CCHE CCES
Canadian Centre for Health Economics
Centre canadien en économie de la santé

Working Paper Series
Document de travail de la série

SHOULD THE GROSSMAN MODEL OF INVESTMENT IN HEALTH CAPITAL RETAIN ITS ICONIC STATUS?

Audrey Laporte

Working Paper No: 2014-04

www.canadiancentreforhealtheconomics.ca

January 8, 2015

Canadian Centre for Health Economics
Centre canadien en économie de la santé
155 College Street
Toronto, Ontario

CCHE/CCES Working Paper No: 2014-04
Revised: January 8, 2015

Should the Grossman model of investment in health capital retain its iconic status?

JEL Classification: I12

Key words: health economics; grossman model; theoretical framework

Corresponding Author:

Audrey Laporte
Institute of Health Policy, Management, and Evaluation
Canadian Centre for Health Economics
University of Toronto
155 College St.
Toronto, Canada
Email: audrey.laporte@utoronto.ca

I Introduction

Ever since 1972, Michael Grossman's model of investment in health capital has been the cornerstone of the way economists model health related behavior both theoretically and empirically (Grossman, 1972). Grossman's model is firmly in the Becker tradition of Human Capital: it assumes that the individual is a forward looking, optimizing individual who, in making decisions today, takes account of their possible future consequences. In Grossman's framework, as the name implies, the individual's underlying level of health is treated as a capital good, to be built up by investment and run down by lack of investment. It is not a commodity that can be acquired instantaneously - an individual who wishes to increase his stock of health capital to some target can only do so over time. Health Capital as conceived here is different from how healthy an individual happens to feel today: having a flu, or even a more serious illness, will not necessarily reduce ones stock of health capital, regardless of how much it might reduce ones instantaneous level of utility. Health capital is best thought of as relating to the individual's ability to resist disease, and to perform what the health care literature refers to as activities of daily life: serious arthritis, which makes it difficult to go upstairs, does represent a reduction in ones stock of health capital.

Grossman's model assumes that the individual makes decisions about how much to invest in his stock of health capital at any instant on the basis of a calculation of the costs and benefits, where both costs and benefits may be distributed over time. The benefits side of the calculation is generally seen as having two components: a consumption benefit, in the sense of the direct utility which an individual receives as a result of being healthier, and an investment benefit, which refers to the impact of the individual's health on their income. This latter benefit can refer to immediate payoff from being healthier, as in the case of someone who is paid on a daily or piecework basis and whose income falls when they are unable, because of poor health, to put in as much work, or it can refer to longer term effects, as in the case where a salaried employee who has a history of sickness-related absences from work might be less likely to receive promotions. It is not unusual for researchers to focus on one or the other of these two benefits and excluding the other, considering the pure consumption or pure investment version of the Grossman model. We shall refer to the version of the model that includes both components as the full Grossman model.

Over the years, a number of criticisms have been aimed at the Grossman model [hereafter, MGM]. The Grossman criticism literature has been well summarized by Zweifel (2012), Galama et al. (2012) and Galama and Kapteyn (2011), so we take these papers as our reference points[1]. Zweifel (2012) is particularly forthright in arguing that the Grossman model is fatally flawed saying (pg. 677): "...the acronym MGM already suggests that the model amounts to something like the Hollywood dream factory Metro-Goldwyn-Mayer: much elegance, very inspiring, but of limited relevance to the real world"; (pg. 679) "In sum, even after 40 years of effort, the main criticisms of the MGM still stand"; (pg. 681) "there is something to be gained by breaking away from the MGM fixation." The criticisms most often cited as fatal to the MGM are that the model: 1) suffers from a fundamental indeterminacy with regard to the optimal level of investment in health (Ehrlich and Chuma, 1990); 2) predicts that there will be a positive relationship between health investment and health status whereas empirically this relationship is typically negative; 3) does not make current health behavior dependent on the past; 4) does not predict that health declines with lower socio-economic status; and 5) does not preclude an individual choosing to live forever.

If these criticisms are indeed fatal, then economists need to develop an entirely new framework for the analysis of health-related behavior. This paper evaluates the merits of these criticisms by

[1] Zweifel's editorial prompted a response by Kaestner (2013) to which Zweifel (2013) responded.

deriving the testable predictions of the MGM within a dynamic optimization framework. In effect we reverse the typical empirical testing procedure by taking empirical results which it is claimed the model cannot predict, and show that in fact it does encompass them. A version of the MGM model that includes both the consumption and investment benefits of health is presented using the tools of optimal control including phase diagram analysis.[2] Using phase diagrams allows the optimal trajectories that the MGM predicts individuals will follow over time to be traced and provides deeper insight into the inherent dynamics of the model. The major criticisms leveled at the MGM are then addressed within this framework where the first set of criticisms are treated by analyzing the comparative statics of the model and the second set of criticisms that relate to whether the MGM adequately reflects inter-temporal effects are then analyzed using phase diagrams. This paper argues that these criticisms are baseless and that when the MGM is properly characterized as a problem in inter-temporal optimization it provides a firm foundation for empirical analysis of health-related behaviors.

II Grossman as an application of dynamic optimization theory

Formally, Grossman's model can be analyzed using any of the tools of inter-temporal optimization: optimal control theory, dynamic programming or Chow's Lagrangian approach to inter-temporal optimization, and looked at as either a discrete or a continuous time problem. The various approaches are ultimately equivalent, so choice among them comes down to convenience and to preferences for the readiness with which certain parameters of interest fall out of the analysis. This paper adopts the continuous time optimal control approach to analyzing the model, since it allows for use of phase diagram representation of the individual's optimal lifetime trajectory of investment in health.

Consider a one-state variable version of the MGM set within a continuous time optimal control framework.[3] The individual aims to maximize her discounted lifetime utility:

$$Max \int_0^T U(C_t, H_t)e^{-\rho t}dt \quad U_C > 0, U_{CC} < 0, U_H > 0, U_{HH} < 0, U_{CH} > 0 \tag{1}$$

where C is non health-related consumption, H is health capital (the state variable in the problem) and ρ is the individual's subjective discount rate. As it is set up, the model includes both the consumption and investment benefits of health, since H enters through the utility that the individual derives from having it and will also be positively associated with income. The individual is assumed to have a finite life T. The stock of health capital, H is assumed to evolve according to the following equation of motion:

$$\dot{H} = G(I) - \delta H, \quad G_I > 0, G_{II} \leq 0, G(0) = 0, 0 < \delta < 1, I \geq 0 \tag{2}$$

[2] Galama and Kapteyn (2011) set the Grossman problem up in an optimal control framework, but work with the present value Hamiltonian. This effectively makes their problem non-autonomous. This paper uses a current value Hamiltonian, which effectively makes the problem autonomous by removing time as an explicit argument and therefore allows for the use of phase diagram techniques. Time is obviously still present but has been subsumed into other parts of the problem.

[3] See Ferguson and Lim (1998) for an introduction to continuous time optimal control and Leonard and Van Long (1992) for a more advanced treatment.

where I is health investment and G(I) is the instantaneous production function for health capital. Here $\dot{H}$ can be thought of as net investment in health, with δ the depreciation rate and G(I) as the gross investment term. I is assumed to have a positive marginal product in G.[4] Note that I does not enter the utility function - it yields benefit to the individual only through the additional health that it generates.

The nature of the gross investment term, G(I), varies across the literature. Here it is assumed that health is produced using only purchased inputs, I. It is not uncommon in the literature for there to be a time element in the input set, either because using the I goods takes time or on the assumption that time devoted to healthy activities can add to health even in the absence of complementary specialized I goods. In this model, time is not included as an input since it is not crucial to the inter-temporal issues dealt with in this paper[5].

One aspect of the form of the health production function, G(I), is the productivity of I in the production of H. As will be discussed below, one long-running strand in the literature focuses on the question of whether the production function can have constant returns to scale (CRS) or whether the model fails unless decreasing returns to scale (DRS) is assumed[6]. How this question is operationalized depends on the form of the production function. In assuming a single, purchased input I, the general case is taken to be that $G_I > 0$, $G_{II} \leq 0$. When we are dealing with issues pertaining to the implications of the assumption of constant returns in the Grossman framework, we will assume that G(I) = I.

The issue of constant versus diminishing marginal productivity of I will be discussed once the model is set out in general form since this plays a key role in some of the debate surrounding the MGM. A non-negativity constraint can be imposed on I, indicating that the individual cannot reduce her stock of health by, in effect, selling it.

In expression (2), δ is the intrinsic rate of depreciation of health capital. Under the assumption that G(0) = 0, the equation of motion for H says that when the individual undertakes no health investment, their stock of health capital will decline at a constant proportional rate δ. Grossman (1972) discussed the case where δ increases with the age of the individual, and indeed made it a key part of his discussion of finite life. Section VII discusses the implications of relaxing the constant δ assumption.

