60 for the first income quartile households (lowest income) and then remains fairly constant out till age 90. The 2nd to 4th income quartiles however show a more gradual decline.

Similarly, Case and Deaton (2005) present several plots of self-reported health (SRH) status from the NHIS as a function of age. Women and men in the bottom income quartile show a rapid deterioration in SRH between ages 20 and 60 after which the SRH curve flattens significantly (see their Figure 2). Again we see no evidence for a flattening of SRH with increasing age for the upper income quartile (in fact we see gradually deteriorating SRH status). This suggests that high SES individuals reach a health threshold much later (their SRH deteriorates slower) than low SES individuals. As a result they see no need to consume medical goods/services even at late ages and their effective health deterioration does not slow with age.

Van Kippersluis et al. (2009) find similar results for the Netherlands using a rich dataset based on the Health Interview Surveys and administrative data from Statistics Netherlands (CBS). The data allows the authors to study SRH as well as mortality, to disentangle the effect of ageing from that of cohort effects and to use actual (not reported) income from tax files. The authors find the pattern of the SES-health gradient over the life cycle in the Netherlands to be remarkably similar to that in the U.S., despite significant differences in the two countries' institutions.

Wagstaff (1993) fits an empirical reformulation of the Grossman model to two data subsets, those aged under 41 and those aged over 41. The author finds that for the over 41s the rate of effective health deterioration $\dot{H}(t)$ is lower than for the under 41s (the estimated relationship is $H_t \propto 0.849 H_{t-1}$ for the over 41s [Table 2b in Wagstaff, 1993] and $H_t \propto 0.687 H_{t-1}$ for the under 41s [Table 2a in Wagstaff, 1993]). Further, the fit is better for the over 41s ($R^2 = 0.595$) than for the under 41s ($R^2 = 0.394$). Since we expect that an older population will have relatively more individuals with health levels at or near the health "threshold" we would expect this population to provide a better fit to the "traditional" solution of the Grossman model.

So, perhaps older individuals, and in particular low income individuals, are slowing their effective health deterioration $\dot{H}(t)$ in late age by consuming medical goods / services as a threshold model would predict.[22]

4. **Effective health deterioration and medical care are negatively correlated**

   "Optimal" stock: According to the structural form equation (2.27) we find $\dot{H}(t) \propto -\epsilon(\beta_3 + \beta_6)H(t)$ (assuming variation in wages $w(t)$, prices $p_m(t)$ and environment $\mathbf{X}(t)$ is slow). Thus, high effective health deterioration requires that $\beta_3 + \beta_6$ is large and/or that health $H(t)$ is large ($\epsilon > 0$ is required to reproduce other empirical findings; see note *a* in Table 2.1). The model then predicts that medical goods/services $m(t) \propto H(t)e^{(\beta_3+\beta_6)t}$ are also high and increase exponentially with age (see equation 2.28). This would produce a positive correlation between effective health deterioration and medical goods/services.

   Health threshold: Measured across healthy and unhealthy individuals we expect to observe that healthy individuals will have rapid health deterioration ($\dot{H}(t) = -d(t)H(t)$) and low demand for medical care ($I(t) = 0$; they do not consume medical goods/services) while unhealthy individuals will be characterized by low effective health deterioration rates ($\dot{H}(t) = I(t) - d(t)H(t)$) and high demand for medical care ($I(t) > 0$). This would produce a negative correlation between effective health deterioration and the consumption of medical goods/services.

   Empirical literature: The discussion under item 3 suggests that individuals may slow their effective health deterioration as they age and begin to consume medical care. Further research is needed to empirically test this prediction.

5. **Medical care increases discontinuously when individuals become unhealthy**

   "Optimal" stock: In the standard solution of the Grossman model health evolves as the "optimal" health stock and individuals consume medical care continuously, i.e. there is no switching of dynamics and we do not expect to see discontinuous changes in medical care.

[22] At these high ages SRH may suffer from selection effects. Unhealthy individuals may have higher mortality and drop out of the sample in higher numbers than healthy individuals. Further, SRH status suffers from framing bias, that is, individuals compare their health with a reference of what constitutes good health in their respective age group. In other words, they may be answering the question "Considering my age I am in good/bad health" instead of "I am in good/bad health". Both effects would either reduce the significance of the observed flattening of SRH or could provide an alternative explanation for the observation.

Health threshold: Healthy people ($H(t) > H_*(t)$) do not consume medical care. When, as a result of health deterioration their health reaches the health threshold $H(t) = H_*(t)$ (i.e., they have become unhealthy by our definition) they begin to consume medical care.

Empirical literature: The literature has, as far as we know, not tested this prediction before. The empirical test is described in Section 2.3. Some moderate support for the notion that the dynamics of healthy and unhealthy individuals are significantly different comes from the following observation. Grossman noted in his original work (Grossman, 1972a; Chapter V, p. 56) that over two thirds of the NORC sample he used in empirical testing of his model, reported no sick days. He notes that *"... Since the characteristics of these two groups [reporting sick days and no sick days] are very similar, it is difficult to explain the behavior of the [group that had no sick days]. Put differently, the two groups essentially represent "two different samples," and problems arise when the data are pooled ..."*[23]

6. **Blue collar workers let their health deteriorate faster and to lower levels than white collar workers**

"Optimal" stock: Blue and white collar workers[24] consume medical care at all times. Blue collar workers (see equation 2.27) have lower levels of health, assuming lower wages $w(t)$, lower levels of education E, higher "natural" deterioration rates $d(t)$ (i.e. higher values of $d_\bullet$, β_3, and $\boldsymbol{\beta_4}$ and assuming $\epsilon > 0$, $0 < k_I < 1$ and $\rho_I > 0$; see equation 2.26). The "traditional" solution of the Grossman model is unclear about the effective health deterioration rate $\dot{H}(t)$ for blue versus white collar workers.[25]

[23]Strictly speaking we distinguish healthy from unhealthy individuals by whether they are above (healthy) or below (unhealthy) the health threshold and whether they do not (healthy) or do (unhealthy) consume medical care. But the number of sick days is assumed to be a function of health, and healthy individuals are expected to report relatively fewer sick days than unhealthy individuals.

[24]Blue collar workers are broadly defined as individuals who generally have 1) lower levels of education, 2) lower wages, and 3) perform "hard" labor (e.g., construction). White collar workers on the other hand generally 1) are more educated, 2) earn higher wages, and 3) perform "light" jobs (e.g., office workers). As a result of "hard" labor and worse working environments blue collar workers are believed to be characterized by higher "natural" health deterioration rates $d(t)$ than white collar workers (e.g., Case and Deaton, 2005; Muurinen and Le Grand, 1985).

[25]Assuming wages $w(t)$, medical prices $p_m(t)$ and environmental variables $\mathbf{X}(t)$ are relatively constant with age t we have $\dot{H}(t) \propto -\epsilon(\beta_3 + \beta_6)H(t)$. From this it is not immediately obvious that the effective health deterioration rate would be different for blue versus white collar workers, though β_3 (the exponential rate of decay of $d(t)$; see equation 2.26), may be higher for blue collar workers.

Health threshold: In scenario A, initially while blue and white collar workers are healthy (health above the "threshold"), a blue collar worker's health deteriorates faster than that of a white collar worker, assuming blue collar workers have higher health deterioration rates $d(t)$ as a result of physically demanding work and working environments that are more detrimental to health (see equation 2.31). The health of blue collar workers deteriorates to lower levels as their health threshold is lower (see discussion above under *"Optimal" stock* and equation 2.27). Once workers reach the health threshold it is unclear what the nature of differences (if any) is for the effective health deterioration rate $\dot{H}(t)$ for blue versus white collar workers (see discussion above under *"Optimal" stock*).

Empirical literature: Case and Deaton (2005) investigate the rate of change in self reported health by occupation using data from the NHIS. The authors find that those who are employed in manual occupations have worse health than those who work in professional occupations and that the health effect of occupation operates at least in part independently of the personal characteristics of the workers. Cutler et al. (2011) present similar results using mortality as an indicator of health. Van Kippersluis et al. (2009) present similar results using the self reported health status of Dutch working males.

Further, as discussed earlier under item 3, the health of women and men in the bottom income quartile deteriorates much faster than that of the top income quartile. It is much harder to assess from the self-reported health measures presented in Smith (2004, 2007), Case and Deaton (2005) and Van Kippersluis et al. (2009) whether blue collar workers let their health deteriorate to lower levels of health, though generally speaking blue collar workers are found to be in worse health than white collar workers (e.g., Case and Deaton, 2005; Smith, 1999, 2004, 2007; Van Kippersluis et al. 2008, 2009; as well as the evidence provided by the aforementioned studies that estimated the demand for health and found health to increase with, e.g., education, wages and to decrease with, e.g., physically demanding work). Similar patterns hold for other measures of socioeconomic status, such as education and wealth and other indicators of health, such as disability, and mortality (e.g., van Doorslaer et al. 2008).

7. **The relationship between education and health is expected to be positive and differs for healthy and unhealthy individuals**

"Optimal" stock: Health and education are positively related if the efficiency of medical care increases with education (equation 2.27 for $\rho_I > 0$). If health deterioration

$d(t)$ decreases with education E, i.e., education is part of the vector $\mathbf{X}(t)$ of environmental variables that affect the deterioration rate, then the education component of $\boldsymbol{\beta_4}$ ($\beta_{4,E}$) is negative and hence higher levels of education E through their affect on the deterioration rate $d(t)$ increase the level of health. The effect is similar to the presumed increased efficiency of medical care usage through education, $\rho_I > 0$, and both effects cannot be separated in the "traditional" solution of the Grossman model (the term in the structural form equation is $\ln H(t) \propto \epsilon(\rho_I - \beta_{4,E})E$). There is no difference between healthy and unhealthy individuals as in the "traditional" solution of the Grossman model this distinction is not made.

Health threshold: In scenario A, initially while individuals are healthy any relationship between health and education (see equation 2.31) works only through the effect (if any) of education on the rate of deterioration $d(t)$ and we have $\ln H(t) \propto -e^{\beta_{4,E}}$ ($\beta_{4,E} < 0$). When individuals have reached the health "threshold" both pathways (through the presumed increased efficiency of medical care usage and through any affect on the rate of deterioration $d(t)$) are relevant and we have the same relationship as for the "optimal" stock: $\ln H(t) \propto \epsilon(\rho_I - \beta_{4,E})E$.

Empirical literature: A positive association between education and health has been established in the empirical literature (see, e.g., the evidence provided by the aforementioned studies that estimated the demand for health and found health to increase with education). To the best of our knowledge the literature has not yet made an attempt to test the interpretation of the Grossman model advocated here, i.e., to distinguish between healthy and unhealthy individuals and test differences in their respective relationships between health and education. The empirical test is described in Section 2.3.

2.5 Discussion

We have presented arguments for a generalized solution of the Grossman model (Grossman, 1972a, 1972b). Our generalized solution of the Grossman model can deal with an important criticism of the model: that the model's prediction that health and medical care are positively related is consistently rejected by the data (e.g., Zweifel and Breyer, 1997, p. 62). We find that this prediction is based on the widely used and unnecessary assumption that the health stock is always at Grossman's solution for "optimal" health. There is no theoretical basis for this assumption and empirical evidence suggests it is not valid. Removing this widely used restriction and allowing for the existence of corner so-

lutions where individuals do not consume medical care, we find that the Grossman model predicts the existence of a health threshold.

We have contrasted the predictions of the generalized solution of the Grossman model advocated here with the empirical literature. Our generalized solution replicates the predictions of the traditional Grossman model (which have largely been verified in the empirical literature) with the exception of the problematic prediction that health and medical care should be positively correlated (which has been rejected in the empirical literature). As with the traditional solution of the Grossman model (a special case of our generalized solution) we broadly expect health to decrease with the cost of medical goods/services and with environmental factors that are detrimental to health (e.g., working conditions) and to increase with education. The effect of income is unclear as different sub models predict a different relation with health. With regard to the demand for medical care, we expect medical care to decrease with the cost of medical goods/services $p_m(t)$ and with education, and to increase with wages and with environmental factors that are detrimental to health.

In addition, our generalized solution of the Grossman model produces a number of predictions that are different from the traditional solution of the Grossman model. First, it replicates the observed negative relation between health and medical care as in our generalized solution of the Grossman model healthy individuals (whose health is above the health threshold) do not consume medical care while the unhealthy (at the threshold) do. Second, we find that individuals do not consume medical care at all times as healthy people do not consume medical care. Basically our generalized solution of the Grossman model predicts the intuitively natural behavior that healthy individuals do not go to the doctor or stay in hospital while the unhealthy do (except for preventive care or as a result of a sudden health shock, both phenomena are currently not part of the Grossman model). Third, we find that effective health deterioration slows as individuals reach the health threshold and begin to consume medical care. Fourth, our generalized solution of the Grossman model predicts that the effective health deterioration rate $\dot{H}(t)$ (the net effect of "aging" and medical care) will be smaller for individuals who consume more medical care. Fifth, we predict that the consumption of medical care increases discontinuously as healthy individuals begin to consume medical care once their health reaches the health threshold. Sixth, our generalized solution of the Grossman model can account for the observation that blue collar workers tend to have faster rates of effective health deterioration $\dot{H}(t)$ than white collar workers (e.g., Case and Deaton, 2005). Lastly, because the model distinguishes between healthy and unhealthy individuals who behave differently, the model allows for a number of tests that are not possible in the traditional

interpretation of the Grossman model. For example, Muurinen (1982) has argued that education improves health through lowering the natural health deterioration rate $d(t)$ (aging) and not just (or perhaps not at all) through improving the efficiency of an individual's consumption of medical care (Grossman, 1972a, 1972b). Since the first pathway (lowering the deterioration rate) operates only for healthy individuals and both pathways operate for unhealthy individuals it should in theory be possible to establish empirically the relative importance of both pathways. Also, while the natural deterioration rate $d(t)$ is not directly observable in the traditional interpretation of the Grossman model, it is directly observable in our interpretation as individuals who are healthy let their health deteriorate at exactly this rate (assuming good empirical measures of health status are available).

A review of the empirical literature suggests that our generalized solution of the Grossman model can account for a greater number of observations than can the traditional solution. Ultimately though, the model needs to be verified in direct empirical testing. To this end we have provided detailed structural and reduced form equations for the pure consumption and pure investment models for both the healthy and unhealthy phases of life. Empirically, the proposed model is a switching regression model with endogenous switching. Once health hits the health threshold, the process governing health and medical care switches.

The corner solutions presented in this work contribute to better describing the behavior of individuals whose health is above the threshold level for parts of the life cycle (the *healthy* and the *very healthy*). However, for those individuals whose health is at the threshold over the life cycle (the *ill*) we have simply adopted the assumption commonly made in the Grossman literature that individuals are able to adjust their health to a desirable level. This assumption may be less severe though in the case of the *ill*. It is, for example, not necessary to assume that the adjustment is instantaneous as individuals will have had ample time to consume medical care before they enter the labor force. There is also naturally an adjustment cost associated with these investments in the sense that such individuals begin their work life with fewer assets as a result of the purchase of medical care in the market before they entered the labor force.

Natural extensions of the model would be to include uncertainty and health shocks (e.g., to address the criticism by Cropper, 1977; Dardanoni and Wagstaff, 1987), to revisit the assumption of complete health repair (e.g., the criticism by Case and Deaton, 2005), to revisit the unrealistic so-called "bang-bang" solutions that the model produces when an individual's health is initially below the threshold (the *ill*; the criticism by Ehrlich and Chuma, 1990), to include length of life as a decision variable (endogenous T; e.g., Ehrlich

and Chuma, 1990), to include healthy and unhealthy behaviors such as unhealthy consumption (e.g., smoking), healthy consumption (e.g., dieting; see Case and Deaton, 2005) and preventive care, and to explore the solutions in which the decision to perform "hard" labor is endogenous (see, e.g., Case and Deaton, 2005). Following Cropper (1981) and Wagstaff (1986a) we have assumed that the natural deterioration rate $d(t)$ is exogenously determined by environmental factors such as, e.g., working conditions, hazardous environment, etc. The model thus assumes that blue collar workers have no choice but to perform hard labor and face worse living, working and schooling environments. But, as Case and Deaton (2005) argue, individuals may accept risky and unhealthy work environments, in exchange for higher pay.

2.6 Appendix

2.6.1 First-order conditions

Associated with the Lagrangian (equation 2.10) we have the following conditions:

$$\begin{aligned} \dot{q}_A(t) &= -\partial\Im(t)/\partial A(t) \Rightarrow \\ \dot{q}_A(t) &= -\delta q_A(t) \Leftrightarrow \\ q_A(t) &= q_A(0)e^{-\delta t}, \end{aligned} \tag{2.44}$$

$$\begin{aligned} \dot{q}_H(t) &= -\partial\Im(t)/\partial H(t) \Rightarrow \\ \dot{q}_H(t) &= q_H(t)d(t) - \frac{\partial U(t)}{\partial s(t)}\frac{\partial s(t)}{\partial H(t)}e^{-\beta t} \\ &- q_A(0)e^{-\delta t}\frac{\partial Y[H(t)]}{\partial s(t)}\frac{\partial s(t)}{\partial H(t)}, \end{aligned} \tag{2.45}$$

$$\begin{aligned} \partial\Im(t)/\partial X(t) &= 0 \Rightarrow \\ \partial U(t)/\partial C(t) &= q_A(0)\left(\frac{p_X(t)}{\partial C(t)/\partial X(t)}\right)e^{(\beta-\delta)t} \\ &\equiv q_A(0)\pi_C e^{(\beta-\delta)t}, \end{aligned} \tag{2.46}$$

$$\begin{aligned} \partial\Im(t)/\partial \tau_C(t) &= 0 \Rightarrow \\ \partial U(t)/\partial C(t) &= q_A(0)\left(\frac{w(t)}{\partial C(t)/\partial \tau_C(t)}\right)e^{(\beta-\delta)t} \\ &\equiv q_A(0)\pi_C e^{(\beta-\delta)t}, \end{aligned} \tag{2.47}$$

$$\begin{aligned} \partial\Im(t)/\partial m(t) &= 0 \Rightarrow \\ q_H(t) + q_I(t) &= q_A(0)\left(\frac{p_m(t)}{\partial I(t)/\partial m(t)}\right)e^{-\delta t} \\ &\equiv q_A(0)\pi_I e^{-\delta t}, \end{aligned} \tag{2.48}$$

$$\begin{aligned} \partial\Im(t)/\partial \tau_I(t) &= 0 \Rightarrow \\ q_H(t) + q_I(t) &= q_A(0)\left(\frac{w(t)}{\partial I(t)/\partial \tau_I(t)}\right)e^{-\delta t} \\ &\equiv q_A(0)\pi_I e^{-\delta t}. \end{aligned} \tag{2.49}$$

Equation (2.46) provides the first-order condition for maximization of (2.1) with respect to consumption, subject to the conditions (2.2) and (2.3). Using (2.48) to obtain

an expression for $\dot{q}_H(t)$ and substituting the results for $q_H(t)$ and $\dot{q}_H(t)$ in (2.45) we find the first-order condition for maximization of (2.1) with respect to health, subject to the conditions (2.2) and (2.3). The resulting first-order conditions are provided by equations (2.11) and (2.13) in section 2.2.

2.6.2 Structural and reduced form: pure investment model

We begin with the first-order condition for optimal health (2.18). We have (using equations 2.22 through 2.25)

$$\begin{aligned}
\pi_I(t) &= \frac{\partial Y(t)}{\partial s(t)}\frac{\partial s(t)}{\partial H(t)}\left[d(t)+\delta-\tilde{\pi}_I(t)\right]^{-1} && (2.50)\\
&= \beta_1\beta_2 w(t)H(t)^{-(\beta_2+1)}[d(t)+\delta-\tilde{\pi}_I(t)]^{-1} && (2.51)\\
&= \frac{p_m(t)}{\partial I(t)/\partial m(t)} = \frac{e^{-\rho_I E}}{\mu_I(t)k_I^{k_I}(1-k_I)^{(1-k_I)}} w(t)^{k_I} p_m(t)^{(1-k_I)}. && (2.52)
\end{aligned}$$

This leads to the structural form equation (2.27).

Now consider the equations for medical health investment (equations 2.2 and 2.24) and using (2.25),

$$\begin{aligned}
\ln I(t) &= \rho_I E + (1-k_I)\ln m(t) + k_I \ln\tau_I(t) + \ln\mu_I(t) && (2.53)\\
&= \rho_I E + \ln m(t) + k_I \ln p_m(t) - k_I \ln w(t)\\
&+ \ln\mu_I(t) + k_I\ln[k_I/(1-k_I)] && (2.54)\\
&= \ln[\dot{H}(t)+d(t)H(t)] && (2.55)\\
&= \ln d(t) + \ln H(t) + \ln[1+\tilde{H}(t)/d(t)]. && (2.56)
\end{aligned}$$

This leads to the structural form equation (2.28).

Using (2.27) and (2.28) we find

$$\begin{aligned}
\ln m(t) &= \beta_8 + [k_I+\epsilon(1-k_I)]\ln w(t) - [k_I+\epsilon(1-k_I)]\ln p_m(t)\\
&- (1-\epsilon)\rho_I E + (1-\epsilon)\ln d_\bullet + (1-\epsilon)(\beta_3+\beta_6)t + (1-\epsilon)\boldsymbol{\beta_4}\mathbf{X}(t)\\
&- \epsilon\ln\{1+d_\bullet^{-1}e^{-\beta_3 t-\boldsymbol{\beta_4}\mathbf{X}(t)}[\delta-k_I\tilde{w}(t)-(1-k_I)\tilde{p_m}(t)-\beta_6]\}\\
&+ \ln[1+\tilde{H}(t)/d(t)], && (2.57)
\end{aligned}$$

where $\beta_8 \equiv \beta_5+\beta_7$.

Combining equations (2.54) and (2.55) we find:

$$\dot{H}(t)+d(t)H(t) = \mu_I(t)[k_I/(1-k_I)]^{k_I} m(t) p_m(t)^{k_I} w(t)^{-k_I} e^{\rho_I E}, \qquad (2.58)$$

the solution of which is

$$H(t) = e^{\rho_I E}[k_I/(1-k_I)]^{k_I} \int_0^t \mu_I(x) m(x) p_m(x)^{k_I} w(x)^{-k_I} e^{-\int_x^t d(s)ds} dx. \quad (2.59)$$

We then have

$$1 + \tilde{H}(t)/d(t) = \frac{\mu_I(t) m(t) p_m(t)^{k_I} w(t)^{-k_I}}{d(t) \int_0^t \mu_I(x) m(x) p_m(x)^{k_I} w(x)^{-k_I} e^{-\int_x^t d(s)ds} dx}. \quad (2.60)$$

Substituting equation (2.60) into equation (2.57) and differentiating the result with respect to time t we find the reduced form expression (2.29).

While the literature largely focuses on the relations for health $H(t)$ and medical goods/services $m(t)$ the model does allow for the derivation of relations for consumption goods $X(t)$ and assets $A(t)$. In the pure investment model we have $\partial U(t)/\partial H(t) = 0$, i.e. utility $U(t)$ is independent of health $H(t)$. We assume a simple functional form for the utility function:

$$U[C(t)] = \frac{C(t)^{1-\rho}}{1-\rho}. \quad (2.61)$$

The first-order condition (equation 2.11) then leads to:

$$C(t)^{-\rho} = q_A(0) \pi_C(t) e^{(\beta-\delta)t}. \quad (2.62)$$

Grossman (1972a, 1972b, 2000) assumes that medical health investment is produced by combining time and medical goods/service according to a Cobb-Douglass constant returns to scale production function (see equation 2.24). A similar assumption can be made that consumption is produced by combining time τ_C and consumption goods $X(t)$ as follows:

$$C(t) = \mu_C(t) X(t)^{1-k_C} \tau_C(t)^{k_C} e^{\rho_C E}, \quad (2.63)$$

where $\mu_C(t)$ is an efficiency factor, $1-k_C$ is the elasticity of consumption $C(t)$ with respect to consumption goods $X(t)$, k_C is the elasticity of consumption $C(t)$ with respect to time input $\tau_C(t)$, and ρ_C determines the extent to which education E improves the efficiency of consumption $C(t)$.

Further the ratio of the marginal product of medical care with respect to medical goods/services $\partial I(t)/\partial m(t)$ and the marginal product of medical care with respect to own-time investment $\partial I(t)/\partial \tau_I(t)$ equals the ratio of the price of medical goods/services $p_m(t)$ to the wage rate $w(t)$ (representing the opportunity cost of time; see equation 2.25). Similarly, the ratio of the marginal product of consumption with respect to consumption goods $\partial C(t)/\partial X(t)$ and the marginal product of consumption with respect to time inputs

$\partial C(t)/\partial \tau_C(t)$ equals the ratio of the price of consumption good $p_X(t)$ to the wage rate $w(t)$ (see equation 2.12). We then have

$$\pi_I(t) = \frac{p_m(t)}{\partial I(t)/\partial m(t)} = \frac{p_m(t)^{1-k_I} w(t)^{k_I} e^{-\rho_I E}}{\mu_I(t) k_I^{k_I} (1-k_I)^{(1-k_I)}}, \tag{2.64}$$

$$\pi_C(t) = \frac{p_X(t)}{\partial C(t)/\partial X(t)} = \frac{p_X(t)^{1-k_C} w(t)^{k_C} e^{-\rho_C E}}{\mu_C(t) k_C^{k_C} (1-k_C)^{(1-k_C)}}. \tag{2.65}$$

Assuming the Cobb-Douglass constant returns to scale production function for medical health investment (equation 2.24) and for consumption (equation 2.63) we obtain the following expressions for consumption goods $X(t)$ and medical goods/services $m(t)$

$$X(t) = (1-k_C)\frac{\pi_C(t)}{p_X(t)} C(t), \tag{2.66}$$

$$m(t) = (1-k_I)\frac{\pi_I(t)}{p_m(t)} [\dot{H}(t) - d(t)H(t)]. \tag{2.67}$$

Using equations (2.62, 2.65, and 2.66) we find

$$\begin{aligned} \ln X(t) &= \beta_{13} - [k_C + (1-k_C)/\rho] \ln p_X(t) + k_C[(\rho-1)/\rho] \ln w(t) \\ &- \rho_C[(\rho-1)/\rho]E - [(\beta-\delta)/\rho]t + \ln q_A(0)^{-1/\rho}, \end{aligned} \tag{2.68}$$

where $\beta_{13} \equiv \ln(1-k_C) - [(\rho-1)/\rho] \ln[k_C^{k_C}(1-k_C)^{(1-k_C)}] - [(\rho-1)/\rho] \ln \mu_C(t)$.

It is straightforward though tedious to derive an expression for the shadow price of wealth $q_A(0)$, using the life-time budget constraint (2.5), the expression for sick time $s[H(t)]$ (equation 2.22), income $Y[H(t)]$ (equation 2.9), consumption good $X(t)$ (the above equation), health $H(t)$ (equation 2.27), and medical goods/services $m(t)$ (equation 2.28). $q_A(0)$ is then found to be a complicated function of life-time wealth (assets, life-time income), wages $w(t)$, prices $p_m(t)$, $p_X(t)$, education E and the health deterioration rate (terms $d_\bullet$, β_3 and $\boldsymbol{\beta_4}\mathbf{X}(t)$). The expression itself is not very insightful and is hence not reproduced here.

2.6.3 Structural and reduced form: pure consumption model

Using the utility specification (2.36), the first-order conditions (2.11) and (2.13), and equation (2.35) we find

$$\begin{aligned}\frac{\partial U[C(t),H(t)]}{\partial C(t)} &= \zeta C(t)^{\zeta-\rho\zeta-1}H(t)^{1-\zeta-\rho+\rho\zeta} \\ &= q_A(0)\pi_C(t)e^{(\beta-\delta)t} \qquad (2.69)\\ \frac{\partial U[C(t),H(t)]}{\partial H(t)} &= (1-\zeta)C(t)^{\zeta-\rho\zeta}H(t)^{-\zeta-\rho+\rho\zeta} \\ &= q_A(0)\left[\pi_H(t)-\varphi_H(t)\right]e^{(\beta-\delta)t}+\left[\dot{q}_I(t)-q_I(t)d(t)\right]e^{\beta t} \\ &= \pi_C(t)^{-1}\left[\pi_H(t)-\varphi_H(t)\right]\frac{\partial U[C(t),H(t)]}{\partial C(t)} \\ &+ \left[\dot{q}_I(t)-q_I(t)d(t)\right]e^{\beta t} \qquad (2.70)\end{aligned}$$

The solution for the health threshold (Grossman's solution for "optimal" health) follows from combining equation (2.69) with (2.70), assuming $\varphi_H(t)=0$ (pure consumption) and using $q_I(t)=\dot{q}_I(t)=0$. We then find:

$$\begin{aligned}\ln H(t) &= \ln C(t)+\ln\left(\frac{1-\zeta}{\zeta}\right)+\ln\pi_C-\ln\pi_I-\ln d(t) \\ &- \ln[1+\delta/d(t)-\tilde{\pi}_I(t)/d(t)]. \qquad (2.71)\end{aligned}$$

Combining equations (2.26, 2.64, 2.65, 2.66 and the above expression) leads to the structural form equation (2.37). Further, combining equations (2.66, 2.69, 2.70 and 2.71) we find:

$$\begin{aligned}H(t) &= q_A(0)^{-1/\rho}\zeta^{1/\rho}\left(\frac{1-\zeta}{\zeta}\right)^{\chi}\pi_C(t)^{-\chi(1/\rho\chi-1)}\pi_I(t)^{-\chi}d(t)^{-\chi} \\ &\times \left[1+\delta/d(t)-\tilde{\pi}_I(t)/d(t)\right]^{-\chi} \qquad (2.72)\end{aligned}$$

which leads to the structural form equation (2.40).

As in the pure investment model one can find and expression for the shadow price of wealth $q_A(0)$ for the pure consumption model, using the life-time budget constraint (2.5), the expression for income $Y[H(t)]$ (equation 2.9), consumption good $X(t)$ (equation 2.66), health $H(t)$ (equation 2.40), and medical goods/services $m(t)$ (equation 2.38). As in the pure investment model the expression is found to be a complicated function of life-time wealth (assets, life-time income), wages $w(t)$, prices $p_m(t)$, $p_X(t)$, education E and the health deterioration rate (terms $d_\bullet$, β_3 and $\boldsymbol{\beta_4}\mathbf{X}(t)$).

Combining equation (2.37) with (2.38) we find:

$$\begin{aligned}\ln m(t) &= \beta_{12}+\ln X(t)+\ln p_X(t)-\ln p_m(t)+\ln[1+\tilde{H}(t)/d(t)] \\ &- \ln\{1+d_\bullet^{-1}e^{-\beta_3 t-\boldsymbol{\beta_4}\mathbf{X}(t)}[\delta-k_I\tilde{w}(t)-(1-k_I)\tilde{p_m}(t)-\beta_6]\}, \qquad (2.73)\end{aligned}$$

where $\beta_{12} \equiv \beta_7 + \beta_9$.

Substituting equation (2.60) into equation (2.73) and differentiating the result with respect to time t we find the reduced form expression (2.39).

Chapter 3

A Health Production Model with Endogenous Retirement

We formulate a stylized structural model of health, wealth accumulation and retirement decisions building on the human capital framework of health and derive analytic solutions for the time paths of consumption, health, health investment, savings and retirement. We argue that the literature has been unnecessarily restrictive in assuming that health is always at the "optimal" health level. Exploring the properties of corner solutions we find that advances in population health decrease the retirement age, while at the same time individuals retire when their health has deteriorated. This potentially explains why retirees point to deteriorating health as an important reason for early retirement, while retirement ages have continued to fall in the developed world, despite continued improvements in population health and mortality. In our model, workers with higher human capital invest more in health and because they stay healthier retire later than those with lower human capital whose health deteriorates faster.

This chapter is based upon:

Galama, T.J., Kapteyn, A., Fonseca, R. and Michaud, P.-C. (2009), "Grossman's Health Threshold and Retirement", *RAND Working Paper*, WR-658.

3.1 Introduction

Models of retirement need to be able to reconcile the counterintuitive observations that a) retirees mention deteriorating health as an important reason for early retirement, b) population health and mortality have continued to improve, but c) the age of retirement has declined for nearly a full century in the developed world (though the decline in retirement age has leveled off and reversed somewhat in the last decade; see, e.g., Blau and Goodstein, 2010). Some of this could be explained by justification bias. Individuals may mention health as a reason to justify the fact that they are retired but in fact retire for other reasons, with health actually playing a minor role in the decision. For example, French (2005) estimates a life cycle model of labor supply, retirement, and savings behavior using the panel study of income dynamics (PSID). He finds that the structure of the Social Security system and of pensions are key determinants of the high observed job exit rates at ages 62 and 65 while Social Security benefit levels, health, and borrowing constraints are less important determinants of job exit at older ages. In line with this result Lazear (1986) finds that pensions are typically actuarially unfair and that sharp decreases in the actuarial value of retirement with continued work are used as a device by employers to induce earlier retirement of workers. Also Bazzoli (1985) finds that economic variables play a more important role than health in retirement decisions. On the other hand, Dwyer and Mitchell (1999) find the opposite: that health problems influence retirement plans more strongly than do economic variables. Specifically, Dwyer and Mitchell find that men in poor overall health retire between one and two years earlier than others. In other words, while there is agreement that health influences retirement there is disagreement about the importance of health in the retirement decision. Regardless of its current importance, the increased uptake of defined-contribution type pension vehicles such as 401(k)'s, which are actuarially fair, may reduce the importance of pension structure as a key determinant of retirement. This may warrant the inclusion of health as a more prominent determinant of future retirement.

While health may influence the decision to retire, it is unclear whether retirement in turn has an impact on health. Retirement may be a taxing event, resulting in the loss of friends and support networks, or retirement may be health preserving as it is work that is taxing, not retirement. Empirical evidence on the health effects of retirement is ambiguous (see for instance the literature review in Dave et al. 2006). Using the Health and Retirement Study (HRS), Dave et al. (2006) find that retirement has a detrimental effect on health. On the other hand, Coe and Zamarro (2010) in a cross country comparison find evidence that retirement may actually be health preserving.

Lacking the possibility of a controlled experiment, establishing the direction of causality is wrought with difficulties. The decision to retire may be motivated by a desire to preserve health and/or by bad health hampering one's ability to be a productive member of the workforce.

With aging populations and trends towards earlier retirement despite significant improvements in the health of populations in the developed world, societies are increasingly burdened by the rising costs of a growing elderly economically inactive population that is supported by a relatively shrinking economically active group. Understanding what policy instruments can be used to reduce this burden is therefore essential and requires the inclusion of health in retirement models. Potential levers are: universal healthcare provision, subsidized healthcare for low income workers (weighing societal versus individual benefits from delayed retirement), promotion of healthy lifestyles etc.

The aim of this paper is to investigate the influence of various conditions, in particular that of an individual's health, on the decision to retire. To this end we formulate a theory of health and retirement.[1] In section 3.2, we formulate a stylized structural model of consumption, leisure, health, health investment, wealth accumulation and retirement decisions using the human capital framework of health provided by Grossman (1972). Health provides utility and healthy individuals have greater earnings causing individuals to invest in their health. Individuals can accumulate savings and/or borrow without restriction, and they are free to decide when to retire. We find that the inclusion of retirement in the formulation complicates matters in that at the age of retirement the "optimal" level of the health stock is discontinuous. This implies that individuals invest an infinite amount (positive or negative) of health investment over an infinitesimally small period of time around the age of retirement. We address this feature of the literature spawned by Grossman by introducing corner solutions in which individuals do not invest in health for periods of time. In section 3.3 we first solve the optimal control problem conditional on retirement age. Specification of a functional form for the utility function allows us to derive analytical solutions for consumption, health, health investment and wealth, conditional on a given retirement age. In section 3.4 we discuss an extension of the model, and in section 3.5 we then maximize the implied indirect utility function with respect to the retirement age. In the model individuals find retirement increasingly attractive as they age as a result of three effects: (1) wage declines as a result of gradual health deterioration reducing income from work with age, (2) increased leisure time during retirement and (3) accumulation of pension wealth with years in the workforce. We

[1] For other models of endogenous health and retirement see Wolfe (1985), French and Jones (2007) and Fonseca et al. (2009).

provide simulations in section 3.6 and find that our model can explain that improvements in population health decrease the retirement age, while at the same time individuals retire when their health has deteriorated. We conclude in section 3.7 and provide detailed derivations in the Appendix.

3.2 General framework: a health production model

A natural framework for our analysis is provided by Grossman (1972a). For an excellent review of the basic concepts of this model see Muurinen and Le Grand (1985). Our formulation is most closely related to Case and Deaton (2005), Wagstaff (1986a), Wolfe (1985) and Ehrlich and Chuma (1990).

Let us assume that a consumer is endowed with an intra-temporal utility function $U[L(t), C(t), H(t)]$ at age t, where leisure $L(t)$, consumption $C(t)$, and health $H(t)$ are all positive quantities. The utility function has diminishing marginal returns and is an increasing function in its arguments $L(t)$, $C(t)$ and $H(t)$. Let leisure during one's working life be equal to L_0 and during retirement equal to τL_0, with $\tau > 1$. Assuming separability of the utility function we can then write utility before retirement as $U_w[C(t), H(t)]$, and after retirement as $U_r[C(t), H(t)]$. Consumers maximize the life time utility function

$$\int_0^R U_w[C(t), H(t)]e^{-\beta t}dt + \int_R^T U_r[C(t), H(t)]e^{-\beta t}dt, \tag{3.1}$$

where T denotes total life time, R is the age of retirement and β is a subjective discount factor. Time t is measured from the time individuals begin employment. The objective function (3.1) is maximized subject to the following constraints:

$$\begin{aligned}
\dot{H}(t) &= \mu(t)m(t) - d(t)H(t) && 0 \leq t \leq T \\
\dot{A}(t) &= \delta A(t) + Y[H(t)] - C(t) - p(t)m(t) && 0 \leq t \leq T \\
Y[H(t)] &= \begin{cases} w_0(t) + \varphi(t)H(t) & 0 \leq t \leq R \\ b & R < t \leq T \end{cases}
\end{aligned} \tag{3.2}$$

Furthermore we have initial and end conditions: $H(0)$, $A(0)$ and $A(T)$ are given.

$\dot{H}(t)$ and $\dot{A}(t)$ denote time derivatives of health $H(t)$ and assets $A(t)$. The first equation of (3.2) shows that an individual can invest in the stock of health $H(t)$ by investing $m(t)$ in medical care and/or other health promoting activities (e.g., exercise, diet, etc)

with an efficiency $\mu(t)$ to improve health and counter the "natural" health deterioration rate $d(t)$. While Ehrlich and Chuma (1990) argue that medical technology should realistically exhibit diminishing returns to scale we use the more commonly used assumption of a medical technology that has constant returns to scale (as in the first equation of 3.2). As Grossman (2000) argues, diminishing returns to scale would greatly complicate the model, while the benefits (certainly for the purpose of our simplified analytical model) may be limited. We further note that we impose diminishing returns of the utility of health, to ensure that infinite medical care is not demanded by consumers.

The second equation is simply the inter-temporal budget constraint, where δ is the interest rate, $Y[H(t)]$ is income, $C(t)$ is consumption and $p(t)$ is the price of health investment at time t. The product $p(t)m(t)$ is out-of-pocket medical expenditures. One way to interpret prices is by defining $m(t)$ as the "true" medical expenditures and $p(t)$ as the co-payment. In such a formulation "prices" vary dramatically depending on insurance status. For uninsured individuals in the U.S. the co-pay may effectively be 100%.

The third equation in (3.2) shows how income $Y[H(t)]$ consists of earnings during working life and pension income during retirement. Earnings are a function of health, with $w_0(t)$ a base wage rate that is age dependent (but independent of health) and the marginal production benefit of health $\partial Y[H(t)]/\partial H(t) = \varphi(t) \geq 0$ determines the extent to which health increases one's wage. Retirement income b is independent of health. Note that the system dynamics change at the age of retirement R (where income, consumption, health investment and prices can be discontinuous and the dynamic equations change).

The essential features of the human-capital model of health are: 1) that the demand for medical care is a "derived" demand in that consumers demand good health, not medical care per se, 2) that health provides consumption benefits (utility is a function of health) and 3) that health provides production benefits (health increases earnings; see equation 3.2).

Integrating the second equation of (3.2) over the life time we obtain the life-time budget constraint

$$\begin{aligned} &\int_0^T C(t)e^{-\delta t}dt + \int_0^T p(t)m(t)e^{-\delta t}dt = A(0) - A(T)e^{-\delta T} + \\ &\int_0^R w_0(t)e^{-\delta t}dt + \tfrac{b}{\delta}(e^{-\delta R} - e^{-\delta T}) + \int_0^R \varphi(t)H(t)e^{-\delta t}dt. \end{aligned} \tag{3.3}$$

The left-hand side of (3.3) represents life-time consumption and life-time health investment, and the right-hand side represents life-time financial resources in terms of (from left to right): use of life-time assets, life-time income from wages and from benefits, and lastly, additional life-time earnings, resulting from good health and health investment.

Using the first equation of (3.2) we can write $H(t)$ as a function of health investment and initial health.

$$H(t) = H(0)e^{-\int_0^t d(s)ds} + \int_0^t \mu(x)m(x)e^{-\int_x^t d(s)ds}dx. \tag{3.4}$$

As the relation suggests, individuals cannot "choose" health optimally. Instead they can invest in health $m(t)$ optimally.

We demand that health investment $m(t) \geq 0$, i.e., that individuals cannot "sell" their health through negative health investment $m(t)$. Health $H(t)$ at time t is path dependent; it is a function of the entire history $0 \leq t' < t$ of health investment $m(t')$ and of initial health $H(0)$. In the optimization problem we thus have to optimize with respect to the entire prior history of health investment $m(t')$.

Thus, we have the following optimal control problem: the objective function (3.1) is maximized with respect to the control functions $C(t)$ and $m(t)$ and subject to the constraints (3.2). The Lagrangean or generalized Hamiltonian (see, e.g., Seierstad and Sydsaeter 1987) of this problem is:

$$\begin{aligned} \Im &= U[C(t), H(t)]e^{-\beta t}dt + p_A(t)\{\delta A(t) + Y[H(t)] - C(t) - p(t)m(t)\} \\ &+ q(t)m(t), \end{aligned} \tag{3.5}$$

where $U[C(t), H(t)] = U_w[C(t), H(t)]$ for $t \leq R$; $U[C(t), H(t)] = U_r[C(t), H(t)]$ for $t > R$; $p_A(t)$ is the adjoint variable associated with the differential equation (3.2) for assets $A(t)$ and $q(t)$ a multiplier associated with the condition that health investment $m(t) \geq 0$. The inclusion of the multiplier $q(t)$ is an essential difference between our formulation and prior formulations of the Grossman model. It allows us to explicitly impose the constraint that medical care is positive $m(t) \geq 0$ at all times. We discuss the implications of this choice and the arguments for making it in detail in section 3.3.

We proceed as follows. First we solve the optimal control problem conditional on retirement age R (i.e., for a fixed exogenous retirement age R) and specify a functional form for the utility function. For given exogenous time varying deterioration rate $d(t)$, prices $p(t)$, efficiency $\mu(t)$, base wage rate $w_0(t)$, benefits b and production benefit $\varphi(t)$, we can then solve for the control variables $C(t)$ and $m(t)$ which in turn provides us with solutions for the state variables $H(t)$ and $A(t)$. We then maximize the resulting indirect utility function with respect to retirement age R. Health, savings and retirement thus are jointly determined in our model.

3.3 Exogenous retirement

The first order conditions for maximization of (3.1) subject to (3.2) are (for details see the Appendix):

$$\begin{aligned} \frac{\partial U_w(t)}{\partial C(t)} &= p_A(0)e^{(\beta-\delta)t} \quad (t \leq R) \\ \frac{\partial U_r(t)}{\partial C(t)} &= p_A(0)e^{(\beta-\delta)t} \quad (t > R), \end{aligned} \tag{3.6}$$

and

$$\begin{aligned} \frac{\partial U_w(t)}{\partial H(t)} &= p_A(0)\left[\pi_H(t) - \varphi(t)\right]e^{(\beta-\delta)t} \\ &+ \frac{e^{\beta t}}{\mu(t)}\dot{q}(t) - \frac{e^{\beta t}}{\mu(t)}\left[\frac{\dot{\mu}(t)}{\mu(t)} + d(t)\right]q(t) \qquad (t \leq R) \\ \frac{\partial U_r(t)}{\partial H(t)} &= p_A(0)\pi_H(t)e^{(\beta-\delta)t} \\ &+ \frac{e^{\beta t}}{\mu(t)}\dot{q}(t) - \frac{e^{\beta t}}{\mu(t)}\left[\frac{\dot{\mu}(t)}{\mu(t)} + d(t)\right]q(t) \qquad (t > R), \end{aligned} \tag{3.7}$$

where

$$\pi_H(t) \equiv \left[p(t)/\mu(t)\right]\left[d(t) + \delta - \dot{p}(t)/p(t) + \dot{\mu}(t)/\mu(t)\right] \tag{3.8}$$

is the the user cost of health capital at the margin (the interest rate δ represents an opportunity cost).

Equations (3.6) and (3.7) are similar to those by Case and Deaton (2005; their equations 5 and 6) for $q(t) = 0$, i.e., $m(t) > 0$. Equation (3.6) requires the marginal benefit of consumption to equal $p_A(0)$ (the shadow price of wealth) times a time varying exponent that either grows or decays with time, depending on the sign of $\beta - \delta$ (the difference between the time preference rate β and the interest rate δ). The marginal benefit of health investment (equation 3.7) equals the product of the marginal benefit of consumption (equation 3.6) and the user cost of health capital at the margin $\pi_H(t)$ (equation 3.8) minus the marginal production benefits of health $\varphi(t)$ if the individual is working.[2]

[2]We impose that the user cost of health capital at the margin exceeds the marginal production benefit of health $\pi_H(t) \equiv \left[p(t)/\mu(t)\right]\left[d(t) + \delta - \dot{p}(t)/(p(t) + \dot{\mu}(t)/\mu(t)\right] > \varphi(t)$. Without this condition, the investment in health would finance itself by increasing wages by more than the user cost of health. As a result of this, consumers would choose infinite health investment paid for by infinite wage increases to reach infinite health.

We can make a number of observations with respect to the first order conditions for consumption and health investment (equations 3.6 and 3.7). For now we discuss the case where $q(t) = 0$, i.e., $m(t) > 0$. As we will discuss in more detail later, this represents a special case in which the evolution of an individual's health follows the solution for the "optimal" health stock. First, increasing life time resources will lower $p_A(0)$ and hence increase health investment and consequently health. Second, while health continous to provide a consumption benefit (utility) health does not provide a production benefit (greater income) after retirement (last equation of 3.2) and retired individuals will reallocate away from health expenditures in the direction of more consumption. Third, a lower price of health investment increases health. This is pertinent in a cross-country comparison, but also when comparing across the life-cycle, for instance if health care is subsidized for certain age groups (like Medicare in the U.S.). Finally, more efficient health investment will lead to more health. Efficiency can explain variations within a country (if for instance individuals with a higher education level are more efficient in their health investment, Goldman and Smith, 2002) or across countries (if health care is more efficient in one country than in another).

In order to derive analytical solutions for consumption, health, health investment and wealth, we specify the following constant relative risk aversion (CRRA) form for the utility function (1):

$$U_w(C,H) = \frac{\left[C^{\zeta}H^{1-\zeta}\right]^{1-\rho}}{1-\rho};\ U_r(C,H) = kU_w(C,H), \tag{3.9}$$

where ζ $(0 \leq \zeta \leq 1)$ is the relative "share" of consumption $C(t)$ versus health $H(t)$ and ρ $(\rho > 0)$ the coefficient of relative risk aversion.

The factor k is the ratio of utility when retired and when working. A simple way to motivate the introduction of the multiplicative factor k is to include leisure in the utility function as follows: $U(C,H,L) = \left[C^{\zeta}H^{1-\zeta}L^{\tau}\right]^{1-\rho}$, where L is leisure and where we have omitted the multiplicative constant $1/(1-\rho)$. Assume that during the working years leisure is equal to L_0 while during retirement leisure is equal to $k_r L_0$ with $k_r > 1$. This implies that the ratio of utility before and after retirement is equal to $k \equiv k_r^{\tau(1-\rho)}$. This specification is consistent with the Stock and Wise (1990) specification in which the utility of consumption in retirement is a multiple of the utility of consumption when working. If $\rho < 1$ (i.e., utility is less concave than logarithmic) the ratio is greater than one. That is, at the same consumption level, utility is higher when retired. For $\rho > 1$ we have $k < 1$. In the latter case it is still the case that for a given consumption level, utility is higher in retirement, since utility is negative for $\rho > 1$.

This formulation can reproduce the drop in consumption observed at retirement (Banks, Blundell and Tanner, 1998; Bernheim, Skinner and Weinberg, 2001). For $k < 1$ and $\rho > 1$ (or for $k > 1$ and $\rho < 1$), and for a given consumption level the marginal utility of consumption is lower in retirement than while working and hence it is optimal to spend more money on consumption before retirement than after retirement.

Hurd and Rohwedder (2003, 2006) review some of the explanations put forward to explain the drop in consumption. The first of these is the occurrence of unanticipated shocks at the time of retirement, where, e.g., retirees are surprised to find that their economic resources are fewer than anticipated and adjust consumption accordingly. This would suggest that agents are insufficiently forward looking and would complicate employment of life-cycle models (used in this paper). However, Hurd and Rohwedder present evidence that the reductions are fully anticipated. In addition there are alternative explanations that are consistent with a life-cycle approach. For example, a second explanation involves uncertainty in the timing of retirement, where, e.g., workers retire because of a health event or unemployment resulting in a reduction in resources. Hurd and Rohwedder (2006) find that an unanticipated decline in lifetime resources caused by early retirement could explain a spending decline for part of the population. But the authors conclude that the empirical importance of health shocks is not great enough to explain fully the recollected declines in consumption. In line with this result, Blau (2008) suggests that a simple life cycle model in which individuals choose when to retire but are subject to shocks can account qualitatively for these stylized facts. However, Blau finds that the magnitude of the drop in consumption among households that experience a decline is too small in a calibrated model compared to the data. Blau concludes that other proposed explanations for the decline in consumption at retirement should continue to be explored in future research based on the life cycle framework. A third explanation is the increase in leisure at retirement, which would be consistent with $k < 1$. The increase in leisure can decrease consumption, e.g., because housekeeping, home repairs etc. are performed by the consumer after retirement and no longer purchased. Hurd and Rohwedder (2006) report that a transition into retirement is associated with approximately a 5.5 hrs increase per week in time spent on home production. Hurd and Rohwedder (2006) conclude that this supports the view that the increased ability to engage in home production or thriftier shopping during retirement is an important reason for the observed spending declines.

In our stylized formulation we employ a life-cycle model (e.g., we assume agents are to a large extent rational, rejecting the first explanation), specify a utility function that allows for a drop in consumption due to increased leisure at retirement (incorporating the

third explanation), but do not incorporate uncertainty (i.e., we do not model the effect of the second explanation).

3.3.1 Model solutions: the "optimal" health stock

We begin analyzing the case where $q(t) = 0$, i.e. $m(t) > 0$. This case is associated with the "optimal" health stock, as utilized in the literature spawned by Grossman. We will denote the solutions for consumption, health investment and health with $C_*(t)$, $m_*(t)$, and $H_*(t)$ for this special case to distinguish from the more general solutions $C(t)$, $m(t)$, and $H(t)$. Solving the first order conditions (3.6) and (3.7) and using the Cobb-Douglas utility specification (3.9), we find the following solutions for the control functions $C_*(t)$ and $m_*(t)$ (for details see the Appendix):

$$C_*(t) = \zeta\Lambda\left[\pi_H(t) - \varphi(t)\right]^{1-\chi} e^{-\left(\frac{\beta-\delta}{\rho}\right)t} \quad (t \leq R) \tag{3.10}$$

$$C_*(t) = k^{\frac{1}{\rho}}\zeta\Lambda\left[\pi_H(t)\right]^{1-\chi} e^{-\left(\frac{\beta-\delta}{\rho}\right)t} \quad (t > R) \tag{3.11}$$

$$\begin{aligned} m_*(t) &= \frac{1}{\mu(t)} e^{-\int_0^t d(s)ds} \\ &\times \frac{d}{dt}\left\{(1-\zeta)\Lambda\left[\pi_H(t) - \varphi(t)\right]^{-\chi} e^{-\left(\frac{\beta-\delta}{\rho}\right)t} e^{\int_0^t d(s)ds}\right\} \quad (t \leq R) \end{aligned} \tag{3.12}$$

$$\begin{aligned} m_*(t) &= \frac{1}{\mu(t)} e^{-\int_0^t d(s)ds} \\ &\times \frac{d}{dt}\left\{k^{\frac{1}{\rho}}(1-\zeta)\Lambda\left[\pi_H(t)\right]^{-\chi} e^{-\left(\frac{\beta-\delta}{\rho}\right)t} e^{\int_0^t d(s)ds}\right\} \quad (t > R), \end{aligned} \tag{3.13}$$

where we have used the following definitions:

$$\chi \equiv \frac{1 + \rho\zeta - \zeta}{\rho}, \tag{3.14}$$

and

$$\Lambda \equiv \zeta^{\frac{1-\rho}{\rho}} \left(\frac{\zeta}{1-\zeta}\right)^{1-\chi} p_A(0)^{\frac{-1}{\rho}}. \tag{3.15}$$

The constant Λ [and hence $p_A(0)$] can be determined by substituting the solutions for health $H(t)$, consumption $C(t)$ and health investment $m(t)$ into the life-time budget constraint (3.3). The result can be written as a fraction $\Lambda \equiv \Lambda_n/\Lambda_d$, where the numerator Λ_n is similar to the expression for life-time resources (right hand side of 3.3). Hence increasing initial assets $A(0)$, base wages $w_0(t)$, retirement benefits b, production benefits of health $\varphi(t)$ or initial health $H(0)$ increases the constant Λ and thereby consumption

$C(t)$, health investment $m(t)$, and health $H(t)$. The denominator Λ_d is a complicated function of the time paths of $d(t)$, $p(t)$, $\mu(t)$, $\varphi(t)$ and various model parameters:

$$\Lambda_d = \Lambda_d[d(t), p(t), \mu(t), \varphi(t), \delta, \beta, R, T, k, \rho, \zeta]. \tag{3.16}$$

The full solutions for Λ are provided in the Appendix for each of six scenarios (equations 3.73, 3.74, 3.75, 3.77, 3.78, 3.79, 3.91, 3.92, 3.93, 3.100, 3.101, and 3.102; for more details on the scenarios see section 3.3.2).

Consumption and health investment (equations 3.10 through 3.13) are functions of various combinations of the user cost of health capital at the margin $\pi_H(t)$ (see equation 3.8), minus the marginal production benefit of health $\varphi(t)$, to the power $1-\chi$ (consumption) or $-\chi$ (health investment).[3]

For constant time paths of $d(t) = d_0$, $p(t) = p_0$, $\mu(t) = \mu_0$, $\varphi(t) = \varphi_0$, consumption and health investment decrease (increase) exponentially with time if the time preference rate β is larger (smaller) than the interest rate δ. For $\beta = \delta$ we have constant time paths for consumption and for health investment except for jumps at retirement $t = R$ (due to the absence of a health production benefit $\varphi(t) = 0$ during retirement and due to the factor k associated with greater leisure time during retirement).

Consumption increases with the user cost of health capital at the margin $\pi_H(t)$ and decreases with the marginal production benefit of health $\varphi(t)$ for $0 < \chi < 1$ (i.e. if $\rho > 1$ and $\zeta < 1$). The opposite pattern is found for $\chi > 1$ (i.e. if $0 < \rho < 1$ and $\zeta < 1$). For $\chi = 1$ (i.e. $\rho = 1$ or $\zeta = 1$) consumption is constant (for $\beta = \delta$), independent of the user cost of health capital at the margin and independent of the marginal production benefit of health.[4] Health investment shows a more complex dependence on the user cost of health capital at the margin and the marginal production benefit of health than consumption does (see equations 3.12 and 3.13).

For the "optimal" health stock $H_*(t)$ we find the following solutions:

[3]Notice that $\min\{1, 1/\rho\} \leq \chi \leq \max\{1, 1/\rho\}$, given that $\rho > 0,\ 0 \leq \varsigma \leq 1$.

