Congestion in a Public Health Service: A Macro Approach*

Mark Kelly,†Michael Kuhn‡

November 2020

Abstract

To study the trade-offs and the macroeconomic repercussions of rising health care demand in a public health service, we develop a continuous-time overlapping generations model with a public health care sector and a realistic aging process. Health care services are provided free of charge at the point of service. Without a price mechanism, the government relies on a queuing rule for allocating its services. We conceptualize this mechanism as congestion that lowers the efficacy of health care. Then, we calibrate the model to match UK data from 2007-2016 and analyze the steady-state, general equilibrium response of the economy of shocks to productivity/income and medical effectiveness. Our analysis suggests that the optimal response to an increase in the demand for health care depends strongly on whether it is due to an increase in income or medical effectiveness. We also show that there is disagreement across age-groups on the preferred policy.

*Financial support for this research by the Austrian Science Fund under grant P 26814-G11 is gratefully acknowledged. Ivan Frankovic and Stefan Wrzaczek contributed greatly to the conception of this research. We are grateful for their input at that stage as well as for ongoing help. We would also like to express our appreciation for the comments provided by the participants and discussants at the 2019 meeting of the Field Committee on Health Economics of the German Economic Association, the 2019 Texas Health Economics Conference, and the 2020 Austrian Economic Association Meeting. All responsibility lies entirely with us.

†Baylor University, United States

‡Wittgenstein Centre (IIASA, OeAW, University of Vienna) and Vienna Institute of Demography, Austria.

1 Introduction

Between 1990-2014 the average annual growth rate of the output share of health care among OECD nations was approximately 1.34%. Moving forward, the rapid rise in the utilization of government-financed health care services threatens fiscal sustainability of many OECD countries. Consequently, most developed nations have pursued cost containment reforms that attempt to slow the growth rate of health care expenditures. Such policies can be broadly classified into one of two categories; price controls and resource rationing. Price controls exist in virtually all public insurance schemes and are implemented in order to directly influence the cost of financing health care by fixing the prices that suppliers can charge for the services they provide.

Resource rationing, on the other hand, is slightly less common and is most effective in health care systems where the government is a major provider of health care services. In this case, the government is attempting to contain the cost of health care indirectly by limiting its supply. There are many different varieties of resource rationing, many of which rely on waiting (see e.g. Siciliani et al. 2014 for a recent survey) as a more or less explicit rationing mechanism. More generally, waiting can be understood as a form of congestion. We should stress at this point that although we broadly frame our paper in the context of waiting, not the least because we use waiting time data for the calibration of our model, our general notion of congestion is broader and includes other more implicit forms of rationing, such as reductions in consultation times or reductions in the quality of care.

Specifically, our study focuses on a rationing regime whereby the policymaker decides on a given level of supply of health care services (as measured in total hours of treatment available) and allocates it, in the event that demand exceeds the fixed supply, via a waiting list, such that the effective utilization (hours demanded net of waiting time) equals the supply. By reducing the effectiveness of health care and increasing the time cost to the individual,

waiting serves as a mechanism to contain demand. By maintaining a larger capacity within the health care system, the policymaker can to some extent control waiting times. This is well-illustrated for the English National Health Service (NHS), where the development of waiting times in NHS hospitals over the time span 1999-2017 is following a trend that is broadly opposite to the trend in the NHS spending share (see figure 1).

[Insert Figure 1 here]

Considering the macroeconomic efficiency of the public provision of health care services subject to congestion, three key issues emerge: (i) Which mechanisms determine the allocation of health care and its outcomes through waiting and what are the macroeconomic repercussions which arise from funding and from health-related changes in longevity and labor supply? (ii) What constitutes an efficient level of public supply when (a) it determines waiting times and, thus, the effectiveness of medical care (both directly and indirectly through its effect on individual demand) and when (b) timely access to health care impacts health outcomes, in particular, longevity, the benefits of which trade-off against the resource costs of the health care sector, including distortions arising from its funding? (iii) What policy recommendations can be made with respect to the public provision of health services and, more specifically, what rules – targeting fiscal sustainability (i.e. maintaining a constant health expenditure share) as opposed to a maximum waiting time – are commendable, in the face of productivity growth and medical progress as two well-known drivers of the joint expansion of health care expenditure and longevity (Hall and Jones 2007, Jones 2016, Kelly 2017, Böhm et al. 2018, Fonseca et al. 2020, Frankovic et al. 2020a,b)?