Essentially (2) is a non-stochastic[7] first order differential equation in H. Health investment, I, is defined as a commodity which can be purchased in the market at a price pI. It is assumed that the individual has an instantaneous budget constraint of the form:

[4] We can easily extend the model to the case where the production function is represented by G(I) - F(S) where S represents goods (consumption of cigarettes, for example) which are harmful to the individual's health. We focus on health goods rather than health "bads". See for example, Jones et al. (2014).

[5] When the time input approach is taken the utility function has to be modified to include leisure time as one of its arguments, and a time constraint, binding at every instant (since one cannot bank or borrow time) added. The time budget in such a model generally takes account of leisure time, healthy work time, time devoted to health investment, and sick time. If there was a choice among a number of different possible health investment strategies, differences in their time requirements and the opportunity cost of time would be a significant factor.

[6] Grossman (1972) assumed CRS in the production of health and made use of some of the implications of CRS in his detailed analysis. The question might well be asked whether CRS is an essential element of the MGM, or merely a simplifying assumption. It would presumably surprise very few people if it were to turn out that empirically, the health production function does not display CRS. The most significant of the criticisms of the CRS assumption from a theoretical perspective is that by Ehrlich and Chuma, which we shall discuss below.

[7] This is a non-stochastic (i.e. deterministic) version of the model, but it is possible to incorporate uncertainty. See for example, Cropper (1977) and Ferguson and Laporte (2007).

$$Y_0 + Y(H) = C_t + p_I I \quad Y_H > 0, Y_{HH} < 0 \tag{3}$$

where Y_0 is the exogenous portion of the individual's income and Y(H) is the portion of the individual's income which depends on her stock of health. If one were working with a model that included time variables as choice variables, then the concept of healthy time could be defined (with healthy time and sick time adding up to total time) and allowance made for the individual's decision as to how to allocate healthy time across work, leisure and health investment activities. This paper tackles the investment aspect of the Grossman framework by making income a function directly of H, with H depending on past decisions about allocating money income between C and I.

The price of non-health related goods is set to 1 for simplicity. This means that the price of I can be regarded as being relative to the price of C and income as measured in terms of real consumption. At this point, a binding instantaneous budget constraint is used as a further simplification from Grossman's original formulation. This assumption is relaxed below to include an asset equation in order to address a fundamental debate in the literature that stems from the paper by Ehrlich and Chuma (1990).

The Lagrangian for this version of the Grossman problem can be written as

$$\mathcal{L} = U(Y_0 + Y(H) - p_I I, H) + \psi[G(I) - \delta H] + \lambda I \tag{4}$$

where the budget constraint is used to substitute out C. In this formulation, ψ is the co-state on H or the shadow price of an additional unit of H. It is the increase in maximized lifetime utility that could be attained at some time t if the individual were to be given unexpectedly at t, an additional unit of H. λI is a Kuhn-Tucker expression, with $\lambda = 0$ when $I > 0$, and $\lambda > 0$ when I = 0.

The first of the Pontryagin necessary conditions for the one state variable problem, on the assumption, for the moment, of an interior solution for I (i.e. $\lambda = 0$) is

$$\frac{\partial \mathcal{L}}{\partial I} = -p_I U_C + \psi G_I(I) = 0 \tag{5}$$

which can be written as

$$\psi G_I(I) = p_I U_C \tag{6}$$

This is a marginal benefit (MB) equals marginal cost (MC) condition for investment in health. The left hand side of this expression is the product of GI, the marginal product of an additional unit of I in terms of production of health times ψ, the shadow price of another unit of health, and hence shows the utility value of another unit of I, or the MB of I in terms of additional utility from H. The right hand side is the MC of an additional unit of I, also in utility terms. One unit of I costs p_I dollars, and since the price of C has been set to 1 that means that a one-unit increase in I must be matched by a reduction of p_I units in C. The marginal utility of a unit of C is U_C, so the cost in terms of utility foregone of an additional unit of I is $p_I U_C$.

This condition plays a key role in two of the debates in the literature on the MGM - whether the model must exhibit decreasing (rather than constant) returns in order for it to have a solution, and whether the fact that $G_I > 0$ means that the model predicts that there must be a positive correlation between I and H when the model is estimated on individual level data.

III Comparative statics relation between I and H

Zweifel (2012) argues that the MGM predicts a positive relation between I and H but that in empirical work a negative relation is generally found; that is, sicker people (lower health status) are, for example, found to visit the doctor more frequently. This criticism is based on the presumption that because the marginal product of I in the health production function is positive i.e. $G_I(I) > 0$ when I is regressed on H, a production function is being estimated. In effect, it treats H as non-durable as opposed to a durable capital good-the only way to get more H now is through more I this period.

To understand the issue involved it is important to remember that the individual's observed choice of I is a derived demand, and that, if the model is correct, it must always satisfy the first order condition in equation (6), that the additional benefit of an additional unit of I must always equal its MC.

Since (6) must always hold, assuming the model is correct, it can be used to derive testable hypotheses about factors affecting I. Totally differentiating (6) yields:

$$G_I\partial\psi + [\psi G_{II} + p_I^2 U_{CC}]\partial I - p_I U_{CC}\partial Y_0 + [p_I U_{CC} - U_C]\partial p - [p_I U_{CC} Y_H + p_I U_{CH}]\partial H = 0 \quad (7)$$

and rearranging (7):

$$\frac{\partial I}{\partial\psi} = \frac{-G_I(I)}{[\psi G_{II} + p_I^2 U_{CC}]} > 0 \quad (8)$$

so that, as might be expected, an increase in the utility value of H increases the optimal level of I. Looking at the relation between I and p_I, gives:

$$\frac{\partial I}{\partial p_I} = \frac{U_C - p_I U_{CC}}{[\psi G_{II} + p_I^2 U_{CC}]} < 0 \quad (9)$$

which shows that the demand curve for I is downward sloping in p$_I$. From (10) it can be seen that investment in health is a normal good (i.e. when looking at the effect of a change in the strictly exogenous part of income Y_0):

$$\frac{\partial I}{\partial Y_0} = \frac{p_I U_{CC}}{[\psi G_{II} + p_I^2 U_{CC}]} > 0 \quad (10)$$

To determine what the MGM predicts about the impact of changes in H on the level of health investment I (holding everything else constant including ψ, the shadow price of H), depends on the magnitude of Y_H since $[\psi G_{II} + p_I^2 U_{CC}] < 0, U_{CC} < 0$ and $U_{CH} > 0$.

$$\frac{\partial I}{\partial H} = \frac{p_I U_{CC} Y_H + p_I U_{CH}}{[\psi G_{II} + p_I^2 U_{CC}]} \lesseqgtr 0 \quad (11)$$

It should be noted that the sign of (11) depends on which version of the MGM is being considered. For example, if $Y_H = 0$ as with a pure consumption version of the MGM (i.e. no investment benefit of health) then (11) can be signed:

$$\frac{\partial I}{\partial H} = \frac{p_I U_{CH}}{[\psi G_{II} + p_I^2 U_{CC}]} < 0 \tag{12}$$

which means that an increase in H reduces the optimal level of I. On the other hand, with the pure investment version, H does not appear in the utility function and $U_{CH} = 0$, then there would be a positive comparative static relation between I and H. Note that the negative relation between I and H is not a consequence of an assumption of CRS, but holds even when $G_{II} < 0$; that is under an assumption of diminishing marginal productivity of I. Given that G_{II} and U_{CC} are both negative, the sign of (12) clearly depends on the sign of U_{CH} which is taken to be positive.

Basically, the higher the initial level of H the lower the value of another unit of I relative to its opportunity cost in terms of foregone consumption of other, non-health related commodities. Again, noting that the value of a unit of I is derived from the value of the additional H which it can produce and hence the lower the optimal level of I.

If $Y_H > 0$ and H appears in the utility function with $U_{CH} > 0$, as with the full consumption plus investment version of the MGM[8], then the sign of (11) will depend on whether $p_I U_{CC} Y_H + p_I U_{CH} \lesseqgtr 0$. If Y_H is sufficiently large, $\frac{\partial I}{\partial H} > 0$, otherwise $\frac{\partial I}{\partial H} < 0$.

In summary, the MGM may allow a positive relation between an individual's stock of health and their investment in health but this is not a consequence of the form of the production function, as asserted by the critics, rather it is a consequence of the magnitude of the return to health in the form of additional income. If Y$_H$ is large when H is low it is not impossible to observe a positive relation between I and H, but, at values of H at which Y$_H$ is small one would expect to observe a negative relation.

In what follows it is assumed that $[U_{CC} Y_H + U_{CH}] > 0$. Clearly, though, the magnitude of Y_H is a question of some importance for empirical work: for an individual who is initially in very poor health, an improvement in health might result in a sufficiently large increase in income that the income effect would lead to further investments in health.