[4]This can be understood as follows. The cost of holding the health stock increases with the user cost of health capital at the margin $\pi_H(t)$ and decreases with the marginal production benefit of health $\varphi(t)$ (see equation 3.7; $q(t) = 0$). Higher cost of holding the health stock would thus result in lower health levels $H_*(t)$. The marginal cost of consumption on the other hand does not change with changes in the user cost of health capital at the margin or with the marginal production benefit of health (see equation 3.6). In other words, the marginal benefit of consumption is also unchanged. The marginal benefit of consumption $\partial U(t)/\partial C(t) \propto H(t)^{\rho(\chi-1)} C(t)^{-\rho\chi}$ (where we have used 3.6 and 3.9 and $\beta = \delta$) increases with health for $\chi > 1$ and decreases with health for $\chi < 1$. In other words, higher costs of holding the health stock result in lower health levels and therefore lower (higher) consumption levels for $\chi < 1$ ($\chi > 1$). For $\chi = 1$ the marginal benefit of consumption is independent of health and hence there is no effect of health changes on the level of consumption.

$$H_*(t) = (1-\zeta)\Lambda\left[\pi_H(t)-\varphi(t)\right]^{-\chi} e^{-\left(\frac{\beta-\delta}{\rho}\right)t} \quad (t \leq R) \tag{3.17}$$

$$H_*(t) = k^{\frac{1}{\rho}}(1-\zeta)\Lambda\left[\pi_H(t)\right]^{-\chi} e^{-\left(\frac{\beta-\delta}{\rho}\right)t} \quad (t > R). \tag{3.18}$$

The trajectory described by equation (3.17) is the path that individuals would follow if initial health $H(0)$ would be exactly on this trajectory and is what is referred to in the literature as the "optimal" health stock (e.g., Grossman, 1972a, 2000). Similarly, equation (3.18) describes the trajectory that health would follow if health at retirement were exactly equal to $H_*(R_+)$.

As many authors have found (e.g., Case and Deaton, 2005), the "optimal" health stock $H_*(t)$ is constant for constant time paths of $d(t) = d_0$, $p(t) = p_0$, $\mu(t) = \mu_0$, $\varphi(t) = \varphi_0$ (i.e., for a constant user cost of health capital) and for $\beta = \delta$, and decreases for an increasing deterioration rate with age $\dot{d}(t) > 0$.

At the age of retirement the solutions ($q(t) = 0$) for the "optimal" level of consumption (equations 3.10, 3.11), "optimal" level of health investment (equations 3.12, 3.13) and "optimal" level of health (equations 3.17, 3.18) are discontinuous. These jumps represent the change in consumption and health investment as a result of differences in utility from more leisure time during retirement (depending on the value of k, leisure is a substitute or a complement of consumption and health) and because health has no effect on income after retirement ($\varphi(t) = 0$).

3.3.2 Model solutions: general case

The literature generally assumes that individuals are capable of ensuring that their health is at the "optimal" level $H_*(t)$ (e.g., Grossman, 1972a, 1972b, 2000; Case and Deaton, 2005; Muurinen, 1982; Wagstaff, 1986a; Zweifel and Breyer, 1997; Ried, 1998). In other words, the literature assumes that either the initial health endowment $H(0)$ is at or very close to the "optimal" health stock $H_*(0)$ or that individuals find this health level desirable and are capable of rapidly dissipating or repairing any "excess" or "deficit" in health.

Unlike most discussion in the literature we argue that initial conditions are likely of importance and that health will in many circumstances not follow the "optimal" health stock. An essential characteristic of the model is that health cannot deteriorate faster than the natural deterioration rate $d(t)$. As equation (3.4) shows, any surplus in health above the equilibrium health path can at most dissipate at the natural rate of health deterioration $d(t)$ (this would correspond to individuals not investing in their health; $m(t) = 0$). As a result initial conditions cannot be dissipated rapidly (and what use

would it be to shed any excess in health which provides utility and increases earnings?). Nor is there any reason to expect the endowment of health to exactly equal the "optimal" health stock (see also Wolfe, 1985).

We allow health to have an initial value $H(0)$ that is different from the "optimal" health stock (see also Wolfe, 1985). To take into account that any "excess" in health cannot dissipate faster than the natural deterioration rate $d(t)$ we explicitly demand that medical care is a positive quantity $m(t) \geq 0$ by introducing the multiplier $q(t)$ in the Lagrangean (equation 3.5). We thus allow for the existence of corner solutions where individuals do not invest in medical care $m(t) = 0$ for certain periods of time. As a result, given initial health $H(0)$, the "optimal" health stock is not the optimal solution.[5] Any situation with "excessive" initial health (initial health $H(0)$ above $H_*(0)$) is preferable: individuals with excess initial health have higher levels of life-time health and consumption and therefore greater life-time utility. In other words, if individuals could choose they would always prefer "excessive" initial health $H(0)$ over the "optimal" health stock $H_*(0)$ (if $H(0) > H_*(0)$). In our formulation individuals use their "excess" health for the consumption and production benefit this "excess" in health provides.

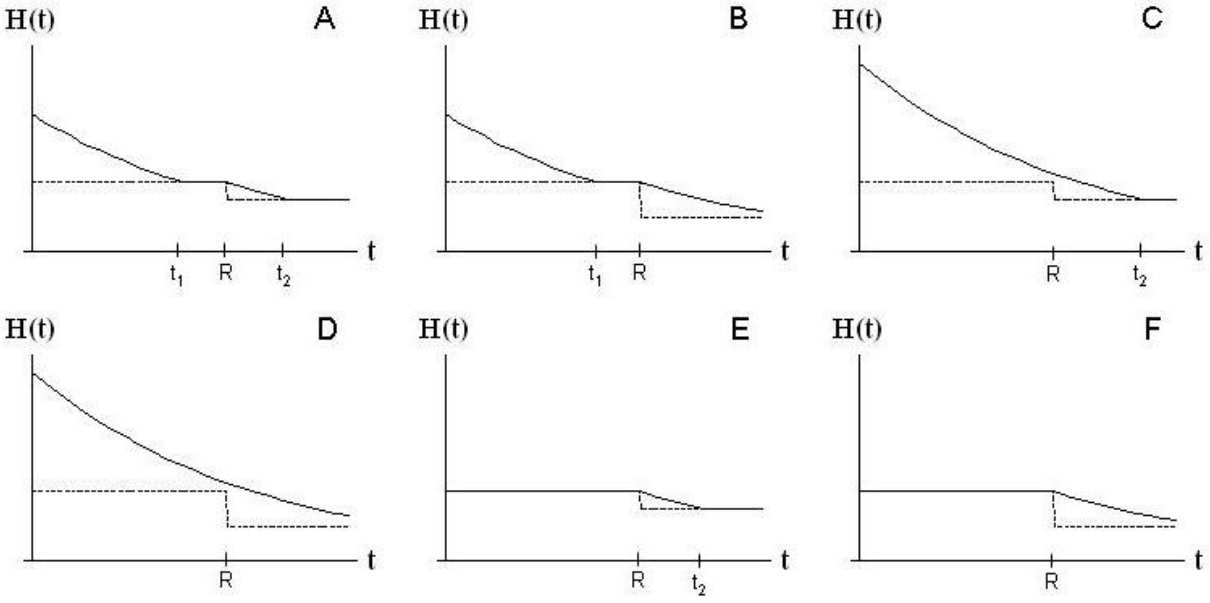
The solution for the "optimal" health stock $H_*(t)$ is instead the *minimum* level individuals "demand" for the productivity benefit and utility that good health provides. Individuals with health endowments $H(0)$ below the "optimal" health stock $H_*(0)$ will invest in medical care (an adjustment cost) to reach the "optimal" health level (see for details section 3.8.7 in the Appendix). For this reason we term the "optimal" health stock $H_*(t)$ the "minimally economically productive" or "minimally productive" health stock. Individuals only invest in health when they are "unhealthy" (health levels below or at the minimally productive level) and not when they are "healthy" (health levels above the minimally productive level). In other words, the minimally productive health level operates as a health threshold. In the following we will refer to what is traditionally called the "optimal" solution for health, as the "health threshold" or as the "minimally productive" level of health.

While the literature spawned by Grossman does not provide a convincing theoretical argument that health should be at or close to the "optimal" health stock $H_*(t)$, the ultimate test of our proposition that this assumption is invalid is to contrast its predictions with data. In separate work (Galama and Kapteyn, 2009 [Chapter 2]) we propose structural and reduced form equations to test our proposition. We also contrast the predictions of our interpretation of the Grossman model (in which solutions where individuals do not invest in health $m(t) = 0$ for certain periods of time are allowed) with the "tradi-

[5] Hence, our use of quotation marks.

tional" interpretation (in which health always follows the "optimal" health stock $H_*(t)$) and with the empirical literature. In a review of the empirical literature we find that the interpretation advocated here provides a better explanation for the observed evolution of health and of medical consumption. Importantly, our interpretation can explain the observation that measures of medical care are negatively correlated with measures of health[6] while the traditional interpretation cannot (the Grossman model has received significant criticism regarding its inability to correctly predict this crucial relationship; see, e.g., Grossman 2000; Zweifel and Breyer, 1997). For more details see Galama and Kapteyn (2009) (Chapter 2).

Figure 3.1: Six scenarios for the evolution of health.



Notes: t_1 and t_2 denote the ages at which health (solid line) has evolved towards the health threshold (dotted line), and R denotes the age of retirement. The health threshold drops at the age of retirement R as a result of differences in utility due to increased leisure time during retirement (depending on the value of k, leisure is a substitute or a complement of consumption and health) and because health has no effect on income after retirement ($\varphi(t) = 0$).

We distinguish six scenarios as shown in Figure 3.1. The health threshold $H_*(t)$ (dotted line) drops at the age of retirement R as a result of the difference in utility due to increased leisure time during retirement (for our choice of parameters $k < 1$ leisure is a substitute of consumption and health) and because health has no effect on income after

[6]Healthy individuals (above the threshold) do not invest in health while unhealthy individuals (at or below the threshold) do.

retirement ($\varphi(t) = 0$). We show the simplest case in which the health threshold $H_*(t)$ is constant with time (e.g., for constant time paths of $d(t) = d_0$, $p(t) = p_0$, $\mu(t) = \mu_0$, $\varphi(t) = \varphi_0$ and for $\beta = \delta$) but the scenarios are valid for more general cases. Scenarios A, B, C and D begin with initial health $H(0)$ greater than the initial health threshold $H_*(0)$ and scenarios E and F begin with initial health $H(0)$ below the initial health threshold $H_*(0)$. In scenarios A and B health $H(t)$ reaches the health threshold $H_*(t)$ before the age of retirement R (at age t_1). In scenario A the health threshold $H_*(t)$ is once more reached at age t_2 before total life time T, but this is not the case in scenario B. In scenario C health $H(t)$ reaches the threshold $H_*(t)$ after the age of retirement R (at age t_2), and in scenario D health $H(t)$ never reaches the threshold $H_*(t)$ during the life of the individual. In scenarios E and F individuals begin working life with health levels $H(0)$ below the initial health threshold $H_*(0)$. Individuals will substitute initial assets $A(0)$ for improved initial health $H(0)$ such that initial health equals the initial health threshold $H(0) = H_*(0)$ (see section 3.8.7 in the Appendix for a more detailed discussion).

The detailed solutions for health $H(t)$, consumption $C(t)$ and health investment $m(t)$ for each of the six scenarios are provided in the Appendix. Assets $A(t)$ can be derived by substituting the solutions for health $H(t)$, consumption $C(t)$ and health investment $m(t)$ as follows:

$$\begin{aligned} A(t) &= \left\{ \int_0^t \left[w_0(x) + \varphi(x)H(x) - C(x) - p(x)m(x) \right] e^{-\delta x} dx \right\} e^{\delta t} \\ &+ A(0)e^{\delta t} \qquad (t \leq R) \end{aligned} \tag{3.19}$$

$$\begin{aligned} A(t) &= \left\{ \int_R^t \left[b - C(x) - p(x)m(x) \right] e^{-\delta x} dx \right\} e^{\delta t} \\ &+ A(R)e^{\delta(t-R)} \qquad (t > R). \end{aligned} \tag{3.20}$$

As a last note, each of the solutions are fully determined, that is by substituting the solutions for health $H(t)$, consumption $C(t)$, health investment $m(t)$ and assets $A(t)$ in the life-time budget constraint (equation 3.3) we can derive the constant Λ (or equivalently the constant $p_A(0)$). For more details see the Appendix.

3.4 Treatment of benefits

We introduce one further level of complexity to the model. In the set-up so far, benefits are independent of work history. Typically benefits are related to how long one has worked and the wages earned during working life. As a stylized representation of this we assume

that a fraction of wages $\alpha w(t)$ are saved for retirement. Benefits accumulate with time and are invested with a return on investment of δ (the interest rate) as follows:

$$b(R) = b_0 + f(R)\alpha \int_0^R w(t)e^{\delta t}dt, \tag{3.21}$$

where the pension accumulation function $f(R)$ describes how benefits accumulate as a function of retirement age R and b_0 represents a base pension benefit.

The base pension benefit b_0 is provided regardless of years worked, e.g., it could represent a first-tier basic pension (OECD, 2005) or a statutory poverty line. The remaining term in (3.21) represents the part of the pension that accumulates with years of work. This could represent a defined contribution (DC) plan or a defined benefit plan (DB) or it could represent an individual's portfolio of DB and DC plans. Pension wealth in retirement thus consists of a base pension b_0 (typically provided by the state), an individual private pension (either DB and/or DC) and accumulated assets $A(R)$ that can be drawn down during retirement.

A particular pension accumulation functional form of interest is $f(R) = \delta/[1 - e^{-\delta(T-R)}]$, which is actuarially fair (accumulated pension wealth is paid out over the number of years in retirement $T - R$). Such a functional form is an approximation of a DC plan where the beneficiary can use his or her accumulated pension investment to purchase a life-time annuity.[7] The function $f(R)$ for a DB plan, on the other hand, would typically consist of an annual contribution rate per year worked and a conversion factor which would depend on R in a way which is not necessarily actuarially fair. As Lazear (1986) finds, the actuarial value of private pensions first rises but then declines as workers continue to work beyond a certain age. Lazear argues that sharp decreases in the actuarial value of retirement with continued work are used as a device by employers to induce earlier retirement of workers. Such a function could be represented by $f(R) = \delta/[1 - e^{-\delta(T-R)}]$ up to a retirement age R_* after which the function flattens to a constant or even slightly declining function of retirement age.

Replacing the assumed flat retirement benefits by (3.21) the previously derived equations remain valid with the following transformation

$$\begin{aligned} w_0(t) &\rightarrow (1-\alpha)w_0(t) \\ b &\rightarrow b_0 + \alpha f(R) \int_0^R w_0(t)e^{\delta t}dt \\ \varphi(t) &\rightarrow \varphi(t)\left[(1-\alpha) + f(R)\tfrac{\alpha}{\delta}\left(e^{-\delta R} - e^{-\delta T}\right)e^{2\delta t}\right] \end{aligned} \tag{3.22}$$

[7] In this example, the annuity is assumed to be actuarially fair. In a world with asymmetric information this assumption clearly needs to be modified.

Thus, even for constant time paths of $d(t) = d_0$, $p(t) = p_0$, $\mu(t) = \mu_0$, $\varphi(t) = \varphi_0$ and for $\beta = \delta$, consumption and health investment are not constant as the transformation for the marginal production benefit of health $\varphi(t)$ is a function of time. A derivation of transformation (3.22) is provided in the Appendix.

3.5 Endogenous retirement

Now let us finally return to the issue of the influence of health on the decision to retire. In our formulation, the decision to retire is determined by three factors. Individuals find retirement increasingly attractive as they age because of: (1) wage declines $w(t) = w_0(t) + \varphi(t)H(t)$ as a result of gradual health deterioration reducing income from work with age, (2) increased leisure time during retirement (factor k boost in utility) and (3) an increasing level of pension benefits $b(R)$ with years in the workforce.

Now consider the case where the age of retirement R can be chosen freely. The optimal R can be determined by inserting the solutions for $C(t)$, $H(t)$ into the "indirect utility function", $V(R)$, and differentiating $V(R)$ with respect to R.

$$V(R) \equiv \int_0^R U_w(t)e^{-\beta t}dt + \int_R^T U_r(t)e^{-\beta t}dt \tag{3.23}$$

Unfortunately the resulting expression for $V(R)$ turns out to be unwieldy for most of the scenarios A through F shown in Figure 3.1 (see the various solutions for $C(t)$ and $H(t)$ in the Appendix; note that we do not show the solution for $V(R)$ given its complexity). After differentiation of $V(R)$ with respect to R we do not find a simple solution for the optimal age of retirement R and therefore have to resort to numerically solving for the optimal retirement age R.

3.6 Simulations

In this section we begin by making some plausible assumptions about the model parameters and initial and terminal conditions. This will provide us with a starting point (our baseline model; section 3.6.1) from which we will subsequently deviate in order to investigate the impact of the various model levers on the decision to retire. For illustrative purposes we graph the solutions for consumption, health investment, health, assets, etc. and contrast the model with some stylized observations from the literature. We then briefly explore model simulations of health inequality (section 3.6.2) and the effect

of health insurance on health and retirement (section 3.6.3). We discuss in detail the sensitivity of retirement age to model parameters (section 3.6.4) and discuss briefly the sensitivity of other model outcomes, such as, life-time consumption, life-time health investment, life-time health, and life-time assets, to changes in model parameters (section 3.6.5).

3.6.1 Calibration baseline model: white collar worker

Individuals begin work at age 20 (corresponding to $t = 0$), and, depending on the solution for the optimal retirement age, retire some 45 years later at an age of about 65 years (corresponding to $R \approx 45$). Individuals die with certainty at 85 years of age (corresponding to $T = 65$).

For simplicity we assume constant time paths of $d(t) = d_0$, $p(t) = p_0$, $\mu(t) = \mu_0$, $\varphi(t) = \varphi_0$, $w_0(t) = w_0$[8] and take $\beta = \delta$. We further assume an annual income of $w(t) \approx$ \$45,000 for healthy "white collar" workers[9] and that healthy workers have a health stock of about 1.5 times that of unhealthy workers (we will discuss "blue collar" workers later). We can then obtain 25% higher earnings for healthy workers[10] for constant marginal production benefits of health $\varphi(t) = \varphi_0 \approx 1.5 w_0 / H_H$ (where H_H is health for a healthy worker), and a constant base wage rate $w_0(t) = w_0 \approx$ \$20,000 per year. A roughly 50% decline in wage between first employment ($t = 0$) and retirement ($t = R$)[6] requires that by the age of retirement health has fallen to one-fourth the level of health at first employment

[8]It is straightforward to use a more realistic wage profile, for example the commonly used earnings function by Mincer (1974) where the log of earnings is a quadratic function of age and linear in years of schooling. However, this would introduce additonal complexity into the model. The overal shape, i.e. height at peak, age at peak and curvature of the earnings function with age would influence the optimal age of retirement. In order to not complicate the interpretation of the effect on the retirement age of parameters that are of greater interest (than the parameters of the wage profile) we have chosen a simple constant base-wage rate $w_0(t) = w_0$.

[9]Median annual earnings for males were \$40,798 and for females \$31,223 in 2004, according to the US Census Bureau.

[10]French (2005) provides hourly wage and annual hours worked profiles for males by age and self-reported health status from the panel study of income dynamics (PSID). French finds that the effect of health on wages is relatively small: the hourly wage is about 10% higher and the annual hours worked are some 10% higher for healthy compared with unhealthy individuals. In our formulation we use annual wages, i.e. the product of hourly wages times the annual hours worked. Thus annual wages would be about 20-25% higher for healthy individuals. The hourly wage profiles show a wide hump (relatively flat between the ages of 40 and 60) for both healthy and unhealthy males with wages peaking near age 55 and a fairly rapid decline after age 60. The annual hours worked profiles show a relatively smooth decline with age, dropping by about 20% from age 30 to age 60 after which the decline accelerates and drops to 50% by age 70 (again compared with age 30).

$H(0)$. We can simulate such results with an initial health $H(0)$ of \$30,000,[11] a constant health deterioration rate $d(t) = d_0$ of 5%, a contribution rate for retirement α of 15% of wages, zero basic benefits $b_0 = 0$, a coefficient of relative risk aversion $\rho = 1.32$, a constant health investment efficiency $\mu(t) = \mu_0 = 0.7\%$ and time preference rate and interest rate $\beta = \delta$ of 3%. We interpret prices $p(t)$ as the co-pay rates, which we take to be constant at $p(t) = p_0 = 20\%$, and $m(t)$ as the total annual medical expenditures – though this could include the cost of other health promoting activities such as exercise, diet, etc.

Hurd and Rohwedder (2003, 2006) find that "on average" consumption drops between 15 and 20% after retirement. We use this observation to determine the value for k by requiring that consumption $C(t)$ drops at retirement to 85% of its value before retirement.[12] Hence we demand that (see for details the Appendix):

$$k^{\frac{1}{\rho\chi}} = 0.85. \tag{3.24}$$

For the values chosen, we have $k = 0.81$.

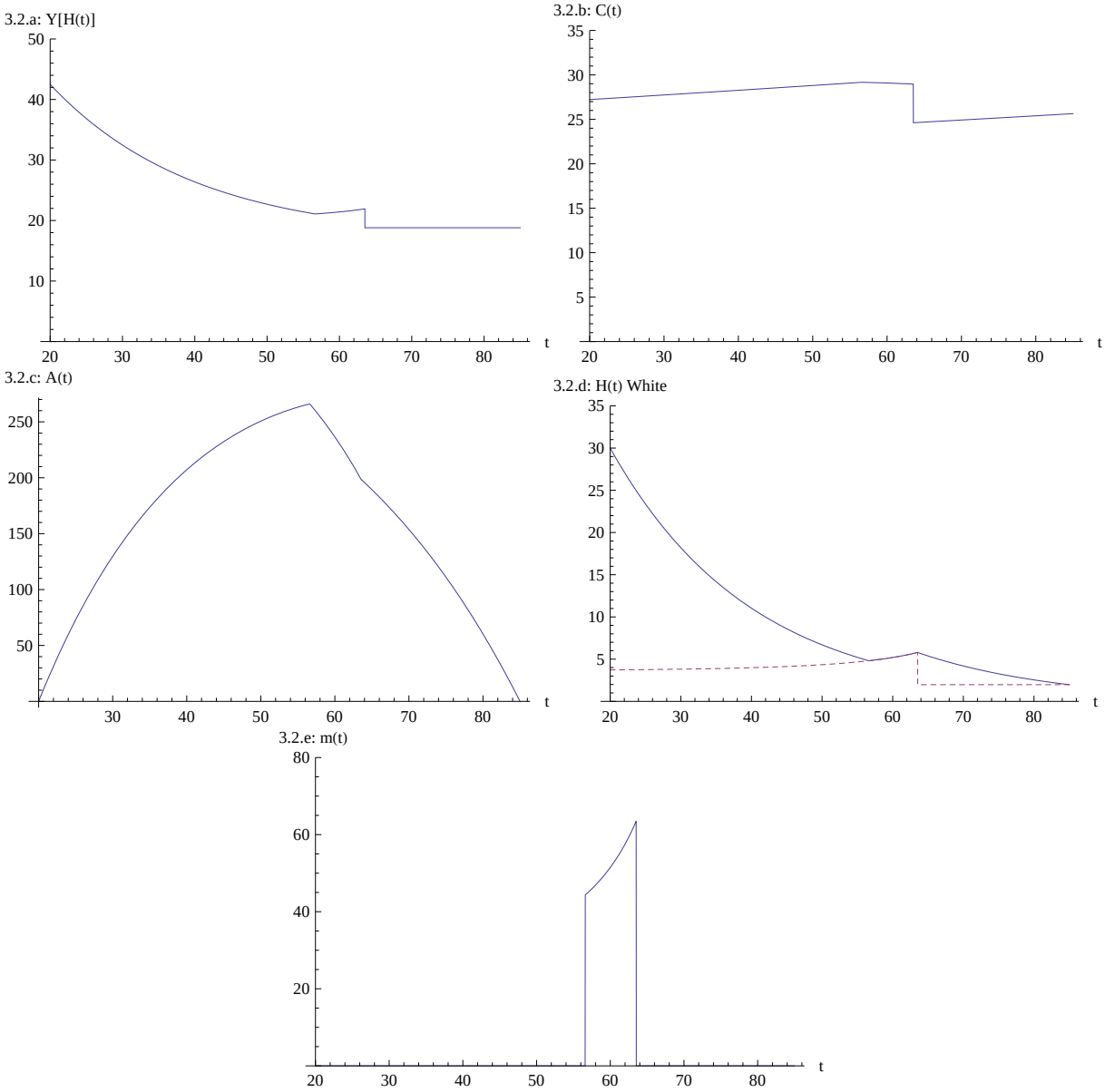
To ensure that health investment is not too far from the observed mean out-of-pocket medical expenditures of around \$3000 per year (corresponding to total medical expenditures of \$15,000) we assume $\varsigma = 0.85$, i.e. that an individual's preferences are significantly skewed towards consumption and away from health. We assume an actuarially fair benefits accumulation function $f(R) = \delta/[1 - e^{-\delta(T-R)}]$, i.e. as approximately in a DC plan. Lastly, we assume that individuals leave no bequests and receive no bequests, i.e. $A(0) = A(T) = 0$. There are likely many other plausible scenarios and parameter values. The current values are only for illustrative purposes.

For this set of parameters and assumptions (see Table 3.1 for a quick overview) we find ourselves in scenario A and determine an optimal age of retirement of 63.52 (corresponding to $R = 43.52$). Figures 3.2.a-3.2.e describe the evolution of income, consumption, assets, health and health investment for the optimal retirement age of 63.52 years.

[11] The dimension of health (dollars) can be understood as follows. Denoting the dimension of health by $[H]$ we have according to the first equation of 3.2 that $[\dot{H}] = [H]/[t] = [m][\mu]$ (where $[t]$ is the dimension of time [e.g., days, seconds etc], $[m]$ is the dimension of medical care [e.g., dollars per unit of time] and $[\mu]$ is the dimension of the efficiency of medical care $\mu(t)$). We then have $[H] = \$\,[\mu]$. For simplicity we assume the efficiency function is dimensionless and hence health is expressed in dollars.

[12] Hurd and Rohwedder (2006) argue that a number of explanations operate together to explain the magnitude of the observed drop in consumption at retirement. The substitution between leisure and consumption is only one such factor. In addition, there are individuals who do not experience a drop in consumption and there are those who experience more substantial drops in consumption. The assumed drop of magnitude 15% is for illustrative purposes only.

Figure 3.2: Income, consumption, assets, health and health investment versus age for a white collar worker.



Notes: Income ($Y[H(t)]$; \$ thousands), consumption ($C(t)$; \$ thousands), assets ($A(t)$ \$ thousands), health ($H(t)$; \$ thousands; total health [solid line], health threshold [dashed line]) and health investment ($m(t)$; \$ thousands per year) versus age for a "white" collar worker.

As Figure 3.2.a shows, earnings $Y[H(t)]$ during working life fall with declining health until the age of retirement when earnings are replaced by an annuity.[13] Consumption $C(t)$

[13]As discussed earlier (see footnote 8) it is relatively easy to introduce more realistic wage age profiles. Because the shape of the wage age profile influences retirement and because we are primarily interested in the effect of health on the optimal retirement age we have chosen an simple wage profile where the

(Figure 3.2.b) is relatively constant over time as individuals smooth life-time consumption through the use of savings $A(t)$[14] (Figure 3.2.c). Consumption shows a sudden drop at retirement to 85% of its level before retirement (this is the direct result of our choice for the value of leisure k; see equation 3.24) as individuals substitute leisure for consumption. For the parameters chosen, individuals build up assets $A(R)$ of $\approx$ \$198,700 at the age of retirement (Figure 3.2.c) and a pension b of \$18,800 per year (representing a present discounted value $(b/\delta)[1 - e^{-\delta(T-R)}]$ of \$222,800). Health $H(t)$ (the solid line in Figure 3.2.d) declines fairly rapidly from a value of \$30,000 to about \$4,800 by age 56.6 ($t_1 = 36.6$) after which the individual starts investing in health (see Figure 3.2.e). Health reaches \$5,800 by the age of retirement R and declines further to about \$2,000 by the end of life T. The dashed line in Figure 3.2.d shows the health threshold. The health threshold increases over time up to the retirement age[15] after which it suddenly drops due to the substitution of health for leisure and the disappearance of production benefit of health φ during retirement. In our formulation and for the parameters chosen, the effect of retirement on an individual's health is negative – retirement is bad for health – as individuals lower their investment in health due to substitution of health for leisure and because health loses its relevance as a means to increase an individual's income.

Because the marginal production benefit of health $\varphi(t)$ is the only term in the transformation (3.22) that is time dependent, and because the model solutions after retirement are not functions of $\varphi(t)$, we see that the health threshold (Figure 3.2.d) is constant over time during retirement (given our choice of constant health deterioration $d(t)$, prices $p(t)$, efficiency $\mu(t)$ and interest rate δ).

3.6.2 Health inequality

Case and Deaton (2005) show that "white collar" workers are in better health and have lower health deterioration rates than "blue collar" workers (based on self-reported health assessments). They, as well as Muurinen and Le Grand (1985), suggest this observation could be explained by the need for blue collar workers to perform more physically demanding work than non-manual occupations, which may not be open to lower educated workers. As a result blue collar workers "wear" their bodies out more quickly. An additional (or alternative) explanation could be that "blue collar" workers have lower health

base wage $w_0(t)$ is constant. Thus we can isolate the direct effect of parameter changes from any indirect effect that operates through the wage age profile.

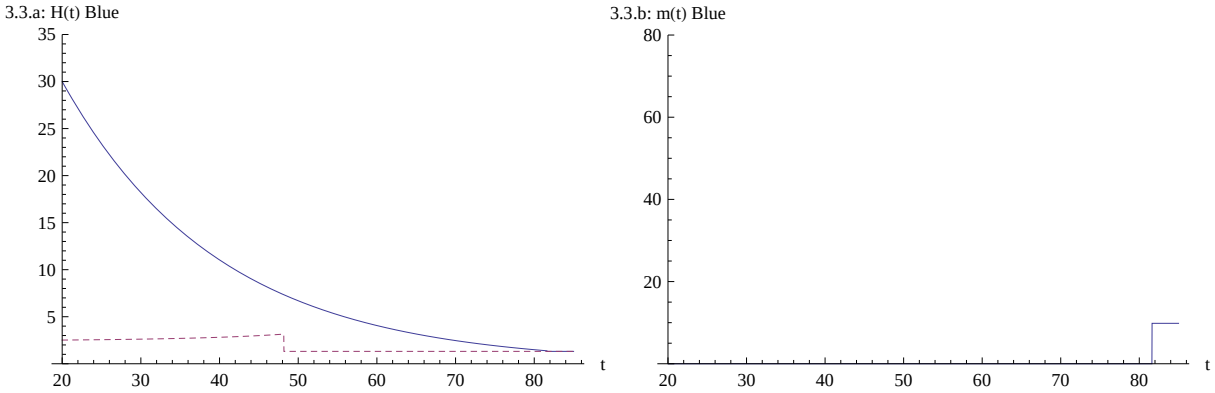
[14]Note that individuals are also allowed to borrow at interest rate δ

[15]This is the result of the time dependence of the marginal benefit of health $\varphi(t)$ as a result of the benefit transformation (equation 3.22).

thresholds (lower levels of minimally productive health) $H_*(t)$ (but essentially the same "natural" health deterioration rate $d(t)$ as "white collar" workers) as a result of access to lower life-time resources. The lower value of $H_*(t)$ induces them to invest less in health.

Figure 3.3.a shows the evolution of health for "blue collar" workers with a base wage rate of w_0=\$10,000 (half that of "white collar" workers; everything else held constant). The lower earnings of "blue collar" workers reduce their life-time income, their health threshold, and induce earlier retirement at age 53.16 ($R = 33.16$). As Figure 3.3.b shows, health investment is lower over the life-time for "blue collar" workers. For these specific values workers do not invest in health during working life but only near retirement (scenario C). As a result health declines to about \$5,700 by the age of retirement 53.16 ($R = 33.16$) and to \$1,400 by age 81.62 when individuals start investing in health ($t_2 = 81.62$). Also, earlier retirement extends the retirement phase of life for "blue collar" workers which is characterized by a lower health threshold (lower level of minimally productive health) and therefore associated with lower levels of health investment and consequently lower health. As a result, at age 82 ($t = 62$), white collar workers are more than 40 percent healthier than blue collar workers.

Figure 3.3: Blue collar health and blue collar health investment versus age.



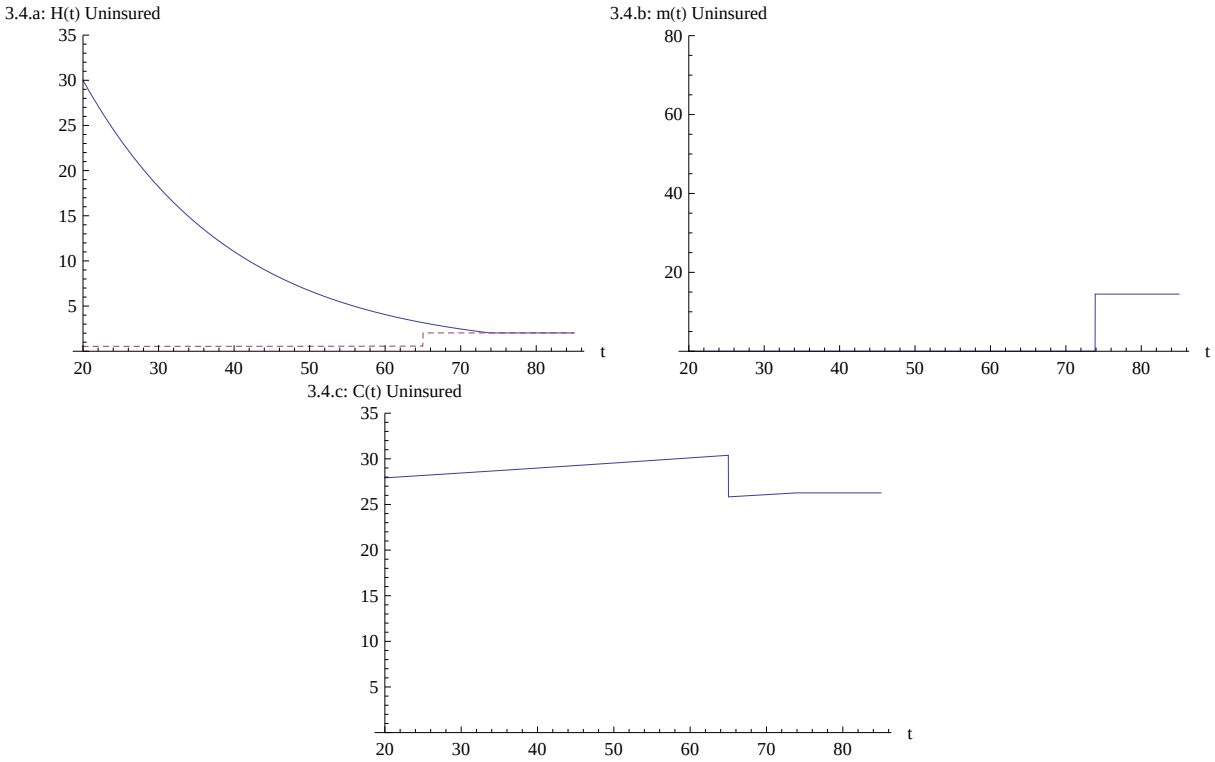
Notes: Blue collar health (3.3.a left-hand side; health [solid line] and health threshold [dashed]; \$ thousands) and blue collar health investment (3.3.b right-hand side; \$ thousands) versus age.

3.6.3 Health insurrance

We now explore the role of health insurance on health, health investment and retirement. Figure 3.4 shows the impact of being uninsured. We use the same parameters as before for a white collar worker (our baseline model) but assume $p(t) = 1.0$ (i.e., health investment is paid for one hundred percent out-of-pocket) before the age of Medicare eligibility.

Afterwards $p(t) = 0.2$ (i.e., we assume that after age 65 the uninsured are covered by a universal health insurance program, such as Medicare). Figures 3.4.a, 3.4.b and 3.4.c show how uninsured individuals invest much less in health (health investment begins at age 73.88 [$t_2 = 53.88$]), therefore have higher effective health deterioration rates and are unhealthier (compare with Figure 3.2). Interestingly, consumption is not significantly affected while the age of retirement now coincides with the age of Medicare eligibility (age 65). Note the significantly lower level of the health threshold (the minimally productive health level) before the Medicare eligibility age of 65, during which health investment is paid one hundred percent out-of-pocket.

Figure 3.4: Health, health investment and consumption for the uninsured versus age.



Notes: Health (3.4.a left-hand side; $ thousands), health investment (3.4.b center; $ thousands) and consumption (3.4.c right-hand side; $ thousands) versus age.

3.6.4 Retirement

We are further interested in the effect of assets, wages, benefits, health, health deterioration rates, and other variables and parameters on the decision to retire. Figures 3.5.a through 3.5.l show the effect of various model parameters on the decision to retire. The

solid, dotted and dashed lines in each of the graphs show how, respectively, optimal retirement age R, t_1 and t_2 change in response to variation in a number of variables and parameters. As variables and parameters are varied, the solutions cycle through the scenarios A through F (see Figure 3.1). For example, Figure 3.5.b shows that as we increase the base wage rate w_0, we transition from scenario D ($t_1 > R$ and $t_2 > T$) for values of w_0 below $\sim$ \$6,000 to scenario C ($t_1 > R$ and $t_2 < T$) for values of w_0 between $\sim$ \$6,000 and $\sim$ \$20,000. For values of w_0 between $\sim$ \$14,000 and $\sim$ \$20,000 the age of retirement R falls slightly as the optimal age of retirement tracks the evolution of t_1, i.e. the solution remains on the boundary between scenarios A and C ($t_1 = R$ and $t_2 < T$). Around $w_0 \sim$ \$20,000 we observe a jump in the age of retirement R as we move to scenario B for the remainder of the graph. Initially the solution remains on the boundary between scenarios B and D ($t_2 = T$) explaining the "flat" initial portion of the retirement graph for \$20,000 $< w_0 <$ \$24,000. For values $w_0 >$ \$24,000 we have $t_2 > T$ and retirement R continues its upward trend with increasing base wage rate w_0 (scenario B). Similar explanations hold for the other graphs in Figure 3.5.

We now concentrate on the variation of the optimal retirement age R with the various variables and parameters (solid line in Figures 3.5.a through 3.5.l). Figure 3.5.a shows how greater initial assets $A(0)$ reduce the retirement age. Wealthy people have less incentive to work as they can fulfill all or part of their consumption needs through inherited wealth. Figure 3.5.b shows that higher wages w_0 increase the age at which individuals retire. Unlike a one-off contribution to life-time resources (such as initial assets $A(0)$), higher wages provide additional resources for as long as the individual works, thereby increasing the age of retirement. Indeed Mitchell and Fields (1984) find that higher earnings result in later retirement.

Figure 3.5.c shows how increasing levels of basic benefits b_0 reduce the retirement age.[16] Indeed, we expect earlier retirement in countries with more generous benefits, as was shown in the cross-country comparison project of Gruber and Wise (1999, 2004, 2010).

Figure 3.5.d shows that the higher the portion α of wages set aside for retirement the earlier an individual retires. Given that retirement in our formulation is completely the result of individual choice (benefits are approximately actuarially fair and the timing of retirement is not constrained) the role of pension wealth and that of regular savings is essentially the same. Lower pension savings will almost exactly be offset by larger accumulated savings (total life-time resources remain the same). In case retirement is

[16]Very early retirement in our model should probably be interpreted as the result of generous unemployment benefits rather than retirement benefits.

not a choice variable (or at least restricted in various ways) lower benefits will decrease life time resources, which will lower consumption and thereby also generate more asset accumulation. Indeed Kapteyn and Panis (2003) find a strong negative relation between wealth at retirement and replacement rates when comparing Italy, the Netherlands, and the U.S.

Figure 3.5.e shows how increasing initial health $H(0)$ reduces the retirement age. Health provides "health capital" as can be seen from the equation for total life-time resources (right-hand side of 3.3). Initial health $H(0)$ thus operates qualitatively similar to assets and we observe a decrease in the age of retirement with increasing initial health.

The age of retirement increases with increasing rates of health deterioration $d(t) = d_0$ (Figure 3.5.f). For one, higher health deterioration over one's life-time reduces the amount of additional life-time earnings resulting from an individual's inherited health $H(0)$ (see the last term in Equation 3.3). This would increase the retirement age as it reduces the "effective" initial health endowment. In addition, the user cost of health capital at the margin $[p_0/\mu_0]\,[d_0 + \delta] - \varphi_0$ is higher, which also leads to delayed retirement.

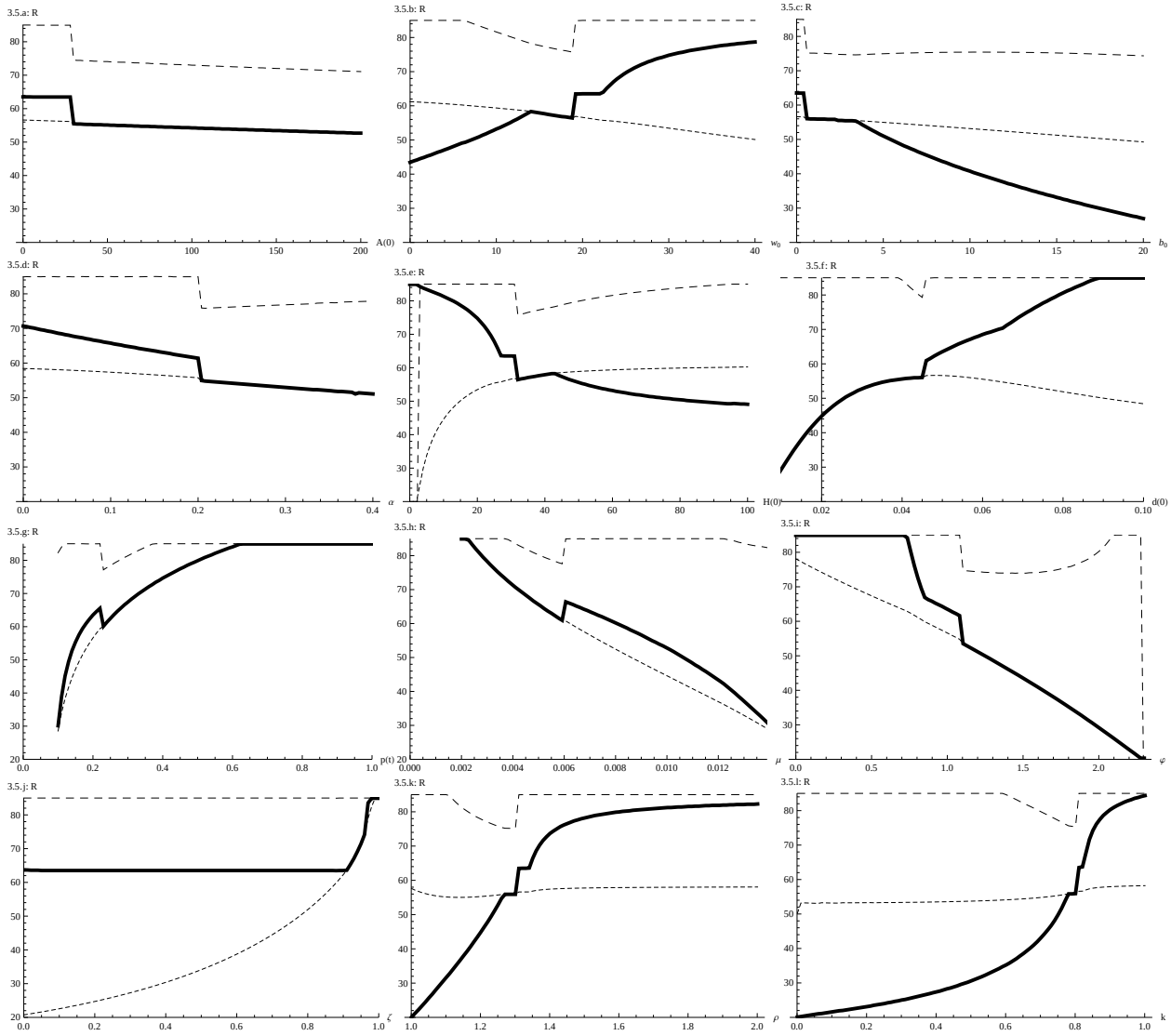
Similarly increasing prices of health care $p(t) = p_0$ (Figure 3.5.g), decreasing health investment efficiency $\mu(t) = \mu_0$ (Figure 3.5.h) and decreasing marginal productivity benefits of health $\varphi(t) = \varphi_0$ (Figure 3.5.i) increase the cost of health capital at the margin and raise the retirement age.

The relationships between prices $p(t)$, health investment efficiency $\mu(t)$, the marginal production benefits of health $\varphi(t)$, the coefficient of relative risk aversion ρ (Figure 3.5.k), and the factor k (Figure 3.5.l; describing the increased utility from leisure during retirement) and retirement are particularly strong in that individuals never work ($R = 0$) or never retire ($R = T$) for certain parameter values. The relative utility weight ς given to consumption versus health has very little impact on the age of retirement (Figure 3.5.j) except near the extreme of $\varsigma \approx 1$ (pure consumption model). Model simulations as well as observations of analytical solutions from simplified versions of our model (a $\varsigma \approx 1$ pure consumption model as in our simulation and $\delta = \beta = 0$) show that the parameters ρ (Figure 3.5.k), k (Figure 3.5.l) and retirement R are strongly related.

3.6.5 Sensitivity analysis

In addition to the effect of the various parameters on retirement it is of interest to understand more generally the sensitivity of the model to the model parameters. Table 3.1 displays the baseline model parameter values P_0 and the sensitivity to changes in each of the model parameters of life-time consumption, life-time health investment, life-time

Figure 3.5: The effect of various variables and parameters on the decision to retire.



Notes: Initial assets $A(0)$, base wage rate w_0, benefits b_0 are shown in \$ thousands, and initial health $H(0)$ in \$ thousands. Values for health deterioration $d(t) < 0.005$, prices $p(t) < 0.088$, health investment efficiency $\mu > 0.016$, and marginal production benefits of health $\varphi > 2.29$ are not shown as they correspond to a user cost of health capital at the margin $[p_0/\mu_0]\,[d_0 + \delta] - \varphi_0$ that is negative. Values of $\rho < 1$ and $k > 1$ are not shown as these require a change in specification; for $\rho = 1$ the utility function switches from being negative ($\rho > 1$) to positive ($\rho < 1$) values. For positive utility, values of $k < 1$ imply disutility from increased leisure, i.e. we need to also switch to values of $k > 1$.

health, life-time assets and the age of retirement (endogenous in the model). The sensitivities were estimated by calculating the relative change in the quantity X of interest (e.g., life-time consumption) in response to a one percent change in model parameter values P_0(e.g., $\partial \ln X / \partial \ln P_0$). For example, a one percent increase in co-payment p_0 increases

the age of retirement by 0.51 percent and life-time consumption by 0.28 percent (see Table 3.1). In other words, retirement and life-time consumption are not very sensitive to co-payment. On the other hand life-time health investment and life-time assets are more responsive to changes in co-payment, showing decreases by 1.12 percent and 1.44 percent, respectively.

Table 3.1: Sensitivity (elasticities) of model outcomes to various variables and parameters.

P	P_0	$\frac{\partial \ln \left[\int_0^T C(t)e^{-\delta t}dt\right]}{\partial \ln P_0}$	$\frac{\partial \ln \left[\int_0^T m(t)e^{-\delta t}dt\right]}{\partial \ln P_0}$	$\frac{\partial \ln \left[\frac{\int_0^T H(t)e^{-\delta t}dt}{\int_0^T e^{-\delta t}dt}\right]}{\partial \ln P_0}$	$\frac{\partial \ln \left[\frac{\int_0^T A(t)e^{-\delta t}dt}{\int_0^T e^{-\delta t}dt}\right]}{\partial \ln P_0}$	$\frac{\partial \ln R}{\partial \ln P_0}$
p_0	0.20	+0.28	-1.12	+0.07	-1.44	+0.51
μ_0	0.7%	-0.25	+2.19	+0.10	+0.96	-0.52
φ	1.0	+0.20	+1.70	+0.06	+1.78	-0.52
d_0	5%	-0.05	+3.07	-0.54	-1.27	+0.68
β	3%	+0.04	-2.63	-0.01	-0.08	-0.17
δ	3%	+0.03	+0.01	+0.01	-0.16	+0.09
ρ	1.32	+0.01	-0.32	+0.01	+0.04	+0.06
ς	0.85	+0.42	-17.66	-0.42	-0.92	+0.00
k	0.81	+0.06	+0.81	+0.02	+0.23	+0.26
w_0	20k$	+0.55	+0.83	+0.04	+0.19	+0.00
α	15%	-0.10	-0.07	+0.01	-0.31	-0.15
H_0	30k$[$\mu$]	+0.45	-1.27	+0.96	+0.81	+0.00

Elasticities greater than one indicate that the model is very sensitive to the particular parameter. Most noticeable is the parameter ς describing the relative "share" of consumption versus health in the utility function. A one percent change in ς decreases life-time health investment by nearly 18 percent. It should be noted though that the results in Table 3.1 are only valid for the particular parameter region close to the model calibration and that sensitivities will be different for different model calibrations.

3.7 Discussion

We have formulated a stylized structural model of consumption, leisure, health, health investment, wealth accumulation and retirement decisions using the human capital framework of health. Specification of a functional form for the utility function and of initial conditions allows us to derive analytic solutions for consumption, health, health investment and wealth, conditional on a given retirement age.

We find that initial conditions are likely of importance and that health will under most circumstances not evolve as the "optimal" health stock $H_*(t)$. An essential characteristic

of the model is that health cannot deteriorate faster than the natural deterioration rate $d(t)$. As a result initial health cannot dissipate rapidly, nor is there any reason to expect the endowment of health $H(0)$ to exactly equal the "optimal" health stock $H_*(0)$ (see also Wolfe, 1985). Wolfe (1985) assumes an initial surplus of health on the grounds that *"... the human species, with its goal of self-preservation, confronts a different problem than the individual who seeks to maximize utility. The evolutionary solution to the former may entail an excessive health endowment in the sense that an individual might prefer to have less health and to be compensated with wealth in a more liquid form ..."* As Wolfe more or less suggests, humans may have been endowed with "excessive" health as a result of our evolutionary history which required good physical condition to hunt and gather food, defend ourselves, survive periods of hunger, etc. Today's demands on human's physical condition are essentially based on the utility of good health and on economic productivity, which in an increasingly knowledge-intensive environment may be significantly smaller than in pre-historic times.

While Wolfe (1985) provides a convincing argument that high initial health endowments are plausible, we simply assume that initial health $H(0)$ can take any positive value (including values below the "optimal" health stock $H_*(0)$). Exploring corner solutions, in which individuals do not invest in medical care ($m(t) = 0$) for periods of time, we find that what is referred to in the literature as the "optimal" health stock (e.g., Grossman, 1972, 2000) should, given initial condition $H(0)$, not be interpreted as an optimal solution but rather as a health threshold (given by the "minimally productive" level of health). Healthy individuals (whose health is above the threshold) do not invest in health, while unhealthy individuals (whose health is at or below the threshold) do. The threshold is the minimum health level individuals "demand" for the productivity benefits and utility that good health provides.[17]

In a review of the empirical literature Galama and Kapteyn (2009; see Chapter 2) find that the interpretation advocated here provides a better explanation for the observed evolution of health and of medical consumption. Importantly, our interpretation can explain the observation that measures of medical care are negatively correlated with measures of

[17]Wolfe (1985), to the best of our knowledge, is the only researcher who has attempted to explore the consequences of corner solutions in some detail. His model and interpretation is however substantially different from ours. Wolfe employs a simplified Grossman model where health does not provide utility. Further, Wolfe interprets the onset of *"... a discontinuous mid-life increase in health investment ..."* with retirement. We however do not associate the discontinuous increase in health investment with retirement but with becoming unhealthy (health levels at or below the health threshold leading to health investment). Retirement in our model is the result of life-time utility maximization.

health while the traditional interpretation cannot (see, e.g., Zweifel and Breyer, 1997, and references therein).

We employ the model to investigate the optimal age of retirement by maximizing the implied indirect utility function with respect to the retirement age. In the model individuals find retirement increasingly attractive as they age as a result of three effects: (1) wage declines as a result of gradual health deterioration reducing income from work with age, (2) increased leisure time during retirement and (3) accumulation of pension wealth (which can only be consumed after retirement) with years in the workforce.

Our model of health and retirement is an improvement over the model presented by Wolfe (1985) in which retirement is defined as the time when an individual begins to invest in health (i.e., when health has deteriorated to the level of the health threshold). We, however, allow the retirement decision to not only be determined by the timing of health investment, but also by wage declines as a result of gradual health deterioration (reducing income from work with age), increased leisure time during retirement and the accumulation of pension wealth with years in the workforce (including the detailed pension structure).

The model can reproduce the observation that the retirement age has continued to fall while retirees point to deteriorating health as an important reason for early retirement at the same time that population health and mortality have continued to improve in the developed world. If advances in population health are largely the result of better nutrition, preventative medicine (through, e.g., vaccination and other means), and better (less taxing) living, working and schooling environments then the overall health endowment $H(0)$ of the population increases and/or the health deterioration rate $d(t)$ decreases. Both effects result in earlier retirement.[18,19] Workers with higher earnings (say white collar workers) invest more in health and because they stay healthier retire later than those with lower earnings (say blue collar workers) whose health deteriorates faster. In other words, health is an important determinant of early retirement. Indeed Dwyer and Mitchell (1999) find that men in poor overall health are expected to retire one to two years earlier, an effect that persists after the authors correct for potential endogeneity of self-rated health problems.

We find that higher income (base wage rate $w_0(t)$) increases the retirement age, while greater wealth (initial assets $A(0)$) and greater pension wealth (base pension benefit b_0 and

[18]If on the other hand advances in medical care or other advances increase the efficiency or lower the cost of health investment then retirement will be postponed.

[19]This prediction crucially depends on the assumption that a significant share of the population has health levels above the health threshold, i.e., that corner solutions are fairly common.

a higher fraction α of wages saved for retirement) decreases the retirement age. Advances in population wealth levels, but not income, could provide an alternative explanation for decreasing retirement ages.

Further, we can explain differences in the observed health deterioration rates between blue and white collar workers by differences in their health thresholds (their minimally productive level of health) and their resulting differences in health investment. We do not need to resort to physical effort or work-type related health effects (e.g., as in Case and Deaton, 2005). Even though we do not find it unreasonable to assume that certain types of jobs result in higher health deterioration rates, we do offer that poorer individuals also invest less in health as their health thresholds (minimally productive levels of health) are lower than for richer individuals.

Our model is nevertheless not without problems. A number of problematic features can be attributed to the standard assumption in the literature spawned by Grossman of constant returns to scale in health investment (an exception is Ehrlich and Chuma, 1990). This leads to a "bang-bang" solution in which the level of health investment is undetermined (e.g., Ehrlich and Chuma, 1990; Wolfe, 1985). And it requires one to assume that individuals are capable of adjusting their health to the "optimal" level instantaneously and without adjustment costs. Grossman (2000) is *"... willing to assume that consumers reach their desired stocks instantaneously in order to get sharp predictions that are subject to empirical testing ..."* . But, because of the degenerate nature of the solutions, the resulting model predictions seem caricatures of real life. For example, in the corner solutions that we have introduced in this work, healthy individuals do not invest in health at all $m(t) = 0$ for periods of time, while in reality most people see the doctor at least once per year. Further, the solution for the "optimal" health stock in the literature spawned by Grossman, is a function of current prices, wages etc (myopic). This is in direct contradiction with the relation (3.4) which suggests that health depends on the initial health stock $H(0)$ and on the subsequent history of health investments (and hence prices, wages etc) made. In keeping with the literature and to allow for comparison with prior work we have assumed constant returns to scale in health investment. Nevertheless there seems to be room for further theoretical extensions in the demand for health literature. Introducing diminishing returns to scale in health investment may be one potential avenue to pursue. Another may be the introduction of some form of adjustment costs.

3.8 Appendix

3.8.1 First-order conditions

The objective function (3.1) is maximized subject to the constraints (3.2). Health can be solved as in (3.4).