In order to analyze these issues, we develop a theoretical model of a public health care system in which the government is the sole provider of health care services. We assume that the government supplies a predetermined level of medical capacity which is financed out of an earnings tax. We further assume that the economy is composed of a continuum of

finitely-lived individuals born into distinct birth cohorts. Each individual derives utility from a non-medical good and leisure time and accumulates health deficits, which can be reduced by the individual consumption of medical services. By slowing down the accumulation of health deficits the individual can expand its longevity in the spirit of Dalgaard and Strulik (2014, 2017). The effectiveness of health care depends on the extent of congestion which leads to waiting. Here, waiting times will fall whenever the supply of health care services grows in excess of the aggregate demand for them. The health care system is embedded in an economy that features a private and competitive final goods production sector besides the health care sector. Aside from choosing time-intensive health care, individuals make decisions on consumption and saving as well as on their retirement, the latter then determining the aggregate supply of labor.

We calibrate the model to match data from the UK from 2007-2016. The British health care system is run by the National Health Service (NHS) which functions as both the primary financier of health care and the primary provider of all medical services, and therefore fits our theoretical framework.[1] We then engage in the analysis of three numerical experiments: (i) a 10% expansion of NHS supply; (ii) a 10% increase in total factor productivity in the production of final goods, which is tantamount to an economic growth impulse; and (iii) a 10% increase in the effectiveness of health care in curbing the accumulation of health care deficits. For the latter two experiments we compare the outcomes under three different policies: (a) a status quo policy, in which the NHS supply is held constant; (b) a policy aimed at maintaining the NHS expenditure share as a fiscal target; and (c) a policy aimed at meeting a waiting list target.

Our key results suggest that for our calibration, a 10% expansion of the NHS is improving instantaneous welfare (across all cohorts) even if this comes at the cost of a lower

[1]The NHS accounted for approximately 86.11% of all medical consumption in the UK from 1994-2013. Households can buy supplemental private health insurance that can be used to bypass the queuing process for certain procedures that are performed in private hospitals and private units of NHS hospitals.

per capita income and consumption. Notably, the expansion of supply leads to a reduction in waiting times despite the increase in demand both at the intensive margin, reflecting an increase in individual demand from each cohort, and at the extensive margin, reflecting the demand from cohorts who now survive to an older age. Thus, the supply expansion has allowed both for greater consumption of health care and, through the reduction in waiting/congestion, has rendered health care more effective. This translates into a sizeable slow down in deficit accumulation and, thus, to an increase in longevity. At the same time, our analysis shows that both the health share and the health care tax increase substantially, with the latter leading to a reduction in labor supply. Overall, the increase in longevity raises instantaneous welfare when measured across all cohorts, but from a cohort perspective the welfare gain is highly skewed towards the elderly. With the tax burden and loss in lifetime consumption predominantly affecting the young. Individuals below age 40 tend to suffer from a reduction in their life-cycle utility, implying that the capacity expansion fails to satisfy Pareto optimality.

Unsurprisingly both productivity growth and medical progress contribute to an increase in welfare and in this case both from an instantaneous cross-cohort perspective as well as from a life-cycle utility perspective. However, the overall impact of these shocks depends on the target that governs the policy response. We find that aiming for the fiscal target tends to boost the instantaneous cross-cohort welfare gain in the presence of productivity growth, whereas aiming for a waiting list target tends to boost the instantaneous cross-cohort welfare gain in the presence of medical progress. While individuals from all age groups benefit from productivity growth and medical progress regardless of the accompanying policy response, there is again no unanimous agreement on the preferred response. Both the pursuit of a fiscal target in the presence of a productivity shock and the pursuit of a waiting list target in the presence of medical progress imply the most expansionary policy with respect to NHS capacity and the associated tax. While these policies are maximizing cross-cohort welfare as

well as the life-cycle utility of the elderly, again the youngest cohorts would prefer the least expansionary policy, i.e. a constant capacity.