In addition, it is worth repeating that Zweifel's argument essentially assumes that H is a non-durable commodity in which the level of H today depends solely on the level of I today. In fact, H is a durable capital good, a fact that plays a very significant role in determining the individual's current optimal value of I. This point will be taken up in section V, where the determinants of the optimal trajectories for I and H implied by the MGM are explored.

IV Does the Grossman model suffer from indeterminacy?

The expression

$$-p_I U_C + \psi G_I = 0 \tag{13}$$

has been referred to as the necessary condition for I in the MGM. It says that the optimal level of I at any instant is that at which the MC of I equals its MB, where the MB is the additional health that an additional unit of I will produce, multiplied by the shadow price of each unit of H.

[8] The term U$_{CH}$ will also = 0 in the consumption version of the model if the utility function is separable in H and C.

This condition is at the root of a long-standing issue in the literature, starting with the paper by Ehrlich and Chuma on optimal length of life in the MGM[9]. In that paper, based on their version of the MGM, Ehrlich and Chuma argue that the MC=MB condition cannot be satisfied for the MGM unless DRS hold, and in particular that under Grossman's assumption of constant returns MC will in general not equal MB and the optimal level of I cannot be determined in the model as Grossman set it up.

In the context of the model presented here, it is easy to show that the argument made by Ehrlich and Chuma does not hold. MB has been identified as being equal to ψG$_I$. MC is p$_I$U$_C$, because given the assumption of an instantaneously binding budget constraint, each additional unit of I which is purchased results in a reduction in C of p$_I$ units, each of which has a utility value equal to U$_C$. As I increases C decreases and U$_C$ increases so the marginal cost curve is positively sloped in I.

The production function expression, G(I), allows for differing degrees of returns to scale in the production of H. The simplest case of constant returns would be G(I) = I, giving G$_I$ = 1. Substituting this into the MB = MC condition gives

$$p_I U_C = \psi \tag{14}$$

which is in general easily satisfied, even if ψ were constant[10]. Ehrlich and Chuma's indeterminacy argument is concerned with the case where ψ is everywhere above $p_I U_C$ so there is no limit to optimal investment in health. If there is a no-intersection problem it would actually be in the case where ψ is everywhere below MC for non-negative I, meaning that there is no interior solution for I, which would arise if the MB of another unit of health was so low that the individual would really like a negative value of I. Here, however, the non-negativity condition on I cuts in and the value of I is determinate, being equal to zero. Thus the Ehrlich-Chuma argument does not apply in this version of the MGM.

Our version does, however, differ from Ehrlich and Chuma's in a number of ways, the most important of which is that theirs is a two-state-variable model. Rather than the instantaneous budget constraint used here, they follow Grossman in assuming a binding lifetime budget constraint, meaning that at any point in time the individual's expenditure might exceed or fall short of their income, although lifetime expenditure cannot exceed lifetime income. To see the E-C critique in this context[11], the Hamiltonian can be set up for their version of the problem.

The Grossman problem in this context is

[9] Ried (1998) criticizes the Ehrlich and Chuma result, raising a different issue but still in an optimal control perspective.

[10] Grossman (2000) makes the point that the term ψ is the shadow price of H. This means that its value at any time t is the value to the individual in terms of maximized lifetime utility from that point on of unexpectedly receiving another unit of H. This value function is itself a utility function, with a particular, additively separable over time structure, and will display a positive and diminishing marginal value of H. Since I yields no utility in itself (indeed, I may even yield disutility in itself) its utility is a derived utility, based on the impact on the value function at t of the additional units of H which will be produced from an additional unit of I. Thus diminishing returns to H in the value function translate into diminishing returns to I, so ψ will not be a constant, or independent of I, as Ehrlich and Chuma argue.

[11] There are some other differences between Ehrlich and Chuma's version of the model and that presented here, but they are not relevant to the issue at hand.

$$Max \int_0^T U(C_t, H_t)e^{-\rho t}dt \quad U_C > 0, U_{CC} < 0, U_H > 0, U_{HH} < 0, U_{CH} \geq 0 \tag{15}$$

subject to

$$\dot{A} = rA + Y_0 + Y_H - C - P_I I \tag{16}$$

and

$$\dot{H} = G(I) - \delta H, \quad G_I > 0, G_{II} \leq 0, G(0) = 0, 0 < \delta < 1, I \geq 0 \tag{17}$$

giving as the Hamiltonian

$$\mathcal{H} = U(C_t, H_t) + \Psi_H[G(I) - \delta H] + \psi_A[rA + Y_0 + Y_H - C - P_I I] \tag{18}$$

where ψ_A is the costate, or shadow price for financial assets, A. Here C_t becomes a choice variable along with I$_t$ since the two are no longer bound by the budget constraint at time t. (Note that it is still assumed that $p_C = 1$.) The necessary conditions for C and I are

$$U_C - \psi_A = 0 \tag{19}$$

and

$$\psi_H G_I - \psi_A P_I = 0 \tag{20}$$

In their presentation of the MGM, Ehrlich and Chuma (following Grossman's original exposition) introduce a cost of investment function, C(I), which has increasing MC when I is produced under conditions of DRS and constant MC when I is produced under CRS. In their formulation of the model, I is itself a produced commodity, produced using market inputs and time. Their equation of motion for H is

$$\dot{H} = I - \delta H \tag{21}$$

Thus Ehrlich and Chuma put the DRS in at the stage of the production of I, and assume that one unit of I is equivalent to one unit of H, i.e. I is the current output of the health production function. Here we assume that I is an input into the production of H, which can be bought directly in the market (medical care, for example) and allow non-constant returns to enter through the production function G(I). To make the two sets of notation equivalent, one could replace Ehrlich and Chuma's I, in their equation of motion for H, with the production function for I, and make the arguments of their production function the choice variables. They characterize CRS in terms of the production function for I; here constant returns are characterized by writing G(I) = I.

In terms of the notation used here, Ehrlich and Chuma's expression for what they term the flow equilibrium for optimal investment in health is

$$C_I(I) = \psi_H/\psi_A \tag{22}$$

where the counterpart to their MC term is $p_I/GI(I)$. Thus equation (22) can be re-written as

$$p_I/G_I(I) = \psi_H/\psi_A \tag{23}$$

This is an expression that is equivalent to Ehrlich and Chuma's flow equilibrium condition. Ehrlich and Chuma argue that the right hand side of (23) is independent of the scale of investment. Their argument with regards to CRS is that if G(I) = I so that G_I(I) = 1, (23) becomes

$$p_I = \psi_H/\psi_A \tag{24}$$

and that there is no guarantee that (24) can be satisfied, or yield a determinate choice of I, since both sides are taken to be independent of I.

Ehrlich and Chuma's introduction of the C(I) function clouds a certain key aspect of the problem - the fact that the decision-maker is a utility-maximizing individual, and this is an individual-level problem. Their use of C(I) (which follows from Grossman's discussion of the equivalence between the individual's problem of the optimal choice of health investment level and the firm's problem of choosing an optimal level of physical capital) draws attention away from the first order conditions of the problem, and has led researchers who followed their approach to pass over the same issue.

To see this, consider equation (19) above: $U_C - \psi_A = 0$. This obviously gives

$$\psi_A = U_C \tag{25}$$

Substituting (25) into (23) and rearranging slightly gives

$$U_C p_I = \psi_H G_I(I) \tag{26}$$

which is the same as the necessary condition which was derived in the absence of an equation of motion for A. In that case, of an instantaneously binding budget constraint, it was argued that an increase in I must reduce C, so the opportunity cost of an increase in I was lost utility from consumption of other, non-health related commodities.

The same argument holds here, even though the budget constraint is binding over the lifetime and not instantaneously. The decision to increase I by one unit is a decision to reduce accumulated assets by p_I units. Each unit of A has a shadow price of ψ_A, and optimality requires that $\psi_A = \mathrm{U}_C$. No matter how the reduction in accumulated financial capital is distributed across consumption over time, an increase in I still has an opportunity cost in the form of reduced utility from other consumption. The opportunity cost will be less in this case than in the case of an instantaneously binding budget constraint because the individual has the option of taking part of the reduction of consumption at some point in the future rather than taking it all today, but there will still be an opportunity cost. Ehrlich and Chuma's introduction of a cost-of-production function for I draws attention to the market element of the cost of investment in health and away from the element which is accounted for by the individual's subjective opportunity cost.

Now, introducing CRS by setting $G_I(I) = 1$, the condition (26) becomes

$$U_C p_I = \psi_H \tag{27}$$

And again there is no reason to doubt that this condition can be satisfied - i.e. no reason to believe that there will necessarily be an indeterminacy problem with ψ_H above the marginal cost of I at all values of I.