We have

$$\dot{p}_A(t) = -\frac{\partial \Im}{\partial A(t)} = -p_A(t)\delta, \tag{3.25}$$

the solution of which is

$$p_A(t) = p_A(0)e^{-\delta t}; \tag{3.26}$$

further $q(t) \geq 0$ and $q(t) = 0$ for $m(t) > 0$.

We now introduce Ł, the integral over time of the Lagrangian $\Im$ (equation 3.5).

$$\begin{aligned} Ł &= \int_0^T \Im dt \\ &= \int_0^R U_w[C(t), H(t)]e^{-\beta t}dt + \int_R^T U_r[C(t), H(t)]e^{-\beta t}dt \\ &+ p_A(0)\int_0^T \{\delta A(t) + Y[H(t)] - C(t) - p(t)m(t)\}e^{-\delta t}dt \\ &+ \int_0^T q(t)m(t)dt. \end{aligned} \tag{3.27}$$

Maximizing Ł with respect to consumption $C(t')$ results in the following first order conditions:

$$\frac{\partial U_w(t')}{\partial C(t')} = p_A(0)e^{(\beta-\delta)t'} \quad t' \leq R \tag{3.28}$$

$$\frac{\partial U_r(t')}{\partial C(t')} = p_A(0)e^{(\beta-\delta)t'} \quad t' > R, \tag{3.29}$$

where we have used

$$\frac{\partial C(t)}{\partial C(t')} = \delta(t - t'), \tag{3.30}$$

where $\delta(t-t')$ is the Dirac delta function. The Dirac delta function is the continuous equivalent of the discrete Kronecker delta function. It has the property $\int_\Omega f(t)\delta(t-t')dt = f(t')\,(t' \in \Omega)$ and can informally be thought of as a function $\delta(x)$ that has the value of infinity for $x = 0$, the value zero elsewhere and has an area of 1 (normalized).

Using the functional form (3.9) of the utility function allows us to write the first order conditions with respect to consumption $C(t')$ (equations 3.28 and 3.29) as follows:

$$\frac{\partial U_w(t')}{\partial C(t')} = \varsigma C(t')^{\varsigma-\rho\varsigma-1} H(t')^{1-\varsigma-\rho+\rho\varsigma} = p_A(0)e^{(\beta-\delta)t'} \quad t' \leq R \tag{3.31}$$

$$\frac{\partial U_r(t')}{\partial C(t')} = k\varsigma C(t')^{\varsigma-\rho\varsigma-1} H(t')^{1-\varsigma-\rho+\rho\varsigma} = p_A(0)e^{(\beta-\delta)t'} \quad t' > R. \tag{3.32}$$

Before we continue with the first order conditions with respect to health investment $m(t')$, it is useful to look at the derivative of $H(t)$ (see equation 3.4) with respect to $m(t')$:

$$\frac{\partial H(t)}{\partial m(t')} = \mu(t')e^{-\int_{t'}^{t} d(s)ds} \quad t' \leq t \tag{3.33}$$

$$\frac{\partial H(t)}{\partial m(t')} = 0 \quad t' > t, \tag{3.34}$$

where once more we have used that

$$\frac{\partial m(t)}{\partial m(t')} = \delta(t - t'). \tag{3.35}$$

Maximizing Ł with respect to health investment $m(t')$ leads to

$$\begin{aligned}
\int_{t'}^{R} \frac{\partial U_w(t)}{\partial m(t')} e^{-\beta t} dt + \int_{R}^{T} \frac{\partial U_r(t)}{\partial m(t')} e^{-\beta t} dt &= \\
\int_{t'}^{R} \frac{\partial U_w(t)}{\partial H(t)} \mu(t') e^{-\int_{t'}^{t} d(s)ds} e^{-\beta t} dt &+ \\
\int_{R}^{T} \frac{\partial U_r(t)}{\partial H(t)} \mu(t') e^{-\int_{t'}^{t} d(s)ds} e^{-\beta t} dt &= \\
p_A(0)p(t')e^{-\delta t'} - p_A(0) \int_{t'}^{T} \frac{\partial Y(t)}{\partial m(t')} e^{-\delta t} dt - q(t') \quad t' \leq R
\end{aligned} \tag{3.36}$$

$$\begin{aligned}
\int_{t'}^{T} \frac{\partial U_r(t)}{\partial m(t')} e^{-\beta t} dt &= \\
\int_{t'}^{T} \frac{\partial U_r(t)}{\partial H(t)} \mu(t') e^{-\int_{t'}^{t} d(s)ds} e^{-\beta t} dt &= \\
p_A(0)p(t')e^{-\delta t'} - p_A(0) \int_{t'}^{T} \frac{\partial Y(t)}{\partial m(t')} e^{-\delta t} dt - q(t') \quad t' > R,
\end{aligned} \tag{3.37}$$

where the lower integration limit t' reflects the fact that the stock of health (which utility and wages are functions of) is a function of past but not future health investment.

Using once more the functional form (3.9) of the utility function, using the Leibniz Integral Rule to differentiate equations (3.36) and (3.37) with respect to t' and substituting the result back into equations (3.36) and (3.37) we find:

$$\begin{aligned}
\frac{\partial U_w(t')}{\partial H(t')} &= (1-\varsigma)C(t')^{\varsigma-\rho\varsigma}H(t')^{-\varsigma+\rho\varsigma-\rho} \\
&= p_A(0)\left[\pi_H(t') - \varphi(t')\right]e^{(\beta-\delta)t'} \\
&- \frac{e^{\beta t'}}{\mu(t')}\left[\frac{\dot{\mu}(t')}{\mu(t')} + d(t')\right]q(t') + \dot{q}(t')\frac{e^{\beta t'}}{\mu(t')} \\
&= \mathbb{A}p_A(0)e^{(\beta-\delta)t'} + \mathbb{B} \qquad (t' \leq R) \qquad (3.38) \\
\frac{\partial U_r(t')}{\partial H(t')} &= k(1-\varsigma)C(t')^{\varsigma-\rho\varsigma}H(t')^{-\varsigma+\rho\varsigma-\rho} \\
&= p_A(0)\pi_H(t')e^{(\beta-\delta)t'} \\
&- \frac{e^{\beta t'}}{\mu(t')}\left[\frac{\dot{\mu}(t')}{\mu(t')} + d(t')\right]q(t') + \dot{q}(t')\frac{e^{\beta t'}}{\mu(t')} \\
&= \mathbb{A}'p_A(0)e^{(\beta-\delta)t'} + \mathbb{B} \qquad (t' > R), \qquad (3.39)
\end{aligned}$$

where $\pi_H(t')$ is the user cost of health capital at the margin (equation 3.8) and the definitions for $\mathbb{A}$, $\mathbb{B}$, and $\mathbb{A}'$ follow directly from equations (3.38) and (3.39).

3.8.2 Solutions for health, consumption and health investment

Solving the first order conditions (equations 3.31, 3.32, 3.38 and 3.39) we find

$$C(t) = H(t)\left\{\frac{\zeta}{1-\zeta}\left[\mathbb{A} + \frac{\mathbb{B}e^{-(\beta-\delta)t}}{p_A(0)}\right]\right\} \quad t \leq R \qquad (3.40)$$

$$C(t) = H(t)\left\{\frac{\zeta}{1-\zeta}\left[\mathbb{A}' + \frac{\mathbb{B}e^{-(\beta-\delta)t}}{p_A(0)}\right]\right\} \quad t > R, \qquad (3.41)$$

and the following solutions for $C(t)$ and $H(t)$:

$$C(t) = \zeta\Lambda\left[\mathbb{A} + \frac{\mathbb{B}}{p_A(0)}e^{-(\beta-\delta)t}\right]^{1-\chi} e^{-\left(\frac{\beta-\delta}{\rho}\right)t} \quad t \leq R \qquad (3.42)$$

$$C(t) = k^{1/\rho}\zeta\Lambda\left[\mathbb{A}' + \frac{\mathbb{B}}{p_A(0)}e^{-(\beta-\delta)t}\right]^{1-\chi} e^{-\left(\frac{\beta-\delta}{\rho}\right)t} \quad t > R \qquad (3.43)$$

$$H(t) = (1-\zeta)\Lambda\left[\mathbb{A} + \frac{\mathbb{B}}{p_A(0)}e^{-(\beta-\delta)t}\right]^{-\chi} e^{-\left(\frac{\beta-\delta}{\rho}\right)t} \quad t \leq R \qquad (3.44)$$

$$H(t) = k^{1/\rho}(1-\zeta)\Lambda\left[\mathbb{A}' + \frac{\mathbb{B}}{p_A(0)}e^{-(\beta-\delta)t}\right]^{-\chi} e^{-\left(\frac{\beta-\delta}{\rho}\right)t} \quad t > R, \qquad (3.45)$$

where once more we have used the definitions for χ (equation 3.14) and for Λ (equation 3.15).

Using equation (3.2) one can then solve for health investment $m(t)$:

$$\begin{aligned} m(t) &= \frac{1}{\mu(t)}(1-\zeta)e^{-\int_0^t d(s)ds} \\ &\times \frac{\partial}{\partial t}\left\{\Lambda\left[\mathbb{A}+\frac{\mathbb{B}}{p_A(0)}e^{-(\beta-\delta)t}\right]^{-\chi} e^{-\left(\frac{\beta-\delta}{\rho}\right)t} e^{\int_0^t d(s)ds}\right\} \quad t \leq R \qquad (3.46) \\ m(t) &= \frac{1}{\mu(t)}k^{1/\rho}(1-\zeta)e^{-\int_R^t d(s)ds} \\ &\times \frac{\partial}{\partial t}\left\{\Lambda\left[\mathbb{A}'+\frac{\mathbb{B}}{p_A(0)}e^{-(\beta-\delta)t}\right]^{-\chi} e^{-\left(\frac{\beta-\delta}{\rho}\right)t} e^{\int_R^t d(s)ds}\right\} \quad t > R. \qquad (3.47) \end{aligned}$$

With solutions for the control functions consumption $C(t)$ and health investment $m(t)$, and for the state variable health $H(t)$ we can find the solution for the state variable assets $A(t)$ using equations (3.20) and (3.20).

For positive health investment $m(t) > 0$ we have $q(t) = 0$ and $H(t) = H_*(t)$ and therefore $\mathbb{B} = 0$. These are the solutions for the health threshold (see equations 3.10, 3.11, 3.12, 3.13, 3.17 and 3.18). On the other hand, for initial conditions $H(0)$ and $H(R_+)$ that are above the health threshold (the minimally productive health level) $H_*(0)$ and $H_*(R_+)$ (see Figure 3.1 scenarios A through F) we have a situation of "excessive" initial health, i.e., the individual is endowed with an initial stock of health that is greater than the level required to be economically productive. In such cases individuals would want to "sell" their health, i.e., chose negative health investment $m(t) < 0$. Since this is not possible (health investment is a positive quantity) we have a corner solution where $m(t) = 0$. We can derive the solutions for consumption $C(t)$ and health $H(t)$ by imposing $m(t) = 0$. We then find a differential equation in $q(t)$ with the following solutions:

$$\begin{aligned} q(t) &= p_A(0)\int_0^t \mu(x)\left[H(0)e^{\left(\frac{\beta-\delta}{\rho}\right)x} e^{-\int_0^x d(s)ds}\frac{1}{\Lambda(1-\zeta)}\right]^{\frac{-1}{\chi}} e^{\int_x^t\left[\frac{\dot{\mu}(s)}{\mu(s)}+d(s)\right]ds} e^{-\delta x}dx \\ &- p_A(0)\int_0^t \mu(x)\left[\pi_H(x)-\varphi(x)\right] e^{\int_x^t\left[\frac{\dot{\mu}(s)}{\mu(s)}+d(s)\right]ds} e^{-\delta x}dx \\ &+ q(0)e^{\int_0^t\left[\frac{\dot{\mu}(s)}{\mu(s)}+d(s)\right]ds}, \qquad (t \leq R) \qquad (3.48) \end{aligned}$$

$$
\begin{aligned}
q(t) &= p_A(0)\int_R^t \mu(x)\left[H(R)e^{\left(\frac{\beta-\delta}{\rho}\right)x}e^{-\int_R^x d(s)ds}\frac{1}{k^{1/\rho}\Lambda(1-\zeta)}\right]^{\frac{-1}{\chi}} e^{\int_x^t\left[\frac{\dot{\mu}(s)}{\mu(s)}+d(s)\right]ds}e^{-\delta x}dx \\
&- p_A(0)\int_R^t \mu(x)\pi_H(x)e^{\int_x^t\left[\frac{\dot{\mu}(s)}{\mu(s)}+d(s)\right]ds}e^{-\delta x}dx \\
&+ q(R)e^{\int_R^t\left[\frac{\dot{\mu}(s)}{\mu(s)}+d(s)\right]ds}. \qquad (t>R) \qquad (3.49)
\end{aligned}
$$

Substituting the above solutions for $q(t)$ into those for consumption $C(t)$ (equations 3.42 and 3.43), health $H(t)$ (equations 3.44 and 3.45) and health investment $m(t)$ (equations 3.46 and 3.47), we find:

$$
\begin{aligned}
H(t) &= H(0)e^{-\int_0^t d(s)ds} \quad (t\leq R) && (3.50)\\
H(t) &= H(R)e^{-\int_R^t d(s)ds} \quad (t>R) && (3.51)\\
C(t) &= \zeta\Lambda^{1/\chi}(1-\zeta)^{\frac{1-\chi}{\chi}}H(0)^{-\left(\frac{1-\chi}{\chi}\right)}e^{\left(\frac{1-\chi}{\chi}\right)\int_0^t d(s)ds}e^{-\left(\frac{\beta-\delta}{\rho\chi}\right)t} \quad (t\leq R) && (3.52)\\
C(t) &= k^{1/\chi\rho}\zeta\Lambda^{1/\chi}(1-\zeta)^{\frac{1-\chi}{\chi}}H(R)^{-\left(\frac{1-\chi}{\chi}\right)}e^{\left(\frac{1-\chi}{\chi}\right)\int_R^t d(s)ds}e^{-\left(\frac{\beta-\delta}{\rho\chi}\right)t} \quad (t>R) && (3.53)\\
m(t) &= 0. && (3.54)
\end{aligned}
$$

A perhaps more intuitive way of arriving at the same result is by simply substituting $m(t)=\xi(t)^2$ and solving the optimization problem for the control variables $\xi(t)$ (instead of $m(t)$) and consumption $C(t)$ (i.e., one then does not have to resort to using the multiplier $q(t)$ associated with the condition that health investment $m(t)\geq 0$ in the Lagrangian 3.5). One then finds the same first order conditions for maximization with respect to consumption (equations 3.28 and 3.29). For the first order conditions for maximization with respect to $\xi(t)$ one finds that either $\xi(t)=0$ (and hence $m(t)=0$) or that the first order conditions conditions equations (3.36 and 3.37) are valid for $q(t)=0$ ($\mathbb{B}=0$).

We now have the material to solve the solutions for each of the scenarios A through F (see Figure 3.1) in detail.

3.8.3 Scenario A

Scenario A: $0\leq t\leq t_1$

Figure 3.1 shows how in scenario A initial health $H(0)$ is above the initial health threshold $H_*(0)$ and individuals do not invest in health $m(t)=0$. As a result health deteriorates with rate $d(t)$ untill age t_1 when health reaches the health threshold $H_*(t_1)$. We have the following condition $[H(t_1)=H_*(t_1)]$:

$$H(0)e^{-\int_0^{t_1} d(s)ds} = (1-\zeta)\Lambda_A \left[\pi_H(t_1) - \varphi(t_1)\right]^{-\chi} e^{-\left(\frac{\beta-\delta}{\rho}\right)t_1}, \tag{3.55}$$

and the following solutions for consumption $C(t)$, health $H(t)$ and health investment $m(t)$:

$$\begin{aligned}
H(t) &= H(0)e^{-\int_0^t d(s)ds} && (3.56)\\
&= (1-\zeta)\Lambda_A \left[\pi_H(t_1) - \varphi(t_1)\right]^{-\chi} e^{-\left(\frac{\beta-\delta}{\rho}\right)t_1} e^{\int_t^{t_1} d(s)ds} && (3.57)\\
C(t) &= \zeta \Lambda_A^{1/\chi} (1-\zeta)^{\frac{1-\chi}{\chi}} H(0)^{-\left(\frac{1-\chi}{\chi}\right)} e^{\left(\frac{1-\chi}{\chi}\right)\int_0^t d(s)ds} e^{-\left(\frac{\beta-\delta}{\rho\chi}\right)t} && (3.58)\\
&= \zeta \Lambda_A \left[\pi_H(t_1) - \varphi(t_1)\right]^{1-\chi} e^{\left(\frac{1-\chi}{\chi}\right)\left(\frac{\beta-\delta}{\rho}\right)t_1} e^{-\left(\frac{1-\chi}{\chi}\right)\int_t^{t_1} d(s)ds} e^{-\left(\frac{\beta-\delta}{\rho\chi}\right)t} && (3.59)\\
m(t) &= 0. && (3.60)
\end{aligned}$$

Scenario A: $t_1 < t \leq R$

Between the age t_1 and retirement R individuals invest in health $m(t) > 0$ and follow the health threshold (the minimally productive health path): $H_*(t)$, $C_*(t)$, and $m_*(t)$.

$$\begin{aligned}
H_*(t) &= (1-\zeta)\Lambda_A \left[\pi_H(t) - \varphi(t)\right]^{-\chi} e^{-\left(\frac{\beta-\delta}{\rho}\right)t}, && (3.61)\\
C_*(t) &= \zeta \Lambda_A \left[\pi_H(t) - \varphi(t)\right]^{1-\chi} e^{-\left(\frac{\beta-\delta}{\rho}\right)t}, && (3.62)\\
m_*(t) &= \frac{1}{\mu(t)} e^{-\int_0^t d(s)ds} \frac{d}{dt}\left((1-\zeta)\Lambda_A \left[\pi_H(t) - \varphi(t)\right]^{-\chi} e^{-\left(\frac{\beta-\delta}{\rho}\right)t} e^{\int_0^t d(s)ds}\right). && (3.63)
\end{aligned}$$

Scenario A: $R < t \leq t_2$

At retirement the health threshold drops to $H_*(R_+)$ and once more individuals do not invest in health ($m(t) = 0$) till age t_2 when health reaches the health threshold $H_*(t_2)$. We have the following condition [$H(t_2) = H_*(R_-)e^{-\int_R^{t_2} d(s)ds} = H_*(t_2)$]:

$$\begin{aligned}
&(1-\zeta)\Lambda_A \left[\pi_H(R) - \varphi(R)\right]^{-\chi} e^{-\left(\frac{\beta-\delta}{\rho}\right)R} e^{-\int_R^{t_2} d(s)ds}\\
=\ & k^{1/\rho}(1-\zeta)\Lambda_A \left[\pi_H(t_2)\right]^{-\chi} e^{-\left(\frac{\beta-\delta}{\rho}\right)t_2}, && (3.64)
\end{aligned}$$

and the following solutions for consumption $C(t)$, health $H(t)$ and health investment $m(t)$:

$$
\begin{aligned}
H(t) &= H_*(R_-)e^{-\int_R^t d(s)ds} && (3.65)\\
&= (1-\zeta)\Lambda_A\left[\pi_H(R)-\varphi(R)\right]^{-\chi} e^{-\left(\frac{\beta-\delta}{\rho}\right)R} e^{-\int_R^t d(s)ds} && (3.66)\\
C(t) &= k^{1/\rho\chi}\zeta\Lambda_A^{1/\chi}(1-\zeta)^{\frac{1-\chi}{\chi}} H_*(R_-)^{-\left(\frac{1-\chi}{\chi}\right)} e^{\left(\frac{1-\chi}{\chi}\right)\int_R^t d(s)ds} e^{-\left(\frac{\beta-\delta}{\rho\chi}\right)t} && (3.67)\\
&= k^{1/\rho\chi}\zeta\Lambda_A\left[\pi_H(R)-\varphi(R)\right]^{1-\chi} e^{\left(\frac{1-\chi}{\chi}\right)\left(\frac{\beta-\delta}{\rho}\right)R} e^{\left(\frac{1-\chi}{\chi}\right)\int_R^t d(s)ds} e^{-\left(\frac{\beta-\delta}{\rho\chi}\right)t} && (3.68)\\
m(t) &= 0. && (3.69)
\end{aligned}
$$

Scenario A: $t_2 < t \leq T$

Between the age t_2 and the end of life T individuals invest once again in health ($m(t) > 0$) and follow the health threshold (the minimally productive health path): $H_*(t)$, $C_*(t)$, and $m_*(t)$.

$$
\begin{aligned}
H_*(t) &= k^{1/\rho}(1-\zeta)\Lambda_A\left[\pi_H(t)\right]^{-\chi} e^{-\left(\frac{\beta-\delta}{\rho}\right)t}, && (3.70)\\
C_*(t) &= k^{1/\rho}\zeta\Lambda_A\left[\pi_H(t)\right]^{1-\chi} e^{-\left(\frac{\beta-\delta}{\rho}\right)t}, && (3.71)\\
m_*(t) &= k^{1/\rho}\frac{1}{\mu(t)} e^{-\int_0^t d(s)ds}\frac{d}{dt}\left((1-\zeta)\Lambda_A\left[\pi_H(t)\right]^{-\chi} e^{-\left(\frac{\beta-\delta}{\rho}\right)t} e^{\int_0^t d(s)ds}\right). && (3.72)
\end{aligned}
$$

Scenario A: determination of Λ_A

Using the life-time budget constraint (3.3) and substituting the solutions for health $H(t)$, consumption $C(t)$ and health investment $m(t)$ we can determine the constant Λ_A. Define:

$$\Lambda_A \equiv \frac{\Lambda_{An}}{\Lambda_{Ad}}, \tag{3.73}$$

where Λ_{An} is the numerator and Λ_{Ad} is the denominator of Λ_A. We find:

$$\begin{aligned}
\Lambda_{An} &= A(0) - A(T)e^{-\delta T} + \int_0^R w_0(x)e^{-\delta x}dx + \int_R^T b(x)e^{-\delta x}dx \\
&+ H(0)\int_0^{t_1} \varphi(x)e^{-\int_0^x d(s)ds}e^{-\delta x}dx \qquad (3.74) \\
\Lambda_{Ad} &= \int_{t_1}^R [\pi_H(x) - \varphi(x)]^{1-\chi} e^{-\kappa x}dx + k^{1/\rho}\int_{t_2}^T [\pi_H(x)]^{1-\chi} e^{-\kappa x}dx \\
&+ \zeta [\pi_H(t_1) - \varphi(t_1)]^{1-\chi} e^{\left(\frac{\beta-\delta}{\rho}\right)\left(\frac{1-\chi}{\chi}\right)t_1}\int_0^{t_1} e^{-\left(\frac{1-\chi}{\chi}\right)\int_x^{t_1} d(s)ds}e^{-\left(\frac{\beta-\delta}{\rho\chi}\right)x}e^{-\delta x}dx \\
&+ \zeta k^{1/\rho\chi} [\pi_H(R) - \varphi(R)]^{1-\chi} e^{\left(\frac{\beta-\delta}{\rho}\right)\left(\frac{1-\chi}{\chi}\right)R}\int_R^{t_2} e^{\left(\frac{1-\chi}{\chi}\right)\int_R^x d(s)ds}e^{-\left(\frac{\beta-\delta}{\rho\chi}\right)x}e^{-\delta x}dx \\
&+ (1-\zeta)\frac{p(R)}{\mu(R)} [\pi_H(R) - \varphi(R)]^{-\chi} e^{-\kappa R} - (1-\zeta)\frac{p(t_1)}{\mu(t_1)} [\pi_H(t_1) - \varphi(t_1)]^{-\chi} e^{-\kappa t_1} \\
&+ (1-\zeta)k^{1/\rho}\frac{p(T)}{\mu(T} [\pi_H(T)]^{-\chi} e^{-\kappa T} - (1-\zeta)k^{1/\rho}\frac{p(t_2)}{\mu(t_2)} [\pi_H(t_2)]^{-\chi} e^{-\kappa t_2}, \qquad (3.75)
\end{aligned}$$

where we have used the following definition:

$$\kappa \equiv \frac{\delta\rho + \beta - \delta}{\rho}. \qquad (3.76)$$

3.8.4 Scenario B

Scenario B: $0 \leq t \leq t_1$

Figure 3.1 shows how similar to scenario A initial health $H(0)$ is above the health threshold $H_*(0)$ and individuals do not invest in health $m(t) = 0$. As a result health deteriorates with rate $d(t)$ untill age t_1 when health reaches the health threshold $H_*(t_1)$. The same condition $[H(t_1) = H_*(t_1)]$ holds as in scenario A (equation 3.55; replace Λ_A with Λ_B). Also the solutions for consumption $C(t)$, health $H(t)$ and health investment $m(t)$ are the same as in scenario A (3.56, 3.57, 3.58, 3.59, and 3.60; replace Λ_A with Λ_B).

Scenario B: $t_1 < t \leq R$

As in scenario A, between the age t_1 and retirement R individuals invest in health $m(t) > 0$ and follow the health threshold (the minimally productive health path): $H_*(t)$, $C_*(t)$, and $m_*(t)$ (see equations 3.61, 3.62, and 3.63; replace Λ_A with Λ_B).

Scenario A: $R < t \leq T$

As in scenario A, at retirement the health threshold drops to $H_*(R_+)$ and once more individuals do not invest in health ($m(t) = 0$). In scenario B (unlike in scenario A)

health, after the retirement age R, does not deteriorate to the health threshold level $H_*(t)$ before the end of life T. The solutions for consumption $C(t)$, health $H(t)$ and health investment $m(t)$ are given by equations 3.65, 3.66, 3.67, 3.68, and 3.69 (replace Λ_A with Λ_B) and are valid for $R < t \leq T$.

Scenario B: determination of Λ_B

Defining

$$\Lambda_B \equiv \frac{\Lambda_{Bn}}{\Lambda_{Bd}}, \tag{3.77}$$

where Λ_{Bn} is the numerator and Λ_{Bd} is the denominator of Λ_B, we find:

$$\begin{aligned}
\Lambda_{Bn} &= A(0) - A(T)e^{-\delta T} + \int_0^R w_0(x)e^{-\delta x}dx + \int_R^T b(x)e^{-\delta x}dx \\
&+ H(0)\int_0^{t_1} \varphi(x)e^{-\int_0^x d(s)ds}e^{-\delta x}dx
\end{aligned} \tag{3.78}$$

$$\begin{aligned}
\Lambda_{Bd} &= \int_{t_1}^R \left[\pi_H(x) - \varphi(x)\right]^{1-\chi} e^{-\kappa x}dx \\
&+ \zeta\left[\pi_H(t_1) - \varphi(t_1)\right]^{1-\chi} e^{\left(\frac{\beta-\delta}{\rho}\right)\left(\frac{1-\chi}{\chi}\right)t_1}\int_0^{t_1} e^{-\left(\frac{1-\chi}{\chi}\right)\int_x^{t_1} d(s)ds}e^{-\left(\frac{\beta-\delta}{\rho\chi}\right)x}e^{-\delta x}dx \\
&+ \zeta k^{1/\rho\chi}\left[\pi_H(R) - \varphi(R)\right]^{1-\chi} e^{\left(\frac{\beta-\delta}{\rho}\right)\left(\frac{1-\chi}{\chi}\right)R}\int_R^T e^{\left(\frac{1-\chi}{\chi}\right)\int_R^x d(s)ds}e^{-\left(\frac{\beta-\delta}{\rho\chi}\right)x}e^{-\delta x}dx \\
&+ (1-\zeta)\frac{p(R)}{\mu(R)}\left[\pi_H(R) - \varphi(R)\right]^{-\chi} e^{-\kappa R} \\
&- (1-\zeta)\frac{p(t_1)}{\mu(t_1)}\left[\pi_H(t_1) - \varphi(t_1)\right]^{-\chi} e^{-\kappa t_1}.
\end{aligned} \tag{3.79}$$

3.8.5 Scenario C

Scenario C: $0 \leq t \leq R$

Figure 3.1 shows how similar to scenarios A and B initial health $H(0)$ is above the initial health threshold $H_*(0)$ and individuals do not invest in health $m(t) = 0$. But unlike scenarios A and B, health reaches the health threshold $H_*(t_2)$ only at age t_2, after the retirement age R. Individuals thus only invest in health during retirement and not during working life. A similar condition $[H(t_2) = H_*(t_2)]$ holds as in scenarios A and B (equation 3.55). We have:

$$H(0)e^{-\int_0^{t_2} d(s)ds} = k^{1/\rho}(1-\zeta)\Lambda_C\left[\pi_H(t_2)\right]^{-\chi} e^{-\left(\frac{\beta-\delta}{\rho}\right)t_2}, \tag{3.80}$$

and the following solutions for consumption $C(t)$, health $H(t)$ and health investment $m(t)$:

$$\begin{aligned} H(t) &= H(0)e^{-\int_0^t d(s)ds} && (3.81)\\ &= k^{1/\rho}(1-\zeta)\Lambda_C\left[\pi_H(t_2)\right]^{-\chi} e^{-\left(\frac{\beta-\delta}{\rho}\right)t_2} e^{\int_t^{t_2} d(s)ds} && (3.82)\\ C(t) &= \zeta\Lambda_C^{1/\chi}(1-\zeta)^{\frac{1-\chi}{\chi}} H(0)^{-\left(\frac{1-\chi}{\chi}\right)} e^{\left(\frac{1-\chi}{\chi}\right)\int_0^t d(s)ds} e^{-\left(\frac{\beta-\delta}{\rho\chi}\right)t} && (3.83)\\ &= k^{-\left(\frac{1-\chi}{\rho\chi}\right)}\zeta\Lambda_C\left[\pi_H(t_2)\right]^{1-\chi} e^{\left(\frac{1-\chi}{\chi}\right)\left(\frac{\beta-\delta}{\rho}\right)t_2} e^{-\left(\frac{1-\chi}{\chi}\right)\int_t^{t_2} d(s)ds} e^{-\left(\frac{\beta-\delta}{\rho\chi}\right)t} && (3.84)\\ m(t) &= 0. && (3.85) \end{aligned}$$

Scenario C: $R < t \leq t_2$

The solutions for consumption $C(t)$, health $H(t)$ and health investment $m(t)$ are:

$$\begin{aligned} H(t) &= H(0)e^{-\int_0^t d(s)ds} && (3.86)\\ &= k^{1/\rho}(1-\zeta)\Lambda_C\left[\pi_H(t_2)\right]^{-\chi} e^{-\left(\frac{\beta-\delta}{\rho}\right)t_2} e^{\int_t^{t_2} d(s)ds} && (3.87)\\ C(t) &= k^{1/\rho\chi}\zeta\Lambda_C^{1/\chi}(1-\zeta)^{\frac{1-\chi}{\chi}} H(0)^{-\left(\frac{1-\chi}{\chi}\right)} e^{\left(\frac{1-\chi}{\chi}\right)\int_0^t d(s)ds} e^{-\left(\frac{\beta-\delta}{\rho\chi}\right)t} && (3.88)\\ &= k^{1/\rho}\zeta\Lambda_C\left[\pi_H(t_2)\right]^{1-\chi} e^{\left(\frac{1-\chi}{\chi}\right)\left(\frac{\beta-\delta}{\rho}\right)t_2} e^{-\left(\frac{1-\chi}{\chi}\right)\int_t^{t_2} d(s)ds} e^{-\left(\frac{\beta-\delta}{\rho\chi}\right)t} && (3.89)\\ m(t) &= 0. && (3.90) \end{aligned}$$

Scenario C: $t_2 < t \leq T$

Between the age t_2 and the end of life T individuals invest once again in health ($m(t) > 0$) and follow the health threshold: $H_*(t)$, $C_*(t)$, and $m_*(t)$. The equations are the same as in scenario A (equations 3.70, 3.71, and 3.72; replace Λ_A with Λ_C).

Scenario C: determination of Λ_C

Defining

$$\Lambda_C \equiv \frac{\Lambda_{Cn}}{\Lambda_{Cd}}, \tag{3.91}$$

where Λ_{Cn} is the numerator and Λ_{Cd} is the denominator of Λ_C, we find:

$$\begin{aligned}
\Lambda_{Cn} &= A(0) - A(T)e^{-\delta T} + \int_0^R w_0(x)e^{-\delta x}dx + \int_R^T b(x)e^{-\delta x}dx \\
&+ H(0)\int_0^R \varphi(x)e^{-\int_0^x d(s)ds}e^{-\delta x}dx \qquad (3.92) \\
\Lambda_{Cd} &= k^{1/\rho}\int_{t_2}^T [\pi_H(x)]^{1-\chi}\, e^{-\kappa x}dx \\
&+ k^{-\left(\frac{1-\chi}{\rho\chi}\right)}\zeta\,[\pi_H(t_2)]^{1-\chi}\, e^{\left(\frac{\beta-\delta}{\rho}\right)\left(\frac{1-\chi}{\chi}\right)t_2}\int_0^R e^{-\left(\frac{1-\chi}{\chi}\right)\int_x^{t_2} d(s)ds}e^{-\left(\frac{\beta-\delta}{\rho\chi}\right)x}e^{-\delta x}dx \\
&+ k^{1/\rho}\zeta\,[\pi_H(t_2)]^{1-\chi}\, e^{\left(\frac{\beta-\delta}{\rho}\right)\left(\frac{1-\chi}{\chi}\right)t_2}\int_R^{t_2} e^{-\left(\frac{1-\chi}{\chi}\right)\int_x^{t_2} d(s)ds}e^{-\left(\frac{\beta-\delta}{\rho\chi}\right)x}e^{-\delta x}dx \\
&+ (1-\zeta)k^{1/\rho}\frac{p(T)}{\mu(T}\,[\pi_H(T)]^{-\chi}\, e^{-\kappa T} - (1-\zeta)k^{1/\rho}\frac{p(t_2)}{\mu(t_2)}\,[\pi_H(t_2)]^{-\chi}\, e^{-\kappa t_2}. \qquad (3.93)
\end{aligned}$$

3.8.6 Scenario D

Scenario D: $0 \leq t \leq R$

Figure 3.1 shows how similar to scenarios A, B and C initial health $H(0)$ is above the initial health threshold $H_*(0)$ and individuals do not invest in health $m(t) = 0$. But unlike scenarios A, B and C health never reaches the health threshold $H_*(t)$ at any point during the individual's life time. Individuals are sufficiently endowed with initial health capital that they never need to invest in health during working life nor during retirement.

The solutions for consumption $C(t)$, health $H(t)$ and health investment $m(t)$ are:

$$\begin{aligned}
H(t) &= H(0)e^{-\int_0^t d(s)ds} \qquad (3.94) \\
C(t) &= \zeta\Lambda_D^{1/\chi}(1-\zeta)^{\frac{1-\chi}{\chi}}H(0)^{-\left(\frac{1-\chi}{\chi}\right)}e^{\left(\frac{1-\chi}{\chi}\right)\int_0^t d(s)ds}e^{-\left(\frac{\beta-\delta}{\rho\chi}\right)t} \qquad (3.95) \\
m(t) &= 0. \qquad (3.96)
\end{aligned}$$

Scenario D: $R < t \leq T$

The solutions for consumption $C(t)$, health $H(t)$ and health investment $m(t)$ are:

$$\begin{aligned}
H(t) &= H(0)e^{-\int_0^t d(s)ds} \qquad (3.97) \\
C(t) &= k^{1/\rho\chi}\zeta\Lambda_D^{1/\chi}(1-\zeta)^{\frac{1-\chi}{\chi}}H(0)^{-\left(\frac{1-\chi}{\chi}\right)}e^{\left(\frac{1-\chi}{\chi}\right)\int_0^t d(s)ds}e^{-\left(\frac{\beta-\delta}{\rho\chi}\right)t} \qquad (3.98) \\
m(t) &= 0. \qquad (3.99)
\end{aligned}$$

Scenario D: determination of Λ_D

Defining

$$\Lambda_D \equiv \frac{\Lambda_{Dn}}{\Lambda_{Dd}}, \tag{3.100}$$

where Λ_{Dn} is the numerator and Λ_{Dd} is the denominator of Λ_D, we find:

$$\begin{aligned}
\Lambda_{Dn}^{1/\chi} &= A(0) - A(T)e^{-\delta T} + \int_0^R w_0(x)e^{-\delta x}dx + \int_R^T b(x)e^{-\delta x}dx \\
&+ H(0)\int_0^R \varphi(x)e^{-\int_0^x d(s)ds}e^{-\delta x}dx \qquad (3.101) \\
\Lambda_{Dd}^{1/\chi} &= \zeta(1-\zeta)^{\frac{1-\chi}{\chi}}H(0)^{-\left(\frac{1-\chi}{\chi}\right)}\int_0^R e^{\left(\frac{1-\chi}{\chi}\right)\int_0^x d(s)ds}e^{-\left(\frac{\beta-\delta}{\rho\chi}\right)x}e^{-\delta x}dx \\
&+ \zeta(1-\zeta)^{\frac{1-\chi}{\chi}}H(0)^{-\left(\frac{1-\chi}{\chi}\right)}k^{1/\rho\chi}\int_R^T e^{\left(\frac{1-\chi}{\chi}\right)\int_0^x d(s)ds}e^{-\left(\frac{\beta-\delta}{\rho\chi}\right)x}e^{-\delta x}dx. \qquad (3.102)
\end{aligned}$$

3.8.7 Scenarios E and F

Figure 3.1 shows scenarios E and F. In these scenarios initial health $H(0)$ is below the initial health threshold $H_*(0)$. The simplified Grossman model that we employ here allows for complete repair. Case and Deaton (2005) point out that employing such technology is not realistic. Indeed wealthy individuals may have high health threshold levels and the ability to afford any kind of health investment, but they may not necessary be able to repair all types of poor health (e.g., cancer, aids, various disabilities such as blindness etc). Simply stated, not every illness has a cure. Further, while health in the formulation cannot deteriorate faster than the deterioration rate $d(t)$ there is no intrinsic constraint on the rate at which health can be repaired. As such, in scenarios E and F individuals will seek to repair their health instantaneously when they enter the workforce at age 20 ($t = 0$), effectively substituting initial assets $A(0)$ for improved initial health $H(0)$ such that initial health equals the initial health threshold (the initial minimally productive level of health) $H(0) = H_*(0)$. An alternative interpretation is that individuals invest in health $m(t)$ well before they enter the workforce at age 20 ($t = 0$) to ensure their health is at the initial health threshold $H_*(0)$ at $t = 0$. Before they enter the workforce individuals don't consume yet (or at least consumption is paid for by their parents / caretakers) and have no assets $A(t)$ yet. In this case the end result is the same as if health investment were made in an infinitesimally small period of time at $t = 0$. We assume that individuals pay for the health investment themselves, i.e. they start with lower initial assets $A_*(0) = A(0) - p(0)m_*(0)$, where $m_*(0)$ is the quantity of health investment needed

to arrive from initial health $H(0)$ to the initial health threshold $H_*(0)$. Approximating this initial health investment by a delta function, $m(t) = m_*(0)\delta(t-0)$ (i.e., mathematically investment takes place at $t = 0$ during an infinitesimally small period of time) we find:

$$H_*(0) = H(0) + \mu(0)m_*(0), \tag{3.103}$$

and

$$\begin{aligned} A_*(0) &= A(0) - p(0)m_*(0) \\ &= A(0) - \frac{p(0)}{\mu(0)}\left[H_*(0) - H(0)\right] \\ &= A(0) - \frac{p(0)}{\mu(0)}(1-\zeta)\Lambda_{E,F}\left[\pi_H(0) - \varphi(0)\right]^{-\chi} + \frac{p(0)}{\mu(0)}H(0), \end{aligned} \tag{3.104}$$

where $\Lambda_{E,F}$ denotes either Λ_E for scenario E or Λ_F for scenario F.

The solution for scenarios E and F can be derived from solutions A and B, respectively, by setting $t_1 = 0$ and replacing initial assets $A(0)$ and initial health $H(0)$ with the above expressions for $A_*(0)$ and $H_*(0)$. We leave this excercise to the reader.

3.8.8 Benefits transformation

Assuming that pension benefits accumulate over time as a fraction α of wages is invested with a return on investment of δ (the interest rate) as in equation (3.21) life-time income

$$\begin{aligned} \int_0^T Y[H(t)]dt &= \int_0^R w(t)e^{-\delta t}dt + \int_R^T be^{-\delta t}dt \\ &= \int_0^R w_0(t)e^{-\delta t}dt + \frac{b}{\delta}(e^{-\delta R} - e^{-\delta T}) \\ &+ \int_0^R \varphi(t)H(t)e^{-\delta t}dt, \end{aligned} \tag{3.105}$$

in the new formulation becomes

$$\begin{aligned} \int_0^T Y[H(t)]dt &= (1-\alpha)\int_0^R w(t)e^{-\delta t}dt + \int_R^T be^{-\delta t}dt \\ &= (1-\alpha)\int_0^R w_0(t)e^{-\delta t}dt \\ &+ \frac{1}{\delta}\left[b_0 + f(R)\alpha\int_0^R w_0(t)e^{\delta t}dt\right](e^{-\delta R} - e^{-\delta T}) \\ &+ \int_0^R \varphi(t)\left[(1-\alpha) + f(R)\frac{\alpha}{\delta}\left(e^{-\delta R} - e^{-\delta T}\right)e^{2\delta t}\right]H(t)e^{-\delta t}dt \end{aligned} \tag{3.106}$$

Comparing (3.105) with (3.106) leads to the identifications made in (3.22). Note further that the transformations in (3.22) also preserve the form of the Lagrangean (3.5 and 3.27) and that the transformations are independent of the control variables $C(t)$ and $m(t)$. Thus the original solutions remain valid with the transformations as long as one includes the derivative of φ when calculating health investment (equations 3.12, 3.13, 3.46 and 3.47),

$$\dot{\varphi} = 2\varphi f(R)\alpha \left(e^{-\delta R} - e^{-\delta T}\right) e^{2\delta t}. \tag{3.107}$$

Chapter 4

A Contribution to Health Capital Theory

I present a theory of the demand for health, health investment and longevity, building on the human capital framework for health and addressing limitations of existing models. I predict a negative correlation between health investment and health, that the health of wealthy and educated individuals declines more slowly and that they live longer, that current health status is a function of the initial level of health and the histories of prior health investments made, that health investment rapidly increases near the end of life and that length of life is finite as a result of limited life-time resources (the budget constraint). I derive a structural relation between health and health investment (e.g., medical care) that is suitable for empirical testing.

This chapter is based upon:

Galama, T.J. (2011) "A Contribution To Health Capital Theory", *RAND Working Paper*, WR-831.

4.1 Introduction

The demand for health is one of the most central topics in Health Economics. The canonical model of the demand for health and health investment (e.g., medical care) arises from Grossman (1972a, 1972b, 2000) and theoretical extensions and competing economic models are still relatively few. In Grossman's human capital framework individuals demand medical care (e.g., invest time and consume medical goods and services) for the consumption benefits (health provides utility) as well as production benefits (healthy individuals have greater earnings) that good health provides. The model provides a conceptual framework for interpretation of the demand for health and medical care in relation to an individual's resource constraints, preferences and consumption needs over the life cycle. Arguably the model has been one of the most important contributions of Economics to the study of health behavior. It has provided insight into a variety of phenomena related to health, medical care, inequality in health, the relationship between health and socioeconomic status, occupational choice, etc (e.g., Cropper, 1977; Muurinen and Le Grand, 1985; Case and Deaton, 2005) and has become the standard (textbook) framework for the economics of the derived demand for medical care.

Yet several authors have identified limitations to the literature spawned by Grossman's seminal 1972 papers[1] (see Grossman, 2000, for a review and rebuttal of some of these limitations). A standard framework for the demand for health, health investment (e.g., medical care) and longevity has to meet the significant challenge of providing insight into a variety of complex phenomena. Ideally it would explain the significant differences observed in the health of socioeconomic status (SES) groups - often called the "SES-health gradient". In the United States, a 60-year-old top-income-quartile male reports to be in similar health as a 20-year-old bottom-income-quartile male (Case and Deaton 2005) and similar patterns hold for other measures of SES, such as education and wealth, and other indicators of health, such as disability and mortality (e.g., Cutler et al. 2011; van Doorslaer et al. 2008). Initially diverging, the disparity in health between low- and high-SES groups appears to narrow after ages 50-60. Yet, Case and Deaton (2005) have argued that health production models are unable to explain differences in the health deterioration rate (not just the level) between socioeconomic groups.

Another stylized fact of the demand for medical care is that healthy individuals do not go to the doctor much: a strong negative correlation is observed between measures of health and measures of health investment. However, Wagstaff (1986a) and Zweifel and

[1]Throughout this paper I refer to this literature as the health production literature.

Breyer (1997) have pointed to the inability of health production models to predict the observed negative relation between health and the demand for medical care.

Introspection and casual observation further suggests that healthy individuals are those that began life healthy and that have invested in health over the life course. Thus one would expect that health depends on initial conditions (e.g., initial health) and the history of health investments, prices, wages, medical technology and environmental conditions. Yet, Usher (1975) has pointed to the lack of "memory" in model solutions. For example the solution for health typically does not depend on its initial value or the histories of health investment and biological aging.

Further, Case and Deaton (2005) note that *"...If the rate of biological deterioration is constant, which is perhaps implausible but hardly impossible, ...people will "choose" an infinite life ..."*. This suggests that complete health repair is possible, regardless of the speed of the process (the rate itself does not matter in causing health to decline) and regardless of the budget constraint, and as a result declines in health status are driven, not by the rate of deterioration of the health stock, but by the rate of increase of the rate of deterioration (Case and Deaton, 2005). Thus a necessary condition in health production models is that the biological aging rate increases with age to ensure that life is finite and health declines and to reproduce the observed rapid increase in medical care near the end of life. Case and Deaton (2005) argue, however, that a technology that can effect such complete health repair is implausible.

Last, Ehrlich and Chuma (1990) have pointed out that under the constant returns to scale (CRTS) health production process assumed in the health production literature, the marginal cost of investment is constant, and no interior equilibrium for health investment exists. Ehrlich and Chuma argue that this is a serious limitation that introduces a type of indeterminacy ("bang-bang") problem with respect to optimal investment and health maintenance choices. The importance of this observation appears to have gone relatively unnoticed: contributions to the literature that followed the publication of Ehrlich and Chuma's work in 1990 have continued to assume a health production function with CRTS in health investment.[2] This may have been as a consequence of the following factors: First, Ehrlich and Chuma's finding that health investment is undetermined (under the

[2]E.g., Bolin et al. (2001, 2003); Case and Deaton (2005); Erbsland, Ried and Ulrich (2002); Jacobsen (2000); Leu and Gerfin (1992); Liljas (1998); Nocera and Zweifel (1998); Wagstaff (1986a); Ried (1996, 1998). To the best of my knowledge the only exception is an unpublished working paper by Dustmann and Windmeijer (2000) who take the model by Ehrlich and Chuma (1990) as their point of departure. Bolin et al. (2002a, 2002b) assume that the health investment function is a decreasing function of health. Thus they impose a relationship between health and health investment to ensure that the level of investment in health decreases with the health stock rather than deriving this result from first principles.

usual assumption of a CRTS health production process) was incidental to their main contribution of introducing the demand for longevity (or "quantity of life") and the authors did not explore the full implications of a DRTS health production process. Second, Ehrlich and Chuma's argument is brief and technical.[3] This has led Reid (1998) to argue that *"... the authors [Ehrlich and Chuma] fail to substantiate either claim [bang-bang and indeterminacy] ..."*, suggesting there is room for further research into the argument made by Ehrlich and Chuma. Third, there was the incorrect notion that Ehrlich and Chuma had changed the structure of the model substantially and that the alleged indeterminacy of health investment did not apply to the original formulation in discrete time (e.g., Reid, 1998). Last, because of the increased complexity of a health production model that includes endogenous length of life (demand for longevity) Ehrlich and Chuma (1990) had to resort to a particular sensitivity analysis, suitable to optimal control problems (Oniki, 1973), in which the directional effect of a parameter change can be investigated. Ehrlich and Chuma's (1990) insightful work is therefore limited to generating directional predictions. This suggested that obtaining insight into the characteristics of a DRTS health production model would require numerical analysis or the kind of sensitivity analysis performed by Ehrlich and Chuma (1990) – while it would not substantially change the nature of the theory. For example, it was thought that introducing DRTS would result in individuals reaching the desired health stock gradually rather than instantaneously (e.g., Grossman, 2000, p. 364) – perhaps not a sufficiently important improvement to warrant the increased level of complexity.

What then is needed to address the above mentioned limitations? I argue that the answer is two-fold: 1) a reinterpretation is needed of the health stock equilibrium condition, one of the most central relations in the health production literature, as determining the optimal level of health investment and not the "optimal" level of the health stock, and 2) one needs to assume DRTS in the health production process as Ehrlich and Chuma (1990) have argued.

In this paper I present a theory of the demand for health, health investment and longevity based on Grossman (1972a, 1972b) and the extended version of this model by Ehrlich and Chuma (1990). In particular, this paper explores in detail the implications of a DRTS health production process. The theory I develop is capable of reproducing the phenomena discussed above and of addressing the above mentioned five limitations.

[3]It involves a reference to a graph with health investment on one axis and the ratio of two Lagrange multipliers on the other. The authors note that the same results hold in a discrete time setting, using a proof based on the last period preceeding death (see their footnote 4).

This paper contributes to this literature as follows. First, I reduce the complexity of a theory with a DRTS health production process (as in Ehrlich and Chuma, 1990) by arguing for a different interpretation of the health stock equilibrium condition, one of the most central relations in the health production literature: this relation determines the optimal level of health investment (not the health stock), conditional on the level of the health stock. The health production literature has thus far not employed the alternative interpretation of the health equilibrium condition and consistently utilizing it allows me to develop the health production literature further than was previously possible. This is because the equilibrium condition for the health stock is of a much simpler form than the condition which is typically utilized to determine the optimal level of health investment. Many of the subsequent contributions this paper makes follow from the alternative interpretation advocated here.

Second, I show that the alternative interpretation allows for an intuitive understanding as to why the assumption of DRTS in the health production function is necessary, or no solution to the optimization problem exists. Essentially, the CRTS process as utilized in the health production literature represents a degenerate case. This is no new result (Ehrlich and Chuma, 1990), but this paper provides more intuitive, less technical and additional arguments as to why health investment is not determined under the assumption of a CRTS health production process. This is important because the implications of the indeterminacy are substantial (e.g., Ehrlich and Chuma, 1990), yet the debate does not appear to have been settled in favor of a DRTS health production process as illustrated by its lack of use in the health production literature.

Third, the alternative interpretation allows for explorations of a stylized representation of the first-order condition which enable an intuitive understanding of the optimal solution for health investment. I find that a unique optimal solution for health investment exists (thus addressing the indeterminacy as Ehrlich and Chuma, 1990, have also shown). Given an optimal level for health investment, and because in this interpretation the health stock is determined by the dynamic equation for health, the stock is found to be a function of the histories of past health investments and past biological aging rates, addressing the criticism of Usher (1975). Further, I find that the optimal level of health investment decreases with the user cost of health capital and increases with wealth and with the consumption and production benefit of health. Thus I show that one does not need to resort to numerical analyses to gain insight into the characteristics of the solution. This is important because, arguably, the Grossman model has been successful, in part, because of its ability to guide empirical analyses through the intuition that simple representations provide (e.g., Wagstaff, 1986b; Muurinen and Le Grand, 1985).

Fourth, the alternative interpretation allows developing relations for the effects of variations in SES (wealth, education) and in health on the optimal level of health investment.[4] These relations complement explorations of stylized representations by allowing one to distinguish first- from second- and third-order effects and to explore the mechanisms (pathways) that combine to produce the final directional outcome, again, without the need to resort to numerical analyses. Under plausible assumptions the theory predicts a negative correlation between health and health investment (in cross-section). This is an important new result that addresses the criticism by Wagstaff (1986a) and Zweifel and Breyer (1997). Further, greater wealth, higher earnings over the life cycle and more education and experience are associated with slower health deterioration, addressing the criticism by Case and Deaton (2005).[5]

Fifth, empirical tests of the health production literature have thus far been based on structural and reduced form equations derived under the assumption of a CRTS health production process. Arguably, health capital theory has not yet been properly tested because these structural and reduced form relations suffer from the issue of the indeterminacy of health investment (and essentially represent a degenerate case). Absent an equivalent relation for a DRTS health production process I once more employ the alternative interpretation to derive a structural relation between health and health investment (e.g., medical care) that is suitable for empirical testing. The structural relation contains the CRTS health production process as a special case, thereby allowing empirical tests to verify or reject this common assumption in the health production literature.

Last, I perform numerical simulations to illustrate the properties of the theory. These simulations show that the model is capable of reproducing the rapid increase in health investment near the end of life and that the optimal solution for length of life is finite for a constant biological aging rate, addressing the criticism by Case and Deaton (2005) that health production models are characterized by complete health repair. In sum, I find that the theory can address each of the five limitations discussed above.

The paper is organized as follows. Section 4.2 presents the model in discrete time and discusses the characteristics of the first-order conditions. In particular this section offers an alternative interpretation of the first-order conditions. Section 4.3 explores the properties of a DRTS health production process, in several ways, by: a) exploring a stylized representation of the first-order condition for health investment to gain an

[4]Employing Oniki's (1973) method as in Ehrlich and Chuma (1990) is somewhat comparable to the analysis performed here. Unfortunately, due to space limitations, the detailed analysis underlying the directional predictions by Ehrlich and Chuma (1990) has not been published, but is available on request from the authors.

[5]These results are also obtained by exploring a stylized representation of the first-order condition.

intuitive understanding of its properties, b) analyzing the effect of differences in health and socioeconomic status (wealth and education) on the optimal level of health investment and consumption, c) developing structural-form relations for empirical testing of the model and d) presenting numerical simulations of health, health investment, assets and consumption profiles and length of life. Section 4.4 summarizes and concludes. The Appendix provides detailed derivations and mathematical proofs.

4.2 The demand for health, health investment and longevity

I start with Grossman's basic formulation (Grossman, 1972a, 1972b, 2000) for the demand for health and health investment (e.g., medical care) in discrete time (see also Wagstaff, 1986a; Wolfe, 1985; Zweifel and Breyer, 1997; Ehrlich and Chuma, 1990).[6] Health is treated as a form of human capital (health capital) and individuals derive both consumption (health provides utility) and production benefits (health increases earnings) from it. The demand for medical care is a derived demand: individuals demand "good health", not the consumption of medical care.

Using discrete time optimal control (e.g., Sydsaeter, Strom and Berck, 2005) the problem can be stated as follows. Individuals maximize the life-time utility function

$$\sum_{t=0}^{T-1} \frac{U(C_t, H_t)}{\prod_{k=1}^{t}(1+\beta_k)}, \tag{4.1}$$

where individuals live for T (endogenous) periods, β_k is a subjective discount factor and individuals derive utility $U(C_t, H_t)$ from consumption C_t and from health H_t. Time t is measured from the time individuals begin employment. Utility increases with consumption $\partial U_t/\partial C_t > 0$ and with health $\partial U_t/\partial H_t > 0$.

The objective function (4.1) is maximized subject to the dynamic constraints:

$$H_{t+1} = f(I_t) + (1-d_t)H_t, \tag{4.2}$$

$$A_{t+1} = (1+\delta_t)A_t + Y(H_t) - p_{X_t}X_t - p_{m_t}m_t, \tag{4.3}$$

the total time budget Ω_t

$$\Omega_t = \tau_{w_t} + \tau_{I_t} + \tau_{C_t} + s(H_t), \tag{4.4}$$

[6] In line with Grossman (1972a; 1972b) and Ehrlich and Chuma (1990) I do not incorporate uncertainty in the health production process. This would unnecessarily complicate the optimization problem and require numerical methods, while it is not needed to explain the stylized facts regarding health behavior discussed in this paper. For a detailed treatment of uncertainty within the Grossman model the reader is referred to Ehrlich (2000), Liljas (1998), and Ehrlich and Yin (2005).

and initial and end conditions: H_0, H_T, A_0 and A_T are given. Individuals live for T periods and die at the end of period $T-1$. Length of life T (Grossman, 1972a, 1972b) is determined by a minimum health level $H_{\min}$. If health falls below this level $H_t \leq H_{min}$ an individual dies ($H_T \equiv H_{\min}$).