The Ehrlich and Chuma indeterminacy argument lies behind many of the recent critiques of the MGM, but it is argued here that their particular critique does not have the force that it has been accorded in the literature. There have been other arguments made relating to the CRS assumption which we turn to in the next section.

V Dynamic predictions from the Grossman model

If as was argued above, the MGM with CRS does not suffer from the indeterminacy of I problem, one can reasonably ask whether the other arguments made about the failures of the CRS version of the model hold up. The critics take Grossman's use of a CRS production function for health as not just an issue of a simplifying assumption but seem to regard it as exposing a fatal intrinsic flaw in the entire MGM framework.

The criticisms which are tied to CRS are now addressed within a model which assumes CRS in the production function for gross investment in health i.e. G(I)=I. Under the CRS consumption plus investment version of the model, $G_I = 1, G_{II} = 0, Y_H > 0$, and $Y_{HH} < 0$.

Two key arguments have been made which purport to expose fatal weaknesses in the MGM. First, that solutions to the current health investment decision lack history in that they do not take account of initial health (Galama et al., 2012) and related to this that the MGM does not reflect the behavior of an individual who has suffered a major illness (Zweifel, 2012). Second, that the MGM does not predict that health will decline faster for individuals of lower socio-economic status (Galama et al., 2012).

Returning to the one-state variable version of the MGM introduced in section II, allows for the use of phase diagrams[12], which in turn allows for a focus on the dynamic aspects of the Grossman approach that seem sometimes to be overlooked but are essential to addressing the arguments raised by the critics. As we have argued in the previous section, the choice between an instantaneously binding and a lifetime budget constraint is not fundamental and the gain in clarity from being able to use the phase diagram technique outweighs any loss from working in a one-state variable framework.

The evolution of H and I over time can be represented in a diagram with I on the vertical and H on the horizontal axis. The equations of motion for H and I will be used to determine how the individual's optimal trajectory evolves over time. The starting point for a phase diagram is the definition of stationary loci for each of I and H. The stationary locus for I has the property that, on the locus there is no intrinsic tendency for I to change. In other words, on the stationary locus for I, $\dot{I} = 0$. Similarly, on the stationary locus for H, $\dot{H} = 0$. Given the stationary loci, the (I,H)

[12] Phase diagrams cannot, in general, be used for two state variable problems.

space can be divided into regions where I and H are increasing or decreasing by finding the phase arrows for I and H[13].

In the CRS version of the one-state variable MGM, the individual's objective remains to maximize (1), but when G(I) is replaced by I, the equation of motion for H becomes

$$\dot{H} = I - \delta H \tag{28}$$

Maintaining the assumption that the budget constraint is satisfied at all values of t, and assuming, following Grossman, that I is non-negative (I $\geq$ 0), instead of the Lagrangian the current value Hamiltonian for the problem can be written as:

$$\mathcal{H} = U(Y_0 + Y(H) - p_I I, H) + \psi[I - \delta H] \tag{29}$$

The Pontryagin necessary condition with respect to I in this case is:

$$\partial\mathcal{H}/\partial I : -p_I U_C(Y_0 + Y(H) - p_I I, H) + \psi = 0 \tag{30}$$

which can be re-written as $\psi = p_I U_C(Y_0 + Y(H)) - p_I I, H)$. The second of Pontryagin's necessary conditions is:

$$\dot{\psi} = \rho\psi - \mathcal{H}_H \tag{31}$$

or

$$\dot{\psi} = \rho\psi - [U_C Y_H + U_H - \delta\psi] \tag{32}$$

This condition is one of the advantages of using the optimal control framework since it draws attention to the fact that the shadow price of H will change as time passes simply because of the passage of time with no change in any of the exogenous variables in the model. As the first of the Pontryagin necessary conditions (30) shows, if the shadow price of H is changing as time passes, the level of I must also change to ensure that MC continues to equal MB.

The stationary locus for H in this case becomes a straight line:

$$I = \delta H \tag{33}$$

Turning to the stationary locus for I, totally differentiating (30) with respect to time will give an equation in $\dot{H}, \dot{I}$ and $\dot{\psi}$. Substituting in (28) and (32) and rearranging yields an equation for $\dot{I}$:

$$\dot{I} = \frac{p_I[U_{CH} + U_{CC}Y_H][I - \delta H] + [U_C Y_H + U_H] - [\rho + \delta]p_I U_C}{p_I^2 U_{CC}} \tag{34}$$

To find the slope of the stationary locus for I, set (34) to 0 which requires that

[13] See Ferguson and Lim (1998) for an introduction to the technique.

$$p_I[U_{CH} + U_{CC}Y_H][I - \delta H] + [U_C Y_H + U_H] - [\rho + \delta]p_I U_C = 0 \tag{35}$$

Differentiating (35) with respect to I gives

$$p_I[U_{CH} + U_{CC}Y_H] - p_I^2[I - \delta H][U_{CHC} + U_{CCC}Y_H] + [\rho + \delta]p_I^2 U_{CC} - p_I[U_{CH} + U_{CC}Y_H] \tag{36}$$

or

$$[\rho + \delta]p_I^2 U_{CC} - p_I^2[I - \delta H][U_{CHC} + U_{CCC}Y_H] \tag{37}$$

The term [I- δ H] which is equal to $\dot{H}$, could be positive or negative depending on which side of the stationary locus for H one happens to be at, and will be zero at the point where the stationary locus for I cuts the stationary locus for H. At that intersection point, the second term in (37) drops out and $[\rho + \delta]p_I^2 U_{CC} < 0$. Elsewhere, the term $[U_{CHC} + U_{CCC}Y_H]$, which is the derivative of U_{CC} with respect to H, has one negative (the first) and one positive (the second) element and cannot be definitively signed. Because of the mix of signs it can probably assumed that it is small in absolute value whatever its sign might be.

Now taking (35) and differentiating it with respect to H, we have

$$\begin{aligned}[U_{CC}Y_H^2 + 2U_{CH}Y_H + U_C Y_{HH} + U_{HH}] - [\rho + 2\delta][U_{CH} + U_{CC}Y_H] \\ +[I - \delta H]p_I[U_{CHH} + 2U_{CHC}Y_H + U_{CCC}Y_H^2 + U_{CC}Y_{HH}]\end{aligned} \tag{38}$$

The first term in square brackets in this expression can be assumed to be negative, on the assumption that the overall marginal utility of health is diminishing when accounting for both its effect on income and its direct consumption effect. The last term contains a mixture of positives and negatives and can presumably be assumed to be small.

The slope of the stationary locus for I is

$$\frac{\partial I}{\partial H} = \frac{[\rho + 2\delta][U_{CH} + U_{CC}Y_H] - [U_{CC}Y_H^2 + 2U_{CH}Y_H + U_C Y_{HH} + U_{HH}] - [I - \delta H]p_I[U_{CHH} + 2U_{CHC}Y_H + U_{CCC}Y_H^2 + U_{CC}Y_{HH}]}{[\rho + \delta]p_I^2 U_{CC} - p_I^2[I - \delta H][U_{CHC} + U_{CCC}Y_H]} \tag{39}$$

In many ways, the most important point at which to be able to sign (39) is at the intersection of the stationary loci for I and H, which is the equilibrium point for the system. At this point (39) becomes

$$\frac{\partial I}{\partial H} = \frac{[\rho + 2\delta][U_{CH} + U_{CC}Y_H] - [U_{CC}Y_H^2 + 2U_{CH}Y_H + U_C Y_{HH} + U_{HH}]}{[\rho + \delta]p_I^2 U_{CC}} \tag{40}$$

which, on the assumptions made above is assumed to be negative. When we put the terms involving [I - δH] back in, because they are mixtures of positive and negative terms and are third derivatives of the utility function, it can probably be assumed that they will not change the sign of the slope of the stationary locus, even if they change its magnitude to some degree.

Then to find the phase arrows for I, $\partial \dot{I}/\partial I$ or $\partial \dot{I}/\partial H$ can be evaluated close to the stationary locus for I:

$$\frac{\partial \dot{I}}{\partial I} = \frac{[\rho+\delta]p_I^2 U_{CC}}{p_I^2 U_{CC}} > 0 \tag{41}$$

$$\frac{\partial \dot{I}}{\partial H} = \frac{[U_{CC}Y_H^2 + 2U_{CH}Y_H + U_C Y_{HH} + U_{HH}] - [\rho+2\delta][U_{CH} + U_{CC}Y_H]}{p_I^2 U_{CC}} > 0 \tag{42}$$

The phase arrows for H can be found from the fact that $\dfrac{\partial \dot{H}}{\partial I} = 1, > 0$ and $\dfrac{\partial \dot{H}}{\partial H} = -\delta, < 0$.

The phase diagram for the CRS case of the one-state variable consumption plus investment version of the MGM is depicted in Figure 1. The phase arrows in the four regions defined by the stationary loci and illustrative trajectories in each region that are consistent with the phase arrows are shown in Figure 1. Obviously only trajectories that have the potential to reflect the individual's lifetime utility maximization decision are of interest. This means that the Pontryagin necessary conditions must be satisfied at every point along any trajectory that is worth considering. Because the phase diagram is derived using the first-order conditions, every trajectory in it is a potential optimal trajectory. The trajectory that is actually chosen by the individual depends on the level of health that she starts with and where she wants to end up.

Like virtually all optimal control problems, this one displays saddle-point dynamics. A saddle-point equilibrium (E in Figure 1) is defined as the point of intersection of the stationary loci (because at that point there is no intrinsic tendency for either I or H to change); there are only two trajectories, referred to as the stable branches that converge to the equilibrium. There are also two trajectories referred to as the unstable branches, which point directly away from the equilibrium. Every other trajectory that could be drawn on the diagram, would initially move toward the equilibrium but eventually turn around and diverge from it. This is why saddle-point dynamics are generally taken to imply the uniqueness of solution trajectories.

The intersection of the stationary loci was referred to above as the equilibrium for the problem. Normally in economic modeling, it is taken for granted that the system will either be at or be converging to its equilibrium. In optimal control problems this is certainly true for most macro economic applications and for models of economic growth. It is not however true for most micro-economic problems. The reason for this is that it takes an infinite amount of time to reach the equilibrium. Optimal control problems include among their necessary conditions what are referred to as terminal transversality conditions. For an infinite horizon problem the transversality conditions require that whatever the initial value of the state variable, the control variable must be selected so that the optimal trajectory is the stable branch to the equilibrium. For a finite horizon problem, like the one considered here, the transversality condition gives a different endpoint, so the equilibrium point of the system will not be part of the individual's optimal trajectory. It makes no sense to assume that the optimizing individual is fully informed and forward looking with the one minor exception that she assumes that she will live forever. Economies may live forever; individuals within them do not.

For a finite horizon problem, there are a number of options for the terminal transversality condition. One is that the stock of the state variable at T equals zero, meaning that the stock of the state variable has been used up at the end of the horizon. This condition is often used in models of accumulation of financial assets in the absence of a bequest motive. In some cases it is not possible for the state variable to reach zero in finite time. In those cases the transversality

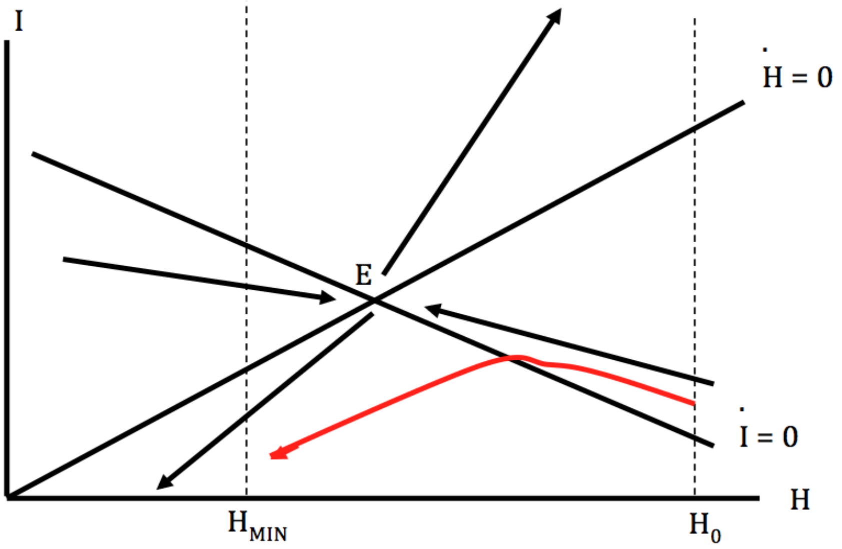

Figure 1: An optimal path within the Grossman model

condition is ψ at T equals zero, meaning that there is no value to having another unit of the state variable at the end of the problem. The third case is what is known as a fixed endpoint problem in which a specific target value is chosen for the state variable at T. This would be the case in a model for example, of financial asset accumulation when the individual has a target bequest motive. In what follows it is assumed that H$_{MIN}$ is a fixed endpoint target. The implications of relaxing this assumption are discussed in section VII.

A Lack of path dependence

One issue that has been raised is whether the MGM properly takes account of the way an individual's initial stock of health capital affects their later health investment decisions along with the related issue of whether it adequately captures path dependency in health (Galama et al. 2012). These are fundamentally dynamic questions dealing with the nature of the joint trajectory of health investment I, and the stock of health H, that the model predicts for an individual.

Figure 1 includes a trajectory which is typical of those found in Grossman problems. The individual is born healthy, in the sense that her initial H_0 is above the level to which an infinite-lived individual would tend over time. H$_{MIN}$ is taken to be a firm endpoint value of H and the length of the time horizon, from t = 0 to t = T, is treated as fixed. The individual's problem, then, is to select a trajectory of I and H that will take her from her initial value of H to her terminal value with an elapsed time of exactly T time units. This individual's stock of health capital will decline throughout her life, although not necessarily at a constant proportional rate. Initially she chooses a low value of I, which allows her H to decline. As time passes and H falls she increases her optimal I, slowing the rate of decline in H but not reversing it, so that her stock of health capital continues to fall. After a certain time has elapsed - indicated on the diagram by the point at which the trajectory cuts the stationary locus for I, she begins to let her I fall again, while H continues to fall towards H$_{MIN}$. Looking at this trajectory in a lifetime perspective, it can be seen that there are intervals during which I and H are moving in the same direction and intervals in which they are moving in opposite directions. This relates to the point made earlier about the observed relationship between I and H being conditioned by the fact that H is a durable good so that the value of H in any period depends not just on the current value of I but also on the amount of H which the individual had going into that period.

It is worth noting that the phase diagram has been drawn holding all of the exogenous variables - p_I and Y_0, notably - unchanged, so that the observed relation between I, H and the exogenous variables will change over time at a non-constant rate. To investigate the determinants of the individual's optimal I and H then, it is important to allow for the intrinsically dynamic nature of the problem and to keep in mind the fact that, to the extent that there is a valid equilibrium concept for this individual, it is represented by *a trajectory*, not a point.

Figure 2 shows the optimal trajectories for two individuals; one born with a high level of initial health (H$_{0HIGH}$), a second individual who is identical to the first with the sole exception that she is born with a much lower initial stock of health (H$_{0LOW}$). The trajectory for the individual who starts at H$_{0HIGH}$ is the one that was already discussed in the context of Figure 1. Given equation (21), the initial value of H, and the terminal value of H, it can be seen that only one trajectory satisfies the optimality conditions and fits the time horizon.

In Figure 2, it can be seen that compared to the first individual, the individual born with a much lower initial stock of health (H$_{0LOW}$), according to the phase arrows, will tend to start with a high value of I (point B) causing her stock of health to increase, and as her health increases she can then reduce her level of I. Her stock of H continues to increase until her trajectory reaches the $\dot{H} = 0$ locus at which point both H and I will decrease until she reaches the end of the horizon. The MGM predicts that it is optimal for the individual born with a low initial stock of health to front-load her health investment. Contrary to the assertion made by Galama et al. (2012) then, in a standard MGM with CRS, current health status at any value of t is indeed a function of the individual's initial level of health and her history of prior health investments made.

Zweifel (2013) says that the MGM predicts "...total investment in health should decrease at least in the case of a serious illness when time to death suddenly becomes short, reducing the present value of returns to investment. Neither prediction is borne out by the data"(pg. 361); meaning that the model predicts a downward jump because of the shortened horizon.

In Figure 3, again consider an individual who is born with a high stock of health. She is initially following the trajectory illustrated in Figure 2 but part way through the trajectory she is struck by a significant unanticipated health shock (at point B) that takes her to H$_{LOW}$. At H$_{LOW}$ it is clear that the original level of health investment will no longer be optimal. In terms of a control theory problem a shock like this to the state variable necessitates re-planning. Re-optimizing over the remaining horizon (which may or may not have been affected by the shock), taking as the initial stock of health for the new plan that stock with which she now finds herself to be endowed at the instant after the illness struck. In Figure 3 the individual responds to this new lower health state by dramatically increasing her level of investment to point C and by following a new optimal trajectory for the remainder of the planning horizon. The amount of the upward jump will depend on how long that remaining horizon is and in particular on whether the illness has also shortened the individual's life expectancy. The shorter the remaining horizon, the smaller the expected jump.