Health (equation 4.2) can be improved through investment in health I_t and deteriorates at the biological aging rate d_t. The relation between the input, health investment I_t, and the output, health improvement $f(I_t)$, is governed by the health production function $f(\cdot)$. The health production function $f(\cdot)$ is assumed to obey the law of diminishing marginal returns in health investment. For simplicity of discussion I use the following simple functional form

$$f(I_t) = I_t^{\alpha}, \tag{4.5}$$

where $0 < \alpha < 1$ (DRTS).[7,8]

Assets A_t (equation 4.3) provide a return δ_t (the rate of return on capital), increase with income $Y(H_t)$ and decrease with purchases in the market of consumption goods and services X_t and medical goods and services m_t at prices p_{X_t} and p_{m_t}, respectively. Income $Y(H_t)$ is assumed to be increasing in health H_t as healthy individuals are more productive and earn higher wages (Currie and Madrian, 1999; Contoyannis and Rice, 2001).

Goods and services X_t purchased in the market and own time inputs τ_{C_t} are used in the production of consumption C_t. Similarly medical goods and services m_t and own time inputs τ_{I_t} are used in the production of health investment I_t. The efficiencies of production are assumed to be a function of the consumer's stock of knowledge E (an individual's human capital exclusive of health capital [e.g., education]) as the more educated may be more efficient at investing in health (see, e.g., Grossman 2000):

$$I_t = I[m_t, \tau_{I_t}; E], \tag{4.6}$$

$$C_t = C[X_t, \tau_{C_t}; E]. \tag{4.7}$$

[7]For $\alpha = 1$ we have Grossman's original formulation of a linear health production process.

[8]Mathematically, equation (4.5) is equivalent to the assumption made by Ehrlich and Chuma (1990) of a dual cost-of-investment function with decreasing returns to scale (their equation 5) and a linear health production process ($\alpha = 1$ in equation 4.5 in this paper). Conceptually, however, there is an important distinction. In principle one could imagine a scenario where the investment function I_t has constant or even increasing returns to scale in its inputs of health investment goods / services m_t and own time τ_{I_t}, but where the ultimate health improvement (through the health production process) has diminishing returns to scale in its inputs m_t and τ_{I_t} as assumed in equation (4.5; this paper). Arguably, it is not the process of health investment but the process of health production (the ultimate effect on health) that is expected to exhibit decreasing returns to scale.

The total time available in any period Ω_t (equation 4.4) is the sum of all possible uses τ_{w_t} (work), τ_{I_t} (health investment), τ_{C_t} (consumption) and $s(H_t)$ (sick time; a decreasing function of health). In this formulation one can interpret τ_{C_t}, the own-time input into consumption C_t as representing leisure.[9]

Income $Y(H_t)$ is taken to be a function of the wage rate w_t times the amount of time spent working τ_{w_t},

$$Y(H_t) = w_t \left[\Omega_t - \tau_{I_t} - \tau_{C_t} - s(H_t)\right]. \tag{4.8}$$

Thus, we have the following optimal control problem: the objective function (4.1) is maximized with respect to the control functions X_t, τ_{C_t}, m_t and τ_{I_t} and subject to the constraints (4.2, 4.3 and 4.4). The Hamiltonian of this problem is:

$$\Im_t \;=\; \frac{U(C_t, H_t)}{\prod_{k=1}^{t}(1+\beta_k)} + q_t^H H_{t+1} + q_t^A A_{t+1}, \quad t = 0, \ldots T-1 \tag{4.9}$$

where q_t^H is the adjoint variable associated with the dynamic equation (4.2) for the state variable health H_t and q_t^A is the adjoint variable associated with the dynamic equation (4.3) for the state variable assets A_t.[10]

The optimal control problem presented so far is formulated for a fixed length of life T (see, e.g., Seierstad and Sydsaeter, 1977, 1987; Kirk, 1970; see also section 4.3.4). To allow for differential mortality one needs to introduce an additional condition to the optimal control problem to optimize over all possible lengths of life T (Ehrlich and Chuma, 1990). One way to achieve this is by first solving the optimal control problem conditional on length of life T (i.e., for a fixed exogenous T), inserting the optimal solutions for consumption C_t^* and health H_t^* (denoted by $*$) into the "indirect utility function"

$$V_T \equiv \sum_{t=0}^{T-1} \frac{U(C_t^*, H_t^*)}{\prod_{k=1}^{t}(1+\beta_k)}, \tag{4.10}$$

and maximizing V_T with respect to T.[11]

[9] Because consumption consists of time inputs and purchases of goods/services in the market one can conceive leisure as a form of consumption consisting entirely or mostly of time inputs. Leisure, similar to consumption, provides utility and its cost consists of the price of goods/services utilized and the opportunity cost of time.

[10] For a CRTS health production function ($f(I_t) \propto I_t$) as employed in the health production literature we have to explicitly impose that health investment is non negative, $I_t \geq 0$ (see Galama and Kapteyn 2009). This can be done by introducing an additional multiplier q_t^I in the Hamiltonian (equation 4.9) associated with the condition that health investment is non negative, $I_t \geq 0$. This is not necessary for a DRTS health production function, where diminishing marginal benefits and choice of suitable functional forms ensure that the optimal solution for health investment I_t is non negative.

[11] This is mathematically equivalent to the condition utilized by Ehrlich and Chuma (1990) (in continuous time) that the Hamiltonian equal zero at the end of life $\Im_T = 0$ (transversality condition).

4.2.1 First-order conditions

Maximization of (4.9) with respect to the control functions m_t and τ_{I_t} leads to the first-order condition for health investment I_t

$$\begin{aligned}\frac{\pi_{I_t}}{\prod_{k=1}^{t}(1+\delta_k)} &= -\sum_{i=1}^{t}\left[\frac{\partial U(C_i,H_i)/\partial H_i}{q_0^A\prod_{j=1}^{i}(1+\beta_j)}+\frac{\partial Y(H_i)/\partial H_i}{\prod_{j=1}^{i}(1+\delta_j)}\right]\frac{1}{\prod_{k=i}^{t}(1-d_k)}\\ &+ \frac{\pi_{I_0}}{\prod_{k=1}^{t}(1-d_k)},\end{aligned} \tag{4.11}$$

where π_{I_t} is the marginal cost of health investment I_t

$$\pi_{I_t} \equiv \frac{p_{m_t}I_t^{1-\alpha}}{\alpha[\partial I_t/\partial m_t]} = \frac{w_t I_t^{1-\alpha}}{\alpha[\partial I_t/\partial \tau_{I_t}]}, \tag{4.12}$$

and the Lagrange multiplier q_0^A is the shadow price of wealth (see, e.g, Case and Deaton, 2005).

An alternative expression is obtained by using the final period $T-1$ as point of reference

$$\begin{aligned}\frac{\pi_{I_t}}{\prod_{k=1}^{t}(1+\delta_k)} &= \sum_{i=t+1}^{T-1}\left[\frac{\partial U(C_i,H_i)/\partial H_i}{q_0^A\prod_{j=1}^{i}(1+\beta_j)}+\frac{\partial Y(H_i)/\partial H_i}{\prod_{j=1}^{i}(1+\delta_j)}\right]\prod_{k=t+1}^{i-1}(1-d_k)\\ &+ \frac{\pi_{I_{T-1}}\prod_{k=t+1}^{T-1}(1-d_k)}{\prod_{k=1}^{T}(1+\delta_k)}.\end{aligned} \tag{4.13}$$

Using either the expression (4.11) or (4.13) for the first-order condition for health investment and taking the difference between period t and $t-1$ we obtain the following expression

$$(1-d_t)\pi_{I_t} = \pi_{I_{t-1}}(1+\delta_t) - \left[\frac{\partial U(C_t,H_t)/\partial H_t\prod_{j=1}^{t}(1+\delta_j)}{q_0^A\prod_{j=1}^{t}(1+\beta_j)}+\frac{\partial Y(H_t)}{\partial H_t}\right], \tag{4.14}$$

or

$$\frac{\partial U(C_t,H_t)}{\partial H_t} = q_0^A\left(\sigma_{H_t}-\varphi_{H_t}\right)\frac{\prod_{j=1}^{t}(1+\beta_j)}{\prod_{j=1}^{t}(1+\delta_j)}, \tag{4.15}$$

where σ_{H_t} is the user cost of health capital at the margin

$$\sigma_{H_t} \equiv \pi_{I_t}\left[(d_t+\delta_t)-\frac{\Delta\pi_{I_t}}{\pi_{I_t}}(1+\delta_t)\right], \tag{4.16}$$

φ_{H_t} is the marginal production benefit of health

$$\varphi_{H_t} \equiv \frac{\partial Y(H_t)}{\partial H_t}, \tag{4.17}$$

and $\Delta\pi_{I_t} \equiv \pi_{I_t} - \pi_{I_{t-1}}$.

Maximization of (4.9) with respect to the control functions X_t and τ_{C_t} leads to the first-order condition for consumption C_t

$$\frac{\partial U(C_t, H_t)}{\partial C_t} = q_0^A \pi_{C_t} \frac{\prod_{j=1}^{t}(1+\beta_j)}{\prod_{j=1}^{t}(1+\delta_j)}, \tag{4.18}$$

where π_{C_t} is the marginal cost of consumption C_t

$$\pi_{C_t} \equiv \frac{p_{X_t}}{\partial C_t / \partial X_t} = \frac{w_t}{\partial C_t / \partial \tau_{C_t}}. \tag{4.19}$$

The first-order condition (4.11) (or the alternative forms 4.13 and 4.15) determines the optimal solution for the control function health investment I_t. The first-order condition (4.18) determines the optimal solution for the control function consumption C_t.[12] The solutions for the state functions health H_t and assets A_t then follow from the dynamic equations (4.2) and (4.3). Length of life T is determined by maximizing the indirect utility function V_T (see 4.10) with respect to T.

4.2.2 An alternative interpretation of the first-order condition

One of the most central relations in the health production literature is the first-order condition (4.15). This relation equates the marginal consumption benefit of health $\partial U_t/\partial H_t$ to the user cost of health capital σ_{H_t} and the marginal production benefit of health φ_{H_t}, and is interpreted as an equilibrium condition for the health stock H_t. It is equivalent to, e.g., equation (11) in Grossman (2000) and equation (13) in Ehrlich and Chuma (1990).[13] An alternative interpretation of relation (4.15) is, however, that it determines the optimal level of health investment I_t. My argument is as follows.

First, the first-order condition (4.15) is the result of maximization of the optimal control problem with respect to investment in health and hence, first and foremost, it

[12] Because the first-order condition for health investment goods / services m_t and the first-order condition for own time inputs τ_{I_t} are identical (see Appendix section 4.5.1) one can consider a single control function I_t (health investment) instead of two control functions m_t and τ_{I_t}. The same is true for consumption C_t. Because of this property, the optimization problem is reduced to two control functions I_t and C_t (instead of four) and two state functions H_t and A_t.

[13] Notational differences with respect to Grossman (2000) are: $q_0^A \to \lambda$, $\pi_{I_t} \to \pi_t$, $\partial U_t/\partial H_t \to [\partial U/\partial h_t][\partial h_t/\partial H_t] = U h_t G_t$ (where h_t is healthy time, a function of health H_t), $\varphi_{H_t} \to W_t G_t$, $\delta_t \to r$, $d_t \to \delta_t$, $\beta_t \to 0$, and $T \to n$. Notational differences with respect to Ehrlich and Chuma (1990), apart from using discrete rather than continuous time, are: $q_0^A \to \lambda_A(0)$, $\pi_{I_t} \to g(t)$, $\partial U_t/\partial H_t \to [\partial U(t)/\partial h(t)][\partial h(t)/\partial H(t)] = U_h(t)\varphi'(H(t))$ (where $h(t)$ is healthy time), $\varphi_{H_t} \to w\varphi'(H(t))$, $\delta_t \to r$, $d_t \to \delta(t)$ and $\beta_t \to \rho$.

determines the optimal level of health investment I_t. Optimal control theory distinguishes between control functions and state functions. Control functions are determined by the first-order conditions and state functions by the dynamic equations (e.g., Seierstad and Sydsaeter, 1977, 1987; Kirk, 1970). The first-order condition (4.15) is thus naturally associated with the control function health investment I_t and the state function health H_t is determined by the dynamic equation (4.2).[14]

Second, in the health production literature the optimal solution for health investment I_t is assumed to be determined by the first-order condition (4.11) (or the alternative form 4.13). It is equivalent to, e.g., equation (9) in Grossman (2000) and equation (8) in Ehrlich and Chuma (1990).[15] However, it can be shown that the first-order conditions (4.11) and (4.15) are mathematically equivalent

$$(4.11) \Leftrightarrow (4.15), \tag{4.20}$$

proof of which is provided in the Appendix (section 4.5.2). Thus if equation (4.11) is the first-order condition for health investment I_t (the interpretation in the health production literature) then equation (4.15) is too (and vice versa).

From a purely mathematical standpoint one could conceive the condition (4.15) as determining the level of the health stock because a direct relation exists between health H_t and health investment I_t, namely the dynamic equation (4.2). Optimizing with respect to health investment entails optimizing with respect to health. Thus, in principle, one ought to be able to reconcile both interpretations. However, the health production literature assumes CRTS in the health production process.[16] In section 4.3.1 I show that under this particular assumption the level of health investment is not determined, i.e. that it represents a special degenerate case. As a result, both approaches cannot be reconciled in this particular case.

[14] Analogously, the first-order condition (4.18) is associated with the control variable consumption C_t and the dynamic equation (4.3) is associated with the state function assets A_t.

[15] One important difference between the results derived by Grossman (equation 9 in Grossman, 2000) and those derived here is the absense in Grossman's derivations of the reference point π_{I_0} in equation (4.11) or the reference point $\pi_{I_{T-1}}$ in equation (4.13). Using optimal control techniques I find these reference points to be required in a discrete time formulation (see equations 4.11 and 4.13). This is also true for a continuous time formulation. To the best of my knowledge this observation has not been made before. It has important implications for the model's interpretation as the begin or end point references allows one to ensure that the solution is consistent with the begin and end conditions for health and assets: H_0, H_T, A_0 and A_T.

[16] I.e., $f(I_t) = I_t^{\alpha}$ with $\alpha = 1$ (equations 4.2 and 4.5) and a Cobb-Douglas (CRTS) relation between investment in medical care I_t and its inputs own time and goods/services purchased in the market.

In the remainder of this paper I will use relations (4.11) and (4.15) as being equivalent. Both conditions determine the optimal level of health investment I_t, conditional on the level of the health stock H_t.

4.3 A DRTS health production process

In this section I explore the properties of a health production process in several ways. In section 4.3.1 I discuss a stylized representation of the first-order condition for health investment to gain an intuitive understanding of its properties. In particular I contrast the characteristics of the solution for health investment under a DRTS health production process with that of a CRTS process. In this section I also provide additional arguments for the claim made by Ehrlich and Chuma (1990) that DRTS in the health production process are necessary to guarantee the existence of a solution to the optimization problem.[17] In section 4.3.2 I explore the effect of differences in health and socioeconomic status (wealth and education) on the optimal level of health investment and consumption. In section 4.3.3 I derive structural form relations for empirical testing of the model. Last, in section 4.3.4 I perform numerical simulations of health, health investment, assets and consumption profiles and length of life.

In the following I assume diminishing marginal utilities of consumption $\partial^2 U_t/\partial C_t^2 < 0$ and of health $\partial^2 U_t/\partial H_t < 0$, and diminishing marginal production benefit of health $\partial \varphi_{H_t}/\partial H_t = \partial^2 Y_t/\partial H_t^2 < 0$. In addition I make the usual assumption of a Cobb-Douglas CRTS relation between the inputs goods/services purchased in the market and own-time and the outputs investment in curative care I_t and consumption C_t. As a result we have $\pi_{I_t} \propto I_t^{1-\alpha}$ and $\partial \pi_{C_t}/\partial C_t = 0$ (see equations 4.81 and 4.84 in Appendix section 4.5.4).

4.3.1 Stylized representation

In this section I contrast the properties of a DRTS health production process[18] (section 4.3.1) with those of a CRTS health production process[19] (section 4.3.1).

[17] Providing further corroboration of their claim is important because the implications are substantial and the debate does not appear to have been settled in favor of a DRTS health production process as illustrated by its lack of use in the health production literature.

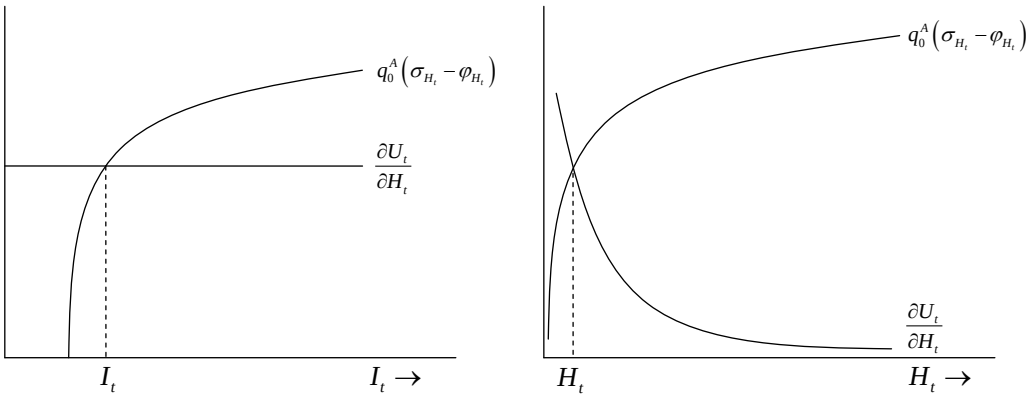
[18] $0 < \alpha < 1$ and a Cobb-Douglas health investment process I_t.

[19] $\alpha = 1$ and a Cobb-Douglas health investment process I_t.

Decreasing returns to scale

Figure 4.1 provides a stylized representation of the first-order condition for health investment I_t (4.15): it graphs the marginal benefit and marginal cost of health as a function of health investment I_t (left-hand side) and as a function of health H_t (right-hand side).[20]

Figure 4.1: Marginal benefit versus marginal cost of health for a DRTS health production process.



Notes: In labeling the curves I have omitted the term $\prod_{j=1}^{t}(1+\beta_j)[\prod_{j=1}^{t}(1+\delta_j)]^{-1}$.

Consider the left-hand figure first. The optimal level of health investment I_t is determined by equating the consumption benefit of health $\partial U_t/\partial H_t$ with the cost of maintaining the health stock $q_0^A(\sigma_{H_t} - \varphi_{H_t})$ (here and in the remainder of the discussion in this section I ommit for convenience of notation the term $\prod_{j=1}^{t}(1+\beta_j)[\prod_{j=1}^{t}(1+\delta_j)]^{-1}$).

Utility is derived from health H_t and consumption C_t but not from health investment I_t (the demand for medical care is a derived demand). Further, the evolution of the health stock H_t is determined by the dynamic equation (4.2) which can be written (using 4.5) as

$$H_t = H_0 \prod_{j=0}^{t-1}(1-d_j) + \sum_{j=0}^{t-1} I_j^{\alpha} \prod_{i=j+1}^{t-1}(1-d_i). \tag{4.21}$$

In other words, health H_t is a function of past health investment I_s but not of current health investment I_t ($s < t$). Thus the consumption benefit of health $\partial U_t/\partial H_t$ is independent of the level of health investment I_t: this is shown as the horizontal solid line labeled $\partial U_t/\partial H_t$.

[20] While in principle one can derive predictions for the level of health investment I_t from the left-hand figure without the need to resort to the right-hand figure, it is useful to consider the right-hand figure in order to illustrate the effect of differences in the health stock H_t on the optimal level of health investment (see section 4.3.2) and to make comparisons with the usual interpretation of this relation as determining the "optimal" health stock (rather than optimal investment; see section 4.3.1).

The cost of maintaining the health stock is a function of the shadow price of wealth q_0^A,[21] the user cost of health capital σ_{H_t}, the production benefit of health φ_{H_t}, and an exponential factor that varies with age t depending on the difference between the time preference rate β_t and the rate of return on capital δ_t. The marginal cost of health investment π_{I_t} and hence the user cost of health capital σ_{H_t} is increasing in the level of investment in health I_t ($\pi_{I_t} \propto I_t^{1-\alpha}$; see equation 4.81 in Appendix section 4.5.4). The marginal production benefit of health φ_{H_t} is not a function of the level of health investment I_t. As a result, the cost of maintaining the health stock is upward sloping in the level of health investment (labeled $q_0^A(\sigma_{H_t} - \varphi_{H_t})$). The intersection of the two curves determines the optimal level of health investment (dotted vertical line labeled I_t).

Now consider the right-hand side of Figure 4.1. The marginal consumption benefit of health $\partial U_t/\partial H_t$ is downward sloping (convex) in health (curve labeled $\partial U_t/\partial H_t$) and the cost of maintaining the health stock $q_0^A(\sigma_{H_t} - \varphi_{H_t})$ is upward sloping (concave) in health (curve labeled $q_0^A(\sigma_{H_t} - \varphi_{H_t})$) due to the diminishing marginal production benefit of health φ_{H_t}. Since health is a stock its level is given (dotted vertical line labeled H_t) and provides a constraint: the two curves have to intersect at this level H_t. It is possible for the two curves to intersect at H_t through endogenous health investment I_t. A higher(/lower) level of health investment I_t increases(/decreases) (ceteris paribus) the marginal cost of health investment and hence the user cost of health capital. As a result the cost of maintaining the health stock (curve labeled $q_0^A(\sigma_{H_t} - \varphi_{H_t})$) shifts upward(/downward) while the marginal benefit of health (curve labeled $\partial U_t/\partial H_t$) remains stationary (it is not a function of the level of health investment).

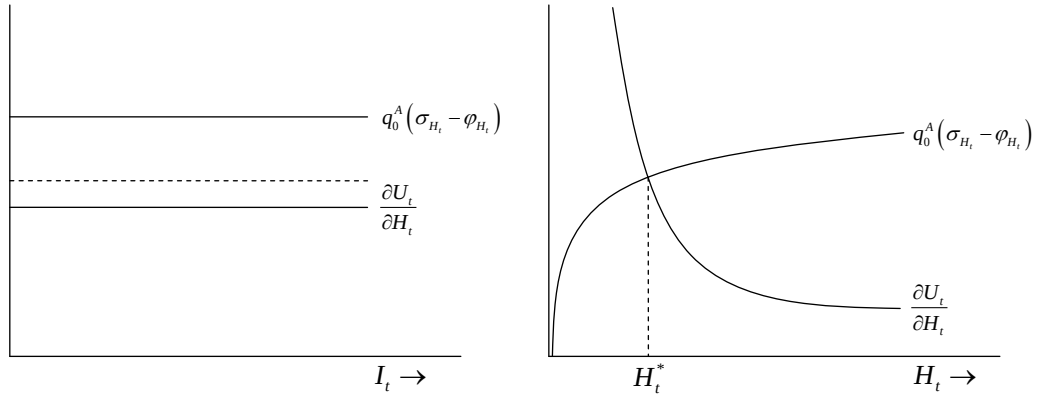
The level of the marginal consumption benefit of health (labeled $\partial U_t/\partial H_t$ on the left-hand side of Figure 4.1) for which the health stock is at H_t (draw a horizontal line from the left-hand to the right-hand side of Figure 4.1) determines the optimal solution for health investment I_t. The optimal level of health investment I_t decreases with the user cost of health capital σ_{H_t} and increases with wealth (lower q_0^A) and with the consumption $\partial U_t/\partial H_t$ and production φ_{H_t} benefit of health. Further, the optimal level of health investment I_t is a direct function of the level of health stock H_t as can be seen from the first-order condition (4.15) and from its stylized representation in Figure 4.1 (more on this in the next sections 4.3.2 and 4.3.3). Hence, for a DRTS health production process a unique solution for health investment I_t exists for every level of the health stock H_t. This addresses the issue of the indeterminacy of health investment (e.g., Ehrlich and Chuma, 1990).

[21] q_0^A is decreasing in initial assets and life-time earnings. See, e.g., Wagstaff (1986a).

Constant returns to scale

Figure 4.2 provides a stylized representation of the first-order condition (4.15) for health investment for a CRTS health production process, as typically assumed in the health production literature: it graphs the marginal benefit and marginal cost of health as a function of health investment I_t (left-hand side) and as a function of the health stock H_t (right-hand side). In the following I follow the discussion in the previous section 4.3.1 and emphasize the differences with respect to a DRTS health production process.

Figure 4.2: Marginal benefit versus marginal cost of health for a CRTS health production process.



Notes: In labeling the curves I have omitted the term $\prod_{j=1}^{t}(1+\beta_j)[\prod_{j=1}^{t}(1+\delta_j)]^{-1}$.

Consider the left-hand side first. Unlike the DRTS process, for a CRTS process the marginal cost of health investment π_{I_t}, and hence the user cost of health capital σ_{H_t}, is independent of the level of health investment I_t ($\pi_{I_t} \propto I_t^{1-\alpha}$ = constant for $\alpha = 1$; see equation 4.81 in Appendix section 4.5.4). Thus, not only the marginal utility of health $\partial U_t/\partial H_t$ but also the net marginal cost is independent of the level of health investment I_t: this is shown as the horizontal solid lines labeled $q_0^A(\sigma_{H_t} - \varphi_{H_t})$ and $\partial U_t/\partial H_t$.

Because individuals cannot adjust their health instantaneously, the level of the health stock H_t at age t is given and provides a constraint for the optimization problem at age t. Generally the constraint provided by H_t will result in different values for the marginal benefit and marginal cost of health: this is depicted by the two horizontal lines having distinct levels (they do not overlap). The intersection of the two solid curves would determine the optimal level of health investment I_t but only in the peculiar case that both lines exactly overlap does such an optimal solution exist. Thus for most values of the health stock no solution for health investment I_t exists.

Now consider the right-hand side of Figure 4.2. The consumption benefit of health $\partial U_t/\partial H_t$ is downward sloping to represent diminishing marginal utility in health. The cost of maintaining the health stock $q_0^A(\sigma_{H_t} - \varphi_{H_t})$ is upward sloping to represent diminishing marginal production benefits of health φ_{H_t}. As the graph shows, a unique level of health H_t^* exists (dashed vertical line) for which the consumption benefit of health equals the cost of maintaining the health stock. The health production literature assumes this unique solution H_t^* describes the "optimal" health path. Turning again to the left-hand side of Figure 4.2, note that for this particular value of the health stock H_t^* the consumption benefit of health $\partial U_t/\partial H_t$ and the cost of maintaining the health stock $q_0^A(\sigma_{H_t} - \varphi_{H_t})$ overlap (they both lie on the dashed horizontal line). Thus a solution for the level of investment in health I_t exists, but any non negative value can be allowed: once more the optimal level of investment in health I_t is not determined.

In order to illustrate that this result does not depend on the equivalence of the first-order conditions (4.11) and (4.15) I show next that this result also holds for (4.11), the relation that is utilized in the health production literature as determining the optimal level of health investment. The first-order condition for health investment (4.11) equates the current marginal monetary cost of investment in health π_{I_t} (left-hand side; LHS) with a function of the current and all *past* values of the marginal utility of health $\partial U_s/\partial H_s$ and the marginal production benefit of health φ_{H_s} $(0 \leq s \leq t)$ (right-hand side; RHS). The LHS of (4.11) is not a function of health investment as the marginal monetary cost of health investment π_{I_t} is independent of the level of investment for a CRTS health production process. The RHS of (4.11) is also not a function of current investment I_t because the marginal utility of health $\partial U_s/\partial H_s$ and the marginal production benefit of health φ_{H_s} are functions of the health stock H_s $(0 \leq s \leq t)$ which in turn is a function of past but not current health investment I_s $(s < t$; see equation 4.21). Thus the first-order condition for health investment (4.11) is not a function of health investment I_t and the level of health investment is not determined.

Ehrlich and Chuma (1990) have reached the same conclusion on the basis of a technical argument. From equation (4.59) or (4.60) it follows that the marginal monetary cost of health investment π_{I_t} is the ratio of two Lagrange multipliers

$$\pi_{I_t} = \frac{q_t^H}{q_t^A}. \tag{4.22}$$

The right-hand side of (4.22) is not a function of health investment I_t by definition.[22,23] For a CRTS health production proces π_{I_t} is also not a function of health investment I_t and hence the level of health investment is not determined by the first-order conditon for health investment.

4.3.2 Variation in health and socioeconomic status

In this section I explore the effects of differences in health and socioeconomic status. I employ the first-order condition (4.15) to explore the effects of differences in initial assets (section 4.3.2) and in initial health (section 4.3.2) on the level of health investment I_t.

Variation in initial assets

Consider two optimal life time trajectories, different only (ceteris paribus) in their initial level of assets, A_0, and, $A_0 + \Delta A_0$, and the resulting difference in the two optimal life

[22]As Isaac Ehrlich pointed out to me in a private communication, the co-state variables (Lagrangian multipliers) cannot be a function of the flow of investment because they measure the value of the *stocks* of health capital and monetary wealth, which are not affected by the *flows* of investment in health and earnings, respectively, although they shift with time in current values. The mathematical proof is part of Pontryagin optimal control theory and the maximum principle.

[23]Grossman (2000) has questioned the argument by Ehrlich and Chuma (1990) noting (in a discrete time setting) that the first-order condition for health investment (4.13) equates the current marginal monetary cost of investment in health π_{I_t} (LHS) with a function of all *future* values of the marginal utility of health $\partial U_s/\partial H_s$ and the marginal production benefit of health φ_{H_s} $(t < s \leq T-1)$ (RHS). These in turn are functions of health and health is a function of all *past* values of health investment I_s $(0 \leq s < t$; see equation 4.21). Thus the RHS of the first-order condition for health investment (4.13) is a function of current health investment I_t (and, in fact, all future and all past values as well) and hence a solution for health investment I_t ought to exist. This apparent discrepancy can be reconciled by noting that implicit in the first-order condition for health investment (4.13) is the use of the final period $t = T-1$ as the point of reference, while the relation (4.21) for the health stock uses the initial period $t = 0$ as the point of reference. Consistently using the initial period $t = 0$ as the point of reference, i.e., using the form (4.11) instead of (4.13) for the first-order condition for health investment, one finds that the RHS of (4.11) is not a function of current investment as the health stock is a function of past but not current health investment I_s $(s < t)$. Likewise, consistently using the final period $t = T$ as the point of reference, i.e., using the alternative expression $H_t = H_T/[\prod_{i=t}^{T-1}(1-d_i)] - \sum_{j=t}^{T-1} I_j^{\alpha}/[\prod_{i=t}^{j}(1-d_i)]$ and comparing this with the first-order condition (4.13) one finds that the first-order condition is independent of current health investment I_t.

time trajectories

$$\begin{aligned} q_0^A &\rightarrow q_0^A + \Delta q_{0,A}^A \\ C_t &\rightarrow C_t + \Delta C_{t,A} \\ I_t &\rightarrow I_t + \Delta I_{t,A} \\ H_t &\rightarrow H_t + \Delta H_{t,A}, \end{aligned} \tag{4.23}$$

where $\Delta q_{0,A}^A$, $\Delta C_{t,A}$, $\Delta I_{t,A}$ and $\Delta H_{t,A}$ denote associated shifts in the shadow price of wealth q_0^A and in the optimal solutions for consumption C_t, health investment I_t and health H_t at each age t. A higher capital endowment lowers the shadow price of wealth (i.e., negative $\Delta q_{0,A}^A$). This in turn affects the level of consumption C_t and health investment I_t over the life cycle. Gradually differences in health investment I_t lead to differences in health H_t.

Using a first-order Taylor expansion of the first-order conditions for health investment (4.15) and for consumption (4.18) and eliminating $\Delta C_{t,A}$ we have (for details see Appendix section 4.5.3)

$$\begin{aligned} & \frac{\frac{\partial \pi_I}{\partial I}\big|_{I_t,\,H_t}(d_t+\delta_t)}{\sigma_H|_{I_t,\,H_t} - \varphi_H|_{H_t}} \Delta I_{t,A} \\ + & \left\{ \frac{\frac{\partial \pi_I}{\partial H}\big|_{I_t,\,H_t}(d_t+\delta_t) - \frac{\partial \varphi_H}{\partial H}\Big|_{H_t}}{\sigma_H|_{I_t,\,H_t} - \varphi_H|_{H_t}} - \frac{\frac{\partial^2 U}{\partial H^2}\big|_{C_t,\,H_t}}{\frac{\partial U}{\partial H}\big|_{C_t,\,H_t}} - \frac{\frac{\frac{\partial^2 U}{\partial C \partial H}\big|_{C_t,\,H_t}}{\frac{\partial U}{\partial H}\big|_{C_t,\,H_t}} \left(\frac{\frac{\partial^2 U}{\partial C \partial H}\big|_{C_t,\,H_t}}{\frac{\partial U}{\partial C}\big|_{C_t,\,H_t}} - \frac{\frac{\partial \pi_C}{\partial H}\big|_{C_t,\,H_t}}{\pi_C|_{C_t,\,H_t}} \right)}{\frac{\frac{\partial \pi_C}{\partial C}\big|_{C_t,\,H_t}}{\pi_C|_{C_t,\,H_t}} - \frac{\frac{\partial^2 U}{\partial C^2}\big|_{C_t,\,H_t}}{\frac{\partial U}{\partial C}\big|_{C_t,\,H_t}}} \right\} \Delta H_{t,A} \\ = & - \left(\frac{\frac{\frac{\partial^2 U}{\partial C \partial H}\big|_{C_t,\,H_t}}{\frac{\partial U}{\partial H}\big|_{C_t,\,H_t}}}{\frac{\frac{\partial \pi_C}{\partial C}\big|_{C_t,\,H_t}}{\pi_C|_{C_t,\,H_t}} - \frac{\frac{\partial^2 U}{\partial C^2}\big|_{C_t,\,H_t}}{\frac{\partial U}{\partial C}\big|_{C_t,\,H_t}}} + 1 \right) \frac{\Delta q_{0,A}^A}{q_0^A}. \end{aligned} \tag{4.24}$$

The relation (4.24) describes the change in the level of health investment at age t of a trajectory with initial assets $A_0 + \Delta A_0$ compared to a trajectory with initial assets A_0. On the RHS the coefficient of the relative change in the shadow price of wealth $\Delta q_{0,A}^A / q_0^A$ consists of a first-order (direct) effect of a change in wealth (the factor 1) and a second-order (indirect) effect operating through the effect that a corresponding change in consumption has on the level of health investment (the remaining term). Assuming the first-order effect dominates, the term on the RHS is positive because an increase in assets (positive ΔA_0) decreases the shadow price of wealth (negative $\Delta q_{0,A}^A$).[24]

[24]The second-order term on the RHS equals the relative change in the marginal utility of health $\partial U/\partial H|_{C_t,H_t}$ resulting from variation in consumption C_t (numerator) divided by the relative change in the marginal cost of consumption $\pi_C|_{C_t,H_t}$ minus the marginal benefit (utility) of consumption $\partial U/\partial C|_{C_t,H_t}$,

On the LHS we have a term in $\Delta I_{t,A}$ and one in $\Delta H_{t,A}$. The coefficient of the term in $\Delta I_{t,A}$ equals the relative change in the cost of maintaining the health stock $\sigma_H|_{I_t,H_t} - \varphi_H|_{H_t}$ resulting from variation in the level of health investment I_t.[25] The marginal cost of health investment π_{I_t} increases with the level of health investment for a DRTS health production process. As a result the coefficient of the term in $\Delta I_{t,A}$ is positive.

Consider the initial period $t = 0$. Because health at $t = 0$ is given by the initial condition H_0 we have $\Delta H_{0,A} = 0$ (differences in health between the trajectories with initial assets $A_0 + \Delta A_0$ and with A_0 occur at later ages). Because $\Delta H_{0,A} = 0$ an increase in assets (which lowers the shadow price of wealth, i.e., negative $\Delta q^A_{0,A}$) increases the level of initial health investment, i.e. positive $\Delta I_{0,A}$ (see equation 4.24).

A simple graph helps to illustrate this result. Figure 4.3 shows a stylized representation of the first-order condition for inital health investment I_0 (4.15) as a function of I_0. A higher initial endowment of capital (positive ΔA_0) lowers the shadow price of wealth (negative $\Delta q^A_{0,A}$), thus shifting the net cost of maintaining the health stock downward (curve labeled $(q^0_A + \Delta q^A_{0,A})(\sigma_H - \varphi_H)|_{I_0+\Delta I_{0,A},H_0}$; first-order effect). A lower shadow price of wealth also increases the initial level of consumption C_0,[26] potentially affecting the marginal utility of health (second-order effect). If consumption and health are complements $\partial^2 U/\partial C \partial H|_{C_t,H_t} > 0$ in utility, the marginal utility of health shifts upward (curve labeled $\partial U/\partial H|_{C_0+\Delta C_{0,A},H_0}$). The net result is a higher level of initial health investment $I_0 + \Delta I_{0,A}$.

A higher initial endowment of capital (positive ΔA_0) initially induces individuals to invest more in health. As a result their health deteriorates slower. This addresses the criticism of Case and Deaton (2005) that health production models do not predict differences in the effective health deterioration rate with wealth.

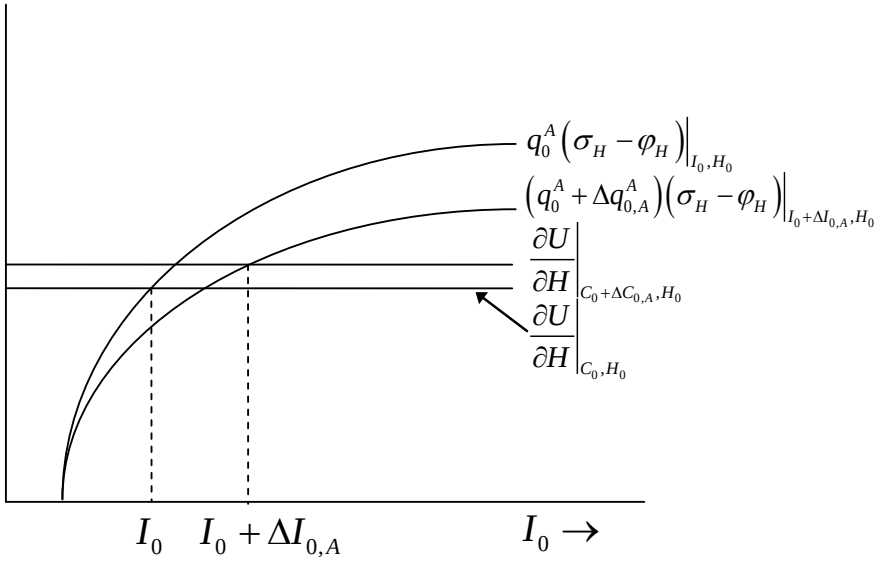
Now consider the next period ($t = 1$). Because of higher health investment $\Delta I_{0,A}$ in the initial period ($t = 0$) health will be higher in the next period $\Delta H_{1,A} > 0$ ($t = 1$). If the

resulting from variation in consumption C_t (denominator). For the usual assumptions of a Cobb-Douglas consumption process and diminishing marginal utility of consumption we have $\partial \pi_C/\partial C|_{C_t,H_t} = 0$ and $\partial^2 U/\partial C^2|_{C_t,H_t} < 0$. In this case the sign of the second-order term on the RHS depends on whether consumption and health are complements $\partial^2 U/\partial C \partial H|_{C_t,H_t} > 0$ or substitutes $\partial^2 U/\partial C \partial H|_{C_t,H_t} < 0$ in utility. Research by Finkelstein, Luttmer and Notowidigdo (2008) suggests that the marginal utility of consumption declines as health deteriorates, i.e. that $\partial^2 U/\partial C \partial H|_{C_t,H_t} > 0$, in which case the second-order term is also positive.

[25]Note that $\partial \sigma_H/\partial I|_{I_t,H_t} - \partial \varphi_H/\partial I|_{H_t} = \partial \pi_I/\partial I|_{I_t,H_t}(d_t + \delta_t) - \partial \Delta \pi_I/\partial I|_{I_t,H_t}(1 + \delta_t) \sim \partial \pi_I/\partial I|_{I_t,H_t}(d_t + \delta_t)$.

[26]See equation (4.77) and note once more that for the usual assumptions of a Cobb-Douglas consumption process and diminishing marginal utility of consumption we have $\partial \pi_C/\partial C|_{C_t,H_t} = 0$ and $\partial^2 U/\partial C^2|_{C_t,H_t} < 0$. Further, $\partial U/\partial C|_{C_t,H_t} > 0$ and, for $t = 0$, $\Delta H_{0,A} = 0$.

Figure 4.3: Differences in initial assets.



Notes: Marginal consumption $\partial U/\partial H$ and marginal production benefit φ_H of health versus the user cost of health capital at the margin σ_H as a function of initial health investment I_0.

level of health investment remains higher in subsequent periods, both health trajectories will start to deviate, i.e. $\Delta H_{t,A}$ would grow over time. How would this affect the level of health investment?

The coefficient of $\Delta H_{t,A}$ consists of a first-order effect (the first and second terms) and a second-order effect (the third term). The first term is equal to the relative change in the cost of maintaining the health stock $\sigma_H|_{I_t,H_t}-\varphi_H|_{H_t}$ resulting from variation in health H_t. The marginal cost of health investment $\pi_I|_{I_t,H_t}$ increases with the wage rate (opportunity cost of investing in health and not working) which potentially increases with health (healthy individuals are more productive), i.e. $\partial\pi_I/\partial H|_{I_t,H_t}>0$. Diminishing marginal benefits of health imply $\partial\varphi_H/\partial H|_{H_t}<0$. Thus the first term is positive. The second term equals the relative change in the marginal consumption benefit (utility) of health $\partial U/\partial H|_{C_t,H_t}$ resulting from variation in health H_t. The second term is also positive for the usual assumption of diminishing marginal utility of health $\partial^2U/\partial H^2|_{C_t,H_t}<0$. Thus both first-order terms are positive.[27] As a result, the difference in the demand for

[27]The third term, describing a second-order effect, contains the same expression as the second-order term in the coefficient of the relative change in the shadow price of wealth $\Delta q_{0,A}^A/q_0^A$ (which, following the earlier discussion in section 4.3.2, is plausible positive) multiplied by the relative change in the marginal utility of health minus the relative change in the marginal cost of consumption in response to a variation in health: $(\partial^2U/\partial C\partial H|_{C_t,H_t}/(\partial U/\partial H)|_{C_t,H_t}-(\partial\pi_C/\partial H)|_{C_t,H_t}/\pi_C|_{C_t,H_t}$. The marginal cost of consumption $\pi_C|_{C_t,H_t}$ increases with the wage rate (opportunity cost of devoting own time to con-

health investment becomes smaller (smaller $\Delta I_{t,A}$) as the deviation in health between the trajectories with initial assets $A_0 + \Delta A_0$ and with A_0 grows (growing $\Delta H_{t,A}$; see equation 4.24). Greater health reduces the demand for health investment (see also the discussions in sections 4.3.2 and 4.3.3). At some age the difference between the level of health investment could vanish ($\Delta I_{t,A} \sim 0$) and the effective health deterioration rate $H_{t+1} - H_t$ converge between the trajectory with initial assets $A_0 + \Delta A_0$ and with A_0.[28] Despite this convergence, given similar initial endowed health H_0 and an initial period of higher levels of health investment, individuals with greater endowed wealth remain healthier.

Other indicators of socioeconomic status such as life-time earnings and education behave qualitatively similar to endowed wealth (initial assets). The exploration of the effect of variations in these measures on health investment and health is outside the scope of this paper (but see section 4.3.3 and Galama and van Kippersluis [2010] for a discussion of the role of life-time earnings and education). The effect of greater earnings over the life cycle on health differs from the effect of greater endowed wealth in that the "wealth" effect is moderated by the higher opportunity cost of time. The effect of education on health is similar to that of greater earnings over the life cycle, but with the additional effect of increasing the efficiency of health investment.

Variation in initial health

Consider two optimal life time trajectories, different only (ceteris paribus) in their initial level of health, H_0, and, $H_0 + \Delta H_0$, and the resulting difference in initial ($t = 0$) health investment I_0

$$\begin{aligned} I_0 &\rightarrow I_0 + \Delta I_{0,H} \\ C_0 &\rightarrow C_0 + \Delta C_{0,H} \\ q_0^A &\rightarrow q_0^A + \Delta q_{0,H}^A, \end{aligned} \tag{4.25}$$

sumption and not working) which potentially increases with health (healthy individuals are more productive). If consumption and health are strong complements in utility $(\partial^2 U/\partial C \partial H)|_{C_t,H_t}/(\partial U/\partial H)|_{C_t,H_t} > (\partial \pi_C/\partial H|_{C_t,H_t})/\pi_C|_{C_t,H_t}$ the third term is positive and results in an elevated level of health investment (compared to a situation where there is weak complementarity or substitutability in utility) in response to a higher health stock (positive $\Delta H_{t,A}$).

[28] Note that if at some age the difference in health investment $\Delta I_{t,A}$ becomes negative, i.e., an individual with greater endowed wealth ($\Delta A_0 > 0$) would spend less on health ($\Delta I_{t,A} < 0$), the health difference in the next period $t+1$ is reduced (smaller $\Delta H_{t+1,A}$), which leads to a less negative or positive difference in the level of health investment $\Delta I_{t+1,A}$, suggesting a process of gradual convergence in the effective rate of health deterioration $H_{t+1} - H_t$ (where we have a relatively constant $\Delta H_{t,A}$ and small $\Delta I_{t,A}$).

where $\Delta I_{0,H}$, $\Delta C_{0,H}$ and $\Delta q^A_{0,H}$ denote associated shifts in the optimal solution for initial health investment I_0, initial consumption C_0 and in the shadow price of wealth q^A_0.

Using a first-order Taylor expansion of the first-order conditions for health investment (4.15) and for consumption (4.18), eliminating $\Delta C_{0,H}$, and (in order to simplify the discussion) omitting second-order effects, we have

$$\begin{aligned} & \frac{\frac{\partial \pi_I}{\partial I}\big|_{I_0,\,H_0}\,(d_0+\delta_0)}{\sigma_H|_{I_0,\,H_0}-\varphi_H|_{H_0}}\Delta I_{0,H} \\ = & -\left\{\frac{\frac{\partial \pi_I}{\partial H}\big|_{I_0,\,H_0}\,(d_0+\delta_0)-\frac{\partial \varphi_H}{\partial H}\big|_{H_0}}{\sigma_H|_{I_0,\,H_0}-\varphi_H|_{H_0}}-\frac{\frac{\partial^2 U}{\partial H^2}\Big|_{C_0,\,H_0}}{\frac{\partial U}{\partial H}\big|_{C_0,\,H_0}}\right\}\Delta H_0, \end{aligned} \tag{4.26}$$

(this relation can be obtained by considering 4.24 for $t=0$ and labeling the variations with H instead of A).[29] The relation (4.26) describes the change in the initial level of health investment I_0 of a trajectory with initial health $H_0+\Delta H_0$ compared to a trajectory with initial health H_0. Under the usual assumptions the first-order relation between health and health investment is negative.[30]

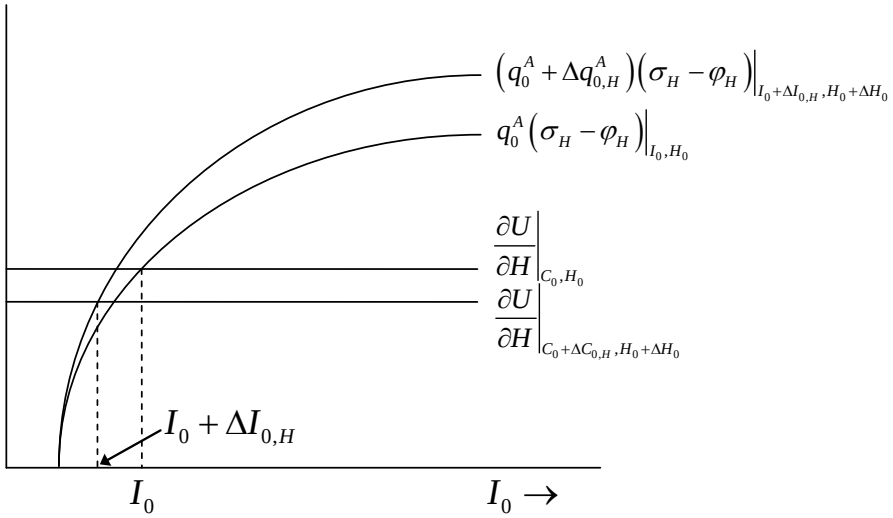
A simple graph helps to illustrate this result. Figure 4.4 shows a stylized representation of the first-order condition for the inital level of health investment I_0 (4.15) as a function of I_0. A higher initial endowment of health (positive ΔH_0) lowers the marginal production benefit of health φ_{H_t} thus shifting the net cost of maintaining the health stock upward (curve labeled $(q^0_A+\Delta q^A_{0,H})(\sigma_H-\varphi_H)|_{I_0+\Delta I_{0,H},H_0+\Delta H_0}$; first-order effect). Further, the marginal utility of health is lower for higher health (the curve labeled $\partial U/\partial H|_{C_0+\Delta C_{0,H},H_0+\Delta H_0}$ shifts downward) as a result of the diminishing marginal utility of health (first-order effect). The net result is a lower level of health investment (negative $\Delta I_{0,H}$).

Greater initial health (positive ΔH_0) reduces the initial demand for health investment (negative $\Delta I_{0,H}$). Because one can start the optimization problem at any age by redefining the initial conditions H_0 and A_0 for that age, this result holds for any age. Thus the theory

[29] A higher level of initial health (positive ΔH_0) enables a higher level of earnings (the production benefit of health), thereby raising life-time earnings and lowering the shadow price of wealth (negative $\Delta q^A_{0,H}$). A higher level of health would thus increase the level of health investment through its effect on wealth. This wealth effect is however a second order-effect in the sense that it operates through the effect of health on wealth, and therefore omitted from (4.26).

[30] The marginal cost of health investment increases with health investment for a DRTS health production process (i.e., $(\partial \pi_I/\partial I)|_{I_0,H_0}>0$) and with the wage rate (opportunity cost of investing in health and not working) which potentially increases with health (healthy individuals are more productive; i.e., $(\partial \pi_I/\partial H)|_{I_0,H_0}>0$). Diminishing marginal production benefit of health implies $(\partial \varphi_H/\partial H)|_{H_0}<0$ and diminishing marginal consumption benefit (utility) of health implies $\partial^2 U/\partial H^2|_{C_0,H_0}<0$.

Figure 4.4: Differences in initial health.



Notes: Marginal consumption $\partial U/\partial H$ and marginal production benefit φ_H of health versus the user cost of health capital at the margin σ_H as a function of initial health investment I_0.

predicts a negative relation between health and health investment in cross-section (see for more details section 4.3.3). This addresses the criticism by Zweifel and Breyer (1997).

4.3.3 Structural equations

Empirical tests of health production models have thus far been based on structural and reduced form equations derived under the assumption of a CRTS health production process.[31] Because these structural and reduced form relations suffer from the issue of the indeterminacy of health investment (see section 4.3.1), I derive in this section structural relations for the DRTS health production process presented in this paper.

Simple functional forms

In order to obtain expressions suitable for empirical testing we have to assume functional forms for model functions and parameters that cannot be observed directly, such as the health investment production process I_t and the biological aging rate d_t.

I specify the following constant relative risk aversion (CRRA) utility function:

$$U(C_t, H_t) = \frac{1}{1-\rho}\left(C_t^{\zeta} H_t^{1-\zeta}\right)^{1-\rho}, \tag{4.27}$$

[31] E.g., Grossman (1972a), Wagstaff (1986a), van Doorslaer (1987), Leu and Gerfin (1992), Nocera and Zweifel (1998), Erbsland, Ried and Ulrich (2002).

where ζ ($0 \leq \zeta \leq 1$) is the relative "share" of consumption versus health and ρ ($\rho > 0$) the coefficient of relative risk aversion. This functional form can account for the observation that the marginal utility of consumption declines as health deteriorates (e.g., Finkelstein, Luttmer and Notowidigdo, 2008) which would rule out strongly separable functional forms for the utility function, where the marginal utility of consumption is independent of health.

I make the usual assumption that sick time is a power law in health

$$s_t = \Omega \left(\frac{H_t}{H_{\min}} \right)^{-\gamma}, \tag{4.28}$$

where $\gamma > 0$ so that sick time decreases with health. This choice of functional form has the properties $\lim_{H_t \to \infty} s_t = 0$ and $\lim_{H_t \downarrow H_{\min}} s_t = \Omega$, where Ω is the total time budget as in (4.4).

Using equation (4.8) we have:

$$\begin{aligned} \varphi_{H_t} &= w_t \gamma \Omega H_{\min}^{\gamma} H_t^{-(1+\gamma)}, \\ &\equiv w_t \Omega^* H_t^{-(1+\gamma)}. \end{aligned} \tag{4.29}$$

Investment in medical care I_t is assumed to be produced by combining own time and goods/services purchased in the market according to a Cobb-Douglas CRTS production function (Grossman, 1972a, 1972b, 2000)

$$I_t = \mu_{I_t} m_t^{1-k_I} \tau_{I_t}^{k_I}, \tag{4.30}$$

where μ_{I_t} is an efficiency factor and $1 - k_I$ and k_I are the elasticities of investment in health I_t with respect to goods and services m_t purchased in the market (e.g., medical care) and with respect to own-time τ_{I_t}, respectively.

Analagously, consumption C_t is assumed to be produced by combining own time and goods/services purchased in the market according to a Cobb-Douglas CRTS production function

$$C_t = \mu_{C_t} X_t^{1-k_C} \tau_{C_t}^{k_C}, \tag{4.31}$$

where μ_{C_t} is an efficiency factor and $1 - k_C$ and k_C are the elasticities of consumption C_t with respect to goods and services X_t purchased in the market and with respect to own-time τ_{C_t}, respectively.

Following Grossman (1972a, 1972b, 2000) I assume that the more educated are more efficient consumers and producers of health investment (based on the interpretation of education as a productivity factor in own time inputs and in identifying and seeking effective care)

$$\mu_{I_t} = \mu_{I_0} e^{\rho_I E}, \tag{4.32}$$

where E is the level of education (e.g., years of schooling) and ρ_I is a constant.

Further, following Galama and van Kippersluis (2010) I assume a Mincer-type wage equation in which the more educated and more experienced earn higher wages (Mincer 1974)

$$w_t = w_E e^{\rho_w E + \beta_x x_t - \beta_{x^2} x_t^2}, \tag{4.33}$$

where education E is expressed in years of schooling, x_t is years of working experience, and ρ_w, β_x and β_{x^2} are constants, assumed to be positive.

Lastly, following Wagstaff (1986a) and Cropper (1981) I assume the biological aging rate d_t to be of the form

$$d_t = d_\bullet e^{\beta_t t + \boldsymbol{\beta_\xi \xi_t}}, \tag{4.34}$$

where $d_\bullet \equiv d_0 e^{-\boldsymbol{\beta_\xi \xi_0}}$ and $\boldsymbol{\xi_t}$ is a vector of environmental variables (e.g., working and living conditions, hazardous environment, etc) that affect the biological aging rate. The vector $\boldsymbol{\xi_t}$ may include other exogenous variables that affect the biological aging rate, such as education (Muurinen, 1982).

Structural relation between health and medical care

A structural relation for the demand for medical goods and services m_t can be obtained from the first-order conditions for health investment (4.15) and for consumption (4.18) and the functional relations defined in the previous section 4.3.3 (see section 4.5.4 in the Appendix for details)

$$b_{it}^1 m_{it}^{1-\alpha} - (1-\alpha) m_{it}^{1-\alpha} \widetilde{m_{it}} \;=\; b_{it}^2 H_{it}^{-1/\chi} + b_{it}^3 H_{it}^{-(1+\gamma)}, \tag{4.35}$$

where I have defined the following functions

$$b_{it}^1 \;\equiv\; \left[d_\bullet e^{\beta_t t_i + \boldsymbol{\beta_\xi \xi_{it}}} + \delta - (1-\alpha k_I)\widetilde{p_{m_{it}}} - \alpha k_I \widetilde{w_{it}} \right], \tag{4.36}$$

$$b_{it}^2 \;\equiv\; b_*^2 \left(q_{0i}^A\right)^{-1/\rho\chi} e^{\alpha \rho_I E_i} p_{m_{it}}^{-(1-\alpha k_I)} w_{it}^{-[k_C(1/\rho\chi - 1) + \alpha k_I]} p_{X_{it}}^{-(1-k_C)(1/\rho\chi - 1)} \left(\frac{1+\beta_i}{1+\delta} \right)^{-t_i/\rho\chi} \tag{4.37}$$

$$b_{it}^3 \;\equiv\; b_*^3 e^{\alpha \rho_I E_i} p_{m_{it}}^{-(1-\alpha k_I)} w_{it}^{1-\alpha k_I}, \tag{4.38}$$

and the following constants

$$b_*^2 \;\equiv\; [(1-\zeta)\Lambda]^{1/\chi} \alpha k_I^{\alpha k_I} (1-k_I)^{1-\alpha k_I} \mu_{I_0}^\alpha \left[k_C^{k_C} (1-k_C)^{1-k_C} \mu_{C_t} \right]^{1/\rho\chi - 1}, \tag{4.39}$$

$$b_*^3 \;\equiv\; \alpha k_I^{\alpha k_I} (1-k_I)^{1-\alpha k_I} \mu_{I_0}^\alpha \Omega_*, \tag{4.40}$$

$$\Lambda \;\equiv\; \zeta^{\frac{1-\rho}{\rho}} \left(\frac{\zeta}{1-\zeta} \right)^{1-\chi}, \tag{4.41}$$

$$\chi \;\equiv\; \frac{1+\rho\zeta-\zeta}{\rho}, \tag{4.42}$$

where the subscript i indexes the i^{th} individual, and where the notation $\widetilde{f_t}$ is used to denote the relative change $\widetilde{f_t} \equiv 1 - \frac{f_{t-1}}{f_t}$ in a function f_t. Further, I have assumed small

relative changes (much smaller than one) in the price of medical care $\widetilde{p_{m_{it}}}$, wages $\widetilde{w_{it}}$ and the efficiency of the health investment process $\widetilde{\mu_{I_{it}}}$ and, for simplicity, assumed a constant discount factor $\beta_t = \beta$ and constant rate of return to capital $\delta_t = \delta$.