Thus from Figure 3 it can be seen that the model can indeed predict an upward jump in investment, I, the size of which will depend on the length of the individual's remaining horizon and also on the magnitude of the cross-partial between health and consumption in the individual's

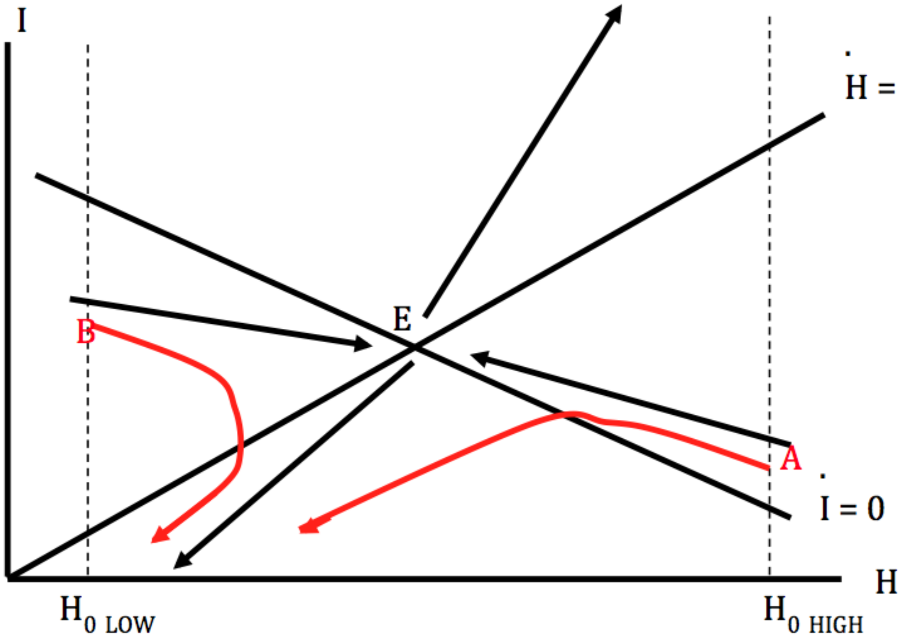

Figure 2: Low versus high initial health

utility function since that will play a key role in determining the opportunity cost of a jump in I.

Even in the context of the CRS version of the MGM it can be seen that individuals respond to their particular health history. In Figure 3 it was assumed that the shock was completely unanticipated. Elsewhere this problem has been investigated as a stochastic control problem within the MGM framework and it was shown that the model can be modified to allow the individual to take account of the probability of a major health shock when she is making her initial health investment plans (Laporte and Ferguson (2007)). Even in that case however, once the shock has occurred the individual re-optimizes[14].

B Socio-economic gradient in health

Galama et al. (2012) argue that the MGM is not capable of predicting a socio-economic gradient in health, specifically that it does not allow health to decline faster for individuals with lower socio-economic status (SES). Here SES is taken to refer to income. Consider the case of two individuals born with the same high level of initial health (H_{HIGH}), one of which has a higher income at each value of t than the other. It is assumed that for each individual, income is constant over time, so this is not the case where health status can affect income, i.e. the Y(H) term is dropped. The two

[14] One reason that it might appear that the model presented here does not allow current decisions to depend on past decisions is that the solutions to control theory problems are expressed as open loop rather than as feedback solutions. As can be seen however, from the necessary conditions, the optimal value of I at any time does depend on H and in the event of a shock to H the individual will re-plan taking account of how her realized H differs from the value she had originally anticipated.

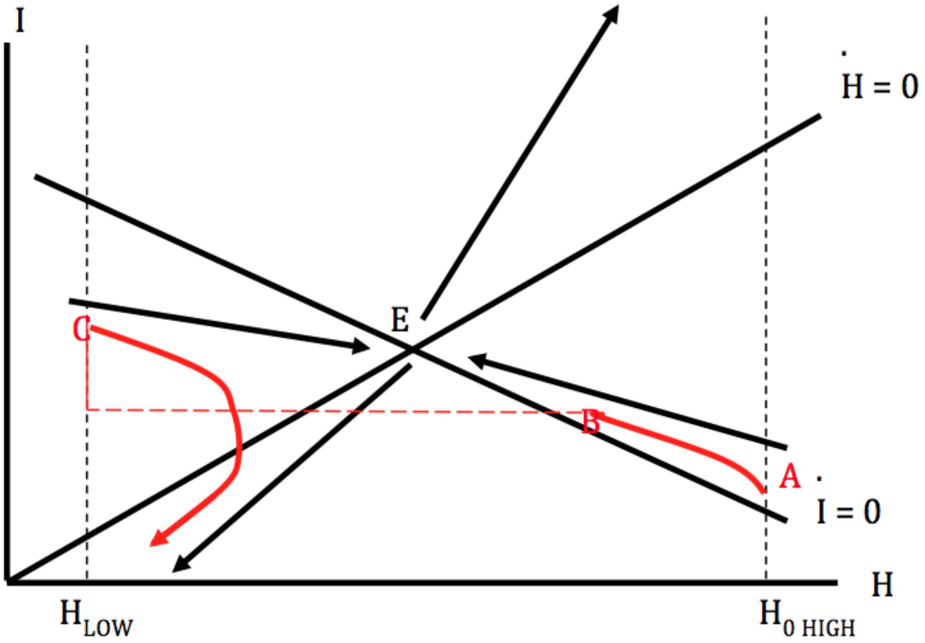

Figure 3: Effect of an unanticipated health shock

individuals also have the same equation of motion for health so income differences do not mean differences in intrinsic productivity when it comes to producing health. This approach is taken to underscore the fact that an SES gradient can emerge even in a CRS pure consumption version of the MGM and is not dependent on a healthier individual gaining an edge.

Recall (33), the condition that defines the stationary locus for H in the CRS case. This expression will clearly not be affected by an increase in Y_0. This is because it is biological rather than behavioural - it says simply that for $\dot{H} = 0$ it must be the case that the level of I is just sufficient to replace the depreciation of H, at what ever the value of H happens to be. To use the terminology of models of investment in physical capital, it must be that the level of gross investment is exactly equal to the level of depreciation so that the level of net investment equals zero.

Equation (34) is the expression for $\dot{I}$ =0. Setting $Y = Y_0$ and at $\dot{H} = 0$, i.e. at the intersection of the stationary loci, (34) becomes:

$$\dot{I} = \frac{(\rho + \delta)p_I U_C - U_H}{p_I^2 U_{CC}} \tag{43}$$

And $\dot{I} = 0$ requires

$$\left[[\rho + \delta] \frac{p_I U_c}{G_I} - U_H \right] \tag{44}$$

Differentiating (44) with respect to I and Y_0, shows how I changes, holding H constant, when Y_0 increases, and hence moves vertically from the (I,H) point at which the stationary locus for I cuts the stationary locus for H. In the CRS case $G_I = 1$ and $G_{II} = 0$ so (44) becomes:

$$[[\rho + \delta]p_I U_c - U_H] \tag{45}$$

And differentiating gives:

$$\frac{\partial I}{\partial Y_0} = \frac{U_{HC} - [\rho + \delta]p_I U_{CC}}{p_I U_{HC} - p_I^2 U_{CC}[\rho + \delta]} = \frac{1}{p_I} > 0 \tag{46}$$

which says that the stationary locus for I, at the point at which it cuts the stationary locus for H, shifts up in response to an increase in Y_0. Thus it can be argued that the stationary locus for I will shift up in response to an increase in Y_0. This in turn will tend to pull up all of the trajectories for the individual whose income has increased upward - this can be seen most clearly in the case of the stable branch to the equilibrium, when comparing the pre- and post-income increase diagrams. Thus comparing two individuals, identical in preferences but one of whom has a higher (exogenous) level of income than the other, one would expect the higher income individual's phase diagram to differ from that for the lower income individual in that the stationary locus for I will be higher (but they will both have the same stationary locus for H) and the higher income individual's potential trajectories will all be shifted up relative to those of the lower income individual. These two individuals are depicted in Figure 4.

Assuming that the two individuals both still have the same, exogenous value of T, and assuming that they are born with the same initial H_0 so that there is no income-related health difference at birth, one would still expect the higher income individual to invest more in her health at every value of t. In economic terms, this is because an increase in Y by increasing the amount of C that can be consumed at any given value of I reduces the opportunity cost of increasing I.

As a result, the higher income individual's H will decline more slowly than that of the lower income individual, since $\dot{H} = I - \delta H$ and I is higher at every H. This means that, contrary to the assertion of Galama et al. (2012), the MGM with CRS is perfectly capable of generating a case in which higher income individuals' health declines more slowly than does that of lower income individuals. Whether this is the explanation of observed income-related gradients in health or not, the claim that the MGM cannot yield such gradients is clearly incorrect.

C The health threshold

Galama and Kapteyn (2011) say that one criticism of the MGM that has not been satisfactorily addressed is the fact that in empirical work I and H are generally found to be negatively related whereas it is said the model predicts a positive relation (Zweifel 2012). They propose modifying the basic MGM to allow for a corner solution in I i.e. I=0. They refer to this as Grossman's missing health threshold. The issue of the relationship between I and H has been addressed in section III. This section sets up the Galama-Kapteyn model in the context of the one-state variable version of the MGM to illustrate the effect of a non-negativity constraint on I using a phase diagram.