A similar expression for own-time inputs $\tau_{I_{it}}$ can be obtained using (4.83). Further, one can substitute the expression (4.33) for the wage rate w_{it} to obtain an expression in terms of years of schooling E_i and years of experience x_{it}.

Pure investment and pure consumption models

Analytical solutions to the Grossman model are usually based on two sub-models (1) the "pure investment" model in which the restriction $\partial U_t/\partial H_t = 0$ is imposed and (2) the "pure consumption" model in which the restriction $\partial Y_t/\partial H_t = 0$ is imposed. In this section I explore the characteristics of these two sub models for the following reasons. First, the two sub models represent two essential characteristics of health: health as a means to produce (investment) and health as a means to provide utility (consumption) and exploring them separately provides insight into these two distinct properties of health. Second, these restrictions allow one to obtain linearized structural expressions. Last, the two sub-models are widely used in the health production literature and exploring them allows for comparisons with previous research.

In the pure investment model health does not provide utility and hence $\zeta = 1$ (see equation 4.27) and $b^2_{it} = 0$, whereas in the pure consumption model health does not provide a production benefit and hence $\varphi_{H_{it}} = 0$ and $b^3_{it} = 0$. We can obtain a structural linear relation for the demand for health investment goods / services m_{it} in the pure investment and pure consumption models as follows.

Pure investment

For small $\widetilde{m_{it}}$ and $b^2_{it} = 0$ we have (see equation 4.35)

$$\begin{aligned}(1-\alpha)\ln m_{it} &\sim \ln b^3_{it} - \ln b^1_{it} - (1+\gamma)\ln H_{it},\\ &= \ln b^3_* - (1+\gamma)\ln H_{it} + \alpha\rho_I E_i - (1-\alpha k_I)\ln p_{m_{it}} + (1-\alpha k_I)\ln w_{it}\\ &- \ln d_\bullet - \beta_{it}t - \boldsymbol{\beta_\xi \xi_{it}} - \ln\left\{1 + \left[\frac{\delta - (1-\alpha k_I)\widetilde{p_{m_{it}}} - \alpha k_I \widetilde{w_{it}}}{d_\bullet e^{\beta_t t_i + \boldsymbol{\beta_\xi \xi_{it}}}}\right]\right\}. \qquad (4.43)\end{aligned}$$

Pure consumption

For small $\widetilde{m_{it}}$ and $b_{it}^3 = 0$ we have (see equation 4.35)

$$\begin{aligned}(1-\alpha)\ln m_{it} &\sim \ln b_{it}^2 - \ln b_{it}^1 - \frac{1}{\chi}\ln H_{it},\\ &= \ln b_*^2 - \frac{1}{\chi}\ln H_{it} - \frac{1}{\rho\chi}\ln q_{0i}^A + \alpha\rho_I E_i - (1-\alpha k_I)\ln p_{m_{it}}\\ &- \left[k_C\left(1/\rho\chi - 1\right) + \alpha k_I\right]\ln w_{it} - \left(1-k_C\right)\left(1/\rho\chi - 1\right)\ln p_{X_{it}} - \ln d_\bullet - \beta_t t_i - \boldsymbol{\beta_\xi \xi_{it}}\\ &- \frac{1}{\rho\chi}\left[\ln(1+\beta_i) - \ln(1+\delta)\right]t_i - \ln\left\{1 + \left[\frac{\delta - (1-\alpha k_I)\widetilde{p_{m_{it}}} - \alpha k_I \widetilde{w_{it}}}{d_\bullet e^{\beta_t t_i + \boldsymbol{\beta_\xi \xi_{it}}}}\right]\right\}. \qquad (4.44)\end{aligned}$$

It is customary to assume that the term $\ln d_\bullet$ in equations (4.43) and (4.44) is an error term with zero mean and constant variance $\xi_1(t) \equiv -\ln d_\bullet$ (as in Wagstaff, 1986a, and Grossman, 1972a, 1972b, 2000) and that the term $\ln[1 + \delta_t/d_t - \widetilde{\pi_{I_t}}/d_t]$ (the last term in equations 4.43 and 4.44) is small or constant (see, e.g., Grossman, 1972a, 2000),[32] or that it is time dependent $\ln[1 + \delta_t/d_t - \widetilde{\pi_{I_t}}/d_t] \propto t$ (e.g., Wagstaff, 1986a).

Reduced form relations

The solution for the health stock H_t follows from the dynamic equation (4.2) and using expressions (4.32) and (4.82)

$$H_t = H_0\prod_{j=0}^{t-1}(1-d_j) + \left(\frac{1-k_I}{k_I}\right)^{-\alpha k_I}\mu_{I_0}^{\alpha}e^{\alpha\rho_I E}\sum_{j=0}^{t-1}p_{m_j}^{\alpha k_I}w_j^{-\alpha k_I}m_j^{\alpha}\prod_{k=j+1}^{t-1}(1-d_k), \qquad (4.45)$$

where I have suppressed the index i for the individual.

The health stock H_t is a function of past levels of consumption of medical goods / services m_j $(j \leq t-1)$ and past biological aging rates d_j $(j \leq t-1)$. In principle one can obtain reduced form expressions for the health stock H_t[33] and for the demand for medical goods / services m_t.[34] This excercise, however, results in complex expressions with arguably limited value for empirical analyses. The reduced form solutions for the health stock H_t and the demand for medical goods / services m_t are functions of the

[32] This would require that the rate of return to capital δ_t and changes in the wage rate w_t and the price p_{m_t} and efficiency μ_{I_t} of health investment goods/services in producing health investment are much smaller than the health deterioration rate d_t or that such changes follow the same pattern as changes in d_t (so that the term is approximately constant).

[33] Substitute the solutions for past consumption of medical goods / services m_j $(j \leq t-1)$ obtained from (4.35) in (4.45) and recursively substitute the expression for the health stock (4.45) for past values of the health stock to obtain an expression for the health stock from which past levels of the health stock H_j $(j \leq t-1)$ and past values of consumption of medical goods / services m_j $(j \leq t-1)$ are removed (with the exception of initial health H_0).

[34] Use (4.35) and recursively substitute the expression for the health stock (4.45) to obtain an expression from which past consumption of medical goods / services m_j $(j \leq t-1)$ and past levels of the health stock H_j $(j \leq t-1)$ are removed (again with the exception of initial health H_0).

initial health stock H_0, wealth q_0^A (endowed assets and life-time earnings) and the *history* of past prices of medical care p_{m_s}, past prices of consumption goods / services p_{X_s}, past wage rates w_s, past biological aging rates d_s and past rates of return to capital δ_s ($s < t$). In addition, the demand for medical goods / services is also a function of the current price of medical care p_{m_t}, the current price of consumption goods / services p_{X_t}, the current wage rate w_t, the current biological aging rate d_t and the current rate of return to capital δ_t (the health stock does not depend on current values).

Discussion

The structural form (4.35) of the first order condition for health investment describes a direct relationship between the demand for health investment goods / services m_t (e.g., medical care), the relative change in the demand for health investment goods / services $\widetilde{m_t}$ and the health stock H_t. For slow changes in the demand for health investment goods / services with time (small $\widetilde{m_t}$), the demand for health investment goods / services m_t falls with the level of health H_t. This is further reflected in the elasticity of health investment goods / services with respect to health H_t, which, for small $\widetilde{m_t}$, is negative (and a function of the health stock H_t)

$$\sigma_{m_t,H_t} = \frac{\partial m_t}{\partial H_t}\frac{H_t}{m_t} = -\frac{1}{1-\alpha}\left[\frac{\frac{1}{\chi}b_t^2 H_t^{-1/\chi} + (1+\gamma)b_t^3 H_t^{-(1+\gamma)}}{b_t^2 H_t^{-1/\chi} + b_t^3 H_t^{-(1+\gamma)}}\right], \tag{4.46}$$

where I have suppressed the index i for the individual. Similarly, the elasticity of health investment goods / services m_t with respect to health H_t (see equation 4.46) for the pure investment model

$$\sigma_{m_t,H_t}^{PI} = -\frac{1+\gamma}{1-\alpha}, \tag{4.47}$$

and the pure consumption model

$$\sigma_{m_t,H_t}^{PC} = -\frac{1}{\chi(1-\alpha)}, \tag{4.48}$$

are negative, where the labels PI and PC refer to the pure investment and pure consumption model, respectively. In other words, I find that the less healthy demand more and the healthy demand less medical goods / services. This prediction from the theoretical model is in line with what has been observed in numerous empirical studies and addresses the criticism by Zweifel and Breyer (1997).

Assuming that both medical goods / services m_t and time input τ_{I_t} increase health investment suggests $0 \leq k_I \leq 1$ (see equation 4.30), and if education E increases the

efficiency of medical care then $\rho_I > 0$ (see equation 4.32). Similarly we have $0 \leq k_C \leq 1$ (see equation 4.31).

For these assumptions and small changes $\widetilde{m_t}$, the demand for health investment goods/services m_t (see relations 4.35 and 4.36) decreases with the biological aging rate d_t (and hence with environmental factors that are detrimental to health $\boldsymbol{\xi_t}$), the rate of return to capital δ_t (an opportunity cost – individuals can invest in health or in the stock market) and increases with price increases $\widetilde{p_{m_t}}$ and wage increases $\widetilde{w}_t$ (it is better to invest in health now when prices p_{m_t} and the opportunity cost of time w_t are higher in the future). In addition, due to the *consumption* aspect of health (health provides utility) the demand for health investment goods/services m_t (see relations 4.35 and 4.37 or 4.44) increases with wealth $q_0^{A^{-1}}$ (the shadow price of wealth is a decreasing function of wealth and life-time earnings),[35] education E (through assumed greater efficiency of health investment with the level of education) and decreases with the price of health investment goods/services p_{m_t}. For $\rho\chi < 1$ the demand for health investment goods/services m_t decreases with the price of consumption goods/services p_{X_t} (for $\rho\chi > 1$ it increases) and with the wage rate w_t (opportunity cost of time) (for $\rho\chi > 1$ the effect of the wage rate w_t is ambiguous)). And, due to the *production* aspect of health (health increases earnings) the demand for health investment goods/services m_t (see relations 4.35 and 4.38 or 4.43) increases with education E (through assumed greater efficiency of health investment with the level of education) and the wage rate w_t (a higher wage rate increases the marginal production benefit of health, and this outweighs the opportunity cost of time associated with health investment) and decreases with the price of health investment goods/services p_{m_t}.

The above discussion masks important effects of earnings and education. In our model of perfect certainty an evolutionary wage change (along an individual's wage profile) does not affect the shadow price of wealth q_0^A as the change is fully anticipated by the individual. Thus comparing panel data for a single individual may reveal a higher wage rate w_t to be associated with a lower demand for medical goods / services m_t due to a higher opportunity cost of time. However, comparing across individuals, those who currently have a higher wage rate will in most cases also have higher life-time earnings and thus

[35] In principle, an expression for the shadow price of wealth q_0^A can be obtained by using the life-time budget constraint (which follows from integrating the dynamic equation 4.3), substituting the solutions for consumption goods/services X_t, health H_t, and health investment goods/services m_t and solving for q_0^A (see, for example, Galama et al. 2008). In practice, the shadow price of wealth q_0^A cannot be solved analytically: it is a very complicated function of the shadow price of health q_{H_t}, wealth (assets), education E and earnings, wages w_t, prices p_{m_t}, p_{X_t}, and health deterioration rates (terms $d_\bullet$, β_t and $\boldsymbol{\beta_\xi}$) over the life cycle.

a lower shadow price of wealth q_0^A. This wealth effect increases the demand for medical goods / services and competes with the opportunity cost of time effect. Similarly, to account for the effect of education it is important not only to consider the possible effect of a higher efficiency of health investment (the parameter ρ_I), as in the structural relations (4.35), (4.43) and (4.44), but also the effect that education has on earnings (opportunity cost of time effect; see equation 4.33) and in turn on wealth (wealth effect). Plausibly, the wealth effect dominates the opportunity cost effect. For example, Dustmann and Windmeijer (2000) and Contoyannis et al. (2004) find a positive effect on health from a permanent wage increase and a negative effect from a transitory wage increase. We expect then that the effect of education and earnings is to increase the demand for health investment goods / services through a wealth effect that may dominate the opportunity cost of time effect associated with higher earnings.[36]

Thus, in testing the theory it will be important to account for wealth. This can be done by employing measures of wealth (endowed assets, life-time earnings) as proxies for the shadow price of wealth q_0^A or, following Wagstaff (1986a), by utilizing an approximation for q_0^A (his equations 15 and 16).

4.3.4 Numerical simulations

In this section I present simulations of the model with a DRTS health production process and a simple step process. I first discuss the step process for fixed length of life (section 4.3.4). I then illustrate the properties of the model with numerical simulations accounting for endogenous length of life (section 4.3.4).

Step process and fixed length of life

We start with the initial condition for health H_0. Initial consumption C_0 then follows from the first-order condition for consumption (4.18), which, for the assumed functional forms in section 4.3.3, can be written as

$$C_t = \left[\frac{q_0^A}{\zeta}\pi_{C_t}\frac{\prod_{j=1}^{t}(1+\beta_j)}{\prod_{j=1}^{t}(1+\delta_j)}\right]^{-1/\rho\chi} H_t^{\frac{\chi-1}{\chi}}, \tag{4.49}$$

[36]Further, one may be tempted to conclude that individuals invest less in health care during middle and old age because of the high opportunity cost of time associated with high earnings at these ages (see equation 4.33). However, as health deteriorates with age the demand for curative care increases (see sections 4.3.2 and 4.3.3). If the latter effect dominates, the model is capable of reproducing the observation that young individuals invest little, the middle-aged invest more and the elderly invest most in curative care.

where π_{C_t} is given by (4.19). Initial consumption C_0 is a function of initial health H_0, the price of goods and services p_{X_0}, the wage rate w_0 (the opportunity cost of not working) and the shadow price of wealth q_0^A. Next, the initial level of health investment I_0 follows from the initial marginal cost of curative care π_{I_0} (see expression 4.12) which is a function of the Lagrange multiplier q_0^H and the shadow price of wealth q_0^A ($\pi_{I_0} = q_0^A/q_0^H$; see equations 4.59 and 4.60). The initial level of health investment I_0 is (through the initial marginal cost of curative care π_{I_0}) a function of the price of goods and services p_{m_0}, the wage rate w_0, education E, and the multipliers q_0^A and q_0^H. Thus, given exogenous education, prices and wage rates, the initial level of health investment I_0 and the initial level of consumption C_0 are functions of the endogenous Lagrange multipliers q_0^A and q_0^H.

Health in the next period H_1 is determined by the dynamic equation (4.2). Assets in the next period A_1 follow from the initial condition for assets A_0 and the dynamic equation for assets (4.3). For the assumed functional forms in section 4.3.3 we have

$$A_{t+1} = (1+\delta_t)A_t + w_t\left[\Omega - \tau_{I_t}^* I_t - \tau_{C_t}^* C_t - s_t\right] - p_{X_t} X_t^* C_t - p_{m_t} m_t^* I_t, \tag{4.50}$$

where s_t, m_t^*, $\tau_{I_t}^*$, X_t^*, $\tau_{C_t}^*$ are defined in (4.28), (4.82), (4.83), (4.85) and (4.86).

Consumption C_1 follows from the first-order condition for consumption (4.49), health investment I_1 follows from the first-order condition for health investment (4.11), (4.13), (4.14) or (4.15), which for the assumed functional forms in section 4.3.3 can be expressed as

$$\pi_{I_t} = \frac{1}{1-d_t}\left[\pi_{I_{t-1}}(1+\delta_t) - \frac{1-\zeta}{q_0^A} C_t^{1-\rho\chi} H_t^{\rho(\chi-1)-1} - w_t \Omega_* H_t^{-(1+\gamma)}\right]. \tag{4.51}$$

Health H_2 and assets A_2 in the next period are determined by the dynamic equations (4.2) and (4.50) and so on. The solutions for consumption C_t, health H_t and health investment I_t for every period t are functions of the two Lagrange multipliers q_0^A and q_0^H. In the final period, the two end conditions for the final level of health $H_T = H_{\min}$ and the final level of assets A_T determine the Lagrange multipliers q_0^A and q_0^H.[37]

Some have argued that length of life is determined in an iterative process by the condition that health at the end of life H_T equal the minimum health stock $H_{\min}$ (e.g., Grossman, 1998; Reid, 1998). These results are however based on a CRTS health production process and are the result of the indeterminacy of health investment. The results do not hold for a DRTS health production process as advocated here. This can be seen as follows. As the preceding discussion shows, the end conditions $H_T = H_{\min}$ and A_T are

[37] Alternatively one could start with the final period $t = T-1$ and use recursive back substitution. Reaching the initial period, the two initial conditions for health H_0 and assets A_0 determine the Lagrange multipliers q_0^A and q_{T-1}^H.

met for fixed length of life T because the solutions for assets, consumption, health and health investment (and having used the initial conditions H_0 and A_0) are functions of the Lagrange multipliers q_0^A and q_0^H. Applying the end conditions A_T and H_T determines the Lagrange multipliers q_0^A and q_0^H for fixed T. Thus, in the health production literature, as pointed out by Ehrlich and Chuma (1990), length of life T is exogenous (fixed) in the absence of the required terminal (transversality) condition.

Simulations with endogenous mortality

In this section I simulate the model for a particular set of parameter values. The purpose of this exercise is to illustrate some properties of the model. Other parameter choices are possible and a full exploration of the model's properties would require exploring a wide range of parameter values. Ultimately one would like to estimate the model with panel data to test its ability to describe human behavior. This is beyond the scope of this paper.

Figure 4.5 shows the results of model simulations using the step process and equations presented in the above section 4.3.4. In the simulations I have used a period step size of one tenth of a year and assumed annual wages of the form

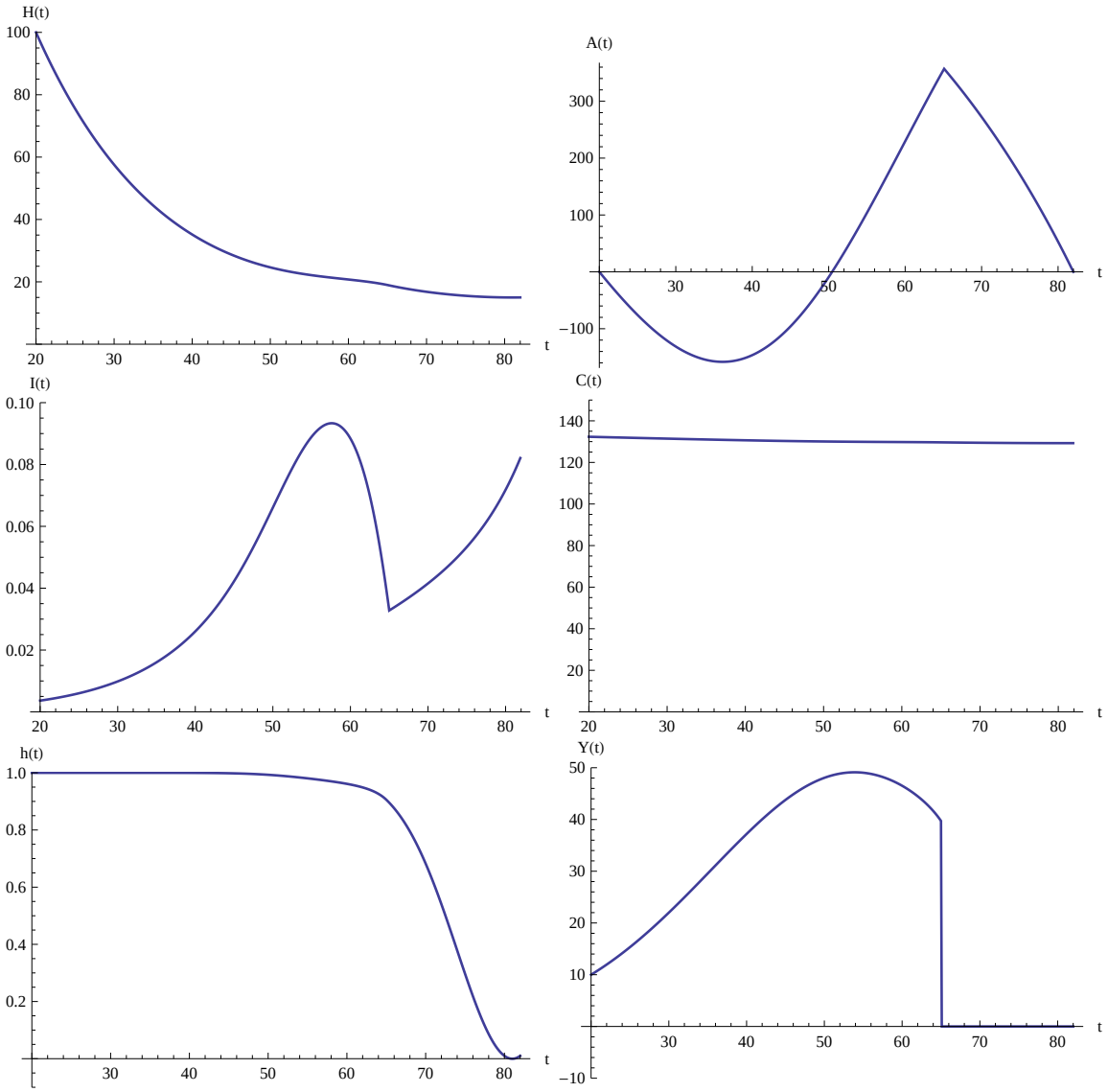
$$w_t \;=\; 10e^{\{1.31383\times 10^{-3}[70(t-20)-(t-20)^2]\}} \;\$ \text{ (thousands)}, \tag{4.52}$$

starting at age $t = 20$ when the individual begins work life until she retires at a fixed retirement age $R = 65$. This Mincer-type wage equation starts with annual wages of \$ 10,000 per year at age $t = 20$ and peaks at \$ 50,000 per year at age $t = 55$ after which it gradualy declines till the age of retirement $R = 65$ after which wages w_t are zero. In addition I use the following parameters: $\alpha = 0.5$, $\gamma = 10$ (sick time increases significantly only upon approaching end of life, i.e., as H_t approaches $H_{\min}$),[38] $H_0 = 100$, $H_T \equiv H_{\min} = 15$, $A_0 = A_T = 0$ \$ (thousands) (no bequests), $\Omega = 0.1$ year (the total time available in a period equals the time step size), $k_I = k_C = 0$ (health investment I_t and consumption C_t consist of purchases in the market, no own time inputs),[39] $p_{m_t} = 0.2$ \$ per medical good/service unit, $p_{X_t} = 0.2$ \$ per consumption good/service unit, $\mu_{I_t} = 0.01$, $\mu_{C_t} = 1$, $\rho = 0.8$, $\zeta = 0.95$ (high relative "weight" of consumption versus health in

[38]Note that this choice of γ allows for a "realistic" relation between health and sick time but does not give rise to large medical expenditures near the end of life. The parameter γ affects sick time, not health, and individuals decide on the level of health investment based on the utility that health provides (note that after retirement there is no production benefit of health).

[39]This simplification helps avoid corner solutions in which the time budget constraint is not satisfied. This is because for this choice healthy time $h_t = \Omega - s_t$ is always positive after retirement, even as s_t approaches Ω as H_t approaches $H_{\min}$. After retirement income Y_t and time spend working τ_{w_t} are zero. Further, for $k_I = k_C = 0$ no time is devoted to health investment $\tau_{I_t} = 0$ or to consumption $\tau_{C_t} = 0$.

Figure 4.5: Simulated profiles for health, assets, health investment, consumption, healthy time and earnings.



Notes: Health ($H(t)$; top-left panel), assets ($A(t)$; \$ thousands; top-right panel), health investment ($I(t)$; center-left panel), consumption ($C(t)$; center-right panel), healthy time ($h(t)$; fraction of total time Ω; bottom-left panel), annual earnings ($Y(t)$; \$ thousands per year; bottom-right panel).

providing utility), a constant aging rate $d_t = d_0 = 0.06$ (per year), a constant return to capital $\delta_t = \delta_0 = 0.03$ (per year) and a constant subjective discount factor $\beta_t = \beta_0 = 0.03$ (per year).

I start with the initial values for health H_0 and assets A_0 and employ the Nelder-Mead method (also called the downhill simplex or amoeba method; Nelder and Mead, 1965) to iteratively determine the shadow price of wealth q_0^A and of health q_0^H that satisfy the

end conditions A_T and H_T. I use the usual values $\alpha_{NM} = 1$, $\gamma_{NM} = 2$, $\rho_{NM} = 0.5$ and $\sigma_{NM} = 0.5$ for the Nelder-Mead reflection, expansion, contraction and shrink coefficients, respectively.

Optimal length of life is determined by maximizing the "indirect utility function" V_T (4.10) with respect to length of life T. I find $T = 82.0$ years.

Health H_t (top-left panel of Figure 4.5) gradually declines with age t and life ends at age $T = 82.0$ years. Health deteriorates somewhat slower during the ages 50 to 65, coinciding with increased levels of health investment I_t (center-left panel of Figure 4.5). The demand for health investment consists of two components. The first component is driven by the production benefit of health and follows a hump shaped pattern similar to the earnings profile Y_t (bottom-right panel of Figure 4.5). Health investment serves to maintain health in order to reduce sick time and hence increase earnings Y_t. Because of the parameter choice $k_I = 0$ there is no opportunity cost of time as the individual does not spend own-time on health investment. As a result, the production benefit of health is roughly proportional to the wage profile (equation 4.52). The second component is driven by the desire of individuals to be healthy (consumption benefit) and to live long lives (increases life-time utility). This component gradually increases with age. Thus the simulation suggests that solutions are feasible in which health investment increases near the end of life.

One possible explanation for the gradual rise in health investment near the end of life is that the simulations suggest that optimal length of life, at least for this set of parameters, coincides with the condition that the change in the health stock with age equal zero at the end of the last period. If the rate of change were positive, health would be below $H_{\min}$ some time before it eventually returned to $H_{\min}$ at the end of the last period $T - 1$, a condition that is not allowed since length of life is defined by the first time an individual's health reaches $H_{\min}$.[40] If the rate of change were negative, adding a period extends life and provides additional utility (again, for this set of parameters). Thus as individuals approach end of life they slow their effective rate of change in health ($H_{t+1} - H_t$ approaches zero) through more and more health investment.

At a price $p_{m_t} = 0.2$ \$ per medical good/services unit her expenditures on health investment goods/services $p_{m_t} m_t$ peak at about \$ 1,800 per year at around age 55. The fact that such humped-shape profiles are generally not observed in medical expenditure data sets, at least not as sizeable as the simulation shows, suggests that the production

[40] To the best of my knowledge the health production literature has failed to observe that the optimization problem allows for solutions where the health stock falls below $H_{\min}$ before the end of life. I explicitly discard such solutions in the numerical simulations.

benefit of health (compared to the consumption benefit of health) may be smaller in real life than is simulated. Again, the simulations are to illustrate the model's characteristics and attempts to estimate the model are left to future research.

The individual's assets A_t (top-right panel in Figure 4.5) initially deplete till about age 50 as she borrows to fund her consumption C_t (center-right panel of Figure 4.5) and health investment I_t needs. She builds up savings between ages 50 and the age of retirement (65) and depletes these savings by end of life. Consumption is relatively constant with age. At a price $p_{X_t} = 0.2$ \$ per consumption good/service unit her expenditures on consumption goods/services $p_{X_t}X_t$ are about \$ 28,000 per year.

Healthy time h_t (bottom-left panel of Figure 4.5) starts to decline rapidly around the age of retirement. While some of this can be explained by a drop in health investment I_t following retirement, this is mostly due to the steep functional relation assumed between health H_t and sick time s_t (equation 4.28 for $\gamma = 10$).

The simulations further show that solutions are feasible for which the biological aging rate is constant, despite the common perception that the biological aging rate needs to increase with age in order to ensure that health falls with age and life is finite (e.g., Grossman, 1972a, 2000; Ehrlich and Chuma, 1990; Case and Deaton, 2005).

4.4 Discussion and conclusions

I have presented a theory of the demand for health, health investment and longevity, building on the human capital framework for health, in particular the work by Grossman (1972a, 1972b) and Ehrlich and Chuma (1990) and related literatures.

My contribution to the health production literature is as follows. First, I argue for a different interpretation of the health stock equilibrium condition, one of the most central relations in the health production literature: this relation determines the optimal level of health investment (not the health stock), conditional on the level of the health stock. Consistently employing the alternative interpretation allows me to simplify the theory and develop the health production literature further than was previously possible. This is because the equilibrium condition for the health stock (4.15) is of a much simpler form than the condition (4.11) which is typically utilized to determine the optimal level of health investment. There are several implications of this interpretation that I discuss in more detail below.

Second, the alternative interpretation of the first-order condition necesitates DRTS in the health production process or no solution for health investment exists. I therefore revisit the debate on the indeterminacy of health investment under the widely used as-

sumption in the health production literature of a CRTS health production process and show that under this assumption the first-order condition for health investment (4.11 or 4.15) is not a function of health investment, and thus health investment is not determined. This widely used assumption represents a degenerate case with problematic properties. While this is no new result (Ehrlich and Chuma, 1990) I provide intuitive, less technical and additional arguments in its support. Revisiting this debate is important because the implications of the indeterminacy are significant and because the debate does not appear to have settled in favor of a DRTS health production process as illustrated by its lack of use in the health production literature. Besides technical reasons that suggest a CRTS health production process is restrictive, the different experiences of developing and developed countries suggest that the economic principle of eventually diminishing returns applies to health production. Quite modest increases in expenditures on health input (food, sanitation) have relatively large impacts on health in the developing world whereas large increases in resources in the developed world have a relatively modest impact on health (e.g., Wagstaff, 1986b).

Third, I explore in detail the implications of the alternative interpretation of the first-order condition and of the properties of a DRTS health production process. In particular the simpler form (4.15) allows me to utilize a stylized representation of the first-order condition for health investment to obtain an intuitive understanding of its properties. I find that for a DRTS health production process and the usual assumptions of diminishing marginal utility and diminishing marginal benefits a unique optimal solution for health investment exists (i.e., the indeterminacy is removed). The optimal level of health investment decreases with the user cost of health capital (i.e., with the price of medical goods / services, the wage rate [the opportunity cost of not working], the biological aging rate and the return to capital [the opportunity cost of investing in, e.g., the stock market rather than in health]) and increases with wealth (endowed assets and life-time earnings) and with the marginal consumption and marginal production benefit of health (because both are decreasing in health, the demand for health investment decreases with health).

Further, I find that for every level of the health stock a unique optimal level of health investment exists. Thus I find no support for the concept of an "optimal" level of the health stock as utilized in the health production literature, in particular the notion that individuals may seek to adjust their health to this "optimal" level in case their health deviates from it. I find that individuals do not aspire to a certain level of health. Instead, given any level of their health stock, individuals decide about the optimal level of health investment. Thus one does not need to assume that any discrepancy between the actual

and the "desired" health stock is dissipated instantaneously and on a continual basis.[41] Theoretically there is no justification for the assumption of a continual adjustment process and empirical work by Wagstaff (1993) suggests non-instantaneous adjustment better describes the health production process.

The simpler form of the first-order condition for health investment (4.15) allows me to investigate the effect of changes in initial conditions such as initial assets, initial health and education on the level of health investment and consumption by studying the effect of variations on the optimal solutions (see section 4.3.2). I find that the wealthy and more educated invest more in health and that their health deteriorates at a slower pace. As a result, given similar initial health endowments, they remain healthier as they age and live longer. Not only does this confirm the directional predictions made earlier by Ehrlich and Chuma's (1990) analysis, but the relations I derive in section 4.3.2 also allow for an understanding of the underlying mechanisms that lead to the predicted outcome. Calbrated simulations by Ehrlich and Yin (2005) of a related model (Ehrlich, 2000) which treats length of life as uncertain, and life expectancy as partly the product of individuals' efforts to self-protect against mortality and morbidity risks also finds that greater endowed wealth and higher wages over the life cycle increase life expectancy.

Further, I find a negative relation (in cross-section) between health and the level of health investment: the healthy demand fewer medical goods / services than the less healthy. This is an important new result that addresses a significant critique of health production models by Zweifel and Breyer (1997; see for more details below).

The simpler form of the first-order condition for health investment (4.15) also allows me to derive structural relations between health and health investment (e.g., medical care) that are suitable for empirical testing. These structural relations contain the CRTS

[41] For the CRTS health production process the model is characterised by a so-called "bang-bang" solution as one has to assume that in the first period individuals adjust their health to its "optimal" level by investing a large positive or negative (depending on the direction of the adjustment) amount of medical care (or other forms of health investment; e.g., Wolfe, 1985; Ehrlich and Chuma, 1990; Grossman, 2000). But even if the health stock is at the "desired" level, health investment (and with it the health stock; see equation 4.2) is still undetermined as any level of investment is allowed (see discussion in section 4.3.1). Further, even if at some age t an individual's health stock is at the "desired" level H_t^*, it is not guaranteed that the health stock will subsequently evolve along this particular health path H_s^* (ages $s > t$) because for a given level of health both the marginal benefit of health $\partial U_t/\partial H_t$ and the cost of maintaining the health stock $q_0^A(\sigma_{H_t} - \varphi_{H_t})$ are determined by exogenous parameters and there is no mechanism to ensure that the two are equal. Thus in a formulation with a CRTS health production process one has to assume that any discrepancy between the actual and the "desired" health stock is dissipated not just once but on a continual basis.

health production process as a special case, thereby allowing empirical tests to verify or disproof this common assumption in the health production literature.

Finally, I show that for a DRTS health production process length of life is not endogenously determined and that an additional condition for optimal length of life is needed (see also, Ehrlich and Chuma, 1990; Seierstad and Sydsaeter, 1987; Kirk, 1970). I numerically solve the model to properly include the role of endogenous length of life. These simulations show that for plausible parameters health investment increases near the end of life and that length of life is finite as a result of limited life-time resources (the budget constraint) and if medical technology cannot fully offset biological aging.

Thus I find that a DRTS health production process addresses five consistent criticisms of the characteristics and predictions of health production models that have been made in the literature. First, as Ehrlich and Chuma (1990) have also shown, by introducing DRTS in the health production process the indeterminacy problem of health investment is addressed.

Second, I have shown that a DRTS health production process is capable of reproducing the observed negative relation between health and the demand for medical care (sections 4.3.2 and 4.3.3), addressing the criticism by e.g. Wagstaff (1986a) and Zweifel and Breyer (1997). This result follows directly from the first-order condition for health investment (as one would expect for such a fundamental feature of the demand for medical care). A CRTS health production process, on the other hand, predicts that the relation between health and investment in health is positive. In other words, the healthy are those that invest more in health (e.g., equation 13 in Wagstaff, 1986a; see also Galama and Kapteyn, 2009). Empirical studies strongly reject this prediction: the negative relationship between health and medical care is found to be the most statistically significant of any relationship between medical care and any of the independent variables in several empirical studies (see, e.g., Grossman, 1972a; Wagstaff, 1986a, 1993; Leu and Doppman, 1986; Leu and Gerfin, 1992; van Doorslaer, 1987; Van de Ven and van der Gaag, 1982; Erbsland, Ried and Ulrich, 2002).[42]

[42]Grossman (2000; pp. 369-370) shows that the model does not always produce the incorrect sign for the relationship between health and investment in medical care. For the pure investment model and assuming that the biological aging rate d_t increases with age (a necessary assumption for the health stock to decline with age in a CRTS formulation), he finds that investment in medical care increases with age while the health stock falls with age if the elasticity of the marginal production benefit of health with respect to health is less than one (Grossman refers to this as the MEC schedule). The requirement that the biological aging rate increase with age is another artifact of the indeterminacy of health investment. I do not have to rely on characteristics of exogenous functions such as the biological aging rate (apart

Third, Case and Deaton (2005) argue that while health production models can explain differences in the *level* of health between socioeconomic status (SES) groups they cannot explain differences in the *rate* of health deterioration between SES groups. In other words, health production models cannot account for the observed widening of disparities in health by SES with age. In section 4.3.2 I show that for a DRTS health production process the wealthy and more educated invest more in health and consume more and that their health deteriorates at a slower pace.[43] As a result, given similar initial health endowments, they remain healthier as they age and live longer. It is plausible that as the disparity in health widens the deteriorating health of low-SES individuals induces them to begin to invest more in health than their high-SES peers (e.g., due to the negative relation between health and investment in health). Thus the model could reproduce both the observed initial widening and the subsequent narrowing of the SES health gradient.

Fourth, Usher (1975) has pointed to the lack of "memory" in health production model solutions (e.g., Usher 1975, p. 220).[44] Casual observation and introspection suggests that our health depends on initial and past conditions: healthy individuals are those that began life healthy and that have invested in health over time. Indeed, in the alternative interpretation of the first-order condition presented here, health is not determined by the condition for "optimal" health (4.15) but by the dynamic equation (4.2), which can be written (using 4.5) in the form (4.21). Thus, the solution for the health stock H_t is a function of the initial health stock H_0 and the history of past health investments I_s and past biological aging rates d_s ($s < t$). As a result I find the health stock to be a complex function of the initial health stock H_0, initial assets A_0, education E and the entire history of prices, wages and environmental conditions (see the discussion in section 4.3.3).

from assuming that aging is detrimental to health, i.e. $d_t > 0$) to obtain a negative relation between health and health investment or to ensure that life is finite (see the criticism by Case and Deaton, 2005).

[43]For a CRTS health production process, however, the effective health deterioration rate $\partial H_t/\partial t$ depends on the rate of the biological aging rate $\partial d_t/\partial t$ (another artifact of the indeterminacy of health investment). Thus if low SES individuals have more rapidly increasing aging rates $\partial d_t/\partial t$ a model with a CRTS health production process could reproduce the observed widening of health disparities by SES. It seems plausible that the aging process d_t is more rapid for low SES individuals through, e.g., environmental factors such as detrimental living and working conditions, but it is not a priori clear that low SES individuals also have faster rates of the biological aging rate $\partial d_t/\partial t$ (e.g., Case and Deaton, 2005).

[44]The structural- and reduced-form solutions for health H_t and the reduced-form solution for health investment I_t and the demand for medical goods / services m_t derived and utilized in the health production literature are functions only of current parameter values (e.g., equations 4-2, 4-6 and 4-7 in Grossman, 1972a; equations 42, 45 and 46 in Grossman, 2000; equations 11, 12 and 14 in Wagstaff, 1986a; see also Galama and Kapteyn, 2009). Initial and past conditions appear to have been "forgotten".

Fifth, Case and Deaton (2005) note that in the health production literature *" . . . if the rate of biological aging is constant, which is perhaps implausible but is hardly impossible, (and if the interest rate is as least as large as the rate of time preference), people will "choose" an infinite life . . . "*[45] Thus, to ensure that life is finite and health falls with age it is necessary to assume that the biological aging rate increases with age ($\partial d_t/\partial t > 0$). In the interpretation of the theory presented here, however, it is not required that the biological aging rate d_t increase with age in order for health to decrease with age and in order for life to be finite. This follows intuitively from the dynamic relation (4.2; or in alternative form: equation 4.21) for health. If medical technology cannot fully repair the health of individuals for certain diseases (e.g., low efficiency of medical care will result in small I_t) then the health stock will decrease with age. Solutions are possible not only for a biological aging rate that increases with age, but also for constant or decreasing biological aging rates with age. The numerical simulations in section 4.3.4 provide an illustration based on a constant biological aging rate with age. This addresses the criticism by Case and Deaton (2005) that health production models are characterized by complete health repair.

In sum, I find health investment to be a decreasing function of health, that the health of wealthy individuals declines more slowly and that they live longer, that current health status is a function of the initial level of health and the histories of prior health investments made, that health investment rapidly increases near the end of life and that length of life is finite as a result of limited life-time resources (the budget constraint) and if medical technology cannot fully offset biological aging. I find no support for the common notion that individuals aspire to a certain "optimal" level of the health stock. Rather, given any level of their health stock, individuals decide about the optimal level of health investment.

Empirical estimation of the model is needed to test the assumptions and the theoretical predictions presented in this work and to contrast these with the predictions of alternative health production models. To this end I have provided structural form relations in section 4.3.3.

[45] For a constant biological aging rate health decreases with age only if the time preference rate β_t exceeds the return to capital δ_t (and increases if the reverse is true). This follows from the first-order condition (4.15) if interpreted as a condition for the "optimal" level of the health stock. See also equation 13 in Grossman, 1972b, equation 11 in Grossman, 2000, or equation 6 in Case and Deaton (2005).

4.5 Appendix

4.5.1 First-order conditions

Associated with the Hamiltonian (equation 4.9) we have the following conditions:

$$
\begin{aligned}
q_{t-1}^A &= \frac{\partial \Im_t}{\partial A_t} \Rightarrow \\
q_{t-1}^A &= (1+\delta_t) q_t^A \Leftrightarrow \\
q_t^A &= \frac{q_0^A}{\prod_{k=1}^t (1+\delta_k)}, && (4.53) \\
q_{t-1}^H &= \frac{\partial \Im_t}{\partial H_t} \Rightarrow \\
q_{t-1}^H &= q_t^H (1-d_t) + \frac{\partial U(C_t,H_t)/\partial H_t}{\prod_{k=1}^t (1+\beta_k)} + q_0^A \frac{\partial Y(H_t)/\partial H_t}{\prod_{k=1}^t (1+\delta_k)} \Leftrightarrow && (4.54) \\
q_t^H &= -\sum_{i=1}^t \left[\frac{\partial U(C_i,H_i)/\partial H_i}{\prod_{j=1}^i (1+\beta_j)} + q_0^A \frac{\partial Y(H_i)/\partial H_i}{\prod_{j=1}^i (1+\delta_j)} \right] \frac{1}{\prod_{k=i}^t (1-d_k)} \\
&+ \frac{q_0^H}{\prod_{k=1}^t (1-d_k)} && (4.55) \\
&= \sum_{i=t+1}^T \left[\frac{\partial U(C_i,H_i)/\partial H_i}{\prod_{j=1}^i (1+\beta_j)} + q_0^A \frac{\partial Y(H_i)/\partial H_i}{\prod_{j=1}^i (1+\delta_j)} \right] \prod_{k=t+1}^{i-1} (1-d_k) \\
&+ q_T^H \prod_{k=t+1}^T (1-d_k) && (4.56) \\
\frac{\partial \Im_t}{\partial X_t} &= 0 \Rightarrow \\
\frac{\partial U(C_t,H_t)}{\partial C_t} &= q_0^A \frac{p_{X_t}}{\partial C_t/\partial X_t} \frac{\prod_{j=1}^t (1+\beta_j)}{\prod_{j=1}^t (1+\delta_j)} \\
&\equiv q_0^A \pi_{C_t} \frac{\prod_{j=1}^t (1+\beta_j)}{\prod_{j=1}^t (1+\delta_j)}, && (4.57) \\
\frac{\partial \Im_t}{\partial \tau_{C_t}} &= 0 \Rightarrow \\
\frac{\partial U(C_t,H_t)}{\partial C_t} &= q_0^A \frac{w_t}{\partial C_t/\partial \tau_{C_t}} \frac{\prod_{j=1}^t (1+\beta_j)}{\prod_{j=1}^t (1+\delta_j)} \\
&\equiv q_0^A \pi_{C_t} \frac{\prod_{j=1}^t (1+\beta_j)}{\prod_{j=1}^t (1+\delta_j)}, && (4.58)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \Im_t}{\partial m_t} &= 0 \Rightarrow \\
q_t^H &= q_0^A \left\{ \frac{p_{m_t} I_t^{1-\alpha}}{\alpha[\partial I_t / \partial m_t]} \right\} \frac{1}{\prod_{j=1}^{t}(1+\delta_j)} \\
&\equiv q_0^A \, \pi_{I_t} \frac{1}{\prod_{j=1}^{t}(1+\delta_j)}, \qquad (4.59) \\
\frac{\partial \Im_t}{\partial \tau_{I_t}} &= 0 \Rightarrow \\
q_t^H &= q_0^A \left\{ \frac{w_t I_t^{1-\alpha}}{\alpha[\partial I_t / \partial \tau_{I_t}]} \right\} \frac{1}{\prod_{j=1}^{t}(1+\delta_j)} \\
&\equiv q_0^A \, \pi_{I_t} \frac{1}{\prod_{j=1}^{t}(1+\delta_j)}, \qquad (4.60)
\end{aligned}
$$

where I have used the following definitions

$$
\sum_{k}^{k-1}(\bullet) \equiv 0,
$$

$$
\prod_{k}^{k-1}(\bullet) \equiv 1.
$$

Combining (4.59) or (4.60) with (4.55) we obtain the first-order condition for health investment (see equations 4.11 and 4.13). The first-order condition for consumption C_t is provided by equation (4.57) or (4.58) (see equation 4.18).

4.5.2 Mathematical equivalency of first-order conditions

Taking the difference between period t and $t-1$ of either expression (4.11) or (4.13) one arrives at (4.14) and (4.15). In other words

$$
(4.11) \Rightarrow (4.15), \qquad (4.61)
$$

$$
(4.13) \Rightarrow (4.15). \qquad (4.62)
$$

Using recursive backward or forward substitution of relation (4.14) (which is equivalent to expression 4.15) one arrives at (4.11) or (4.13). Thus we have

$$
(4.11) \Leftarrow (4.15), \qquad (4.63)
$$

$$
(4.13) \Leftarrow (4.15). \qquad (4.64)
$$

Naturally, this result is also true in a continuous time formulation. In this case the first-order condition for health investment I_t can be written as

$$\begin{aligned} & \pi_I(t)e^{-\int_0^t \delta(u)du} - \pi_I(0)e^{\int_0^t d(u)du} \\ = & -\int_0^t \left[\frac{1}{q_A(0)}\frac{\partial U(s)}{\partial H(s)}e^{-\int_0^s \beta(u)du} + \varphi_H(s)e^{-\int_0^s \delta(u)du}\right] e^{\int_s^t d(u)du}ds, \end{aligned} \tag{4.65}$$

or, in terms of the terminal point $t = T$

$$\begin{aligned} & \pi_I(t)e^{-\int_0^t \delta(u)du} - \pi_I(T)e^{-\int_0^T \delta(u)du}\, e^{-\int_t^T d(u)du} \\ = & \int_t^T \left[\frac{1}{q_A(0)}\frac{\partial U(s)}{\partial H(s)}e^{-\int_0^s \beta(u)du} + \varphi_H(s)e^{-\int_0^s \delta(u)du}\right] e^{-\int_t^s d(u)du}ds. \end{aligned} \tag{4.66}$$

Differentiating (4.65) or (4.66) with respect to t one obtains

$$\frac{\partial U(t)}{\partial H(t)} = q_A(0)\left[\sigma_H(t) - \varphi_H(t)\right] e^{\int_0^t [\beta(s)-\delta(s)]ds}. \tag{4.67}$$

Notation follows the discussion in section 4.2.

Using the Leibniz integral rule to differentiate (analogous to taking the difference between two time periods in discrete time) the first-order condition for health investment (4.65) or the alternative expression (4.66) with respect to t one obtains the alternative expression (4.67). In other words

$$(4.65) \Rightarrow (4.67), \tag{4.68}$$

$$(4.66) \Rightarrow (4.67). \tag{4.69}$$

From (4.67) we obtain a first-order differential equation in $\pi_I(t)$

$$\frac{\partial \pi_I(t)}{\partial t} = \pi_I(t)[d(t) + \delta(t)] - \varphi_H(t) - \frac{\partial U(t)}{\partial H(t)}\frac{1}{q_A(0)}e^{-\int_0^t [\beta(u)-\delta(u)]du}, \tag{4.70}$$

which can be solved (analogous to backward or forward substitution in discrete time)

$$\begin{aligned} \pi_I(t) = & \pi_I(t')e^{\int_{t'}^t [d(u)+\delta(u)]du} \\ & - \int_{t'}^t \left[\frac{1}{q_A(0)}\frac{\partial U(s)}{\partial H(s)}e^{-\int_0^s [\beta(u)-\delta(u)]du} + \varphi_H(s)\right] e^{\int_s^t [d(u)+\delta(u)]du}ds. \end{aligned} \tag{4.71}$$

For $t' = 0$ we obtain (4.65) and for $t' = T$ we obtain (4.66). Thus we have

$$(4.65) \Leftarrow (4.67), \tag{4.72}$$

$$(4.66) \Leftarrow (4.67). \tag{4.73}$$

4.5.3 Variation in initial assets

Consider a variation in initial assets as in (4.23). The first order condition for health investment (4.15)

$$\begin{aligned} & \left.\frac{\partial U}{\partial H}\right|_{C_t+\Delta C_{t,A},\, H_t+\Delta H_{t,A}} \\ = \; & \left\{ \pi_{I_t}|_{I_t+\Delta I_{t,A},\, H_t+\Delta H_{t,A}} (d_t+\delta_t) - \Delta\pi_{I_t}|_{I_t+\Delta I_{t,A},\, H_t+\Delta H_{t,A}} (1+\delta_t) - \varphi_{H_t}|_{H_t+\Delta_{t,A}} \right\} \\ \times \; & \left(q_0^A + \Delta q_{0,A}^A\right) \frac{\prod_{j=1}^{t}(1+\beta_j)}{\prod_{j=1}^{t}(1+\delta_j)}, \end{aligned} \tag{4.74}$$

leads, using a first-order Taylor expansion, to the following expression

$$\begin{aligned} \frac{\left.\frac{\partial^2 U}{\partial C \partial H}\right|_{C_t,\, H_t}}{\left.\frac{\partial U}{\partial H}\right|_{C_t,\, H_t}} \Delta C_{t,A} \; = \; & \frac{\Delta q_{0,A}^A}{q_0^A} + \left[\frac{\left.\frac{\partial \pi_I}{\partial H}\right|_{I_t,\, H_t} (d_t+\delta_t) - \left.\frac{\partial \varphi_H}{\partial H}\right|_{H_t}}{\sigma_H|_{I_t,\, H_t} - \varphi_H|_{H_t}} - \frac{\left.\frac{\partial^2 U}{\partial H^2}\right|_{C_t,\, H_t}}{\left.\frac{\partial U}{\partial H}\right|_{C_t,\, H_t}} \right] \Delta H_{t,A} \\ & + \; \frac{\left.\frac{\partial \pi_I}{\partial I}\right|_{I_t,\, H_t} (d_t+\delta_t)}{\sigma_H|_{I_t,\, H_t} - \varphi_H|_{H_t}} \Delta I_{t,A}, \end{aligned} \tag{4.75}$$

where I have omitted second-order terms.[46]

Similarly, from the first order condition for consumption (4.18)

$$\begin{aligned} & \left.\frac{\partial U}{\partial C}\right|_{C_t+\Delta C_{t,A},\, H_t+\Delta H_{t,A}} \\ = \; & \left(q_0^A + \Delta q_{0,A}^A\right) \pi_C|_{C_t+\Delta C_{t,A},\, H_t+\Delta H_{t,A}} \frac{\prod_{j=1}^{t}(1+\beta_j)}{\prod_{j=1}^{t}(1+\delta_j)}, \end{aligned} \tag{4.76}$$

we have

$$\begin{aligned} & \left(\frac{\left.\frac{\partial \pi_C}{\partial C}\right|_{C_t,\, H_t}}{\pi_C|_{C_t,\, H_t}} - \frac{\left.\frac{\partial^2 U}{\partial C^2}\right|_{C_t,\, H_t}}{\left.\frac{\partial U}{\partial C}\right|_{C_t,\, H_t}} \right) \Delta C_{t,A} \\ = \; & - \left(\frac{\left.\frac{\partial \pi_C}{\partial H}\right|_{C_t,\, H_t}}{\pi_C|_{C_t,\, H_t}} - \frac{\left.\frac{\partial^2 U}{\partial C \partial H}\right|_{C_t,\, H_t}}{\left.\frac{\partial U}{\partial C}\right|_{C_t,\, H_t}} \right) \Delta H_{t,A} - \frac{\Delta q_{0,A}^A}{q_0^A}. \end{aligned} \tag{4.77}$$

Last, combining (4.75) with (4.77) to eliminate $\Delta C_{t,A}$ we obtain (4.24).

[46] Such as, e.g., terms in $(\partial \Delta \pi_I / \partial I)|_{I_t,\, H_t} \Delta I_{t,A}$ and $\Delta q_{0,A}^A (\partial \pi_I / \partial H)|_{I_t,\, H_t} \Delta H_{t,A}$.

4.5.4 Structural relations for empirical testing

From the utility function (4.27) and the first-order conditions (4.15) and (4.18) it follows that

$$C_t = \frac{\zeta}{1-\zeta}\frac{\sigma_{H_t}-\varphi_{H_t}}{\pi_{C_t}}H_t, \tag{4.78}$$

and

$$H_t = (1-\zeta)\Lambda(q_0^A)^{-1/\rho}\left(\sigma_{H_t}-\varphi_{H_t}\right)^{-\chi}\pi_{C_t}^{\chi-1/\rho}\frac{\prod_{j=1}^{t}(1+\beta_j)^{-1/\rho}}{\prod_{j=1}^{t}(1+\delta_j)^{-1/\rho}}, \tag{4.79}$$

$$C_t = \zeta\Lambda(q_0^A)^{-1/\rho}\left(\sigma_{H_t}-\varphi_{H_t}\right)^{1-\chi}\pi_{C_t}^{\chi-1/\rho-1}\frac{\prod_{j=1}^{t}(1+\beta_j)^{-1/\rho}}{\prod_{j=1}^{t}(1+\delta_j)^{-1/\rho}}, \tag{4.80}$$

where Λ and χ are defined in (4.41) and (4.42).

Using equations (4.12) and (4.30) we have

$$\pi_{I_t} = \frac{p_{m_t}^{1-k_I}\,w_t^{k_I}}{\alpha\,k_I^{k_I}(1-k_I)^{1-k_I}\,\mu_{I_t}}I_t^{1-\alpha} \equiv \pi_{I_t}^* I_t^{1-\alpha}, \tag{4.81}$$

$$m_t = \left(\frac{1-k_I}{k_I}\right)^{k_I}\mu_{I_t}^{-1}p_{m_t}^{-k_I}w_t^{k_I}I_t \equiv m_{I_t}^* I_t, \tag{4.82}$$

$$\tau_{I_t} = \left(\frac{1-k_I}{k_I}\right)^{-(1-k_I)}\mu_{I_t}^{-1}p_{m_t}^{1-k_I}w_t^{-(1-k_I)}I_t \equiv \tau_{I_t}^* I_t. \tag{4.83}$$

Using equations (4.19) and (4.31) we have

$$\pi_{C_t} = \frac{p_{X_t}^{1-k_C}\,w_t^{k_C}}{k_C^{k_C}(1-k_C)^{1-k_C}\,\mu_{C_t}}, \tag{4.84}$$

$$X_t = \left(\frac{1-k_C}{k_C}\right)^{k_C}\mu_{C_t}^{-1}p_{X_t}^{-k_C}w_t^{k_C}C_t \equiv X_{C_t}^* C_t, \tag{4.85}$$

$$\tau_{C_t} = \left(\frac{1-k_C}{k_C}\right)^{-(1-k_C)}\mu_{C_t}^{-1}p_{X_t}^{1-k_C}w_t^{-(1-k_C)}C_t \equiv \tau_{C_t}^* C_t. \tag{4.86}$$

From (4.16), (4.29), (4.79), (4.81) and (4.84) follows

$$a_t^1 I_t^{1-\alpha} - (1-\alpha)I_t^{1-\alpha}\widetilde{I}_t = a_t^2 H_t^{-1/\chi} + a_t^3 H_t^{-(1+\gamma)}, \tag{4.87}$$

where

$$a_t^1 \equiv \left[d_t + \delta_t - (1-k_I)\widetilde{p_{m_t}} - k_I\widetilde{w}_t + \widetilde{\mu_{I_t}}\right], \tag{4.88}$$

$$a_t^2 \equiv \left[(1-\zeta)\Lambda(q_0^A)^{-1/\rho}\right]^{1/\chi}(\pi_{I_t}^*)^{-1}(\pi_{C_t})^{1-1/\rho\chi}\frac{\prod_{j=1}^{t}(1+\beta_j)^{-1/\rho\chi}}{\prod_{j=1}^{t}(1+\delta_j)^{-1/\rho\chi}}, \tag{4.89}$$

$$a_t^3 \equiv w_t(\pi_{I_t}^*)^{-1}\Omega_*, \tag{4.90}$$

where the notation $\widetilde{f}_t$ is used to denote the relative change $\widetilde{f}_t \equiv 1 - \frac{f_{t-1}}{f_t}$ in a function f_t and we have assumed small relative changes in the price of medical care $\widetilde{p_{m_t}}$, wages $\widetilde{w}_t$ and the efficiency of the health investment process $\widetilde{\mu_{I_t}}$.

Using (4.82) and (4.87) and the functional relations defined in section 4.3.3 we obtain a structural relation (4.35) between health investment goods and services m_t purchased in the market and the stock of health H_t.

Chapter 5

A Theory of Socioeconomic Disparities in Health

Understanding of the substantial disparity in health between low and high socioeconomic status (SES) groups is hampered by the lack of a sufficiently comprehensive theoretical framework to interpret empirical facts and to predict yet untested relations. We present a life-cycle model that incorporates multiple mechanisms explaining (jointly) a large part of the observed disparities in health by SES. In our model, lifestyle factors, working conditions, retirement, living conditions and curative care are mechanisms through which SES, health and mortality are related. Our model predicts a widening and possibly a subsequent narrowing with age of the gradient in health by SES.

This chapter is based upon:

Galama, T.J. and van Kippersluis, H. (2010), "A Theory of Socioeconomic Disparities in Health Over the Life Cycle", *RAND Working Paper*, WR-773.

5.1 Introduction

Disparities in health across socioeconomic status (SES) groups – often called the SES health gradient – are substantial. For example, Case and Deaton (2005) show how in the United States, a 20 year old low-income (bottom quartile of family income) male, on average, reports to be in similar health as a 60 year old high-income (top quartile) male. In Glasgow, U.K., life expectancy of men in the most deprived areas is 54 years, compared with 82 years in the most affluent (Hanlon et al. 2006). In cross sectional data the disparity in health between low and high SES groups appears to increase over the life cycle until ages 50-60, after which it narrows. Similar patterns hold for other measures of SES, such as education and wealth and other indicators of health, such as onset of chronic diseases, disability and mortality (e.g., Adler et al. 1994; Marmot, 1999; Smith, 1999). This pattern is remarkably similar between countries with relatively low levels of protection from loss of work and health risks, such as the U.S., and those with stronger welfare systems, such as the Netherlands (House et al. 1994; Kunst and Mackenbach, 1994; Preston and Elo, 1995; Smith 1999; 2004; 2007; Case and Deaton, 2005; van Kippersluis et al. 2010).

Recent significant contributions to the understanding of socioeconomic disparities in health have concentrated on the identification of causal effects, but have stopped short of uncovering the underlying mechanisms that produce the causal relationships. For example, education is found to have a causal protective effect on health (Lleras-Muney, 2005; Oreopoulos, 2006; Silles, 2009) but it is not known exactly how the more educated achieve their health advantage.

Understanding of the relative importance of underlying mechanisms responsible for the observed relationships is hampered by the lack of a sufficiently comprehensive theory. Case and Deaton (2005) argue that it is extremely difficult to understand the relationships between health, education, income, wealth and labor-force status without some guiding theoretical framework. Integrating the roles of proposed mechanisms and their long-term effect into a theoretical framework allows researchers to disentangle the differential patterns of causality and assess the interaction between mechanisms. Such understanding is essential in designing effective policies to reduce disparities (Deaton, 2002). It is no surprise then that several authors (e.g., Case and Deaton, 2005; Cutler et al. 2011) have pointed to the absence of a theory of SES and health over the life cycle and have emphasized the importance of developing one.

A suitable framework in which multiple mechanisms and their cumulative long-term effects can be studied is a structural model of SES and health over the life cycle. Case and

Deaton (2005) have attempted to develop a model for the role of work and consumption behavior in explaining the SES-health gradient. Their starting point is the canonical life cycle model of the demand for health and medical care, due to Grossman (1972a; 1972b). Case and Deaton (2005) present a simplified Grossman model and extend the model to include the detrimental effect of hard/risky labor and of unhealthy consumption behavior on health. However, the authors conclude that the model is not able to explain a number of the most salient features of the SES health gradient. For example, Case and Deaton (2005) argue that while the model can explain differences in the *level* of health between low and high SES groups, it cannot explain differences in the *rate* of health decline. In other words, it cannot account for the widening of the SES health gradient with age through late middle age or early late life, as is observed in empirical studies. Other problems with some of the predictions and properties of health production models have been pointed out in the literature (see Grossman, 2000, for a review and rebuttal of these).

The aim of this paper is to develop a conceptual framework for health and socioeconomic status over the life cycle. The framework includes simplified representations of major mechanisms, which allows us to improve our understanding of their operational roles in explaining the SES health gradient and make predictions. Our starting point is the health production literature spawned by Grossman (Grossman, 1972a; 1972b) and the extensions presented by Ehrlich and Chuma (1990) and Case and Deaton (2005). Our contribution is as follows. First, we address a number of issues identified with this strand of the literature by noting that what is generally interpreted as the equilibrium condition for health can alternatively be interpreted as the first-order condition for health investment (as in Galama, 2011 [Chapter 4]). This interpretation necessitates the assumption of decreasing-returns-to-scale (DRTS) in the health production function (as in Ehrlich and Chuma, 1990), and addresses (i) the indeterminacy problem ("bang-bang" solution) for investment in medical care (Ehrlich and Chuma, 1990), (ii) the inability to reproduce the observed negative relation between health and the demand for medical care (e.g., Zweifel and Breyer, 1997),[1] (iii) the lack of history in the model solutions (e.g., Usher, 1975) and (iv) the lack of capacity to explain differences in the rate of health decline between different socioeconomic groups (Case and Deaton, 2005). With these essential issues addressed our formulation can account for a greater number of observed empirical

[1]It is not entirely correct to assert that health production models always produce the incorrect sign for the relationship between health and investment in curative care. For the pure investment model and assuming that the biological aging rate increases with age, investment in curative care increases with age while the health stock falls if the elasticity of the marginal production benefit of health with respect to health is less than one (Grossman refers to this as the MEC schedule; Grossman, 2000 [p. 369]). This produces a negative correlation between health and medical care.

patterns and suggests that the Grossman model provides a suitable foundation for the development of a life-cycle model of the SES-health gradient.

Yet, utilization of medical services and access to care explain only part of the association between SES and health (e.g., Adler et al. 1993). Our second contribution is therefore to incorporate many potential mechanisms in the model that could explain disparities in health by SES and to include a multitude of potential bi-directional pathways between health and dimensions of SES. One important concept in our work is "job-related health stress", which can be interpreted broadly and can range from physical working conditions (e.g., hard labor) to the psychosocial aspects of work (e.g., low status, limited control, repetitive work, etc). The notion here is that job-related health stress can include any aspect of work that is detrimental to health and as such is associated with a wage premium (a compensating wage differential). Other important features of the model are lifestyle factors (preventive care, healthy and unhealthy consumption), curative (medical) care, labor force withdrawal (retirement) and mortality. The model integrates a life cycle approach, and the concepts of financial, education and health capital (Muurinen and Le Grand, 1985). The focus is on understanding the SES-health gradient as the outcome of rational (constrained) individual behavior, and the framework applies to individuals who have completed their education and participate (or have participated) in the labor-force.

We explore the characteristics of the first-order conditions for a fixed retirement age and a fixed age of death. We find that greater initial wealth, permanently higher earnings (over the life cycle) and a higher level of education induce individuals to invest more in curative and in preventive care, shift consumption toward healthy consumption, and enable individuals to afford healthier working environments (associated with lower levels of physical and psychosocial health stresses) and living environments. The mechanism through which initial wealth, permanent income and education operates is by increasing the demand for curative care and raising the marginal cost of curative care. A higher marginal cost of curative care, in turn, increases the health benefit of (and hence demand for) preventive care and healthy consumption, and the health cost of (and hence reduced demand for) unhealthy working and living environments, and unhealthy consumption. Jointly these behavioral choices gradually lead to growing health advantage with age. Further, the model predicts an initial widening and potentially a subsequent narrowing of the SES-health gradient, as low SES individuals increase their health investment and improve their health-related behavior faster as a result of their worse health. Results from earlier studies (Ehrlich and Chuma, 1990; Ehrlich, 2000; Galama et al. 2008 [Chapter 3]) suggest that the more rapidly worsening health of low SES individuals could lead to early withdrawal from the labor force, potentially widening the gradient in early and mid

age, and shorter life spans, potentially narrowing the gradient in late age. Our model thus holds promise in explaining empirical health patterns. Such a model has not been available before and economists have highlighted the significance of its development (e.g., Cutler et al. 2011; Case and Deaton, 2005).

The paper is organized as follows. Section 5.2 briefly reviews the literature on health disparities by SES to determine the essential components required in a theoretical framework. The relation between SES and health is complex and developing a theory requires simplification and a focus on the essential mechanisms relating SES and health. We highlight potential explanations for the SES health gradient that a) explain a large part of the gradient and b) are relatively straightforward to include in our theoretical framework. Based on these findings we develop our theoretical formulation. Section 5.3 presents and discusses first-order conditions and the characteristics of the model solutions for a fixed age of retirement and a fixed age of death. The section also highlights potential mechanisms through which SES and health influence each other. In section 5.4 we summarize and conclude.

5.2 Components of a model capturing the SES-health gradient

In this section we review the literature to determine the essential components of a theory of health disparities by SES. Based on these findings we extend and refine prior work and present our theoretical formulation.

5.2.1 Background

A significant body of research across multiple disciplines (including epidemiology, sociology, demography, psychology, evolutionary biology and economics) has been devoted to documenting and explaining the substantial disparity in health between low and high socioeconomic status (SES) groups. Progress has been made in recent years in characterizing the relationships between the various dimensions of SES and health over the life cycle and in understanding the relative importance and directions of causal pathways. Epidemiological research has used longitudinal studies to examine the role of behavioral, material, psychosocial and healthcare related pathways in explaining SES-health associations (House et al. 1990; 1994; Lynch et al. 1997; Marmot et al. 1997a; Lantz et al. 1998; Yen and Kaplan, 1999; van Oort et al. 2005; Skalicka et al. 2009). Economists have

recently re-emphasized the importance of the reverse impact of health on SES through ability to work (Smith, 1999; 2004; 2007; Case and Deaton, 2005).

These studies suggest that education is the key dimension of SES for which there appears to be robust evidence of a substantial causal protective effect on health. Secondly, an important part of the health differences by financial indicators of SES can be explained by the fact that bad health impinges on the ability to work, thereby reducing income. Further, these studies highlight the importance of health behaviors (such as smoking, drinking and exercise), curative and preventive care, psychosocial and environmental risk factors, neighborhood social environment, acute and chronic psychosocial stress, social relationships and supports, sense of control, fetal and early childhood conditions, and physical, chemical, biological and psychosocial hazards and stressors at work.

Below we provide more detail on the potential role of the working environment and lifestyle factors and the role of various potential pathways between health and SES and vice versa.

Working environment and lifestyle factors:
Low SES individuals more often perform risky, manual labor than high SES individuals, and their health deteriorates faster as a consequence (Marmot et al. 1997b; Schrijvers et al. 1998; Borg and Kristensen, 2000). Case and Deaton (2005) find that those who are employed in manual occupations have worse health than those who work in professional occupations and that the health effect of occupation operates at least in part independently of the personal characteristics of the workers. Cutler et al. (2011) present similar results using mortality as an indicator of health. Schrijvers et al. (1998) use Dutch cross-sectional data to study the impact of working conditions on the association between occupational class and self-reported health. Hazardous physical working conditions are more prevalent in lower occupational classes, and this explains a substantial part (for males up to 83 percent) of the association between health and occupational (social) class. Extensive research further suggests an important role of lifestyle factors, particularly smoking, in explaining SES disparities in health (Mackenbach et al. 2004; Khang et al. 2009). Fuchs (1986) even argues that in developed countries, it is personal lifestyles that cause the greatest variation in health. Using three different datasets from the U.K. and the U.S., Marmot et al. (1997a) find that features of the psycho-social working environment, social circumstances outside work, and health behavior jointly account for much of the social gradient in health (see also House et al. 1994). Some epidemiological studies estimate that around two thirds of the social gradient in health deterioration could be

explained by working environment and life style factors alone (Borg and Kristensen, 2000).

A multitude of potential pathways between health and SES and vice versa: As Cutler et al. (2011) note, the mechanisms linking the various dimensions of SES to health are diverse. Some cause health, some are caused by health and some are jointly determined with health.

- *Education on health:* Education is found to have a causal effect on health and mortality (Lleras-Muney, 2005; Oreopoulos, 2006; Smith, 2007; Silles, 2009). However, Cutler et al. (2011) note that the mechanisms by which education affects health are not well understood. While consumption behavior and curative and preventive care can partly explain the effect of education on health, it remains largely unclear why more educated individuals behave in a healthier manner (Cutler et al. 2011). Education increases earnings (e.g., Mincer, 1974) and thereby enables the purchase of health investments (though higher earnings may also increase the opportunity cost of time). Education potentially increases the efficiency of curative and preventive care usage and time inputs into the production of health investment (Grossman, 1972a; 1972b). It appears that the higher educated are better able at managing their diseases (Goldman and Smith, 2002), and high SES individuals appear to benefit more from new knowledge and new technology (Lleras-Muney and Lichtenberg, 2005; Glied and Lleras-Muney, 2008).

- *Health on education:* The existence of an effect of early childhood health on educational attainment has been established in studies from developed as well as developing countries. Studies from the U.S., U.K., and Norway show convincingly that low birth weight individuals have worse schooling outcomes (Behrman and Rosenzweig, 2004; Case et al. 2005; Black et al. 2007; Royer, 2009). Another piece of evidence is derived from the 1918 influenza epidemic in the U.S., and the hookworm eradication from the American South, where adverse conditions in childhood caused a lower educational attainment of the affected cohorts (Almond, 2006; Bleakley, 2007). From developing countries similar evidence is presented by, e.g., Miguel and Kremer (2004).

- *Income or wealth on health:* Income and wealth enable purchases of curative and preventive care and thereby potentially allow for better health maintenance. The impact of financial resources on health is likely to depend on the manner of health care provision in a country. In the case of market provision, income, wealth and employment may determine access to health care, whereas in the case of universal

health care provision these factors may be less important. On the other hand, higher wages are associated with greater opportunity costs, which would reduce the amount of time devoted to health maintenance. Further, more affluent workers may choose safer working (associated with a lower level of job-related health stress) and living environments since safety is a normal good (Viscusi, 1978; 1993).

Smith (2007) finds no effect of financial measures of SES (income, wealth and change in wealth) on changes in health in the U.S. Financial indicators of SES do not seem to cause the onset of health problems at any age (Smith, 2007). Cutler et al. (2011) provide an overview of empirical findings and conclude that the evidence points to no or a very limited impact of income or wealth on health (see also Michaud and van Soest, 2008). Yet, this view is not unequivocally accepted. Replication is still needed and controversy remains on the extent to which these findings apply uniformly to different population segments. For example, Lynch et al. (1997) suggest that accumulated exposure to economic hardship causes bad health, and Herd et al. (2008) argue that there might be causal effects of financial resources on health at the bottom of the income or wealth distribution.

- *Health on income and wealth:* Healthy individuals are more productive, earn higher wages and are able to accrue greater wealth (Currie and Madrian, 1999; Contoyannis and Rice, 2001). Studies have shown that perhaps the most dominant causal relation between health and dimensions of SES is the causal impact that health has on one's ability to work and hence produce income and wealth (e.g., Smith, 2004; 2007; Case and Deaton, 2005).

- *Joint determination:* Fuchs (1982; 1986) (see also Barsky et al. 1997) has argued that the strong correlation between education and health may be due to differences in the time preferences of individuals, which affects investments in both education and health and helps to explain variations in cigarette smoking, diet, and exercise. Other third factors of interest that may produce a spurious correlation between SES and health are intelligence, cognitive ability, and non-cognitive skills (Auld and Sidhu, 2005; Deary, 2008; Chiteji, 2010). In a review of the literature on the relationship between education and health, Cutler and Lleras-Muney (2008) argue that differences in individual preferences (risk aversion and discount rates) appear to explain only a small portion of the SES health gradient (see also Elo and Preston, 1996). But the authors also note that few studies have attempted to investigate the role of individual preferences, that preferences are difficult to measure, and

that preferences with respect to health may differ from preferences with respect to finance, measures of which are usually employed in such studies.

5.2.2 Theoretical formulation

In this section we formalize the above discussion on the features of a theoretical framework for the SES health gradient over the life cycle. The aim is to understand the SES-health gradient as the outcome of rational constrained individual behavior.

A natural starting point for a theory of the relation between health and SES is a model of life cycle utility maximization. The model is based on the Grossman model of the demand for health (Grossman, 1972a; 1972b; 2000) in continuous time (see also Wolfe, 1985; Wagstaff, 1986a; Ehrlich and Chuma, 1990; Zweifel and Breyer, 1997) with seven essential additional features.

First, we assume decreasing-returns-to-scale (DRTS) in the health production function (as in Ehrlich and Chuma, 1990).

Second, individuals choose their level of undesirable job characteristics which potentially have health consequences, denoted as "job-related health stress". The concept of job-related health stress can be interpreted broadly and can range from physical working conditions (e.g., hard or risky labor) to psychosocial aspects of work (e.g., low social status, lack of control, repetitive work, etc). The decision to engage in unhealthy labor is governed by the relative benefit of a possible wage premium – a compensating wage differential (Smith, 1776; Viscusi, 1978; 1979) – versus the cost in terms of a higher health deterioration rate. Evidence is strong that there is a wage premium for jobs with higher mortality risk (Smith, 1978), and also for less serious, non-fatal, health risks (e.g. Viscusi, 1978; Olson, 1981; Duncan and Holmlund, 1983). Thus we introduce the notion that individuals may accept risky and/or unhealthy work environments, in exchange for higher pay (Muurinen, 1982; Case and Deaton, 2005), and explore solutions in which the decision to rapidly "wear one's body down" (i.e., to perform "hard" labor or engage in work with psychosocial health risks) is endogenous.

Third, individuals engage in preventive care (such as check up doctor visits) to slow the biological aging rate. Hence, we explicitly model health investment as consisting of two components: (i) curative care (as in Grossman, 1972a; 1972b), and (ii) a new concept of preventive care. Fourth, consumption may affect the biological aging rate (Case and Deaton, 2005; see also Forster, 2001). We distinguish healthy consumption (such as the consumption of healthy foods, sports and exercise) from unhealthy consumption (such

as smoking, excessive alcohol consumption).[2] Preventive care and healthy consumption are associated with health benefits in that they lower the biological aging rate. Healthy consumption also provides direct utility whereas preventive care is assumed to solely provide health benefits (similar to curative care, individuals demand preventive care solely for the health benefits it provides).[3] We interpret healthy consumption broadly to include decisions regarding housing and neighborhood.[4] Unhealthy consumption provides consumption benefits (utility) but increases the biological rate of aging.

Fifth, we include the decision to withdraw from the labor force (Galama et al. 2008 [Chapter 3]).

Sixth, an essential component of the disparity in health by SES is the observed difference in mortality between SES groups. Further, length of life might be an important determinant of the age of retirement and the level of consumption and health investment over the life-course. Individuals optimize length of life as in Ehrlich and Chuma (1990).[5]

Last, the causal effect of education on income is included in a straightforward manner by assuming a Mincer-type wage relation, in which earnings are increasing in the level of education and the level of experience of workers (e.g., Mincer, 1974).

With the exception of the above seven additional features, the discussion below follows the usual formulation (e.g., Grossman, 1972a; 1972b; 2000; Wagstaff, 1986a; Zweifel and Breyer, 1997). Health is treated as a form of human capital (health capital) and individuals derive both consumption (health provides utility) and production benefits (health increases earnings) from it. Health is modeled as a stock that deteriorates over the life cycle and its deterioration can be counteracted by health investment. The demand for health investment (broadly interpreted as curative and/or preventive care) is a derived demand: individuals demand "good health", not the consumption of curative or preventive care.

[2]It is useful to interpret the endogenous functions as bundles of goods and services (e.g., various consumption goods/services) or composite environmental factors (e.g., various physical and psychosocial health stresses).

[3]The distinction between healthy consumption and preventive care could in practice be difficult for some activities and could differ across individuals (e.g., some individuals exercise because they derive utility from it, whereas others solely exercise because it is healthy).

[4]Living in an affluent neighborhood is an expensive, yet health-promoting and utility-generating choice of individuals. However, the choice of neighborhood (housing) is a constrained choice: low SES individuals cannot afford to live in more affluent areas.

[5]However, to allow qualitative exploration of the characteristics of the solutions we treat mortality and retirement as exogenous (fixed) in this work.

Individuals maximize the life-time utility function

$$\int_0^T U(t)e^{-\beta t}dt, \tag{5.1}$$

where T denotes the life span, β is a subjective discount factor and individuals derive utility $U(t) \equiv U[C_h(t), C_u(t), H(t)]$ from healthy consumption $C_h(t)$, unhealthy consumption $C_u(t)$ and from health $H(t)$. Time t is measured from the time an individual has completed her education and joined the labor force (e.g., around age 25 or so). Utility increases with healthy consumption $\partial U(t)/\partial C_h(t) \geq 0$, unhealthy consumption $\partial U(t)/\partial C_u(t) \geq 0$ and with health $\partial U(t)/\partial H(t) \geq 0$.

The objective function (5.1) is maximized subject to the following dynamic equations,

$$\dot{H}(t) = I_m(t)^\alpha - d(t)H(t), \tag{5.2}$$

$$\dot{A}(t) = \delta A(t) + Y(t) - p_{X_h}(t)X_h(t) - p_{X_u}(t)X_u(t) - p_m(t)m_m(t) - p_p(t)m_p(t), \tag{5.3}$$

the total time budget Ω,

$$\Omega = \tau_w(t) + \tau_{I_m}(t) + \tau_{I_p}(t) + \tau_{C_h}(t) + \tau_{C_u}(t) + s[H(t)], \tag{5.4}$$

and we have initial and end conditions: $H(0)$, $H(T)$, $A(0)$ and $A(T)$ are given.[6]

$\dot{H}(t)$ and $\dot{A}(t)$ in equations (5.2) and (5.3) denote time derivatives of health $H(t)$ and assets $A(t)$. Health (equation 5.2) deteriorates at the biological aging rate $d(t) \equiv d[t, C_h(t), C_u(t), z(t), I_p(t); \xi(t)]$ and can be improved through investment in curative (medical) care $I_m(t)$. The health production function $I_m(t)^\alpha$ is assumed to exhibit DRTS ($0 < \alpha < 1$).[7] The biological aging rate depends endogenously on healthy consumption $C_h(t)$, unhealthy consumption $C_u(t)$, job-related health stress $z(t)$, and investment in preventive care $I_p(t)$ and on a vector of exogenous functions $\xi(t)$. Consumption can be healthy ($\partial d(t)/\partial C_h(t) \leq 0$; e.g., healthy foods, healthy neighborhood) or unhealthy ($\partial d(t)/\partial C_u(t) > 0$; e.g., smoking). Preventive care is modeled analogous to curative care as an activity that provides no utility ($\partial U(t)/\partial I_p(t) = 0$) but is demanded for its health

[6] In Grossman's original formulation (Grossman, 1972a; 1972b) length of life T is determined by a minimum health level $H_{\min}$. If health falls below this level $H(t) \leq H_{min}$ an individual dies, hence $H(T) \equiv H_{\min}$.

[7] Mathematically, this assumption is equivalent to assuming a linear process ($\alpha = 1$) and DRTS in the relation between the inputs of health investment goods / services $m_m(t)$ and own time $\tau_{I_m}(t)$ (as in Ehrlich and Chuma, 1990). Conceptually, however, there is an important distinction. In principle, one could imagine a scenario where the investment function $I_m(t)$ has constant or even increasing returns to scale in its inputs of health investment goods / services $m_m(t)$ and own time $\tau_{I_m}(t)$, but where the resulting health improvement (through the health production process) exhibits diminishing returns to scale in its inputs.

benefits ($\partial d(t)/\partial I_p(t) < 0$). Greater job-related health stress $z(t)$ accelerates the "aging" process ($\partial d(t)/\partial z(t) > 0$).

Assets $A(t)$ (equation 5.3) provide a return δ (the interest rate), increase with income $Y(t)$ and decrease with purchases in the market of healthy consumption goods $X_h(t)$, unhealthy consumption goods $X_u(t)$, curative care $m_m(t)$ and preventive care $m_p(t)$ at prices $p_{X_h}(t)$, $p_{X_u}(t)$, $p_m(t)$ and $p_p(t)$, respectively. Income $Y(t) \equiv Y[H(t), z(t); E, x(t)]$ is assumed to be an increasing function of health $H(t)$ ($\partial Y(t)/\partial H(t) > 0$) and an increasing function of job-related health stress $z(t)$ ($\partial Y(t)/\partial z(t) > 0$; Case and Deaton, 2005). Further, income depends exogenously on the consumer's stock of knowledge (an individual's human capital exclusive of health capital), usually assumed to be a function of years of schooling E and years of working experience $x(t)$ (e.g., Mincer, 1974). Last, we assume that individuals face no borrowing constraints.[8]

Goods and services $m_m(t)$ and $m_p(t)$ as well as own time inputs $\tau_{I_m}(t)$ and $\tau_{I_p}(t)$ are used in the production of curative care $I_m(t)$ and preventive care $I_p(t)$, respectively. Similarly, goods $X_h(t)$ and $X_u(t)$ purchased in the market and own time inputs $\tau_{C_h}(t)$ and $\tau_{C_u}(t)$ are used in the production of healthy and unhealthy consumption, $C_h(t)$ and $C_u(t)$, respectively.[9] The efficiencies of production $\mu_{I_m}(t; E)$, $\mu_{I_p}(t; E)$, $\mu_{C_h}(t; E)$ and $\mu_{C_u}(t; E)$ are assumed to be a function of the consumer's stock of knowledge E as the more educated are assumed to be more efficient consumers and producers of curative (medical) and preventive care (based on the interpretation of education as a productivity factor in own time inputs and in identifying and seeking effective care; Grossman, 1972a; 1972b),

$$I_m(t) \equiv I_m[m_m(t), \tau_{I_m}(t), \mu_{I_m}(t; E)], \tag{5.5}$$

$$I_p(t) \equiv I_p[m_p(t), \tau_{I_p}(t), \mu_{I_p}(t; E)], \tag{5.6}$$

$$C_h(t) \equiv C_h[X_h(t), \tau_{C_h}(t), \mu_{C_h}(t; E)], \tag{5.7}$$

$$C_u(t) \equiv C_u[X_u(t), \tau_{C_u}(t), \mu_{C_u}(t; E)]. \tag{5.8}$$

Further, we implicitly assume that curative care $I_m(t)$, preventive care $I_p(t)$ and job-related health stress $z(t)$ are non negative. We do so by assuming DRTS of the health production function in investment in curative care (see equation 5.2) and diminishing

[8]Imperfect capital markets itself could be a cause of socioeconomic disparities in health if low income individuals face more borrowing constraints than higher income peers, and as such cannot optimally invest in their health.

[9]Because consumption consists of time inputs and purchases of goods/services in the market one can conceive leisure as a form of consumption consisting entirely or mostly of time inputs. Leisure, similar to consumption, provides utility and its cost consists of the price of goods/services utilized and the opportunity cost of time.

marginal benefits for job-related health stress and for investment in preventive care. The notion here is that one cannot "sell" ones health through negative curative care (see Galama and Kapteyn, 2009 [Chapter 2]) or negative preventive care nor can one "buy" health through negative job-related health stress.

The total time available in any period Ω is the sum of all possible uses $\tau_w(t)$ (work), $\tau_{I_m}(t)$ (curative care), $\tau_{I_p}(t)$ (preventive care), $\tau_{C_h}(t)$ (healthy consumption), $\tau_{C_u}(t)$ (unhealthy consumption) and $s[H(t)]$ (sick time). The resulting time budget constraint is shown in equation (5.4).

We follow Grossman (1972a; 1972b; 2000) and assume that income $Y(t)$ is equal to the wage rate $w(t)$ times the amount of time spent working $\tau_w(t)$,

$$Y(t) = w(t)\left\{\Omega - \tau_{I_m}(t) - \tau_{I_p}(t) - \tau_{C_h}(t) - \tau_{C_u}(t) - s[H(t)]\right\}. \tag{5.9}$$

After the age of retirement R we have $\tau_w(t) = 0$ and $Y(t) = b(t)$, where $b(t)$ is a pension benefit function (potentially accrued over time as in Galama et al. 2008 [Chapter 3]).

The wage rate $w(t) \equiv w[t, z(t); E, x(t)]$ is a function of job-related health stress $z(t)$

$$w(t) = w_*(t)[1 + z(t)]^{\gamma_w}, \tag{5.10}$$

where $\gamma_w \geq 0$ and $w_*(t) \equiv w_*[E, x(t)]$ represents the "stressless" wage rate, i.e., the wage rate associated with the least job-related health stress $z(t) = 0$.[10] The stressless wage rate $w_*(t)$ is a function of the consumer's education E and experience $x(t)$ (e.g., Mincer, 1974),

$$w_*(t) = w_E e^{\rho_E E + \beta_x x(t) - \beta_{x^2} x(t)^2}, \tag{5.11}$$

where education E is expressed in years of schooling, $x(t)$ is years of working experience, and ρ_E, β_x and β_{x^2} are constants, assumed to be positive.

Thus, we have the following optimal control problem: the objective function (5.1) is maximized with respect to the control functions $X_h(t)$, $\tau_{C_h}(t)$, $X_u(t)$, $\tau_{C_u}(t)$, $m_m(t)$, $\tau_{I_m}(t)$, $m_p(t)$, $\tau_{I_p}(t)$ and $z(t)$ and subject to the constraints (5.2, 5.3 and 5.4). The Hamiltonian (see, e.g., Seierstad and Sydsaeter, 1977; 1987) of this problem is:

$$\Im = U(t)e^{-\beta t} + q_H(t)\dot{H}(t) + q_A(t)\dot{A}(t), \tag{5.12}$$

where $q_H(t)$ is the adjoint variable associated with the differential equation (5.2) for health $H(t)$ and $q_A(t)$ is the adjoint variable associated with the differential equation (5.3) for assets $A(t)$.

[10] Our model concerns individuals who participate in the labor force. Given that our frame of reference is the labor force we associate $z(t) = 0$ with the least amount of job-related health stress possible in employment, and since there is no obvious scale to job-related health stress we employ the simple relationship shown in equation (5.10).

The conditions for the optimal retirement age R and the optimal length of life T are for the Hamiltonian $\Im$ to equal zero at these ages

$$\Im(R) = 0, \tag{5.13}$$

$$\Im(T) = 0. \tag{5.14}$$

5.3 Solutions

In this section we discuss the first-order conditions for optimization (section 5.3.1), the characteristics of the solutions (section 5.3.2), the effect of SES on behavior (section 5.3.3), and the effect of health on behavior (section 5.3.4). Throughout the discussion we assume that an interior solution to the optimization problem exists.

5.3.1 First-order conditions

The first-order condition for maximization of (5.12) with respect to the control function health investment is

$$\frac{\partial U(t)}{\partial H(t)} = q_A(0)\left[\sigma_H(t) - \varphi_H(t)\right] e^{(\beta-\delta)t}, \tag{5.15}$$

where the Lagrange multiplier $q_A(0)$ is the shadow price of wealth (see, e.g, Case and Deaton, 2005), $\sigma_H(t) \equiv \sigma_H[t, I_m(t), C_h(t), C_u(t), z(t), I_p(t); E, x(t), \xi(t)]$ is the user cost of health capital at the margin

$$\sigma_H(t) \equiv \pi_{I_m}(t)\left[d(t) + \delta - \widetilde{\pi_{I_m}}(t)\right], \tag{5.16}$$

$\pi_{I_m}(t) \equiv \pi_{I_m}[t, I_m(t), z(t); E, x(t)]$ is the marginal monetary cost of curative care $I_m(t)$

$$\pi_{I_m}(t) \equiv \frac{p_m(t) I_m(t)^{1-\alpha}}{\alpha[\partial I_m(t)/\partial m_m(t)]} = \frac{w(t) I_m(t)^{1-\alpha}}{\alpha[\partial I_m(t)/\partial \tau_{I_m}(t)]}, \tag{5.17}$$

and $\widetilde{\pi_{I_m}}(t) = \pi_{I_m}(t)^{-1}\left(\partial \pi_{I_m}(t)/\partial t\right)$.[11] The marginal monetary cost of curative care (equation 5.17) is a function of the price of medical goods and services purchased in the market $p_m(t)$ and the opportunity cost of time $w(t)$ (hence monetary). Note that the marginal monetary cost of investment in curative care $\pi_{I_m}(t)$ increases with the level of investment in curative care $I_m(t)$ due to decreasing-returns-to-scale of the health production function

[11] In the remainder of this paper the symbol $\sim$ is used to denote the relative time derivative of a function: $\widetilde{f(t)} \equiv \frac{\partial f(t)}{\partial t} f(t)^{-1}$.

$I_m(t)^\alpha$ ($0 < \alpha < 1$; see equation 5.2). Further, $\varphi_H(t) \equiv \varphi_H[t, H(t), z(t); E, x(t)]$ is the marginal production benefit of health

$$\varphi_H(t) \equiv \frac{\partial Y(t)}{\partial H(t)}, \tag{5.18}$$

reflecting the notion that health increases earnings $Y(t)$.

The first-order condition for maximization of (5.12) with respect to the control function healthy consumption is

$$\frac{\partial U(t)}{\partial C_h(t)} = q_A(0)\left[\pi_{C_h}(t) - \varphi_{dC_h}(t)\right] e^{(\beta-\delta)t}, \tag{5.19}$$

where $\pi_{C_h}(t) \equiv \pi_{C_h}[t, C_h(t), z(t); E, x(t)]$ is the marginal monetary cost of healthy consumption $C_h(t)$

$$\pi_{C_h}(t) \equiv \frac{p_{X_h}(t)}{\partial C_h(t)/\partial X_h(t)} = \frac{w(t)}{\partial C_h(t)/\partial \tau_{C_h}(t)}, \tag{5.20}$$

and $\varphi_{dC_h}(t) \equiv \varphi_{dC_h}[t, H(t), I_m(t), C_h(t), C_u(t), z(t), I_p(t); E, x(t), \xi(t)]$ is the marginal health benefit of healthy consumption

$$\varphi_{dC_h}(t) \equiv -\pi_{I_m}(t)\frac{\partial d(t)}{\partial C_h(t)}H(t). \tag{5.21}$$

The marginal monetary cost of healthy consumption $\pi_{C_h}(t)$ (equation 5.20) is a function of the price of healthy consumption goods and services $p_{X_h}(t)$ and the opportunity cost of time $w(t)$, and represents the *direct* monetary cost of consumption. The marginal health benefit of healthy consumption $\varphi_{dC_h}(t)$ (equation 5.21), is equal to the product of the marginal monetary cost of investment in curative care $\pi_{I_m}(t)$ and the "amount" of health saved $[\partial d(t)/\partial C_h(t)]H(t)$, and represents the marginal monetary value of health saved.[12] Similarly, the first-order condition for maximization of (5.12) with respect to the control function unhealthy consumption is

$$\frac{\partial U(t)}{\partial C_u(t)} = q_A(0)\left[\pi_{C_u}(t) + \pi_{dC_u}(t)\right] e^{(\beta-\delta)t}, \tag{5.22}$$

where $\pi_{C_u}(t) \equiv \pi_{C_u}[t, C_u(t), z(t); E, x(t)]$ is the marginal monetary cost of unhealthy consumption $C_u(t)$ (*direct* monetary cost)

$$\pi_{C_u}(t) \equiv \frac{p_{X_u}(t)}{\partial C_u(t)/\partial X_u(t)} = \frac{w(t)}{\partial C_u(t)/\partial \tau_{C_u}(t)}, \tag{5.23}$$

[12]The marginal health benefit can be understood intuitively as the reduced need for health investment because of a lower health deterioration rate. While the health benefit is expressed in terms of the marginal cost of curative care, this is essentially arbitrary, as the monetary value of health saved could also be expressed in terms of the reduced need for other types of health investments such as preventive care or healthy consumption.

and $\pi_{dC_u}(t) \equiv \pi_{dC_u}[t, H(t), I_m(t), C_h(t), C_u(t), z(t), I_p(t); E, x(t), \xi(t)]$ is the marginal health cost of unhealthy consumption (marginal monetary value of health lost)

$$\pi_{dC_u}(t) \equiv \pi_{I_m}(t)\frac{\partial d(t)}{\partial C_u(t)}H(t). \tag{5.24}$$

The first-order condition for maximization of (5.12) with respect to the control function job-related health stress is

$$\pi_{dz}(t) \;\; = \;\; \varphi_z(t), \tag{5.25}$$

where $\pi_{dz}(t) \equiv \pi_{dz}[t, H(t), I_m(t), C_h(t), C_u(t), z(t), I_p(t); E, x(t), \xi(t)]$ is the marginal health cost of job-related health stress (marginal monetary value of health lost)

$$\pi_{dz}(t) \equiv \pi_{I_m}(t)\frac{\partial d(t)}{\partial z(t)}H(t), \tag{5.26}$$

and $\varphi_z(t) \equiv \varphi_z[t, H(t), z(t); E, x(t)]$ is the marginal production benefit of job-related health stress

$$\varphi_z(t) \equiv \frac{\partial Y(t)}{\partial z(t)}, \tag{5.27}$$

reflecting the notion that job-related health stress is associated with a compensating wage differential (greater earnings).

Lastly, the first-order condition for maximization of (5.12) with respect to the control function preventive care is

$$\pi_{I_p}(t) \;\; = \;\; \varphi_{dI_p}(t), \tag{5.28}$$

where $\pi_{I_p}(t) \equiv \pi_{I_p}[t, z(t), I_p(t); E, x(t)]$ is the marginal monetary cost of preventive care $I_p(t)$

$$\pi_{I_p}(t) \equiv \frac{p_p(t)}{\partial I_p(t)/\partial m_p(t)} = \frac{w(t)}{\partial I_p(t)/\partial \tau_{I_p}(t)}, \tag{5.29}$$

and $\varphi_{dI_p}(t) \equiv \varphi_{dI_p}[t, H(t), I_m(t), C_h(t), C_u(t), z(t), I_p(t); E, x(t), \xi(t)]$ is the marginal health benefit of preventive care (marginal monetary value of health saved)

$$\varphi_{dI_p}(t) \equiv -\pi_{I_m}(t)\frac{\partial d(t)}{\partial I_p(t)}H(t). \tag{5.30}$$

The five first-order equations (5.15, 5.19, 5.22, 5.25 and 5.28) and the transversality conditions (5.13) and (5.14) define the dynamics of the problem we are interested in. Solving the first-order equations provides solutions for the time paths of the control functions $I_m(t)$, $C_h(t)$, $C_u(t)$, $z(t)$ and $I_p(t)$. The state functions health $H(t)$ and assets $A(t)$ can subsequently be obtained through the dynamic equations (5.2) and (5.3). Lastly, the

optimal retirement age R and the optimal length of life T follow from the transversality conditions (5.13) and (5.14).

We have thus arrived at a life cycle model that incorporates labor force participation, healthy and unhealthy consumption (including housing, neighborhood social environment), health, curative (medical) and preventive care, job-related physical and psychosocial health stresses, wealth and mortality.

The Grossman model (Grossman, 1972a; 1972b) is a special case of our model and is defined by the first-order equations (5.15) and (5.19) for an exogenous biological aging rate $d(t)$. The first-order conditions (5.19), (5.22) and (5.25) are similar (but not identical) to those presented by Case and Deaton (2005). Ehrlich and Chuma (1990) have extended the Grossman model with the transversality condition (5.14) for optimal length of life T. The inclusion of endogenous retirement follows Galama et al. (2008; see Chapter 3).

5.3.2 Characteristics of the solutions

In the remainder of this paper we qualitatively explore the properties of the solutions for health $H(t)$, investment in curative care $I_m(t)$, investment in preventive care $I_p(t)$, healthy consumption $C_h(t)$ and unhealthy consumption $C_u(t)$ and job-related health stress $z(t)$. We do so by assessing the effects of parameter changes on the endogenous functions of interest, utilizing stylized representations (graphs) of the first-order conditions. However, stylized representations are less useful in assessing the nature of the transversality conditions for retirement R (5.13) and length of life T (5.14); this requires numerical approaches to solving the model. Thus, in practice, we explore the characteristics of the model conditional on retirement age R and length of life T (i.e., for fixed R and T).[13]

Assumptions

In the remainder, we assume:

1. Diminishing returns to scale (DRTS) in the health production function $I_m(t)^\alpha$ $(0 < \alpha < 1)$,

[13]Treating retirement R and length of life T as exogenous (fixed) does not significantly affect our qualitative results regarding the formation of the SES health gradient (discussed in this work). Optimizing the age of retirement R and length of life T affects the overall level of health investment and consumption over the life cycle, as the transversality conditions (5.13) and (5.14) in combination with the initial and end conditions ($A(0)$, $A(T)$, $H(0)$ and $H(T)$), determine the parameters $q_A(0)$ and $q_H(0)$ in equations (5.31) and (5.32), but have limited effect on the direction of changes in the level of health investment and consumption.

2. Diminishing marginal utilities of healthy $C_h(t)$ and unhealthy consumption $C_u(t)$ and of health $H(t)$,

$$\frac{\partial^2 U(t)}{\partial C_h(t)^2} < 0, \quad \frac{\partial^2 U(t)}{\partial C_u(t)^2} < 0, \quad \frac{\partial^2 U(t)}{\partial H(t)^2} < 0;$$

3. Diminishing marginal production benefit of health $\varphi_H(t)$, diminishing marginal production benefit of job-related health stress $\varphi_z(t)$, diminishing marginal health benefit of healthy consumption $\varphi_{dC_h}(t)$, and diminishing marginal health benefit of investment in preventive care $\varphi_{dI_p}(t)$

$$\begin{aligned}
\frac{\partial \varphi_H(t)}{\partial H(t)} &= \frac{\partial^2 Y(t)}{\partial H(t)^2} < 0, \\
\frac{\partial \varphi_z(t)}{\partial z(t)} &= \frac{\partial^2 Y(t)}{\partial z(t)^2} < 0, \\
\frac{\partial \varphi_{dC_h}(t)}{\partial C_h(t)} &= -\pi_{I_m}(t)\frac{\partial^2 d(t)}{\partial C_h(t)^2} H(t) < 0, \\
\frac{\partial \varphi_{dI_p}(t)}{\partial I_p(t)} &= -\pi_{I_m}(t)\frac{\partial^2 d(t)}{\partial I_p(t)^2} H(t) < 0;
\end{aligned}$$

4. Constant returns to scale (CRTS) in the marginal health cost of unhealthy consumption $\pi_{dC_u}(t)$ and in the marginal health cost of job-related health stress $\pi_{dz}(t)$[14]

$$\begin{aligned}
\frac{\partial \pi_{dC_u}(t)}{\partial C_u(t)} &= \pi_{I_m}(t)\frac{\partial^2 d(t)}{\partial C_u(t)^2} H(t) = 0, \\
\frac{\partial \pi_{dz}(t)}{\partial z(t)} &= \frac{\partial}{\partial z(t)}\left[\pi_{I_m}(t)\frac{\partial d(t)}{\partial z(t)}\right] H(t) = 0;
\end{aligned}$$

5. CRTS in the inputs (goods/services purchased in the market and own-time) for investment in curative care $I_m(t)$, healthy consumption $C_h(t)$, unhealthy consumption

[14]While it seems plausible that the health benefits of investment in curative care, healthy consumption and investment in preventive care exhibit diminishing returns to scale, it is unclear whether the health costs of unhealthy consumption and job-related health stress exhibit decreasing or increasing returns to scale. For example, Forster (2001) assumes decreasing returns to scale for healthy consumption and increasing returns to scale for unhealthy consumption. In simple terms: escalating risky behavior (e.g., illicit drug use) or more hours of dangerous work can lead to rapid health deterioration, whereas after a certain point more investment in curative or preventive care, more exercise or more consumption of healthy foods does not prevent eventual aging. Since it is unclear a priori whether the effect of unhealthy consumption and the effect of job-related health stress on the biological aging rate $d(t)$ exhibits in- or decreasing returns to scale, we assume CRTS for simplicity.

$C_u(t)$ and preventive care $I_p(t)$.[15] As a result we have (see equations 5.17, 5.20, 5.23 and 5.29):

$$\pi_{I_m}(t) \propto I_m(t)^{1-\alpha}, \;\; \frac{\partial \pi_{C_h}(t)}{\partial C_h(t)} = 0, \;\; \frac{\partial \pi_{C_u}(t)}{\partial C_u(t)} = 0, \;\; \frac{\partial \pi_{I_p}(t)}{\partial I_p(t)} = 0;$$

6. Complementarity in utility of consumption $C_h(t)$, $C_u(t)$ and health $H(t)$[16]

$$\frac{\partial^2 U(t)}{\partial C_h(t) \partial H(t)} > 0, \;\; \frac{\partial^2 U(t)}{\partial C_u(t) \partial H(t)} > 0;$$

7. Substitutability in utility of healthy $C_h(t)$ and unhealthy $C_u(t)$ consumption[17]

$$\frac{\partial^2 U(t)}{\partial C_h(t) \partial C_u(t)} < 0;$$

Assumptions 1 through 5 ensure that solutions to the optimal control problem exist.[18] The remaining assumptions are made to illustrate the potential of the model to describe a wide range of behaviors.

Stylized representations

Figures 5.1 and 5.2 provide stylized representations of the first-order conditions for health investment $I_m(t)$ (equation 5.15), healthy consumption $C_h(t)$ (equation 5.19), unhealthy consumption $C_u(t)$ (equation 5.22), job-related health stress $z(t)$ (equation 5.25) and investment in preventive care $I_p(t)$ (equation 5.28). In Section 5.3.3 we consider individuals a and b who differ in one particular SES indicator, but are otherwise identical, and in Section 5.3.4 we consider individuals a and c who differ in their health, but are otherwise identical. Figures 5.1 and 5.2 therefore show each of the five first-order conditions for

[15] A priori it is not clear whether the relationships between the inputs (good/services and own time) and investment exhibit decreasing- or increasing-returns-to-scale. Hence we assume CRTS in these relations for simplicity.

[16] Indeed, Finkelstein et al. (2008) find evidence that the marginal utility of consumption declines as health deteriorates. This would rule out strongly separable functional forms for the utility function where the marginal utility of consumption is independent of health and forms where the marginal utility of consumption would decrease in health.

[17] The substitutability in utility of healthy $C_h(t)$ and unhealthy $C_u(t)$ consumption allows us to model substitution from unhealthy to healthy consumption (or vice versa).

[18] Optimal solutions for the state functions $A(t)$, $H(t)$ and the control functions $X_h(t)$, $\tau_{C_h}(t)$, $X_u(t)$, $\tau_{C_u}(t)$, $m_m(t)$, $\tau_{I_m}(t)$, $m_p(t)$, $\tau_{I_p}(t)$ and $z(t)$ exist if the Hamiltonian $\Im$ (see equations 5.2, 5.3 and 5.12) is concave in each of the state and control functions and differentiable w.r.t. the state and control functions (see, e.g., Seierstad and Sydsaeter, 1977; 1987).

individuals a and b or c. In this section we do not yet vary SES or health indicators and focus on the curves labeled with a.

Investment in curative care:
The solution for investment in curative care $I_m(t)$ is determined by the first-order condition for health investment (5.15), conditional on the level of the health stock $H(t)$.[19] The evolution of the health stock $H(t)$ then follows from the initial condition $H(0)$ and the health investment $I_m(s)$ and biological aging $d(s)$ histories ($s < t$) through the dynamic equation (5.2).

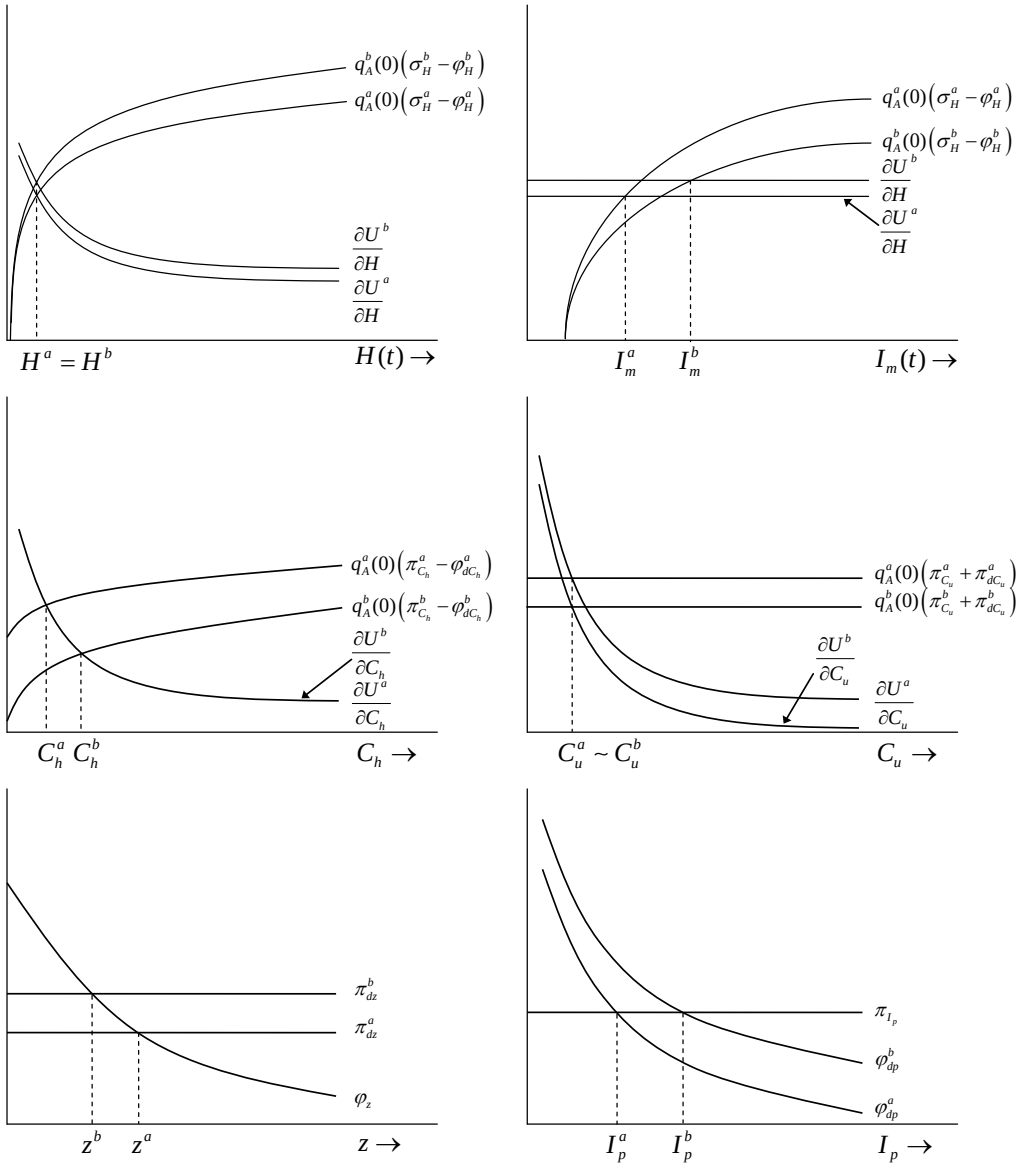
The first-order condition for health investment (5.15) equates the consumption benefit of health $\partial U(t)/\partial H(t)$ with the cost of maintaining the health stock $q_A(0)[\sigma_H(t) - \varphi_H(t)]e^{(\beta-\delta)t}$. Figure 5.2 shows a simple stylized representation of this relation as a function of health $H(t)$ (top left-hand panel) and as a function of investment in curative care $I_m(t)$ (top right-hand panel).

Consider the top left-hand panel and individual a first. The marginal utility of health (labeled $\partial U^a/\partial H$) is diminishing in health (assumption 2). The user cost of health capital $\sigma_H(t) = \pi_{I_m}(t)[d(t) + \delta - \widetilde{\pi_{I_m}}(t)]$ is independent of health and the marginal production benefit of health $\varphi_H(t) = \partial Y(t)/\partial H(t)$ (increased earnings) exhibits DRTS in health (assumption 3). The resulting curve (labeled $q_A(0)(\sigma_H^a - \varphi_H^a)$) is upward sloping in health. Since health is a state function its level is given and provides a constraint: the two curves have to intersect at H^a.

Now consider the top right-hand panel of Figure 5.2. The marginal monetary cost of curative care $\pi_{I_m}(t)$ and hence the user cost of health capital $\sigma_H(t)$ is increasing in the level of curative care $I_m(t)$ ($\pi_{I_m}(t) \propto I_m(t)^{1-\alpha}$; see equations 5.16 and 5.17; assumptions 1 and 5). The marginal production benefit of health $\varphi_H(t)$ (see equations 5.9 and 5.18) is

[19] Note that the first-order condition (5.15) is interpreted in the health production literature spawned by Grossman (1972a; 1972b) as the condition for optimal health, and not as the condition for optimal health investment. However, this condition was derived by optimizing the optimal control problem with respect to health investment (it follows directly from relations 5.32, 5.37 and 5.38) and hence an alternative interpretation is that it determines the optimal level of the control function health investment $I_m(t)$. Health $H(t)$ is a state function and is determined by the dynamic relation (5.2). At a given time t an individual cannot decide about its level (hence conditional). This seemingly subtle difference in interpretation of the first-order condition (together with the assumption of DRTS in the health production function) addresses a number of issues with the health production literature and allows us to accommodate a wider range of health behaviors than existing health production models (see Galama, 2011 [Chapter 4]). Importantly, the first-order condition (5.15) is of a simpler form than the condition used in the health production literature, allowing us to develop a better understanding of the characteristics of the optimal solution for health investment.

Figure 5.1: Differences in SES



Notes: Marginal consumption $\partial U/\partial H$ and marginal production benefit φ_H of health versus the user cost of health capital at the margin σ_H as a function of health (top left) and as a function of health investment (top right). Marginal utility of healthy consumption $\partial U/\partial C_h$ versus the marginal monetary cost π_{C_h} and the health benefit φ_{dC_h} of healthy consumption $C_h(t)$ (center left); Marginal utility of unhealthy consumption $\partial U/\partial C_u$ versus the marginal monetary cost π_{C_u} and the marginal health cost φ_{dC_u} of unhealthy consumption $C_u(t)$ (center right); Marginal health cost π_{dz} versus the marginal production benefit φ_z of job-related health stress $z(t)$ (bottom-left); Marginal monetary cost π_{I_p} versus the marginal health benefit φ_{dI_p} of investment in preventive care $I_p(t)$ (bottom-right). In labeling the curves we have omitted the time varying term with exponent $(\beta - \delta)t$.

independent of the level of investment in curative care $I_m(t)$. The resulting curve is upward sloping (labeled $q_A(0)(\sigma_H^a - \varphi_H^a)$). Further, the marginal utility of health $\partial U(t)/\partial H(t)$ is independent of the level of investment in curative care $I_m(t)$ (horizontal line labeled $\partial U^a/\partial H$). Its level is determined by the level of the health stock H^a (draw a horizontal line from the top left-hand to the top right-hand panel of Figure 5.2). The intersection of the two curves determines the optimal level of investment in curative care I_m^a.