Galama and Kapteyn(2011) also set the problem up in the form of an optimal control problem. They incorporate explicitly a non-negativity constraint on I. It is worth noting that in his original 1972 paper, Grossman discussed such a non-negativity constraint in some detail but his calculus

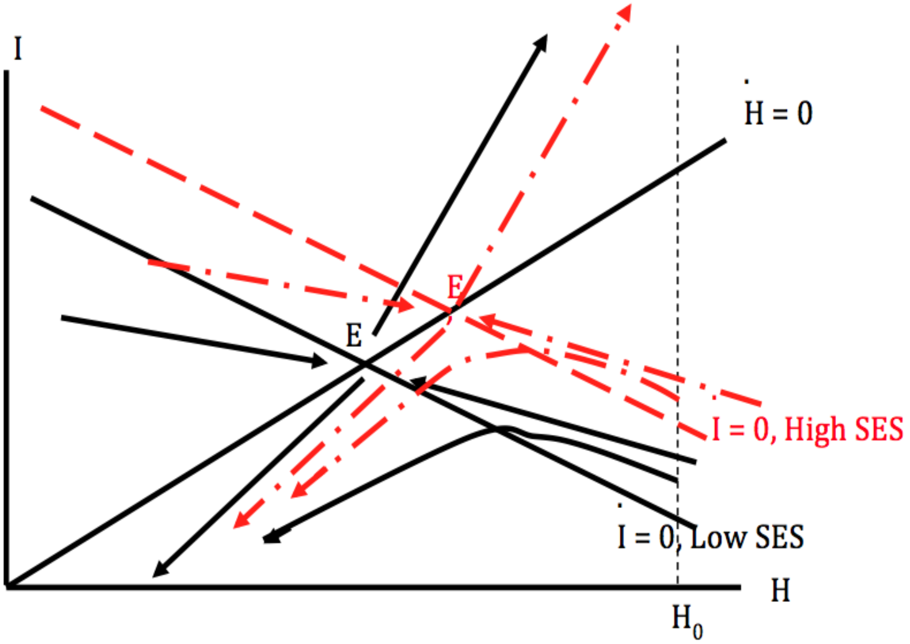

Figure 4: High versus low SES

of variations based approach did not lend itself to a formal demonstration of the effects of this assumption.

Allowance for the possibility of a non-negativity constraint on I in the MGM was outlined at the outset of the paper, in which the λI term was introduced. When I is positive, λ will be equal to 0, so to this point an interior solution has been assumed. To see the effect of allowing the non-negativity constraint on I to be binding (13) can be modified as follows:

$$-p_I U_C + \psi + \lambda = 0 \tag{47}$$

and rearranged to obtain:

$$\psi = p_I U_C - \lambda \tag{48}$$

This necessary condition can be interpreted as MB equals MC where ψ (the shadow price of an additional unit of H) is the MB term. When I is positive and λ is zero, the right hand side of (48) is the MC of the investment necessary to yield an additional unit of H. When I=0, two things should be noted about the right-hand side of (48). One is that C will now be equal to income, and the other is that λ will be positive. The fact that λ must be subtracted from the RHS for the equality to hold means that even when I=0, the LHS, which is MB of an additional unit of H, is less than the MC. In other words, I=0 when the MB of another unit of health is so low that the first order condition cannot be satisfied with equality without subtracting the λ term from the RHS. Given the definition of ψ, there are two cases where this might arise. One at the beginning of the planning horizon, for someone who is born with what one might call perfect health, so that the benefit from investing in health is virtually zero, which is the case Galama and Kapteyn(2011) appear to have in mind. The other is the case where ψ is very low because the remaining time horizon is short so that the payoff period to investing in health is too short to make it worth doing. This would occur at the end of the planning horizon. Thus, it is quite possible that the optimal trajectory for an individual could look like Figure 5.

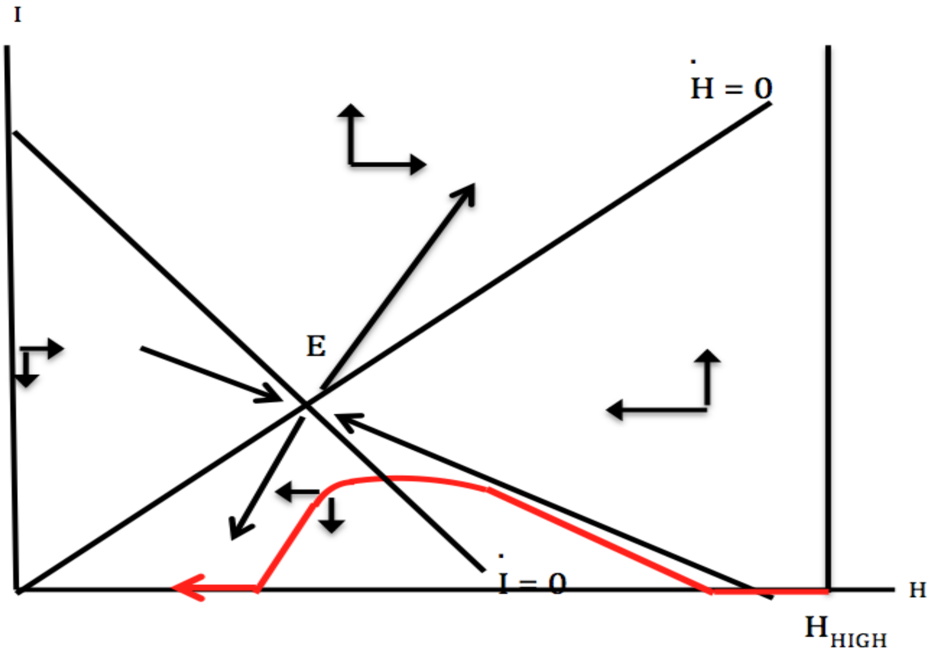

Figure 5: The health threshold effect

VI Other considerations on the form of the production function

Zweifel (2012) raises an additional criticism that the MGM assumes “a fixed ratio between individuals health care expenditure and the cost of their own health enhancing efforts regardless of their state of health” (pg. 677). This is a feature of the Grossman 1972 paper where he explicitly assumes a CRS production function for health, although Zweifel discusses it in the context of a Cobb-Douglas production function. In practice it would be a finding that was associated with any homothetic health production function that had more than one input. Zweifel’s criticism then is that the homotheticity of this isoquant map is unaffected by changes in the individual’s level of H. This is an empirical rather than a theoretical issue. CRS is a common simplifying assumption but is not a fundamental prediction of the MGM. It seems perfectly sensible that when the health production function has more than one input, their relative productivities will vary across isoquants.

There is another consideration that has received rather limited attention in the literature despite the fact that Grossman gave quite a detailed discussion of it in his *Journal of Political Economy* paper. While it relates to the nature of the production function and the productivity of I, it is not a matter of returns to scale in the traditional sense. When talking about returns to scale in a production function what is really referred to is an engineering, and not an economic issue. It deals with the technical point of the rate at which output can increase by increasing all inputs in the same proportion. Under CRS, if all inputs are doubled, output will double. The matter is probably not quite as clear-cut as that summary suggests: it is probably safe to assume that output will double if inputs are doubled by replication of existing production facilities. This is the argument that underlies the common assumption that the industry-level supply function is horizontal in the long run. On the other hand it is not always going to be the case that increasing all inputs by ten per cent will increase output by ten per cent - this is the engineering issue of scalability. Whatever conclusion is reached, however, the degree of returns to scale is a matter that relates to the impact of increases in inputs on output.