The top-right hand panel of Figure 5.2 further illustrates that the level of investment in curative care $I_m(t)$ increases with the consumption benefit of health $\partial U(t)/\partial H(t)$, the production benefit of health $\varphi_H(t)$, and with wealth (lower $q_A(0)$), and decreases with the user cost of health capital $\sigma_H(t)$. Further, the level of health investment $I_m(t)$ is a function of the health stock $H(t)$ (more details are provided in section 5.3.4).

Healthy and unhealthy consumption:
The center-left panel of Figure 5.2 shows the first-order condition for healthy consumption $C_h(t)$ (equation 5.19) which equates the marginal utility of healthy consumption (solid line labeled $\partial U^a/\partial C_h$) to the net marginal cost of healthy consumption (solid line labeled $q_A(0)(\pi_{C_h}^a - \varphi_{dC_h}^a)$). The marginal utility of healthy consumption is diminishing in the level of consumption (assumption 2). The net marginal cost of healthy consumption increases with the marginal monetary cost of healthy consumption $\pi_{C_h}(t)$ (equation 5.20; CRTS [assumption 5]) and decreases with the marginal health benefit $\varphi_{dC_h}(t) = -\pi_{I_m}(t)[\partial d(t)/\partial C_h(t)]H(t)$ (DRTS [assumption 3]). Hence, the net marginal cost of healthy consumption is upward sloping. The point of intersection defines the optimal solution for healthy consumption C_h^a (vertical dashed line).

The center-right panel of Figure 5.2 shows the first-order condition for unhealthy consumption $C_u(t)$ (equation 5.22). The first-order condition is similar to the condition for healthy consumption described in the preceding paragraph. The difference lies in the marginal health cost (rather than health benefit) of unhealthy consumption $\pi_{dC_u}(t) = \pi_{I_m}(t)[\partial d(t)/\partial C_u(t)]H(t)$ (CRTS [assumption 4]) which has to be added rather than subtracted from the marginal monetary cost of unhealthy consumption $\pi_{C_u}(t)$ (equation 5.23; CRTS [assumption 5]). The net marginal cost of unhealthy consumption is represented by the solid horizontal line labeled $q_A(0)(\pi_{C_u}^a + \pi_{dC_u}^a)$. The point of intersection defines the optimal solution for unhealthy consumption C_u^a (vertical dashed line).

The level of healthy $C_h(t)$ and unhealthy $C_u(t)$ consumption increases with the marginal utility of consumption ($\partial U(t)/\partial C_h(t)$ and $\partial U(t)/\partial C_u(t)$), increases with wealth (lower $q_A(0)$), decreases with the marginal monetary costs of consumption ($\pi_{C_h}(t)$ and $\pi_{C_u}(t)$), increases with the marginal health benefit of healthy consumption $\varphi_{dC_h}(t)$, and decreases

with the marginal health cost of unhealthy consumption $\pi_{dC_u}(t)$.

Job-related health stress:
The bottom-left panel of Figure 5.2 shows the first-order condition for job-related health stress $z(t)$ (equation 5.25) which equates the production benefit $\varphi_z(t) = \partial Y(t)/\partial z(t)$ (increased earnings; DRTS [assumption 3]) to the marginal health cost of job-related health stress $\pi_{dz}(t) = \pi_{I_m}(t)[\partial d(t)/\partial z(t)]H(t)$ (CRTS [assumption 4]). The optimal solution for job-related health stress z^a is indicated by the vertical dashed line. The optimal level increases with the marginal production benefit $\varphi_z(t)$ and decreases with the marginal health cost $\pi_{dz}(t)$.

Investment in preventive care:
The bottom-right panel of Figure 5.2 represents the first-order condition for investment in preventive care $I_p(t)$ (equation 5.28) which equates the marginal health benefit of investment in preventive care $\varphi_{dI_p}(t) = -\pi_{I_m}(t)[\partial d(t)/\partial I_p(t)]H(t)$ (DRTS [assumption 3]) to the marginal monetary cost $\pi_{I_p}(t)$ (equation 5.29; CRTS [assumption 5]). The optimal solution for investment in preventive care I_p^a is indicated by the vertical dashed line. The optimal level of investment in preventive care $I_p(t)$ increases with the marginal health benefit $\varphi_{dI_p}(t)$ and decreases with the marginal monetary cost $\pi_{I_p}(t)$ of investment in preventive care.

5.3.3 SES and its effect on behavior

In this section we explore the (cumulative) effect on health over the life cycle of choices made in curative care, in life style and in working environment. Our emphasis is on exploring differences in constraints (e.g., in wealth, skills, experience, education and prices).

Common measures of SES employed in empirical research are wealth, earnings (income) and education. In the following subsections we discuss the relations between wealth and health, earnings and health and education and health. We consider two individuals a and b who differ in one particular SES indicator, but are otherwise identical. Both individuals have the same initial level of health $H(t)$, are of the same age t, face the same environments (e.g., same interest rate δ), and have the same preferences (i.e., same utility function $U[C_h(t), C_u(t), H(t)]$ and same time preference β).[20] We are interested in the

[20]Part of the SES health gradient may be explained by differences in individuals preferences. A lower rate of time preference β operates in a similar manner to wealth, earnings and education. However in contrast to SES, differences between low and high discounting individuals grow larger with time (the discount factor $e^{(\beta-\delta)t}$ grows slower with age t for an individual with a low discount rate). A lower rate

predictions of our model for the subsequent evolution of health for these two individuals, given a ceteris paribus change in one SES indicator.

Wealth and health: pure "asset" effect

Consider two individuals a and b who differ in life-time wealth $q_A(0)$. Individual b has greater life-time wealth, i.e., $q_A^b(0) < q_A^a(0)$, but is otherwise identical. Because of the similarities between the two individuals the difference in life-time wealth is to be interpreted as due to differences in endowed physical capital (e.g., assets $A(0)$).[21]

Investment in curative care:

Figure 5.1 shows a stylized representation of the first-order condition for investment in curative care as a function of health $H(t)$ (top left-hand panel) and as a function of investment in curative care $I_m(t)$ (top right-hand panel). Consider the top right-hand panel first. As a result of greater endowed wealth ($q_A^b(0) < q_A^a(0)$) the net marginal cost of maintaining the health stock shifts downward (curve labeled $q_A^b(0)(\sigma_H^b - \varphi_H^b)$). As a result, the optimal level of investment in curative care is higher $I_m^b > I_m^a$.

An indirect effect operates through consumption. Greater endowed wealth allows individual b to consume more consumption goods and services (see discussion below for further detail). A higher level of consumption increases (or at a minimum leaves unchanged) the marginal utility of health $\partial U(t)/\partial H(t)$ (assumption 6). Thus the marginal utility of health shifts upward (or is unchanged) in both the top left-hand panel and in the top right-hand panel of Figure 5.1 (curves labeled $\partial U^b/\partial H$). This reinforces the effect on curative care, and wealthier individuals invest more in curative care than less affluent peers $I_m^b > I_m^a$.

Now turn to the top-left panel of Figure 5.1. Because the health stock of individuals a and b is the same $H^b = H^a$ the curves need to intersect at the same level of health. The net result is an upward shift in the marginal cost of maintaining the health stock (curve labeled $q_A^b(0)(\sigma_H^b - \varphi_H^b)$), as a result of the upward shift in the marginal utility of health

of time preference may also lead to greater investment in education (not part of our theory) and hence lead to joint determination of health and education (e.g., Fuchs, 1982; 1986).

[21] Endowments need not necessarily be available to the individual at age $t = 0$, but could also be received at later ages. Conceptually there is no distinction between early and late endowments: an endowment at later ages also lowers life-time wealth $q_A(0)$. Our model is deterministic and the individual knows with certainty about the timing and amount of physical assets she will receive.

$\partial U/\partial H$.[22,23]

Healthy and unhealthy consumption:
The center-left panel of Figure 5.1 shows the shift in the level of healthy consumption $C_h(t)$. The product $q_A(0)\pi_{C_h}(t)$ shifts downward as a result of greater endowed physical capital ($q_A^b(0) < q_A^a(0)$) and because the marginal monetary cost of healthy consumption $\pi_{C_h}(t)$ (equation 5.20) is unchanged. Essentially, greater endowed assets enable more purchases of healthy consumption goods. Further, endowed assets increase the health benefit of healthy consumption, $\varphi_{dC_h}(t) = -\pi_{I_m}(t)[\partial d(t)/\partial C_h(t)]H(t)$ (for $H^b = H^a$ and $\pi_{I_m}^b > \pi_{I_m}^a$ [see earlier discussion in "Investment in curative care"]). The resulting net marginal cost of healthy consumption (solid line labeled $q_A^b(0)(\pi_{C_h}^b - \varphi_{dC_h}^b)$) is lower in level (the "wealth" effect) and steeper in slope (the "savings in care" effect; $\varphi_{dC_h}^b > \varphi_{dC_h}^a$). In the example the marginal utility of healthy consumption (curve labeled $\partial U^b/\partial C_h$; center-left panel of Figure 5.1) is shown as unchanged (i.e., we have assumed the level of unhealthy consumption $C_u(t)$ has not changed). The optimal solution for healthy consumption C_h^c (vertical dashed line) of an individual with greater endowed wealth is higher than that of a poorer individual $C_h^b > C_h^a$.

The center-right panel of Figure 5.1 shows the shift in the level of unhealthy consumption $C_u(t)$. As with healthy consumption, greater endowed wealth shifts the product of the shadow price of wealth $q_A(0)$ (lowered) and the marginal monetary cost of unhealthy consumption $\pi_{C_u}(t)$ (unchanged; equation 5.23) downward. Unlike healthy consumption this shift is countered by an increase in the marginal health cost of unhealthy consumption $\pi_{dC_u}(t) = \pi_{I_m}(t)[\partial d(t)/\partial C_u(t)]H(t)$ (for $H^c = H^a$ and $\pi_{I_m}^b > \pi_{I_m}^a$; see earlier discussion in "Investment in curative care"). Greater endowed wealth allows purchasing more unhealthy consumption goods, but also increases the marginal health cost. Further, the marginal utility of unhealthy consumption $\partial U(t)/\partial C_u(t)$ shifts downward as a result of the higher level of healthy consumption $C_h(t)$ (assumption 7; see previous paragraph). The optimal level of unhealthy consumption C_u^b (vertical dashed line) is shown as un-

[22]One can obtain the same result as follows. Since, $\partial U^b/\partial H \geq \partial U^a/\partial H$ we have $q_A^b(0)[\sigma_H^b(t) - \varphi_H^b(t)] \geq q_A^a(0)[\sigma_H^a(t) - \varphi_H^a(t)]$ (equation 5.15) and the net result is an upward shift of the curve labeled $q_A^b(0)(\sigma_H^b - \varphi_H^b)$.

[23]While the net marginal cost of maintaining the health stock (curve labeled $q_A^b(0)(\sigma_H^b - \varphi_H^b)$) shifts downward as a result of greater endowed wealth ($q_A^b(0) < q_A^a(0)$), it shifts upward due to a higher user cost of health capital $\sigma_H(t)$, since a higher optimal level of investment in curative care ($I_m^b > I_m^a$) increases the marginal monetary cost of curative care $\pi_{I_m}(t)$ ($\pi_{I_m}(t) \propto I_m(t)^{1-\alpha}$) (see equation 5.16 and assumption 5).

changed $C_u^b \sim C_u^a$.[24]

Job-related health stress and investment in preventive care:
The first-order conditions for job-related health stress $z(t)$ (equation 5.25; bottom-left corner of Figure 5.1) and for investment in preventive care (equation 5.28; bottom-right corner of Figure 5.1) do not depend on life-time wealth $q_A(0)$. However, there is an indirect effect of greater endowed wealth. Both the marginal health cost of job-related health stress $\pi_{dz}(t)$ (equation 5.26) and the marginal health benefit of preventive care $\varphi_{dI_p}(t)$ (equation 5.30) are proportional to the marginal monetary cost of investment in curative care $\pi_{I_m}(t)$. Higher endowed wealth (individual b) implies $\pi_{I_m}^b(t) > \pi_{I_m}^a(t)$ (see earlier discussion in "Investment in curative care"). Thus wealthier individuals have greater marginal health cost of job-related health stress $\pi_{dz}(t)$ and greater marginal health benefit of preventive care $\varphi_{dI_p}(t)$. Consequently the optimal level of job-related health stress is lower $z^b < z^a$ and the optimal level of investment in preventive care is higher $I_p^b > I_p^a$ for individuals with greater endowed wealth, compared to less-affluent peers.

Income and health: pure "wage" effect

Again, consider two individuals a and b but this time the difference is in their wage rate. Individual b has a higher "stressless" wage rate than individual a ($w_*^b(t) > w_*^a(t)$) and hence has a higher level of earnings over the life cycle $Y^b(t) > Y^a(t)$.[25] It is important to distinguish between an *evolutionary wage change* (differences along the wage path of an individual) and *differences in life-time wage profiles* (between individuals).

Evolutionary wage change: In our model of perfect certainty a change in wage does not affect the parameter $q_A(0)$ (life-time wealth) as the change is fully anticipated by the individual. Such a response is referred to as an evolutionary wage change (along an individual's wage profile). An evolutionary increase in the wage rate $w(t)$ increases

[24]Note that the marginal utility of healthy consumption $\partial U(t)/\partial C_h(t)$ is unchanged if the level of unhealthy consumption $C_u(t)$ is unchanged (as was assumed).

[25]Earnings $Y(t)$ are a function of the wage rate $w(t)$ times the amount of time spent working $\tau_w(t)$ (see equation 5.9). A higher wage rate $w_*(t)$ implies that individual b has higher earnings $Y(t)$ than individual a because the direct effect of higher wages is to increase earnings (the wage rate multiplied by the time spent working). A secondary effect operates through time spent working, where individuals may work more because of the higher opportunity cost of not working (substitution effect). On the other hand individuals may work fewer hours to spend their increased income on care or consumption (income effect). Empirical studies suggest that the substitution and income effects are of the same magnitude (e.g., Blundell and MaCurdy, 2000) and hence that the direct effect of a wage increase is to increase earnings.

the marginal production benefit of health $\varphi_H(t) = \partial Y(t)/\partial H(t)$ and of job-related health stress $\varphi_z(t) = \partial Y(t)/\partial z(t)$ (see equations 5.9 and 5.10).[26] It also increases the opportunity cost of time.[27] As a result, the various marginal costs and benefits of the functions of interest increase, and the net effect on the level of investment in curative care, healthy consumption, job-related health stress, and preventive care is ambiguous. An exception is the level of unhealthy consumption, which is lower since both the marginal monetary cost and the marginal health cost of unhealthy consumption increase with the wage rate. Thus, an evolutionary wage increase could be either good or bad for health.

However, consider the case where the marginal production benefit of health $\varphi_H(t)$ is small compared to the user cost of health capital $\sigma_H(t)$.[28] Since individuals a and b possess the same health stock, and $\sigma_H(t) - \varphi_H(t) \sim \sigma_H(t)$, it follows from the first-order condition for health investment (equation 5.15) that $\sigma_H(t)$ is unchanged and hence $\pi_{I_m}(t)$ is unchanged. Yet a higher wage rate increases the opportunity cost of time, and consequently the level of investment in curative care $I_m(t)$ is lower.[29] Further, the health benefit of healthy consumption $\varphi_{dC_h}(t)$ and of preventive care $\varphi_{dI_p}(t)$ (equations 5.21 and 5.30), and the health cost of unhealthy consumption $\pi_{dC_u}(t)$ and of job-related health stress $\pi_{dz}(t)$ (equations 5.24 and 5.26) are unchanged (because $\pi_{I_m}(t)$ is unchanged). The marginal monetary cost of healthy consumption $\pi_{C_h}(t)$ (equation 5.20) and of unhealthy consumption $\pi_{C_u}(t)$ (equation 5.23) however increase with the wage rate $w(t)$ (reflecting the higher opportunity cost of time) and the level of healthy $C_h(t)$ and unhealthy $C_u(t)$ consumption is lower. In addition, the marginal production benefit of job-related health stress $\varphi_z(t)$ increases with the wage rate (equation 5.27) as does the marginal monetary cost of investment in preventive care $\pi_{I_p}(t)$ (equation 5.29). As a result, the level of job-related health stress $z(t)$ is higher and the level of investment in preventive care $I_p(t)$ lower. Thus, on balance, if the production benefit of health is small, an evolutionary

[26] In our formulation the marginal benefit of job-related health stress is increasing in the wage rate. Case and Deaton (2005) in their narrower definition of $z(t)$ as manual, risky labor (i.e., not including the psychosocial aspects of work), assume that the marginal benefit of additional manual labor is lower among those with higher wages.

[27] The wage rate might not be the most appropriate measure of the opportunity cost of time since time is not always mutually exclusive. Sick time is usually used in the production of curative care, and often institutional arrangements make it possible to continue earning wages while seeking curative care (De Serpa, 1971; Muurinen, 1982).

[28] Note that it is always true that $\sigma_H(t) > \varphi_H(t)$, otherwise the investment in curative care would finance itself through negative net marginal costs of maintaining the health stock and individuals would achieve infinite health.

[29] This can be seen from equation (5.17): if $w(t)$ increases, $I_m(t)$ has to decrease to maintain $\pi_{I_m}(t)$ at the same level.

wage change is bad for health. An exception to this pattern is the level of unhealthy consumption, which is lower.

Differences in life-time wage profiles: Now consider again two individuals a and b. Individual b earns higher wages over the life cycle, i.e., person b has greater life-time wealth (and hence $q_A^b(0) < q_A^a(0)$). Thus the net result of higher earnings over the life cycle would be similar to the "pure" asset effect described in section 5.3.3, except that apart from the life-time wealth effect ($q_A^b(0) < q_A^a(0)$) there is also a competing effect of the greater opportunity cost of time (see discussion above).

Education and health: the additional "efficiency" effect

Consider two individuals a and b who differ in their level of education E. Individual b has obtained more education but is otherwise identical. As a result, individual b has a higher wage rate $w_*(t)$ (equation 5.11). Thus the effect of education is similar to the effect of higher earnings over the life cycle and the discussion presented in section 5.3.3 applies here as well.

But education potentially also improves the efficiencies $\mu(t;E)$ of investment in curative and preventive care, and to a lesser extent healthy and unhealthy consumption (equations 5.5 to 5.8).[30] The marginal cost of investment in curative care $\pi_{I_m}(t)$ is determined by the first-order condition for health investment (equation 5.15) and is unchanged. Since the marginal cost of investment in curative care $\pi_{I_m}(t)$ increases in the level $I_m(t)$ and decreases in the efficiency $\mu_{I_m}(t;E)$ of investment in curative care (see equation 5.17), a higher efficiency due to education implies a higher level of investment in curative care compared to the pure "wage" effect described in section 5.3.3.

A higher efficiency of investment in preventive care $\mu_{I_p}(t;E)$ lowers the marginal monetary cost of preventive care $\pi_{I_p}(t)$ (equation 5.29) while the marginal benefit $\varphi_{dI_p}(t) \propto \pi_{I_m}(t)$ (equation 5.30) is unchanged. Thus the optimal level of investment in preventive care is higher compared with the pure "wage" effect.

If the efficiencies of healthy and unhealthy consumption do not (or only moderately) respond to education then the levels of healthy and unhealthy consumption are unchanged compared to the pure "wage" effect.

[30] Grossman (1972a; 1972b) assumes that the higher educated are more efficient producers and consumers of curative care. We extend his definition to preventive care. However, it is less clear whether the higher educated are more efficient producers and consumers of consumption goods and services.

Table 5.1: The effect of greater endowed wealth and an evolutionary wage increase on behavior.

	Endowed wealth	Evolutionary wage change
$I_m(t)$	$+$	$-$
$C_h(t)$	$+$	$-$
$C_u(t)$	$+/-$	$-$
$z(t)$	$-$	$+$
$I_p(t)$	$+$	$-$

Notes: The effect of greater endowed wealth $q_A^b(0) < q_A^a(0)$ (left column) and of an evolutionary wage increase (right column) on behavior. The results for greater endowed wealth (left column) are also valid for the effect of greater earnings over the life cycle and for the effect of a higher level of education.

Summary and discussion – the effect of SES on behavior

The left column of Table 5.1 provides a brief overview of the effect of greater *endowed wealth* on behavior. The direct effect of endowed wealth (through $q_A(0)$) is to enable a higher level of investment in curative care $I_m(t)$, healthy consumption $C_h(t)$ and unhealthy consumption $C_u(t)$. In addition, associated with a higher level of investment is a higher marginal monetary cost of curative care $\pi_{I_m}(t)$ (assumption 1). As a result, individuals derive greater marginal health benefit from healthy consumption $\varphi_{dC_h}(t)$ and from preventive care $\varphi_{dI_p}(t)$ because of the greater monetary value represented by the amount of health saved. Similarly, the marginal health cost of unhealthy consumption $\pi_{dC_u}(t)$ and of job-related health stress $\pi_{dz}(t)$ are greater because of the greater monetary value represented by the amount of health lost.

Wealthier individuals invest more in curative $I_m(t)$ and preventive care $I_p(t)$ and their level of healthy consumption $C_h(t)$ is higher. Wealthy individuals also engage in work that is more conducive to health: jobs associated with lower levels of job-related health stress $z(t)$. Wealth protects health by encouraging healthy life styles and enabling individuals to work and live in healthy environments. The net effect is ambiguous only for the level of unhealthy consumption as the direct effect of endowed wealth is to enable a higher level of unhealthy consumption $C_u(t)$, whereas the indirect effect is an increase in the marginal health cost of unhealthy consumption $\pi_{dC_u}(t)$.

With regards to consumption, consider a situation where the severity of the health detriment $\partial d(t)/\partial C_u(t)$ resulting from unhealthy consumption is greater than in the example shown in the right-hand panel of Figure 5.1. In this case greater marginal health cost of

unhealthy consumption $\pi_{dC_u}(t) \propto \partial d(t)/\partial C_u(t)$ shifts the net marginal cost of unhealthy consumption ($q_A^b(0)(\pi_{C_u}^b + \pi_{dC_u}^b)$; center-right panel of Figure 5.1) upward. This lowers the level of unhealthy consumption C_u^b.[31] Because the marginal health cost of unhealthy consumption increases in the severity of the health detriment ($\pi_{dC_u}(t) \propto \partial d(t)/\partial C_u(t)$) we expect to observe a pattern in which wealthy individuals consume more of moderately unhealthy consumption goods (e.g., moderate alcohol consumption) and less of more severe unhealthy consumption goods (e.g., cigarettes, high alcohol consumption, illicit drugs) when compared to less wealthy individuals.

Differences between individuals in *life-time earnings* (comparing different individuals with different life cycle wage profiles) operate similar to an increase in endowed wealth. Our model suggests that the health benefit of a "pure" asset endowment would be larger than the effect of a "comparable" change in life-time earnings (similar change in the shadow-price of wealth $q_A(0)$) due to the competing effect of the increased opportunity cost of time. There are reasons to believe that the wealth effect may dominate the effect of the opportunity cost of time (higher current wages). First, this is consistent with the result by Dustmann and Windmeijer (2000) and Contoyannis et al. (2004) that a transitory wage increase affects health negatively while a permanent wage change affects health positively. Second, it is consistent with the rich literature on SES and health that consistently finds that high income individuals are generally in better health than low income individuals.

A higher level of *education* operates similar to greater earnings over the life cycle. But education has an independent effect on health, over and above generating greater life-time earnings and wealth, through enhancing the efficiency of curative and preventive care. This leads to a higher demand for both curative and preventive care.

The right-hand column of Table 5.1 summarizes the effect of an *evolutionary* increase in the wage rate if the marginal production benefit of health $\varphi_H(t)$ is small compared to the user cost of health capital $\sigma_H(t)$. On average, as a result of the increased opportunity cost of time and the greater marginal production benefit of job-related health stress, an evolutionary increase in the wage rate is bad for health.

If there is no complementarity in utility of consumption and health (assumption 6) the predictions would remain the same. If the relation were instead one of substitutability (worse health improves the utility of consumption), solutions are possible in which greater

[31] This shift is exacerbated due to substitutability in utility of healthy $C_h(t)$ and unhealthy $C_u(t)$ consumption (assumption 7), as the marginal utility of healthy consumption $\partial U(t)/\partial C_h(t)$ increases for a lower level of unhealthy consumption $C_u(t)$ and the marginal utility of unhealthy consumption $\partial U(t)/\partial C_u(t)$ decreases for a higher level of healthy consumption $C_h(t)$.

SES leads to less investment in curative care. This is generally not observed. If there were no substitutability in utility of healthy and unhealthy consumption (assumption 7) higher SES would generally not lead to a reduction in unhealthy consumption, except for extremely unhealthy consumption goods, but would still be associated with a shift toward healthy consumption (i.e., a smaller fraction of a larger budget is devoted to unhealthy consumption).

5.3.4 Health and its effect on behavior

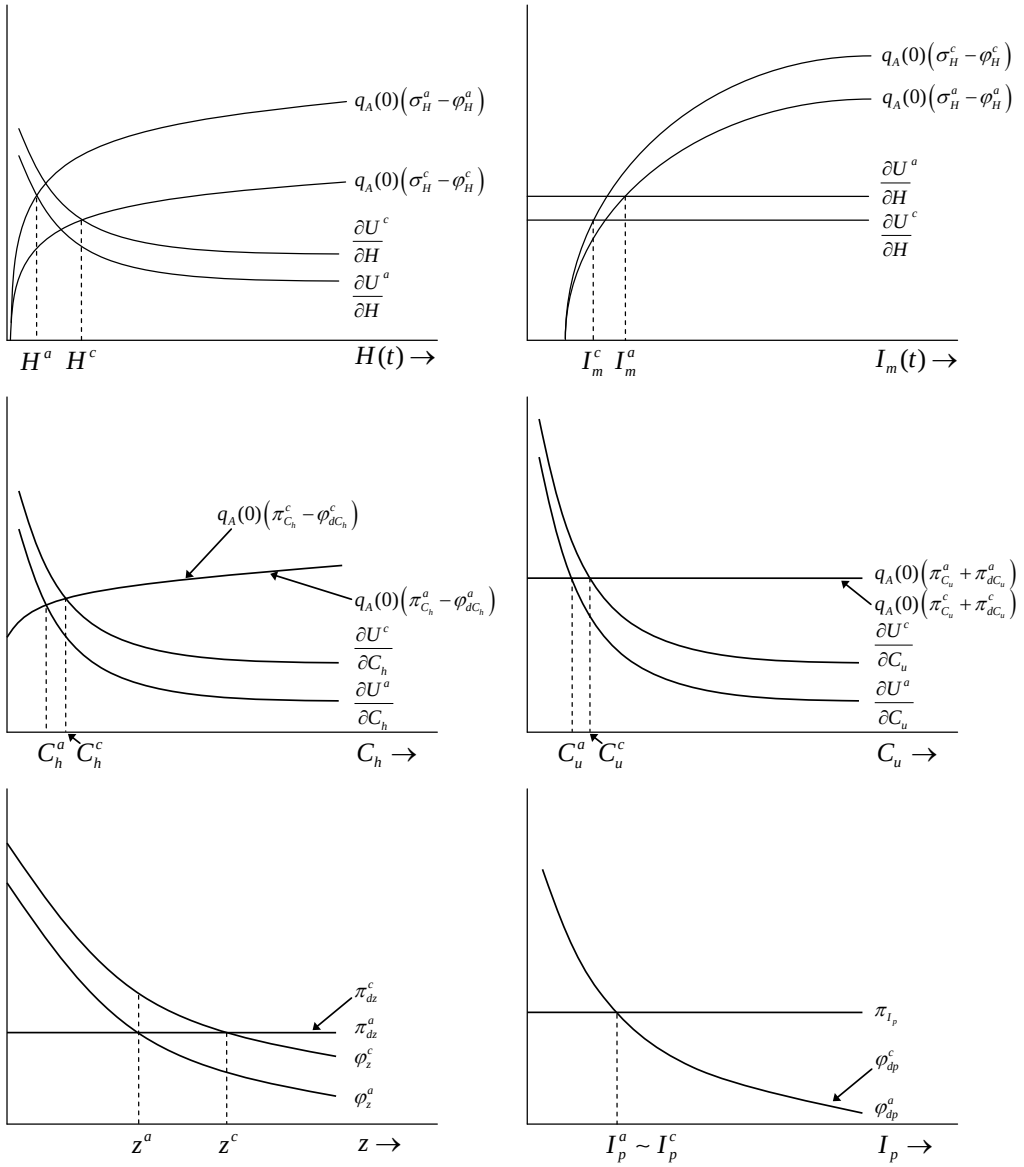
In this section we consider two identical individuals a and c that differ only in their health. Individual c is in better health than individual a ($H^c > H^a$), but is otherwise identical to individual a, i.e., all exogenous variables and functions are assumed to be the same as for person a.

Investment in curative care:
Consider the top right-hand panel of Figure 5.2 first. There is no direct effect of a higher level of health on the user cost of health capital at the margin $\sigma_H(t) = \pi_{I_m}(t)[d(t) + \delta - \widetilde{\pi_{I_m}}(t)]$.[32] However, the production benefit of health $\varphi_H(t) = \partial Y(t)/\partial H(t)$ is lower (DRTS [assumption 3]), and the resulting curve shifts upward (labeled $q_A(0)(\sigma_H^c - \varphi_H^c)$). Further, the marginal utility of health is lower (curve labeled $\partial U^c/\partial H$; diminishing marginal utility [assumption 2]). An indirect effect operates through consumption, is assumed to be smaller than the direct effects, and is discussed below. These shifts are associated with a lower optimal level of investment in curative care $I_m^c < I_m^a$.

Now turn to the top left-hand panel of Figure 5.2 which shows the associated shifts as a function of health. Assuming a lower level of investment in curative care $I_m(t)$, the user cost of health capital $\sigma_H(t)$ is smaller (assumption 5) and the net marginal user cost of health capital shifts downward (labeled $q_A(0)(\sigma_H^c - \varphi_H^c)$). Further, an indirect effect on the marginal utility of health $\partial U(t)/\partial H(t)$ operates through consumption. Higher health increases the marginal utility of consumption and hence increases the level of healthy $C_h(t)$ and unhealthy $C_u(t)$ consumption (assumption 6). This in turn increases the marginal utility of health and the curve shifts upward (labeled $\partial U^c/\partial H$). Thus higher health is

[32] There is however an indirect effect on the biological aging rate $d(t)$ because health affects choices made in working environment and in life style (operating through $C_h(t)$, $C_u(t)$, $z(t)$ and $I_p(t)$). This secondary effect is assumed to be small, which would be the case if $\partial d(t)/\partial C_h(t)$, $\partial d(t)/\partial C_u(t)$, $\partial d(t)/\partial z(t)$ and $\partial d(t)/\partial I_p(t)$ are small. However, as we will see, even under this assumption, as time passes lower levels of healthy consumption, curative and preventive care and higher levels of unhealthy consumption and job-related health stress lead to increasing health disadvantage over the life cycle.

Figure 5.2: Differences in Health



Notes: Marginal consumption $\partial U/\partial H$ and marginal production benefit φ_H of health versus the user cost of health capital at the margin σ_H as a function of health (top left) and as a function of health investment (top right). Marginal utility of healthy consumption $\partial U/\partial C_h$ versus the marginal monetary cost π_{C_h} and the marginal health benefit φ_{dC_h} of healthy consumption $C_h(t)$ (center left); Marginal utility of unhealthy consumption $\partial U/\partial C_u$ versus the marginal monetary cost π_{C_u} and the marginal health cost π_{dC_u} of unhealthy consumption $C_u(t)$ (center right); Marginal health cost π_{dz} versus the marginal production benefit φ_z of job-related health stress $z(t)$ (bottom-left); Marginal monetary cost π_{I_p} versus the marginal health benefit φ_{dI_p} of investment in preventive care $I_p(t)$ (bottom-right). In labeling the curves we have omitted the time varying term with exponent $(\beta - \delta)t$.

associated with two competing shifts: (i) a shift *of* the curve through higher consumption, and (ii) a shift *along* the curve because of diminishing marginal utility (assumption 2). Again, health provides a constraint and both curves have to intersect at H^c. The marginal utility at the point of intersection $\partial U^c/\partial H$ determines the marginal utility of health in the top right-hand panel of Figure 5.2 (draw a horizontal line from the top left-hand to the top right-hand panel in Figure 5.2). There, the intersection of the two curves determines the optimal level of health investment, and $I_m^c < I_m^a$ (consistent with our assumption).[33] Hence, a larger health stock $H^c > H^a$ reduces the marginal monetary cost of curative care $\pi_{I_m}^c < \pi_{I_m}^a$ (see equation 5.16).

Healthy and unhealthy consumption:
Changes in health and in the marginal monetary cost of curative care have no direct effect on the marginal monetary cost of healthy consumption $\pi_{C_h}(t)$ (equation 5.20) and of unhealthy consumption $\pi_{C_u}(t)$ (equation 5.23), hence both are shown as unchanged in the center-left and center-right panels of Figure 5.2. A higher level of health increases the health benefit of healthy consumption $\varphi_{dC_h}(t)$ (and the health cost of unhealthy consumption $\pi_{dC_u}(t)$), yet at the same time the health benefit (cost) is reduced through a lower marginal monetary cost of curative care as healthy individuals demand less curative care ($\varphi_{dC_h}(t) \propto \pi_{I_m}(t)H(t)$ and $\pi_{dC_u}(t) \propto \pi_{I_m}(t)H(t)$; see equations 5.21 and 5.24). The net effect is ambiguous. To reflect this, we show both the net marginal cost of healthy consumption (solid line labeled $q_A(0)(\pi_{C_h}^c - \varphi_{dC_h}^c)$; center-left panel of Figure 5.2) and the net marginal cost of unhealthy consumption (solid line labeled $q_A(0)(\pi_{C_u}^c + \pi_{dC_u}^c)$; center-right panel of Figure 5.2) as being unchanged (i.e., $\pi_{I_m}^c H^c \simeq \pi_{I_m}^a H^a$; we will return to this point later).

Now turn to the marginal utility of healthy consumption $\partial U(t)/\partial C_h(t)$ and of unhealthy consumption $\partial U(t)/\partial C_u(t)$. The direct effect of higher health $H^c > H^a$, due to complementarity in utility of consumption and health (assumption 6) is to shift both the marginal utility of healthy consumption $\partial U(t)/\partial C_h(t)$ (curve labeled $\partial U^c/\partial C_h$, center-left panel of Figure 5.2) and of unhealthy consumption $\partial U(t)/\partial C_u(t)$ (curve labeled $\partial U^c/\partial C_u$,

[33]Note that solutions are possible in which higher health leads to greater investment in health, i.e. a positive correlation between health and curative care (which is generally not observed). This requires the indirect effect on the marginal utility of health, operating through a higher level of consumption (as a result of greater health), to be quite substantial. Such solutions cannot be ruled out but appear less plausible.

center-right panel of Figure 5.2) upward.[34] The optimal solution for healthy consumption as well as for unhealthy consumption is higher: $C_h^c > C_h^a$ and $C_u^c > C_u^a$.

In the scenario discussed above it was assumed that the net marginal cost of healthy and of unhealthy consumption remain unchanged. Now consider two alternative scenarios. In scenario 1, the direct effect of higher health $H^c > H^a$ exceeds the indirect effect of changes in the marginal cost of curative care (as a result of higher health), i.e., $\pi_{I_m}^c H^c > \pi_{I_m}^a H^a$, and in scenario 2 we explore the opposite, i.e., $\pi_{I_m}^c H^c < \pi_{I_m}^a H^a$.[35] In scenario 1 both the marginal health benefit of healthy consumption $\varphi_{dC_h}(t)$ and the marginal health cost of unhealthy consumption $\pi_{dC_u}(t)$ are higher for individual c.[36] This further increases the level of healthy consumption C_h^c, but lowers the level of unhealthy consumption C_u^c (compared with the example shown in the center-left hand panel of Figure 5.2).[37] In scenario 2 we expect to observe the opposite pattern: higher health $H^c > H^a$ decreases the level of healthy consumption C_h^c, and further increases the level of unhealthy consumption C_u^c (compared with the example shown in the center-right hand panel of Figure 5.2).

Job-related health stress and investment in preventive care:
Greater health is potentially associated with a greater marginal production benefit of job-related health stress $\varphi_z(t) = \partial Y(t)/\partial z(t)$ (curve labeled φ_z^c; bottom-left panel of Figure 5.2) as healthy individuals have higher earnings $Y(t)$ (see equation 5.9). The marginal monetary cost of preventive care $\pi_{I_p}(t)$ is independent of the level of health (equation

[34] An indirect effect operates through consumption. Because of substitutability in utility of healthy $C_h(t)$ and unhealthy $C_u(t)$ consumption (assumption 7), both the marginal utility of healthy consumption $\partial U(t)/\partial C_h(t)$ and the marginal utility of unhealthy consumption $\partial U(t)/\partial C_u(t)$ shift downward. Assuming that the direct effect dominates the indirect effect the net result is nevertheless an upward shift.

[35] Scenario 1 corresponds to a small elasticity of health investment with respect to health and scenario 2 corresponds to a high elasticity. Assume $I_m(t) \propto H^{-\gamma}$, where γ is the elasticity of health investment with respect to health. Scenario 1 ($\pi_{I_m}^c H^c > \pi_{I_m}^a H^a$) then implies $\gamma < 1/(1-\alpha)$ while scenario 2 implies $\gamma > 1/(1-\alpha)$.

[36] Although these shifts are not depicted it is useful to use Figure 5.2 for reference. Scenario 1 implies a downward shift in the net marginal costs of healthy consumption with respect to the curve shown $q_A(0)(\pi_{C_h}^c - \varphi_{dC_h}^c)$ in the center-left panel through increased φ_{dC_h}. Scenario 1 also implies an upward shift in the net marginal cost of unhealthy consumption with respect to the curve shown $q_A(0)(\pi_{C_u}^c + \pi_{dC_u}^c)$ in the center-right panel through increased π_{dC_u}.

[37] These shifts are exacerbated due to substitutability in utility of healthy $C_h(t)$ and unhealthy $C_u(t)$ consumption (assumption 7), as the marginal utility of healthy consumption $\partial U(t)/\partial C_h(t)$ increases for a lower level of unhealthy consumption $C_u(t)$ and the marginal utility of unhealthy consumption $\partial U(t)/\partial C_u(t)$ decreases for a higher level of healthy consumption $C_h(t)$.

Table 5.2: The effect of greater health on behavior.

Scenario	1	2
$I_m(t)$	$-$	$-$
$C_h(t)$	$+$	$+/-$
$C_u(t)$	$+/-$	$+$
$z(t)$	$+/-$	$+$
$I_p(t)$	$+$	$-$

5.29). The effect on the marginal health cost of job-related health stress and the marginal health benefit of investment in preventive care is once more ambiguous (see equations 5.26 and 5.30) and both are shown as unchanged (i.e., $\pi^c_{I_m} H^c \simeq \pi^a_{I_m} H^a$). The resulting optimal level of job-related health stress is higher $z^c > z^a$ (bottom-left panel of Figure 5.2) and the level of investment in preventive care is unchanged $I^c_p \sim I^a_p$ (bottomright panel of Figure 5.2).

In scenario 1 (scenario 2) the marginal health cost of job-related health stress π_{dz} and the marginal health benefit of investment in preventive care φ_{I_p} are higher (lower), and the level of job-related health stress decreases (increases) and the level of investment in preventive care increases (decreases) with respect to the case shown in the bottom-left and bottom-right panels of Figure 5.2.

Summary and discussion – the effect of health on behavior:
Table 5.2 provides a brief overview of the effect of greater health on behavior. Regardless of the scenario, individuals in better health invest less in curative care $I_m(t)$. In scenario 1 individuals consume more healthy consumption $C_h(t)$ and invest more in preventive care $I_p(t)$, while the effect on unhealthy consumption $C_u(t)$ and job-related health stress is ambiguous. In scenario 2 individuals consume more unhealthy consumption $C_u(t)$, engage more in job-related health stress $z(t)$, and invest less in preventive care $I_p(t)$, while the effect on healthy consumption $C_h(t)$ is ambiguous.

If there is no complementarity in utility of consumption and health (assumption 6) the predictions would remain the same, except that the effect of greater health on healthy consumption is negative in scenario 2 and the effect of greater health on unhealthy consumption is positive in scenario 1 (i.e., not ambiguous as shown in Table 5.2). If the relation were instead one of substitutability (worse health improves the utility of consumption), solutions are possible in which greater health leads to lower levels of consumption and more investment in curative care. This is generally not observed.

5.4 Discussion and conclusions

The aim of this paper is to provide a contribution toward a theory of the relation between health and socioeconomic status (SES) over the lifecycle. Our life-cycle model incorporates multiple mechanisms that could explain (jointly) a large part of the observed disparities in health by SES. In our model, lifestyle factors (preventive care, healthy and unhealthy consumption), working conditions (physical and psychosocial health stresses), living conditions (housing, neighborhood social environment), curative care and the constraining effect of health on work are mechanisms through which SES (endowed wealth, life-time earnings and education) and health are related.

The main mechanism through which SES translates into health is by increasing the marginal cost of and the demand for curative care. This in turn increases the health benefit of (and hence demand for) preventive care and healthy consumption, and the health cost of (and hence reduced demand for) unhealthy working and living environments, and unhealthy consumption.

Even without the inclusion of additional potential mechanisms responsible for the SES health gradient (beside utilization of curative care), the theory predicts differences in the "effective" rate of health decline $\dot{H}(t)$ between high- and low-SES individuals due to differences in the level of investment in curative care $I_m(t)$. This addresses the criticism leveled by Case and Deaton (2005). But greater SES also induces healthy lifestyles, encourages investment in preventive care and protects individuals from the health risks of physical working conditions (e.g., hard labor) and/or psychosocial aspects of work (e.g., low status, limited control, repetitive work, etc) that are detrimental to health.

Endowed wealth, life time earnings and education each operate in distinct ways. The effect of greater earnings over the life cycle on health differs from the effect of greater endowed wealth in that the "wealth" effect is moderated by the higher opportunity cost of time. Plausibly, however, the effect of greater earnings over the life cycle dominates the opportunity cost effect. For example, Dustmann and Windmeijer (2000) and Contoyannis et al. (2004) find a positive effect on health from a permanent wage increase and a negative effect from a transitory wage increase. The effect of education on health is similar to that of greater earnings over the life cycle, but with the additional effect of increasing the efficiency of the production and consumption of curative and preventive care.

Irrespective of the SES indicator, for individuals who are initially equally healthy, the health trajectories of high and low SES individuals will begin to diverge. In addition, the higher the health stock, the greater the earnings (e.g., see equation 5.9) such that reverse causality (from health to SES) could further reinforce the SES health gradient.

Results from earlier studies (Ehrlich and Chuma, 1990; Ehrlich, 2000; Galama et al. 2008 [Chapter 3]) suggest that the more rapidly worsening health of low SES individuals could lead to early withdrawal from the labor force and shorter life spans. Early withdrawal from the labor force may contribute to further increasing disadvantage (widening of the SES health gradient) as the associated loss of income disproportionally affects low SES individuals. Mortality selection, i.e. lower SES people are more likely to die early, may result in an apparently healthier surviving disadvantaged population, partially explaining the narrowing of the gradient in late age.

Further, depending on the elasticity of investment in curative care with respect to health, the predicted divergence in health trajectories between low and high SES individuals could be further reinforced (scenario 1; small elasticity) or mitigated (scenario 2; large elasticity). In scenario 2 (see Table 5.2 [opposite signs for lower health] in section 5.3.4), over time the rate of divergence slows as subsequent lower levels of health encourages low SES individuals to invest more in health and engage in healthier behavior. Thus the theory predicts an initial widening and potentially a subsequent narrowing of the SES-health gradient. In scenario 1 less healthy individuals engage in less healthy behavior (with the exception of investment in curative care), and the theory predicts a continued widening of the gradient with age (or a weaker narrowing process).

Scenario 1 thus predicts a process of *cumulative advantage* for high SES individuals. The cumulative advantage hypothesis states that health inequalities emerge by early adulthood and subsequently widen as economic and health advantages of higher SES individuals accumulate (House et al. 1994; Ross and Wu, 1996; Lynch, 2003). Any apparent narrowing of SES inequalities in late life is largely attributed to mortality selection. In contrast, scenario 2 predicts an economic variant of the *age-as-leveller* hypothesis (House et al. 1994; Elo and Preston, 1996; Beckett, 2000). The age-as-leveler hypothesis maintains that deterioration in health is an inevitable part of aging irrespective of SES with the result that the SES-health gradient narrows at later ages. Relative to the disadvantaged, economically advantaged people may be better able to postpone, but not prevent, declining health status.

Our theory can explain additional stylized empirical facts. The model predicts that individuals in better health invest less in curative care $I_m(t)$. This finding is supported by casual observation (the healthy do not go to the doctor) and by numerous empirical studies that find a strong negative correlation between measures of health and measures of curative (medical) care usage (see Galama and Kapteyn, 2009 [Chapter 2], for an overview of the empirical literature). Further, as our health declines with age, the demand for curative care increases. If the effect of deteriorating health on investment in curative care

dominates the effect of the opportunity cost of time,[38] the model is capable of reproducing the observation that young individuals invest little in curative care, the middle-aged more and the elderly the most.

Another prediction of the theory is a pattern in which high SES individuals consume more of moderately unhealthy consumption goods (e.g., moderate alcohol consumption) and less of severely unhealthy consumption goods (e.g., cigarettes, high alcohol consumption, illicit drugs) when compared to lower SES individuals. Greater wealth permits more consumption but also increases the marginal monetary value of health lost. This could provide a plausible explanation for the observation that high SES individuals are less likely to smoke cigarettes (bad for health) but are more likely to be moderate drinkers (moderately bad for health) than low SES individuals (e.g., Cutler and Lleras-Muney, 2008; Stringhini et al. 2010).

Our theory suggests that the SES health gradient could be strong in countries with universal health care coverage and low deductions, where the price of curative care is low and health care is affordable for everyone, as well as in countries with large uninsured populations. The marginal cost of curative care is largely determined by life-time wealth ($q_A(0)$, i.e., SES) and by the health stock $H(t)$ (see the first-order condition 5.15). Thus, a low price of curative care $p_m(t)$ does not influence the marginal monetary cost of curative care but increases the demand for curative care $I_m(t)$ (see equation 5.17 and keeping $\pi_{I_m}(t)$ unchanged). Further, because the marginal cost of curative care is not sensitive to price, the marginal health benefit of healthy consumption and preventive care and the health cost of unhealthy consumption and job-related health stress are unchanged. Thus the price of curative care does not affect choices in consumption, preventive care and in living and working environments directly, and also in countries with universal health care coverage and low deductibles there will be a significant SES health gradient. This is particularly true if medical care is not a large determinant of the SES health gradient (e.g., Adler et al. 1993) and could explain why the observed SES health gradient over the life cycle is strikingly similar between countries with relatively low levels of protection from loss of work and health risks, such as the U.S., and those with stronger welfare systems, such as the Netherlands (e.g., Smith, 1999; 2004; 2007; Case and Deaton 2005; van Kippersluis et al. 2010).

[38] Low at young ages and high in middle and old age as a result of the typical hump-shaped wage profile with age (e.g., Mincer, 1974).

The predictions regarding the effects of SES on health depend on the notion that health has both intrinsic as well as instrumental value.[39] Differences in endowed wealth $q_A(0)$ have no effect on health if health does not provide utility (e.g., in the pure investment model, Grossman, 1972a; 1972b). In this case, the effect of greater earnings over the life cycle would be ambiguous as a higher wage rate increases both the user cost of health capital $\sigma_H(t)$ and the production benefit of health $\varphi_H(t)$. The effect of education would mostly operate through greater efficiency of medical and curative care.

Thus, if health is mostly valued for its production benefit (generating earnings) this could explain the absence of strong evidence for a causal effect of financial indicators of SES on health (e.g., Cutler et al. 2011). Another possible explanation of this finding is that the effect of SES on health accumulates over time. The effect of financial indicators of SES on health is typically estimated contemporaneously (or with a small delay) and the wealth effect may be countered by the opportunity cost of time. Education, on the other hand, is obtained early in life (and hence its effect has had ample time to accumulate) and education potentially increases the efficiency of the production and consumption of curative and preventive care. This may provide an explanation for the strong effect of education on health outcomes observed in empirical studies (e.g., Grossman, 2000; Lleras-Muney, 2005; Silles, 2009). It also suggests that the protective effect of SES on health, in particular education, increases with age (Ross and Wu, 1996; Lynch, 2003).

In order to illustrate the theory and to derive predictions, we have made assumptions about the nature of the relations between functions of interest. The assumptions (1 to 5) of diminishing or constant returns to scale are commonly made in economics. If there is no complementarity in utility of consumption and health (assumption 6) the predictions would remain the same. If the relation were instead one of substitutability (worse health improves the utility of consumption), solutions are possible in which greater SES leads to less investment in curative care. This is generally not observed. If there were no substitutability in utility of healthy and unhealthy consumption (assumption 7) higher SES would not lead to a reduction in unhealthy consumption (except for severaly unhealthy consumption) but would still be associated with a shift toward healthy consumption (a smaller fraction of a larger budget is devoted to unhealthy consumption).

Our model includes major mechanisms identified in a review of the literature as explaining (jointly) a large part of the observed disparities in health by SES. Given the complexity (e.g., Cutler et al. 2011) of the various relations between SES and health, we

[39] As recognized by, e.g., Mushkin (1962) who noted that *" Health services ... are partly investment and partly consumption ... An individual wants to get well so that life for him may be more satisfying. But also when he is well he can perform more effectively as a producer"*

have focused on potential explanations that a) explain a large part of the gradient and b) are relatively straightforward to include in our theoretical framework.

Compared to Grossman (1972a; 1972b), Ehrlich and Chuma (1990) and Case and Deaton (2005) the model presented in this paper contains several improvements and extensions: (i) A distinguishing feature is our interpretation of the relation (5.15) as being the first-order condition for optimal health investment $I_m(t)$, conditional on the level of the health stock $H(t)$, rather than the first-order condition for optimal health $H(t)$. This interpretation necessitates the assumption of decreasing-returns-to-scale (DRTS) in the health production function (as in Ehrlich and Chuma, 1990; see also Dustmann and Windmeijer, 2000; and Liljas, 2000), and addresses the indeterminacy problem ("bang-bang" solution) for investment in curative care (Ehrlich and Chuma, 1990), ensures that investment in curative (medical) care is non-negative (for the usual assumptions of functional forms), reproduces the observed negative relation between health and the demand for medical care, finds the health stock to be a function of initial health, past biological aging and past health investments made, and explains differences in the level of health as well as the rate of health decline between low and high SES groups (see Galama, 2011 [Chapter 4]). Addressing these issues has been crucial: unlike alternative life-cycle models of health, medical care, and SES, our formulation can explain the formation of disparities in health by SES with age. (ii) We have included the concept of healthy consumption (as well as unhealthy consumption as in Case and Deaton, 2005) and allow the demand for consumption to be governed both by the direct monetary price of consumption as well as the indirect health benefit (healthy consumption) or indirect health cost (unhealthy consumption). Case and Deaton (2005) on the other hand consider an unhealthy consumption good whose price is only paid in terms of health. (iii) We have broadened the concept of "job-related health stress" to include not only hard/risky labor (as in Case and Deaton, 2005) but also psychosocial aspects of work that are detrimental to health. (iv) We have argued that the effect of housing and neighborhood social environment can be included by extending the definition of healthy consumption as well as exogenous environmental factors to include relevant aspects of housing and neighborhood characteristics. (v) We have introduced the concept of preventive care.

Numerical methods are required to solve the full model, including endogenous retirement decisions and mortality. With regard to mortality, the model provides a natural way to include length of life. In Grossman's original formulation (Grossman, 1972a; 1972b) length of life is determined by a minimum health level $H_{\min}$, below which an individual dies. Endogenous length of life can be incorporated as in Ehrlich and Chuma (1990) and Ehrlich (2000) and simulated and calibrated as in Ehrlich and Yin (2005). With regard to

retirement, as emphasized by Smith (2004) and Case and Deaton (2005), reverse causality from health to income through labor force participation could be an important mechanism explaining the SES-health gradient. In our model, this could be incorporated by an endogenous retirement age (as in Galama et al. 2008 [Chapter 3]).

Another important extension of our model would be to incorporate insights from the literature on socioeconomic differences in the evolution of child health (e.g., Case et al. 2002; Currie and Stabile, 2003; Currie et al. 2007; Murasko, 2008), and from the literature on the impact of fetal and early-childhood conditions on health in adulthood (e.g., Barker et al. 1993; Case et al. 2005; van den Berg et al. 2006).[40] This might be feasible by including the production of health by the family (including the health of the child) similar to, e.g., Jacobson (2000) and Bolin et al. (2001; 2002a; 2002b).

We do not explicitly take into account the influence of the wider social context and social relationships of the family or neighborhood on health (House et al. 1988; Robert, 1998; Kawachi and Berkman, 2003). Less affluent areas are more polluted, have lower quantity and quality of municipal services, have higher crime rates, and are associated with unhealthy lifestyles (Robert, 1998). Also, the social isolation induced by poor quality and quantity of social contacts is an important risk factor for health (House et al. 1988). In our model this is partly captured by the exogenous part of the biological aging rate (exogenous environmental factors $\xi(t)$). However, it is likely that social factors are partly endogenous to socioeconomic status (Robert, 1998). The role of the wider social context, social relationships, and other psycho-social risk factors (House et al. 1988; 1994; Robert, 1998; Kawachi and Berkman, 2003) can partially be captured in our model by extending the definition of healthy consumption to include choice of housing / neighborhood social environment. This might be further extended by including social capital similar to, e.g., Bolin et al. (2003).

We have not explicitly included racial and gender disparities in health. Racial categories importantly capture differences in power, status, and resources (Williams, 1999). Differences in SES between racial groups account for most of the observed racial disparities in health (Williams and Collins, 1995; Lillie-Blanton et al. 1996). Yet, racial differences in health and mortality persist even at "equivalent levels" of SES, and race/ethnicity has an

[40] The potential influence of childhood health on education is not included in our formulation as education is treated as being predetermined by the time individuals join the labor-force. Childhood conditions can be accounted for by treating the health status of an individual joining the labor force and investment in human capital prior to adulthood as initial conditions, i.e., we take initial health $H(0)$ and years of schooling E as given. Our model is therefore limited to explaining the formation of disparities in health from early adulthood till old age but not during childhood or the fetal period. As a result, the formulation cannot model the possible joint determination of education and health.

independent effect beyond indicators of socioeconomic status (e.g., House and Williams, 2000). To the extent that racial/ethnic influences act independently of SES, race/ethnicity can be included in our formulation through the exogenous component of the biological aging rate. The same holds for gender disparities in health, if operating independently of SES (Luchenski et al. 2008). However, it has been argued that gender and race potentially moderate the relation between SES and health. It could be that discrimination makes it difficult to translate high SES into good health, or that employer discrimination makes minorities in poor health particularly likely to lose their jobs. The literature is inconclusive to what extent race/ethnicity and gender moderate the relationship between SES and health (Matthews et al. 1999; House and Williams, 2000; Luchenski et al. 2008).

Lastly, insights from the behavioral-economic and psychological literature regarding myopia and lack of self-control (e.g., Blanchflower et al. 2009) might be incorporated following Laibson (1998). Uncertainty (e.g., health shocks) could be included similar to, e.g., Cropper (1977), Dardanoni and Wagstaff (1990), Liljas (1998) and Ehrlich (2000). Joint determination of health and socioeconomic status due to factors such as intelligence, cognitive ability and non-cognitive skills may be incorporated by allowing these factors to raise the efficiency of household production in a similar way as education (e.g., Chiteji, 2010).

Empirical estimation of the model is needed to test the assumptions and the theoretical predictions presented in this work and to assess the relative importance of mechanisms, study interactions between mechanisms, and disentangle the different patterns of causality. This will require developing structural- and reduced-form relations. Model estimates may contribute to improving our understanding of the operational roles of major mechanisms in explaining the SES health gradient, and to simulating the long-term effects of policy interventions.

5.5 Appendix

5.5.1 First-order conditions

Associated with the Lagrangian (equation 5.12) we have the following conditions:

$$\begin{aligned}
\dot{q}_A(t) &= -\frac{\partial \Im(t)}{\partial A(t)} \Rightarrow \\
\dot{q}_A(t) &= -\delta q_A(t) \Leftrightarrow \\
q_A(t) &= q_A(0)e^{-\delta t}, \qquad (5.31)
\end{aligned}$$

$$\begin{aligned}
\dot{q}_H(t) &= -\frac{\partial \Im(t)}{\partial H(t)} \Rightarrow \\
\dot{q}_H(t) &= q_H(t)d(t) - \frac{\partial U(t)}{\partial H(t)}e^{-\beta t} - q_A(0)\frac{\partial Y(t)}{\partial H(t)}e^{-\delta t} \\
&= q_H(t)d(t) - \frac{\partial U(t)}{\partial H(t)}e^{-\beta t} - q_A(0)\varphi_H(t)e^{-\delta t} \Leftrightarrow \\
q_H(t) &= q_H(0)e^{\int_0^t d(x)dx} - \int_0^t \left[\frac{\partial U(s)}{\partial H(s)}e^{-\beta s} - q_A(0)\varphi_H(s)e^{-\delta s}\right] e^{\int_s^t d(x)dx}ds, \qquad (5.32)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \Im(t)}{\partial X_h(t)} &= 0 \Rightarrow \\
\frac{\partial U(t)}{\partial C_h(t)} &= q_A(0)\frac{p_{X_h}(t)}{\partial C_h(t)/\partial X_h(t)}e^{(\beta-\delta)t} + q_H(t)\frac{\partial d(t)}{\partial C_h(t)}H(t)e^{\beta t} \\
&\equiv q_A(0)\,\pi_{C_h}(t)e^{(\beta-\delta)t} - q_H(t)\frac{\varphi_{dC_h}(t)}{\pi_{I_m}(t)}e^{\beta t}, \qquad (5.33)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \Im(t)}{\partial \tau_{C_h}(t)} &= 0 \Rightarrow \\
\frac{\partial U(t)}{\partial C_h(t)} &= q_A(0)\frac{w(t)}{\partial C_h(t)/\partial \tau_{C_h}(t)}e^{(\beta-\delta)t} + q_H(t)\frac{\partial d(t)}{\partial C_h(t)}H(t)e^{\beta t} \\
&\equiv q_A(0)\,\pi_{C_h}(t)e^{(\beta-\delta)t} - q_H(t)\frac{\varphi_{dC_h}(t)}{\pi_{I_m}(t)}e^{\beta t}, \qquad (5.34)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \Im(t)}{\partial X_u(t)} &= 0 \Rightarrow \\
\frac{\partial U(t)}{\partial C_u(t)} &= q_A(0)\frac{p_{X_u}(t)}{\partial C_u(t)/\partial X_u(t)}e^{(\beta-\delta)t} + q_H(t)\frac{\partial d(t)}{\partial C_u(t)}H(t)e^{\beta t} \\
&\equiv q_A(0)\,\pi_{C_u}(t)e^{(\beta-\delta)t} + q_H(t)\frac{\pi_{dC_u}(t)}{\pi_{I_m}(t)}e^{\beta t}, \qquad (5.35)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \Im(t)}{\partial \tau_{C_u}(t)} &= 0 \Rightarrow \\
\frac{\partial U(t)}{\partial C_u(t)} &= q_A(0)\frac{w(t)}{\partial C_u(t)/\partial \tau_{C_u}(t)}e^{(\beta-\delta)t} + q_H(t)\frac{\partial d(t)}{\partial C_u(t)}H(t)e^{\beta t} \\
&\equiv q_A(0)\,\pi_{C_u}(t)e^{(\beta-\delta)t} + q_H(t)\frac{\pi_{dC_u}(t)}{\pi_{I_m}(t)}e^{\beta t}, \qquad (5.36)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \Im(t)}{\partial m_m(t)} &= 0 \Rightarrow \\
q_H(t) &= q_A(0)\left\{\frac{p_m(t)I_m(t)^{1-\alpha}}{\alpha[\partial I_m(t)/\partial m_m(t)]}\right\}e^{-\delta t} \\
&\equiv q_A(0)\,\pi_{I_m}(t)e^{-\delta t}, \qquad (5.37)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \Im(t)}{\partial \tau_{I_m}(t)} &= 0 \Rightarrow \\
q_H(t) &= q_A(0)\left\{\frac{w(t)I_m(t)^{1-\alpha}}{\alpha[\partial I_m(t)/\partial \tau_{I_m}(t)]}\right\}e^{-\delta t} \\
&\equiv q_A(0)\,\pi_{I_m}(t)e^{-\delta t}, \qquad (5.38)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \Im(t)}{\partial z(t)} &= 0 \Rightarrow \\
0 &= q_H(t)\frac{\partial d(t)}{\partial z(t)}H(t) - q_A(0)\frac{\partial Y(t)}{\partial z(t)}e^{-\delta t} \\
&\equiv q_H(t)\frac{\pi_{dz}(t)}{\pi_{I_m}(t)} - q_A(0)\,\varphi_z(t)e^{-\delta t}, \qquad (5.39)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \Im(t)}{\partial m_p(t)} &= 0 \Rightarrow \\
0 &= q_H(t)\frac{\partial d(t)}{\partial I_p(t)}H(t) + q_A(0)\frac{p_p(t)}{\partial I_p(t)/\partial m_p(t)}e^{-\delta t} \\
&\equiv -q_H(t)\frac{\pi_{dI_p}(t)}{\pi_{I_m}(t)} + q_A(0)\,\pi_{I_p}(t)e^{-\delta t}, \qquad (5.40)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \Im(t)}{\partial \tau_{I_p}(t)} &= 0 \Rightarrow \\
0 &= q_H(t)\frac{\partial d(t)}{\partial I_p(t)}H(t) + q_A(0)\frac{w(t)}{\partial I_p(t)/\partial \tau_{I_p}(t)}e^{-\delta t}, \\
&\equiv -q_H(t)\frac{\pi_{dI_p}(t)}{\pi_{I_m}(t)} + q_A(0)\,\pi_{I_p}(t)e^{-\delta t}, \qquad (5.41)
\end{aligned}$$

Equation (5.33) or (5.34) combined with (5.37) or (5.38) provide the first-order condition for maximization of (5.12) with respect to healthy consumption (equation 5.19). Similarly, equation (5.35) or (5.36) combined with (5.37) or (5.38) provide the first-order condition for maximization of (5.12) with respect to unhealthy consumption (equation 5.22). Using (5.37) or (5.38) to obtain an expression for $\dot{q}_H(t)$ and substituting the results for $q_H(t)$ and $\dot{q}_H(t)$ in (5.32) we find the first-order condition for maximization of (5.12) with respect to investment in curative care (equation 5.15). Combining equations (5.37) or (5.38) and (5.39) to eliminate $q_H(t)$ we find the first-order condition for maximization of (5.12) with respect to job-related health stress (equation 5.25). Lastly, combining equations (5.37) or (5.38) and (5.40) or (5.41) to eliminate $q_H(t)$ we find the first-order condition for maximization of (5.12) with respect to preventive care (equation 5.28).

Nederlandse Samenvatting

Inleiding en motivatie

Een van de opmerkelijkste bevindingen in bevolkingsgezondheid is het sterke verband tussen gezondheid en sociaaleconomische status (SES). Figuur 1.1 (Hoofdstuk 1) toont de belangrijkste eigenschappen van de SES gezondheidsgradiënt in de V.S. (linkerkant) en Nederland (rechtse kant) door op elke leeftijd de fractie mensen die in slechte of matige gezondheid zijn (zelfrapportage) in kaart te brengen per kwartiel van gezinsinkomen (kwartiel 1 vertegenwoordigt de laagste en kwartiel 4 de hoogste gezinsinkomens; kwartielen zijn leeftijdsgebonden). Op elke leeftijd is een lager inkomen geassocieerd met slechtere gezondheid.

De gezondheidsverschillen met inkomen zijn groot. Bijvoorbeeld, de fractie in slechte of matige gezondheid in de V.S. rond 60 jarige leeftijd in het hoogste gezinsinkomenskwartiel, met ongeveer 8 procent, is zowat 35 procent kleiner dan de fractie in het laagste inkomenskwartiel, met ongeveer 44 procent (linkerkant van Figuur 1.1). Case en Deaton (2005) laten zien hoe in de V.S. een 20 jarige man met laag inkomen (laagste kwartiel van familie inkomen) dezelfde gemiddelde gezondheid rapporteert als een 60 jarige man met hoog inkomen (hoogste kwartiel). In Glasgow, Verenigd Koninkrijk, is de levensverwachting van mensen in de armste gebieden 54 jaar, vergeleken met 82 in de meest rijke (Hanlon et al. 2006).

Niet alleen hebben individuen met lage SES een slechtere gezondheid maar hun gezondheid gaat ook sneller achteruit dan de gezondheid van individuen met hoge SES. De ongelijkheid in gezondheid tussen lage en hoge SES groepen lijkt te stijgen over de levenscyclus tot aan leeftijden van ongeveer 50-60 jaar, waarna zij lijkt te versmallen (zie, bijvoorbeeld, Figuur 1.1). Vergelijkbare verbanden worden gevonden voor andere indicatoren van SES, zoals opleiding en vermogen, en andere indicatoren van gezondheid, zoals aanvang van chronische ziekten, invaliditeit en mortaliteit (zie bijvoorbeeld Adler et al. 1994; Marmot, 1999; Smith, 1999).

Velen vinden deze grote verschillen in gezondheid tussen SES groepen sociaal onrechtvaardig. Het idee is dat dergelijke verschillen in gezondheid te vermijden zijn en het gevolg zijn van de omstandigheden waarin mensen opgroeien, leven, werken en ouder worden, en van de ziekenverzorgingsystemen (zie bijvoorbeeld CSDH, 2008). Met dit in gedachten, heeft de Commissie van de Wereldgezondheidsorganisatie (de WGO) over de Sociale Determinanten van Gezondheid (CSDH) opgeroepen tot wereldwijde actie betreffende de sociale determinanten van gezondheid met het doel gezondheidsgelijkheid te bereiken binnen een generatie (CSDH, 2008).

Dit edele doel, echter, wordt belemmerd door het feit dat de oorzaken van sociaaleconomische gezondheidsongelijkheden niet goed worden begrepen. Onderzoek in meerdere disciplines (met inbegrip van epidemiologie, sociologie, demografie, psychologie, evolutiebiologie en economie) laat zien dat er meerdere verklaringen mogelijk zijn, dat er gebrek aan consensus is betreffende het relatieve belang van de verschillende mechanismen (in de verklaring van het fenomeen) en dat het moeilijk is om causaliteit, laat staan de onderliggende mechanismen, vast te stellen (zie bijvoorbeeld Cutler et al. 2011). Bijvoorbeeld, een causaal beschermend effect van opleiding op gezondheid is vastgesteld (lleras-Muney, 2005; Oreopoulos, 2006; van Kippersluis et al. 2011) maar het is niet precies duidelijk hoe hoger opgeleiden hun gezondheidsvoordeel verkrijgen.

Sommige voorgestelde mechanismen impliceren dat SES de gezondheid beïnvloedt (veroorzaakt), andere mechanismen nu juist dat gezondheid SES bepaalt, en weer andere mechanismen dat SES en gezondheid gezamenlijk worden bepaald, zonder een direct oorzakelijk verband. Sommige mechanismen kunnen in alle drie de categorieën vallen. Mogelijke verklaringen voor de sociaaleconomische gradiënt in gezondheid omvatten: toegang tot medische zorg, de rol die gezondheid speelt in arbeidsparticipatie, gezondheidsgedrag (bijvoorbeeld roken, drinken, sport), psychosociale en milieurisicofactoren, het sociale milieu van de buurt, sociale verhoudingen en sociale steun, mate van controle, omstandigheden in de foetale fase en de vroege kinderjaren, en fysieke, chemische, biologische en psychosociale factoren op het werk. Zogenaamde "derde factoren"-verklaringen poneren dat individuele verschillen, bijvoorbeeld, in tijdsvoorkeur en in de mate waarin een individu in staat is om zelfcontrole uit te oefenen, SES en gezondheid op een vergelijkbare manier kunnen beïnvloeden met als gevolg de waargenomen sociaaleconomische gezondheidsgradiënt. Mogelijk kunnen meerdere verklaringen elk een stukje aan het oplossen van het raadsel bijdragen (voor een overzicht zie Galama en van Kippersluis, 2010 [Hoofdstuk 5]).

Kennis van het relatieve belang van de verschillende causale mechanismen verantwoordelijk voor de waargenomen relaties tussen SES en gezondheid wordt belemmerd door het

gebrek aan een voldoende uitvoerige theorie. Zonder kennis van de onderliggende mechanismen is het moeilijk om beleid te ontwikkelen dat in staat is de ongelijkheid effectief en efficiënt te verminderen (Deaton, 2002). Daarom is het integreren van de rollen van voorgestelde mechanismen en hun effect op de lange termijn in een theoretisch kader een essentiële eerste stap voor het ontwikkelen van en de evaluatie van doeltreffend beleid. Zo'n kader staat onderzoekers in meerdere disciplines toe om het relatieve belang van elk voorgesteld mechanisme en de interactie tussen mechanismen vast te stellen en de differentiële patronen van causaliteit te ontwaren. Case en Deaton (2005) beargumenteren dat het uiterst moeilijk is om de verbanden tussen gezondheid, opleiding, inkomen en arbeidparticipatie te begrijpen zonder een of ander leidend theoretisch kader. Het is daarom geen verrassing dat meerdere auteurs (zie bijvoorbeeld Case en Deaton, 2005; Cutler et al. 2011) hebben gewezen op het ontbreken van een theorie van SES en gezondheid over de levenscyclus en het belang hebben benadrukt deze te ontwikkelen.

Het doel van dit proefschrift is om een bijdrage te leveren aan een theorie van sociaaleconomische verschillen in gezondheid over het leven. De beperkte vooruitgang tot dusver is mogelijk het gevolg van de volgende factoren. Een aantal van de voorgestelde mechanismen hebben directe gevolgen op korte termijn, meerdere werken echter op de langere termijn, bijvoorbeeld, door een vrij klein maar blijvend effect op het verouderingsproces of op de snelheid van vermogensopbouw. Ongelijkheden in gezondheid, evenals verschillen in SES (bijvoorbeeld vermogen) bouwen op over de levenscyclus, en zijn aanzienlijk groter op late leeftijd. Met andere woorden, om de bijdrage van elke individuele verklaring volledig te kunnen beoordelen is het essentieel om processen over de gehele levenscyclus te modelleren. Een geschikt kader waarin de veelvoudige mechanismen en hun cumulatieve effect op lange termijn kunnen worden bestudeerd is een structureel model van SES en gezondheid over de levenscyclus. Structurele economische levenscyclusmodellen, waarin individuen hun nut (utility) maximaliseren over het leven, hebben waardevol inzicht in economisch gedrag zoals consumptie, spaargedrag, en arbeidsparticipatie geleverd. Echter, tot zeer recent leden levenscyclusmodellen van gezondheid, medische zorg, en sociaaleconomische status onder ernstige technische problemen.

De hoofdstukken 2, 3 en 4 van dit proefschrift zijn daarom gericht op het behandelen van deze technische kwesties. Hoofdstuk 5 stelt vervolgens een theorie van sociaaleconomische ongelijkheid in gezondheid over de levenscyclus voor.

Overzicht van dit promotieonderzoek

Het onderzoek gepresenteerd in dit proefschrift begon met een eenvoudig idee: om een theorie van gezondheid en pensionering op te stellen. Economen hebben aangetoond dat een belangrijk deel van de gezondheidsverschillen met financiële indicatoren van SES kan worden verklaard door het feit dat een slechte gezondheid de mogelijkheid om te werken beperkt, waardoor het inkomen verminderd (Smith, 1999, 2004, 2007). Pensionering is dus een essentieel onderdeel van een theorie van SES en gezondheid.

Onze benadering was om het pensioneringsbesluit te integreren in de formulering van het standaard model van de vraag naar gezondheid en gezondheidsinvestering (Grossman, 1972a, 1972b). In het model van Grossman is de vraag naar medische zorg bepaald door de consumptievoordelen (gezondheid verstrekt nut) en de productievoordelen (gezonde individuen hebben een hoger inkomen) die een goede gezondheid oplevert. Het model is mogelijk één van de belangrijkste bijdragen van de Economie tot de studie van gezondheidsgedrag geweest. Het model is het standaard kader voor het bestuderen van de vraag naar gezondheid en medische zorg geworden, en er bestaan nog relatief weinig theoretische uitbreidingen en alternatieve economische modellen.

Het integreren van het pensioneringsbesluit in de gezondheidsproductieliteratuur (de literatuur die naar aanleiding van de artikelen van Grossman in 1972 ontwikkeld is) was echter niet zo makkelijk. Een belangrijk artefact van de oplossing voor gezondheid was een discontinuïteit bij de leeftijd van pensionering (zie Galama et al., 2008 [Hoofdstuk 3]). De theorie voorspelde dat onmiddellijk na pensionering de gezondheid een lagere waarde zou hebben (of een hogere waarde) ten gevolge van de substitutie van gezondheid voor vrije tijd en de afwezigheid van een productievoordeel (tijdens pensionering levert gezondheid geen productievoordeel op aangezien gepensioneerden geen loon verdienen). Dit kan niet correct zijn. Gezondheid is een voorraad en in tegenstelling tot stromen (zoals gezondheidsinvestering en consumptie) kan een voorraad niet onmiddellijk worden aangepast. De gezondheid kan slechts geleidelijk aan door gezondheidsinvestering en biologische veroudering veranderen.

Hoofdstuk 2 (Galama en Kapteyn, 2009) onderzoekt een algemene oplossing van het Grossman model, waarin we de gebruikelijke aanname verwerpen dat individuen hun gezondheidsvoorraad instantaan aan een “optimaal” niveau kunnen aanpassen zonder aanpassingskosten. Het model voorspelt het bestaan van een gezondheidsdrempel waarboven individuen geen vraag naar medische zorg hebben (een hoekoplossing). Wij vinden dat de algemene oplossing een groter aantal empirische waarnemingen kan verklaren dan de traditionele oplossing.

Hoofdstuk 3 (Galama et al. 2008) formuleert vervolgens een gestileerd structureel model van gezondheid, vermogensopbouw en het pensioneringsbesluit, waarbij we de algemene oplossing gebruiken die in Hoofdstuk 2 werd ontwikkeld. We leiden analytische oplossingen af voor de tijdspaden van consumptie, gezondheid, gezondheidsinvestering, vermogensopbouw en pensionering. We vinden dat verbetering in bevolkingsgezondheid de pensioneringsleeftijd vermindert, terwijl tegelijkertijd individuen pensioneren wanneer hun gezondheid verslechterd. Dit verklaart mogelijk waarom gepensioneerden verslechterende gezondheid als belangrijke reden voor vroege pensionering geven, terwijl de pensioneringsleeftijd in de ontwikkelde wereld, ondanks de voortdurende verbetering van de bevolkingsgezondheid en de levensduur, is blijven dalen. Het model voorspelt verder dat individuen met veel menselijk kapitaal meer in gezondheid investeren en, omdat zij gezonder blijven, later met pensioen gaan dan individuen met minder menselijk kapitaal, waarvan de gezondheid sneller verslechtert.

Terwijl de hoekoplossingen die in Hoofdstukken 2 en 3 worden gebruikt aanvankelijk veelbelovend leken, bleven er problemen bestaan met de eigenschappen van de oplossingen voor gezondheid en gezondheidsinvestering. Bijvoorbeeld, de voorspellingen van het model zijn karikaturen van de werkelijkheid: in de hoekoplossing investeren gezonde individuen in zijn geheel niet in gezondheid, terwijl in werkelijkheid de meeste mensen de dokter minstens één keer per jaar zien. Een bestudering van de literatuur liet zien dat er minstens vijf belangrijke beperkingen van gezondheidsproductie modellen geïdentificeerd waren (zie Hoofdstuk 4). In het kort zijn dit: a) het probleem dat de oplossing voor het niveau van investering in gezondheid niet bepaald is (Ehrlich en Chuma, 1990), b) het onvermogen van het model om de waargenomen negatieve relatie tussen gezondheid en de vraag naar medische zorg te voorspellen (Wagstaff, 1986a; Zweifel and Breyer, 1997), c) het onvermogen van het model om verschillen in de snelheid van gezondheidsverslechtering (niet alleen het niveau) tussen sociaaleconomische groepen te verklaren (Case en Deaton, 2005), d) het gebrek aan "geheugen" in de modeloplossingen (Usher, 1975) en e) de noodzaak om de aanname te maken dat het biologisch verouderingsproces versnelt met leeftijd zodat het leven eindig is, zodat gezondheid verslechtert naarmate men ouder wordt, en zodat men de waargenomen toename van de vraag naar medische zorg aan het einde van het leven kan reproduceren (Case en Deaton, 2005).

Ehrlich en Chuma (1990) wijzen erop dat de algemene aanname in de literatuur van een lineaire relatie tussen de productie van gezondheid en de consumptie van medische zorg er toe leidt dat er geen oplossing voor het optimale niveau van gezondheidsinvestering bestaat. De auteurs merken op dat dit een belangrijke beperking is van het model.

Dit suggereert dat het wellicht de moeite waard is deze aanname eens te herzien door een meer flexibele functionele relatie toe te laten tussen de productie van gezondheid en de consumptie van medische zorg. Dit bleek een belangrijke stap in dit onderzoek te zijn. Echter, het was in dat stadium van mijn promotieonderzoek verre van duidelijk dat dit daadwerkelijk iets zou opleveren. Ten eerste was het resultaat van Ehrlich en Chuma betwist (Reid, 1998; Grossman, 2000). Ten tweede, werd er algemeen verondersteld dat een meer flexibel gezondheidsproductieproces de complexiteit van het probleem substantieel zou vergroten waardoor theoretische en econometrische analyse zeer moeilijk zou worden (Grossman, 2000, p. 364). Deze veronderstelling was wellicht versterkt door het feit dat Ehrlich en Chuma (1990) hun toevlucht moesten nemen tot vergelijkende dynamica om de eigenschappen van het model te illustreren. Deze techniek (Oniki, 1973) is een gevoeligheidsanalyse waarin het richtingseffect van een parameterverandering bepaald kan worden. Ehrlich en Chuma (1990) konden dus niet meer dan de richting (maar bijvoorbeeld niet de grootte) van een verandering in een parameterwaarde voorspellen. Ten derde, was het niet duidelijk dat een meer flexibel gezondheidsproductieprocess de aard van het model wezenlijk zou veranderen. Bijvoorbeeld, er werd algemeen verondersteld dat het optimale gezondheidsniveau dan geleidelijk aan bereikt zou worden i.p.v. ogenblikkelijk (Grossman, 2000, p. 364) – wellicht niet een voldoende belangrijke verbetering van het model om de grotere complexiteit te rechtvaardigen. Mogelijk als gevolg van de bovengenoemde factoren heeft men in de literatuur, ondanks het werk van Ehrlich en Chuma (1990), nooit een flexibel gezondheidsproductieproces ingevoerd.[41] Dus het verder ontwikkelen van een model met een flexibel gezondheidsproductieproces was nog niet serieus gedaan.

Hoofdstuk 4 (Galama, 2011) presenteert een theorie van de vraag naar gezondheid, gezondheidsinvestering en levensduur, gebaseerd op het werk van Grossman (1972a, 1972b) en Ehrlich en Chuma (1990). In dit hoofdstuk lever ik verscheidene bijdragen aan de literatuur. Ten eerste stel ik een nieuwe interpretatie van de evenwichtsvoorwaarde van de gezondheidsvoorraad voor dan gebruikelijk is in de literatuur. Dit is één van de meest centrale relaties in de literatuur van de gezondheidsproductie: deze relatie bepaald het optimale niveau van gezondheidsinvestering (en niet de gezondheidsvoorraad zoals algemeen wordt verondersteld). Ten tweede toon ik aan dat deze alternatieve interpretatie een meer flexibel gezondheidsproductieproces vereist (anders bestaat er geen oplossing voor het optimaliseringprobleem; Ehrlich en Chuma, 1990). Ten derde onderzoek ik in detail de gevolgen van mijn nieuwe interpretatie en van het toelaten van een flexibel gezondheidsproductieproces, en toon aan dat dit de vijf technische problemen in deze li-

[41] De enige uitzondering is wellicht een ongepubliceerd artikel van Dustmann en Windmeijer (2000) die het model van Ehrlich en Chuma (1990) als punt van vertrek namen.

teratuur kan oplossen. In tegenstelling tot de gezondheidsproductieliteratuur voorspel ik een negatieve correlatie tussen gezondheidsinvestering en gezondheid, dat de gezondheid van rijke en hoger opgeleide individuen langzamer daalt en dat zij langer leven, dat de huidige gezondheidsstatus een functie van het aanvankelijke niveau van gezondheid en van de historie van vroeger gemaakte gezondheidsinvesteringen is, dat gezondheidsinvestering snel stijgt naarmate het leven eindigt en dat de lengte van het leven eindig is. Ten vierde leid ik structurele relaties tussen gezondheid en gezondheidsinvestering af (bijvoorbeeld medische zorg) die geschikt zijn voor het empirisch testen van de voorspellingen van het model. Deze structurele relaties bevatten het lineaire gezondheidsproductieproces als speciaal geval waardoor zij toelaten om deze algemene veronderstelling in de literatuur te verifiëren of te verwerpen. Ten vijfde merk ik op dat de theorie niet het algemene begrip steunt dat individuen een bepaald "optimaal" niveau van de gezondheid nastreven. Individuen beslissen over het optimale niveau van gezondheidsinvestering maar kiezen niet een gewenst niveau van gezondheid.

Met deze essentiële aanpassingen kan onze formulering een groter aantal waargenomen empirische patronen verklaren. De resultaten suggereren verder dat het Grossman model een geschikte basis vormt voor de ontwikkeling van een levenscyclusmodel van de SES-gezondheidsgradiënt. Hoofdstuk 5 (Galama en van Kippersluis, 2010) voltooit dit promotieonderzoek en presenteert een levenscyclusmodel dat veelvoudige mechanismen bevat die (gezamenlijk) mogelijk een groot deel van de waargenomen ongelijkheden in gezondheid met SES kunnen verklaren. Het theoretisch kader omvat vereenvoudigde wiskundige representaties van belangrijke mechanismen, hetgeen ons toestaat ons begrip van hun operationele rol te verbeteren in het verklaren van de SES gezondheidsgradiënt en voorspellingen te maken. Ons uitgangspunt is de gezondheidsproductieliteratuur die naar aanleiding van het werk van Grossman (1972a; 1972b) ontwikkeld is en de uitbreidingen van dit model door Ehrlich en Chuma (1990) en Case en Deaton (2005). Onze bijdrage is als volgt.

Ten eerste gebruiken wij de alternatieve interpretatie van de evenwichtsvoorwaarde voor gezondheid (zoals in Galama, 2011 [Hoofdstuk 4]). Dit lost de vijf technische problemen in deze literatuur op.

Echter gebruik van medische diensten en toegang tot zorg verklaart slechts een deel van de relatie tussen SES en gezondheid (Adler et al. 1993). Onze tweede bijdrage is daarom vele potentiële mechanismen in het model op te nemen die mogelijk de ongelijkheden in gezondheid tussen SES groepen kunnen verklaren. Een belangrijk concept in ons werk is "gezondheidsstressoren ten gevolge van werkomstandigheden". Dit concept kan ruim worden geïnterpreteerd als de gezondheidsgevolgen van fysieke werk omstan-

digheden (bijvoorbeeld zware arbeid) maar ook van de psychosociale aspecten van werk (bijvoorbeeld lage status, beperkte controle, veel herhaling, enz.). Het idee is dat aspecten van werk die schadelijk zijn voor de gezondheid geassocieerd zijn met een loonpremie (een compenserend loonverschil). Andere belangrijke eigenschappen van het model zijn levensstijlfactoren (preventieve zorg, gezonde en ongezonde consumptie), curatieve (medische) zorg, pensionering en mortaliteit.

Wij vinden dat groter aanvankelijk vermogen, permanent hoger inkomen (over de levenscyclus) en een hoger niveau van opleiding, individuen er toe aanzet om meer te investeren in curatieve en in preventieve zorg, gezonder te consumeren, en gezondere werkomgevingen en woonmilieus te kiezen. Aanvankelijk vermogen, permanent inkomen en hogere opleiding, verhoogt de vraag naar curatieve zorg. De marginale kosten van curatieve zorg worden daardoor verhoogd. Hogere marginale kosten van curatieve zorg verhogen het gezondheidsvoordeel van preventieve zorg en gezonde consumptie, en verhogen de gezondheidskosten van ongezonde werk- en woonmilieus, en ongezonde consumptie. Gezamenlijk leiden deze gedragskeuzen geleidelijk aan tot een substantieel gezondheidsvoordeel met leeftijd. Verder voorspelt het model een aanvankelijke verwijding en mogelijk een latere vernauwing van de SES-gezondheidsgradiënt, aangezien lage SES individuen hun gezondheidsinvestering sneller verhogen en hun gedrag sneller verbeteren als gevolg van hun snellere achteruitgang in gezondheid. Resultaten van eerdere studies (Ehrlich en Chuma, 1990; Ehrlich, 2000; Galama et al. 2008 [Hoofdstuk 3]) suggereren dat de sneller verslechterende gezondheid van lage SES individuen tot vroege terugtrekking uit de arbeidsmarkt kan leiden (en daarmee mogelijk de verwijding van de gradiënt op vroege en middelbare leeftijd verklaart), en tot kortere levens leidt (en daarmee mogelijk de latere versmalling van de gradiënt verklaart). Ons model is dus in staat een aantal empirische gezondheidspatronen te verklaren. Een dergelijk model bestond nog niet eerder en economen hebben het belang benadrukt van het ontwikkelen van zo'n theoretisch kader voor het begrijpen van de complexe relaties tussen indicatoren van SES en gezondheid over de levenscyclus (Cutler et al. 2011; Case en Deaton, 2005).

References

Adler, N.E., Boyce, T., Chesney, M., Folkman, S. and Syme, L. (1993), "Socioeconomic inequalities in health: no easy solution", Journal of the American Medical Association, 269: pp. 3140-3145.

Adler, N.E., Boyce, T., Chesney, M.A., Cohen, S., Folkman, S., Kahn, R.L., Syme, S.L. (1994), "Socioeconomic status and health: the challenge of the gradient", American Psychologist, 49(1): pp. 15-24.

Almond, D. (2006), "Is the 1918 influenza pandemic over? Long-term effects of in utero influenza exposure in the post-1940 U.S. population", Journal of Political Economy, 114: pp. 562-712.

Auld, M.C. and Sidhu, N. (2005), "Schooling, cognitive ability, and health", Health Economics, 14(10): pp. 1019-1034.

Banks, J., Blundell, R. and Tanner, S. (1998), "Is there a retirement-savings puzzle?", American Economic Review, 88(4): pp. 769-788.

Barker, D., Gluckman, P.D., Godgrey, K.M., Harding, J.E., Owens, J.A. and Robinson, J.S. (1993), "Fetal nutrition and cardiovascular disease in adult life", Lancet, 341(8850): pp. 938-941.

Barsky, R.B., Juster, F.T., Kimball, M.S. and Shapiro, M.D. (1997), "Preference parameters and behavioral heterogeneity: an experimental approach in the health and retirement study", Quarterly Journal of Economics, 112(2): pp. 537-579.

Bazzoli, G.J. (1985), "The early retirement decision: new empirical evidence on the influence of health", The Journal of Human Resources, 20(2): pp. 214-234.

Beckett, M. (2000), "Converging health inequalities in later life: an artifact of mortality selection?", Journal of Health and Social Behavior, 41: pp. 106-119.

Behrman, J.R. and Rosenzweig, M.R. (2004), "Returns to birth weight", Review of Economics and Statistics, 86(2): pp. 586-601.

Bernheim, B.D., Skinner, J. and Weinberg, S. (2001), "What accounts for the variation in retirement wealth among U.S. households?", American Economic Review, 91(4): pp. 832-857.

Black, S.E., Devereux, P.J. and Salvanes, K.G. (2007), "From the cradle to the labor market? The effect of birth weight on adult outcomes", Quarterly Journal of Economics, 122(1): pp. 409-439.

Blanchflower, D.G., Oswald, A.J., van Landeghem, B. (2009), "Imitative obesity and relative utility", Journal of the European Economic Association, 7: pp. 528-538.

Blau, D.M. (2008), "Retirement and consumption in a life cycle model", Journal of Labor Economics, 26: pp 35-71.

Blau, D.M. and Goodstein, R. (2010), "Can social security explain trends in labor force participation of older men in the United States?", Journal of Human Resources, 45(2): pp. 328-363.

Bleakley, H. (2007), "Disease and development: evidence from hookworm eradication in the American South", Quarterly Journal of Economics, 122(1): pp. 73-117.

Blundell, R. and MaCurdy, T. (2000), "Labor Supply", In: Ashenfelter, O. and Card, D. (eds.), *Handbook of Labor Economics*, Amsterdam: North-Holland: pp. 1559-1695.

Bolin, K., Jacobson, L. and Lindgren, B. (2001), "The family as the producer of health - when spouses are Nash bargainers", Journal of Health Economics, 20: pp. 349-362.

Bolin, K., Jacobson, L. and Lindgren, B. (2002a), "Employer investments in employee health – implications for the family as health producer", Journal of Health Economics, 21: pp. 563-583.

Bolin, K., Jacobson, L. and Lindgren, B. (2002b), "The family as the health producer – when spouses act strategically", Journal of Health Economics, 21: pp. 475-495.

Bolin, K., Lindgren, B., Lindstrom, M., and Nystedt, P. (2003), "Investments in social capital – implications of social interactions for the production of health", Social Science and Medicine, 56(12): pp. 2379-2390.

Borg, V. and Kristensen, T.S. (2000), "Social class and self-rated health: can the gradient be explained by differences in life style or work environment?", Social Science and Medicine, 51: pp. 1019-1030

Case, A. and Deaton, A. (2005). "Broken down by work and sex: how our health declines", In: Wise, D.A. (ed.), *Analyses in the Economics of Aging*, The University of Chicago Press, Chicago, pp. 185-212.

Case, A., Lubotsky, D. and Paxson, C. (2002), "Economic status and health in childhood: the origins of the gradient", American Economic Review, 92: pp. 1308-1334.

Case A., Fertig, A. and Paxson, C. (2005), "The lasting impact of childhood health and circumstance", Journal of Health Economics, 24(2): pp. 365-389.

Chiteji, N. (2010), "Time preference, non-cognitive skills and well being across the life course: do non-cognitive skills encourage healthy behavior?", American Economic Review: Papers and Proceedings, 100: pp. 200-204.

Cochrane, A.L., St Leger, A.S, and Moore, F. (1978), "Health service 'input' and mortality 'output' in developed countries", Journal of Epidemiology and Community Health, 32: pp. 200-205

Coe, N.B. and Zamarro, G. (2010), "Retirement effects on health in Europe", Journal of Health Economics (forthcoming).

Contoyannis, P. and Rice, N. (2001), "The impact of health on wages: evidence from the British Household Panel Survey", Empirical Economics, 26: pp. 599-622.

Contoyannis, P., Jones, A.M. and Rice, N. (2004), "The dynamics of health in the British Household Panel Survey", Journal of Applied Econometrics, 19(4): pp. 473-503.

Cropper, M.L. (1977), "Health, investment in health, and occupational choice", Journal of Political Economy, 85: pp. 1273-1294.

Cropper, M.L. (1981), "Measuring the benefits from reduced morbidity", The American Economic Review, *Papers and Proceedings of the Ninety-Third Annual Meeting of the American Economic Association*, 71(2): pp. 235-240.

CSDH (2008), "Closing the gap in a generation: health equity through action on the social determinants of health", Final Report of the Commission on Social Determinants of Health, Geneva, World Health Organization.

Currie, J. and Madrian, B.C. (1999), "Health, health insurance and the labour market", In: Ashenfelter, O., Card, D. (eds.), *Handbook of labour economics*, 3: pp. 3309-3415, Amsterdam: Elsevier Science B.V.

Currie, J. and Stabile, M. (2003), "Socioeconomic status and child health: why is the relationship stronger for older children?", American Economic Review, 93: pp. 1813-1823.

Currie, A., Shields, M. and Price, W. (2007), "The child health/family income gradient: evidence from England", Journal of Health Economics, 26: pp. 213-232.

Cutler, D.M. and A. Lleras-Muney (2008), "Education and health: evaluating theories and evidence", in Robert F. Schoeni et al. (eds.), *Making Americans Healthier: Social and Economic Policy As Health Policy*, National Poverty Center Series on Poverty and Public Policy: pp 29-60

Cutler, D.M., Lleras-Muney, A. and Vogl, T. (2011), "Socioeconomic status and health: dimensions and mechanisms", Oxford Handbook of Health Economics, forthcoming.

Dardanoni, V. and Wagstaff, A. (1987), "Uncertainty, inequalities in health and the demand for health", Journal of Health Economics, 6: pp. 283-290.

Dardanoni, V. and Wagstaff, A. (1990), "Uncertainty and the demand for medical care", Journal of Health Economics, 9: pp. 23-38.

Dave, D., Rashad, I., and Spasojevic, J. (2006), "The effects of retirement on physical and mental health outcomes", NBER Working Paper 12123.

Deary, I. (2008), "Why do intelligent people live longer?" Nature 456: pp. 175-176.

Deaton, A. (2002), "Policy implications of the gradient of health and wealth", Health Affairs, 21(2): pp. 13-30.

De Serpa, A.C. (1971), "A theory of the economics of time", The Economic Journal, 81(324): pp. 828-846.

Duncan, G.J. and Holmlund, B. (1983), "Was Adam Smith right after all? Another test of the theory of compensating wage differentials", Journal of Labor Economics, 1(4): pp. 366-379.

Dustmann, C., and Windmeijer, F. (2000), "Wages and the demand for health – a life cycle analysis", IZA Discussion Paper 171, Germany.

Dwyer, D.S. and Mitchell, O.S. (1999), "Health problems as determinants of retirement: are self-rated measures endogenous?," Journal of Health Economics, 18(2): pp. 173-193.

Ehrlich, I. and Chuma, H. (1990), "A model of the demand for longevity and the value of life extension", Journal of Political Economy, 98(4): pp. 761-782.

Elo, I.T. and S.H. Preston (1996), "Educational differentials in mortality: United States, 1979-85", Social Science and Medicine, 42: pp. 47-57.

Erbsland, M., Ried, W. and Ulrich, V. (2002), "Health, health care, and the environment: econometric evidence from German micro data", In: *Econometric Analysis of Health Data*,

John Wiley & Sons, Ltd, pp. 25-36.

Finkelstein, A., Luttmer, E.F.P. and Notowidigdo, M.J. (2008) "What good is wealth without health? The effect of health on the marginal utility of consumption", NBER Working Paper 14089.

Forster, M. (2001), "The meaning of death: some simulations of a model of healthy and unhealthy consumption, Journal of Health Economics, 20: pp: 613-638.

French, E. and J.B. Jones (2007), "The effects of health insurance and self-insurance on retirement behavior", MRRC working paper 2007-170

Fonseca, F., Michaud, P.-C., Galama, T. and Kapteyn, A. (2009) "On the rise of health spending and longevity", RAND Working Paper, WR-722.

Fuchs, V.R. (1982), "Time preferences and health: an exploratory study", in Victor Fuchs (ed.), *Economic Aspects of Health*, Chicago: U. Chicago Press: pp. 93-120.

Fuchs, V.R. (1986), *The Health Economy*, first ed. Harvard University Press, Cambridge, MA.

Galama, T.J., Kapteyn, A., Fonseca, F., Michaud, P.C. (2008), "Grossman's health threshold and retirement", RAND Working Paper, WR-658.

Galama, T.J. and Kapteyn, A. (2009), "Grossman's missing health threshold", RAND Working Paper, WR-684.

Galama, T.J. and van Kippersluis, H. (2010), "A theory of socioeconomic disparities in health over the life cycle", RAND Working Paper, WR-773.

Galama, T.J. (2011), "A contribution to health capital theory", RAND Working Paper, WR-831.

Gerdtham, U.G., Johannesson, M., Lundberg, L. and Isacson D. (1999), "The demand for health: results from new measures of health capital", European Journal of Political

Economy, 15(3): pp. 501-521.

Gerdtham, U.G., and Johannesson, M. (1999), "New estimates of the demand for health: results based on a categorical health measure and Swedish micro data", Social science and medicine, 49(10): pp. 1325-1332.

Glied, S. and Lleras-Muney, A. (2008), "Technological innovation and inequality in health", Demography, 45(3): pp. 741-761.

Goldman, D.P. and Smith, J.P. (2002), "Can patient self-management help explain the SES health gradient?", Proceedings of the National Academy of Science, 99: pp. 10929-10934.

Grossman, M. (1972a), "The demand for health - a theoretical and empirical investigation", New York: National Bureau of Economic Research.

Grossman, M. (1972b), "On the concept of health capital and the demand for health", Journal of Political Economy, 80(2): pp. 223-255.

Grossman, M. (1998), "On optimal length of life", Journal of Health Economics, 17: pp 499-509.

Grossman, M. (2000), "The human capital model", In Culyer, A.J. and Newhouse, J.P. (Eds.), *Handbook of Health Economics*, Volume 1, pp. 347-408, Elsevier Science.

Gruber, J., and D. Wise (eds) (1999), *Social security and retirement around the world*, University of Chicago Press: Chicago.

Gruber, J., and D. Wise (eds) (2004), *Social security programs and retirement around the world: micro-estimation*, University of Chicago Press: Chicago.

Gruber, J., and D. Wise (eds) (2010), *Social security programs and retirement around the world: the relationship to youth employment*, University of Chicago Press: Chicago.

Hanlon, P., Walsh D., and Whyte, B. (2006), "Let Glasgow flourish", Glasgow: Glasgow Centre for Population Health.

Herd, P., Schoeni, R.F. and House, J.S. (2008), "Upstream solutions: does the supplemental security income program reduce disability in the elderly?", Milbank Quarterly, 86(1): pp. 5-45.

House, J.S., Landis, K.R., and Umberson, D. (1988), "Social relationships and health", Science, 241(4865): pp. 540-545.

House, J.S., Kessler, R.C., Regula Herzog, A. (1990), "Age, socioeconomic status, and health", Milbank Quarterly, 68(3): pp. 383-411.

House, J.S., Lepkowski, J.M., Kinney, A.M., Mero, R.P., Kessler, R.C., Regula Herzog, A. (1994), "The social stratification of aging and health", Journal of Health and Social Behavior, 35(3): pp. 213-234.

House, J.S. and D.R. Williams (2000), "Understanding and reducing socioeconomic and racial/ethnic disparities in health", In: Smedley, B.D. and Syme, S.L. (eds.), *Promoting Health: Intervention Strategies from Social and Behavioral Research*, Washington, DC: 284 National Academy Press: pp. 81-124.

Hurd, M. and Rohwedder, S. (2003), "The retirement-consumption puzzle: anticipated and actual declines in spending at retirement", NBER Working Paper 9586.

Hurd, M. and Rohwedder, S. (2006), "Some answers to the retirement-consumption puzzle", NBER Working Paper 12057.

Jacobson, L. (2000), "The family as producer of health – an extension of the Grossman model", Journal of Health Economics, 19: pp. 611-637.

Kapteyn, A. and Panis, C. (2005), "Institutions and saving for retirement: comparing the United States, Italy, and the Netherlands" in: David A. Wise (ed.), *Analyses in the economics of aging*, The University of Chicago Press, Chicago, pp. 281-316.

Kawachi, I., and Berkman, L. (Eds.) (2003), *Neighborhoods and health*, Oxford, U.K.: Oxford University Press.

Kirk, D. (1970), *Optimal control theory: an introduction*, Prentice-Hall.

Khang, Y.H., Lynch, J.W., Yang, S., Harper, S., Yun, S.C., Jung-Choi, K., et al. (2009), "The contribution of material, psychosocial, and behavioral factors in explaining educational and occupational mortality inequalities in a nationally representative sample of South Koreans: relative and absolute perspectives", Social Science and Medicine, 68(5): pp. 858-866.

Kunst, A.E. and J.P. Mackenbach (1994), "The size of mortality differences associated with educational level in nine industrialized countries", American Journal of Public Health, 84: pp. 932-937.

Laibson, D. (1998), "Life-cycle consumption and hyperbolic discount functions", European Economics Review, 42: pp. 861-871.

Lantz, P.M., House, J.S., Lepkowski, J.M., Williams, D.R., Mero, R.P., Chen, J. (1998), "Socioeconomic factors, health behaviors, and mortality – results from a nationally representative prospective study of U.S. adults", Journal of the American Medical Association, 279(21): pp. 1703-1708.

Leu, R.E. and Doppman, R.J. (1986), "Gesundheitszustandsmessung und Nachfrage nach Gesundheitsleistungen", In: Wille, E. (Hrsg.) *Informations- und Planungsprobleme in offentlichen Aufgabenbereichen*, Frankfurt am Main/Bern/New York: Lang 1986, pp. 1-90.

Leu, R.E. and Gerfin, M. (1992), "Die nachfrage nach gesundheit - ein empirisher test des Grossman-modells (Demand for health - an empirical test of the Grossman model)", In: Oberender, P. (Ed.), *Steuerungsprobleme im Gesundheitswesen*, Baden-Baden: Nomos, pp. 61-78.

Liljas, B. (1998), "The demand for health with uncertainty and insurance", Journal of Health Economics, 17(2): pp. 153-170.

Liljas, B. (2000), "Insurance and imperfect financial markets in Grossman's demand for health model - a reply to Tabata and Ohkusa", Journal of Health Economics, 19(5): pp. 821-827.

Lillie-Blanton, M., Parsons, P.E., Gayle, H., and Dievler, A. (1996), "Racial differences in health: not just black and white, but shades of gray", Annual Review of Public Health, 17: pp. 411-448.

Lleras-Muney, A. and Lichtenberg, F. (2005), "The effect of education on medical technology adoption: are the more educated more likely to use new drugs", Annales d'Economie et Statistique, special issue in memory of Zvi Griliches, No. 79/80: pp. 671-696.

Lleras-Muney, A. (2005), "The relationship between education and adult mortality in the United States", Review of Economic Studies, 72(1): pp. 189-221.

Luchenski, S., Quesnel-Vallee, A, Lynch, J. (2008), "Differences between women's and men's socioeconomic inequalities in health: longitudinal analysis of the Canadian population, 1994-2003", Journal of Epidemiology and Community Health, 62: 1036-1044.

Lynch, J.W., Kaplan, G.A., Shema, S.J. (1997), "Cumulative impact of sustained economic hardship on physical, cognitive, psychological, and social functioning", New England Journal of Medicine, 337(26): pp. 1889-1895.

Lynch, S.M. (2003), "Cohort and life-course patterns in the relationship between education and health: a hierarchical approach", Demography, 40(2): pp. 309-331.

Mackenbach, J.P., Huisman, M., Andersen, O., Bopp, M., Borgan, J.K., Borrell, C., et al. (2004), "Inequalities in lung cancer mortality by the educational level in 10 European populations", European Journal of Cancer, 40(1): pp. 126-135.

Marmot, M., Bosma, H., Hemmingway, H., Brunner, E., Stansfeld, S. (1997a), "Contribution of job control and other risk factors to social variations in coronary heart disease incidence", Lancet, 350(9073): pp. 235-239.

Marmot, M., Ryff, C.D., Bumpass, L.L., Shipley, M., Marks, N.F. (1997b), "Social inequalities in health: next questions and converging evidence", Social Science and Medicine,

44(6): pp. 901-910.

Marmot, M. (1999), "Multi-level approaches to understanding social determinants", in Lisa Berkman and Ichiro Icawachi (eds.), *Social Epidemiology*, Oxford: Oxford University Press: pp 349-367.

Matthews, S., Manor, O., and Power, C. (1999), "Social inequalities in health: are there gender differences?", Social Science and Medicine, 48(1): pp. 49-60.

Michaud, P.-C. and A. van Soest (2008), "Health and wealth of elderly couples: causality tests using dynamic panel data models", Journal of Health Economics, 27: pp. 1312-1325.

Miguel, E. and Kremer, M. (2004), "Worms: identifying impacts on education and health in the presence of treatment externalities", Econometrica, 72(1): pp. 159-217.

Mincer, J.A. (1974), *Schooling, experience, and earnings*, Columbia University Press, ISBN: 0-870-14265-8.

Murasko, J.E. (2008), "An evaluation of the age-profile in the relationship between household income and the health of children in the United States", Journal of Health Economics, 27(6): pp. 1489-1502.

Mushkin, S.J. (1962), "Health as an investment", Journal of Political Economy 70(5): pp. 129-157.

Muurinen, J-M. (1982), "Demand for health: a generalized Grossman model", Journal of Health Economics, 1: pp. 5-28.

Muurinen, J-M., and Le Grand, J. (1985), "The Economic analysis of inequalities in health", Social Science and Medicine, 20(10): pp. 1029-1035.

Nelder, J.A. and Mead, R. (1965), "A simplex method for function minimization", Computer Journal, 7: pp. 308-313.

Nocera, S. and Zweifel, P. (1998), "The demand for health: an empirical test of the Grossman model using panel data", In: Zweifel, P. (Ed.), *Health, the medical profession and*

regulation, Kluwer academic publishers, Boston/Dordrecht/London, pp. 35-49.

OECD (2005), "Pensions at a glance: public policies across OECD countries".

Olson, C.A. (1981), "An analysis of wage differentials received by workers on dangerous jobs", Journal of Human Resources, 16(2): pp. 167-185.

Oniki, H. (1973), "Comparative dynamics (sensitivity analysis) in optimal control theory", Journal of Economic Theory, 6: pp. 265-283.

Oreopoulos, P. (2006), "Estimating average and local average treatment effects of education when compulsory schooling laws really matter", American Economic Review, 96(1): pp. 152-175

Preston S.H. and Elo, I.T. (1995), "Are educational differentials in adult mortality increasing in the United States?", Journal of Aging and Health, 7: pp. 476-496.

Ried, W. (1996), "Willingness to pay and cost of illness for changes in health capital depreciation", Health Economics, 5: pp. 447-468.

Ried, W. (1998), "Comparative dynamic analysis of the full Grossman model", Journal of Health Economics, 17: pp. 383-425.

Robert, S.A. (1998), "Community-level socioeconomic status effects on adult health", Journal of Health and Social Behavior, 39(1): pp. 18-37.

Ross C.E. and Wu, C.L. (1996), "Education, age, and the cumulative advantage in health", Journal of Health and Social Behavior, 37: pp. 104-120.

Royer, H. (2009), "Separated at girth: U.S. twin estimates of the effects of birth weight", American Economic Journal: Applied Economics, 1(1): pp. 49-85.

Schrijvers, C.T.M., van de Mheen, H.D., Stronks, K., Mackenbach, J.P. (1998), "Socioeconomic inequalities in health in the working population: the contribution of working conditions", International Journal of Epidemiology, 27: pp. 1011-1018.

Seierstad, A. and Sydsaeter, K. (1977), "Sufficient conditions in optimal control theory", International Economic Review, 18(2): pp. 367-391.

Seierstad, A. and K. Sydsaeter (1987), *Optimal control theory with economic applications*, Advanced textbooks in economics, Volume 24, Elsevier, North Holland.

Silles, M. (2009), "The causal effect of education on health: evidence from the United Kingdom", Economics of Education Review, 28(1): pp. 122-128.

Skalicka, V., van Lenthe, F., Bambra, C., Krokstad, S. And Mackenbach, J. (2009), "Material, psychosocial, behavioural and biomedical factors in the explanation of relative socio-economic inequalities in mortality: evidence from the HUNT study", International Journal of Epidemiology, 38(5): pp. 1272-1284

Smith, A. (1776), *An inquiry into the nature and causes of the wealth of nations*, Reprinted, Roy H. Campbell and Andrew S. Skinner, eds. Oxford: Clarendon Press, 1976.

Smith, R.S. (1978), "Compensating wage differentials and public policy: a review", Industrial and Labor Relations Review, 32(3): pp. 339-352.

Smith, J.P. (1999), "Healthy bodies and thick wallets", Journal of Economic Perspectives, 13(2): pp. 145-166.

Smith, J.P. (2004), "Unraveling the SES-health connection", Population and Development Review, 30, Supplement: *Aging, Health, and Public Policy*, pp. 108-132.

Smith, J.P. (2007), "The impact of socioeconomic status on health over the life course", Journal of Human Resources, 42(4): pp. 739-764.

Stock, J.H. and Wise, D.A. (1990), "Pensions, the option value of work, and retirement," Econometrica, 58: pp. 1151-1180.

Stringhini, S., Sabia, S., Shipley, M., Brunner, E., Nabi, H., Kivimaki, M. and Singh-Manoux, A. (2010), "Association of socioeconomic position with health behaviors and mortality", Journal of the American Medical Association, 303(12): pp. 1159-1166.

Sydsaeter, K., Strom, A. and Berck, P. (2005), *Economists' mathematical manual*, ISBN-10 3-540-26088-9, 4th ed., Springer Berlin, Heidelberg, New York.

Usher, D. (1975), "Comments on the correlation between health and schooling", In: N.E. Terleckyj (ed.), *Household Production and Consumption*, (Columbia University Press for the National Bureau of Economic Research, New York): pp. 212-220.

Van den Berg G., Lindeboom M. and Portrait, F. (2006), "Economic conditions early in life and individual mortality", American Economic Review, 96(1): pp. 290-302.

Van de Ven, W. and van der Gaag, J. (1982), "Health as an unobservable: a MIMIC-model of the demand for health care", Journal of Health Economics, 1: pp. 157-183.

Van Doorslaer, E.K. (1987), *Health, knowledge and the demand for medical care*, Assen: Van Gorcum, 171, ISBN 90-232-2335-7.

Van Doorslaer, E.K., Van Kippersluis, H., O'Donnell, O., Van Ourti, T. (2008), "Socioeconomic differences in health over the life cycle: evidence and explanations", Netspar Panel Paper 12, December 2008.

Van Kippersluis, H., Van Ourti, T., O'Donnell, O. and Van Doorslaer, E. (2008), "Health and income across the life cycle and generations in Europe", Tinbergen Institute discussion paper series 08-009/3, Erasmus University Rotterdam.

Van Kippersluis, H., O'Donnell, O., van Doorslaer, E., and Van Ourti, T. (2009), "Socioeconomic differences in health over the life cycle in an egalitarian country", Tinbergen Institute discussion paper series 09-006/3, Erasmus University Rotterdam.

Van Kippersluis, H., O'Donnell, O., van Doorslaer, E., Van Ourti, T., (2010), "Socioeconomic differences in health over the life cycle in an egalitarian country", Social Science and Medicine, 70(3): pp. 428-438.

Van Kippersluis, H., O'Donnell and van Doorslaer, E. (2011), "Long run returns to education: does schooling lead to an extended old age?", Journal of Human Resources, forthcoming.

Van Oort, F.V., van Lenthe, F.J. and Mackenbach, J.P. (2005), "Material, psychosocial, and behavioural factors in the explanation of educational inequalities in mortality in the Netherlands", Journal of Epidemiology and Community Health, 59(3): pp. 214-220.

Viscusi, K.W. (1978), "Wealth effects and earnings premiums for job hazards", Review of Economics and Statistics, 60: pp. 408-416.

Viscusi, K.W. (1979), *Employment hazards: An investigation of market performance*, Cambridge: Harvard U. Press.

Viscusi, K.W. (1993), "The value of risks to life and health", Journal of Economic Literature, 31(4): pp. 1912-1946.

Wagstaff, A. (1986a), "The demand for health: some new empirical evidence", Journal of Health Economics, 5: pp. 195-233.

Wagstaff, A. (1986b), "The demand for health: theory and applications", Journal of Epidemiology and Community Health, 40: pp. 1-11

Wagstaff, A. (1993), "The demand for health: an empirical reformulation of the Grossman model", Health Economics, 2: pp. 189-198.

Wolfe, J.R. (1985), "A model of declining health and retirement", Journal of Political Economy, 93(6): pp. 1258-1267.

Williams, D.R., and Collins, C. (1995), "U.S. socioeconomic and racial differences in health: Patterns and explanations", Annual Review of Sociology, 21: pp. 349-386.

Williams, D.R. (1999), "Race, socioeconomic status, and health: the added effects of racism and discrimination", Annals of the New York Academy of Sciences, 896: pp. 173-188.

Yen, I.H. and Kaplan, G.A. (1999), "Neighborhood social environment and risk of death: multilevel evidence from the Alameda county study", American Journal of Epidemiology, 149(10): pp. 898-907.

Zweifel, P., Breyer, F. (1997), *Health Economics*, Oxford University Press, New York.