In the case of the production function for health capital, though, there is an additional consideration. This can best be summarized by referring to it as the issue of perfect health. Assume, as Grossman does, that there is some technologically or physiologically determined upper value of H, H*, which cannot be exceeded. Set aside the issue of whether H* has been increasing over the centuries or whether H* has always been fixed and society has been doing a better job of approaching it at both population and individual levels. Importantly for empirical implications, the question of the relation, if any, between H* and some maximum possible length of life can also be set aside. For purposes of the discussion here it can simply be assumed that at any time there exists a maximum possible H and that if that H* changes it does so so slowly that it is reasonable to treat it as exogenous as far as individuals in the population are concerned. As a first approximation, one could introduce the issue of perfect health by writing the health production function G as:

$$G(I)[1 - H/H^*], H \leq H^* \tag{49}$$

Here the marginal product of I in the production of H becomes

$$G_I(I)[1 - H/H^*] \tag{50}$$

so that as H approaches H* the marginal productivity of I declines at all levels of I. Again, this has nothing to do with returns to scale in the sense in which that term is used in microeconomics in general precisely because the effect being considered is independent of the level of I. It does, however, come into the individual's decision about the optimal level of I to choose. Under this assumption, (13) the necessary condition for I becomes:

$$\psi G_I[1 - H/H^*] = p_I U_C \tag{51}$$

and totally differentiating (51) gives:

$$\begin{aligned}[\psi G_{II}[1 - H/H^*] + p_I^2 U_{CC}]\partial I + G_I[1 - H/H^*]\partial\psi - [p_I U_{CH} + \psi G_I(I)/H^*]\partial H \\ - p_I U_{CC}\partial Y - [U_C - p_I U_{CC}]\partial p = 0\end{aligned} \tag{52}$$

Rearranging (52) gives:

$$\frac{\partial I}{\partial H} = \frac{\frac{G_I}{H^*}\psi + p_I U_{CH}}{[1 - \frac{H}{H^*}]\psi G_{II} + p_I^2 U_{CC}} \tag{53}$$

which is still negative but whose magnitude now depends on the size of H relative to H*. It is clear from (53) that the closer H is to H* the smaller the MB of another unit of I relative to its MC, regardless of the particular level of I: i.e. regardless of the assumptions made about the curvature of the G(I) term.

In talking about a perfect, or upper, level of health the focus is on the production function, not on the utility function. This does not imply a satiation level of H in the utility function and nor does it imply a value of the co-state ψ (that as H approaches H*) at which the MB of another unit of I falls. This does however, identify another factor that would lead to lower investment in I as H increases.

VII Death in the Grossman model

One key issue in the literature is whether the MGM would actually allow an individual to choose to live forever. If it does, this would have to be regarded as a major theoretical weakness since nobody as yet has successfully made that choice. This is an issue which goes all the way back to Grossman's original 1972 paper and which he said in Grossman (2000) had been of concern to him at the time. Grossman's solution in the 1972 paper was to have the rate of depreciation of health capital increase at an increasing rate so that at a certain point even with CRS in the production function, it became impossible to maintain H above H$_{MIN}$. As the MGM has been written here, δ has been constant. Grossman's approach is not at all unreasonable if it is assumed that the human body simply breaks down at some point at a rate faster than it is possible to repair it[15]. That would seem to suggest that technological improvement in the health production function would still open the possibility of immortality. However, to use the MGM as the basis for empirical investigation of human health behavior it does not seem unreasonable to do what has been done in this paper and assume the individual knows that their life is finite. In other words, to work with a fixed finite horizon form of the problem.

Zweifel (2012) has argued that the assumption that the length of life is known is unreasonable. It is certainly the case that people die at different ages. One possible approach is to define a maximum possible length of life and allow people to choose their length of life up to that upper limit. The transversality condition for an endogenous horizon is that the Hamiltonian equals zero at the optimal end of life. This can clearly be a result of different choices made by different individuals, conditional on their endowments. One can also consider a model in which there is an absolute upper limit to T and the actual age at death is stochastic with the probability of death being a function of the individual's stock of health capital. This moves into the realm of stochastic control theory and Ito's Lemma, techniques that have been little used in health economics despite the fact that health behaviours reflect decisions about inter-temporal optimization under uncertainty[16].

[15] For a different take on the issue of why death happens, see Robson and Hillard (2007).

[16] Cropper(1977) presents a version of the MGM with uncertainty in relation to health where the health shocks are characterized as being minor and having no effect on the stock of health capital (pg. 1277).

The key issue associated with the way to make life finite is the implication of the method chosen for end of life consumption of health care. This is a case however, where it is important to separate the effects of two assumptions, one pertaining to the finite nature of the horizon and the other to the implications of the assumption that is made about the rate of depreciation of health capital. The most obvious combination of these assumptions would appear to be the case where there is an upper bound to the length of life and the depreciation rate also increases at an increasing rate late in life.

From a purely technical point of view making δ increase with age, given that age increases at exactly the same rate as time passes, turns the optimal control problem from a one state variable to a two state variable problem. To this point phase diagram analysis has been used to illustrate the arguments being made about the workings of the MGM; unfortunately, it is extremely difficult, and in many cases impossible to draw a phase diagram for a two-state variable problem. The nearest one could come would be to use a two-stage optimal control problem in which delta was low during the first stage and high during the second stage. This would not result in a discrete jump in I at the point of passing from the first to the second stage since the transversality conditions for a two-stage problem would make such a jump sub-optimal[17]. In terms of the phase diagram it is clear that an increase in δ would rotate the stationary locus for H upward. That by itself would be relatively easy to illustrate. Unfortunately, a change in δ will also shift the stationary locus for I. The result of these two effects is that the overall effect on the optimal trajectory is ambiguous. The most productive approach to introducing age-dependent depreciation rates would appear to be theoretical simulation (see for example, Koka et al., 2014).

VIII Conclusion

This paper argued that the criticisms leveled in the literature at Grossman's 1972 model do not in fact constitute a serious indictment of its theoretical structure. It was shown that the MGM even under the CRS assumption does not suffer from an indeterminacy problem and does indeed take account of an individual's history. The paper underscores the importance of distinguishing between comparative static effects and dynamic effects. It can be seen that in empirical data, when looking at changes in I and changes H between two consecutive points in time, the correlation between changes in I and changes in H could be either positive or negative depending on where one is in the individual's lifetime trajectory.

In empirical work it is important to use estimating equations whose functional form takes explicit account of the intrinsic dynamics of I and H since failing to do so is likely to bias estimates of the comparative statics of the determinants of I. The key message of the Grossman framework is that since individuals are forward looking, it is important that researchers think in terms not of a point in time but rather in terms of an individual's optimal trajectory.

In short, we argue that the criticisms that have been directed at the Grossman model are themselves incorrect, and are the result of looking at a dynamic model through static eyes. It is the contention of this paper that when the techniques of dynamic economic analysis which are standard in other areas of economics, are applied to the Grossman model its status as the workhorse of modeling individual health related behaviours is well justified.

[17] Wagstaff (1993) attempted to implement a two-stage process empirically with less than satisfactory results.

References

Cropper, M.L. (1977) “Health, Investment in Health, and Occupational Choice”, *Journal of Political Economy* 85(6): 1273-1294.

Ehrlich, I., and H. Chuma (1990) “A model of the demand for longevity and the value of life extensions”, *Journal of Political Economy* 98(4): 761-782.

Ferguson, B.S. and G.C. Lim (1998) Introduction to dynamic economic models, Manchester University Press.

Galama, T.J. and A. Kapteyn, (2011) “Grossman’s missing health threshold,” *Journal of Health Economics* 30(5): 1044-1056.

Galama, T.J., P. Hullegie, E. Meijer, and S. Outcault (2012), “Is there empirical evidence for decreasing returns to scale in a health capital model?”, *Health Economics*, 21(9): 1080-1100.

Grossman, M. (1972) “On the concept of health capital and the demand for health”, *Journal of Political Economy* 80(2): 223-255.

Grossman, M. (2000) “The Human Capital Model” in Handbook of Health Economics, Volume 1, edited by A.J. Culyer, and J.P. Newhouse, Elsevier Science B.V.

Grossman, M. (1998) “On optimal length of life”, *Journal of Health Economics* 17(4): 499-509.

Jones, A., Laporte A., Rice N., and Zucchelli, E. (2014) “A synthesis of the Grossman and Becker-Murphy models of health and addiction: theoretical and empirical implications”, Centre for Health Economics, University of York Working paper.

Kaestner, R. (2013) “The Grossman model after 40 years: a reply to Peter Zweifel”, *The European Journal of Health Economics* 14(2): 357-360.

Koka, K., A. Laporte and B.S. Ferguson (2014) “Theoretical simulation in health economics: An application to Grossman’s model of investment in health capital” Canadian Centre for Health Economics, Working paper series, No. 2014-11, June.

Laporte, A. and B.S. Ferguson (2007) “Investment in health when health is stochastic” *Journal of Population Economics* 20(2): 423-444.

Leonard, D. and N. Van Long (1992) Optimal Control Theory and Static Optimization in Economics, Cambridge University Press.

Reid, W.(1998) “Comparative dynamic analysis of the full Grossman model” Journal of Health Economics 17(4):383-426.

Wagstaff (1993) “The Demand for Health: An Empirical Reformulation of the Grossman Model”, *Health Economics* 2(2):189-198.

Robson, Arthur J., and Hillard S. Kaplan. 2007. “Why do We Die? Economics, Biology and Aging.” *American Economic Review* 97(2): 492-495

Zweifel, P. (2012) “The Grossman model after 40 years” *The European Journal of Health Economics* 13(6): 677-682.

Zweifel, P. (2013) “The Grossman model after 40 years: response to Robert Kaestner” *The European Journal of Health Economics*14(2): 361-362.