While shadowing a microeconomic literature on waiting times in the public provision of health care (see Siciliani and Iversen 2012 for a recent survey) the domain of our paper lies more with the macroeconomic modeling of health care. As such it is most closely related to Gaudette (2014) who consider a similar macroeconomic set-up of overlapping generations of individuals who are consuming public health care in order to improve health and survival over their life-cycle while being subject to a waiting time price. While Gaudette (2014) also studies the role of various payment and tax policies aimed at internalizing the waiting time externalities and optimizing welfare, waiting is depicted in a very abstract way as a parameter that raises a patient's time cost in a way that equalizes the aggregate cost of private health care provision with the public health care budget. Furthermore, his model does not feature an explicit production function for either of the two sectors (i.e. final goods and health care). Altogether, this rules out an analysis of how changes to the (physical) supply of public health care interact with the waiting time and, thus, the (physical) demand for health care, which our analysis shows to be crucial for understanding the macroeconomic effects. Furthermore, the lack of an explicit modeling of production and factor markets does not allow for the analysis of general equilibrium repercussions, which again we show to be of relevance.

A second paper that is relatively closely in line with our study, if only as it is also based on a calibration for the English NHS, is Böhm et al. (2018). This study features a general equilibrium model with overlapping generations of individuals who are subject to deficit accumulation and examines how the rationing of public health care bears on welfare if public health care investments induce medical innovations. While our model lacks much of the dynamics present in Böhm et al. (2018) it provides a more thorough modeling of waiting/congestion, which is not an issue in Böhm et al. (2018). Other papers featuring

a public health care sector are Kuhn and Prettner (2016) who study the role of publicly provided health care in the presence of R&D-driven economic growth and Grossmann and Strulik (2019) who employ a model of deficit accumulation to study the role of social security reform in Germany.

More distantly, our paper ties in with a large literature centering on the role of health care reform and/or medical progress in calibrated macroeconomic models of the US economy (e.g. Zhao 2014; Jung and Tran 2016; Kelly 2017, 2020; Conesa et al 2018; Frankovic and Kuhn 2019; Frankovic et al. 2020b), the big differences being that health care rationing in that context is through price rather than waiting times, implying also that the size of the health care sector is determined by market forces, with only indirect scope for policy making.

The rest of the paper is laid out as follows. The model and its solution is detailed in section 2. Section 3 describes the data and calibration procedure. Section 4 contains the numerical analysis covering three sets of numerical experiments. Section 5 concludes.

2 The Model

2.1 Individuals

We assume that NHS services are free to the patient at the point of delivery. Consequently, the NHS will have to rely on queuing to allocate its services, implying that patients will have to spend some time on a waiting list before they become eligible to consume health care services. Let $\omega(z,t)$ and $h(z,t)$ be the amount of time an individual aged z at time t devotes to waiting for and consuming NHS services, respectively.[2]Further, we can define $m(z,t)$, the individual's current demand for health care measured in units of time, as the sum of $\omega(z,t)$ and $h(z,t)$, so that $m(z,t) = \omega(z,t) + h(z,t)$. Defining $\widehat{\omega}(t)$ as the share of total health care time $m(z,t)$ that the individual spends waiting on average, and assum-

[2]Note the implied relationship to the birth year $t_0 = t - z$.

ing this share to be uniform across the economy, we have that $\omega(z,t) = \widehat{\omega}(t)m(z,t)$ and $h(z,t) = (1-\widehat{\omega}(t))m(z,t)$.

Following Dalgaard and Strulik (2014, 2017), we model the aging process as the gradual accumulation of adverse health conditions. We will refer to these conditions as "deficits." Each deficit diminishes bodily function, ultimately resulting in death once the individual's total number of accumulated deficits reaches some maximum survivable threshold $\bar{D}$. Let $D(z,t)$ be the individual's total number of accumulated health deficits at age z and time t.The deficit accumulation function is defined as

$$\dot{D}(z,t) = \mu(D(z,t) - a - f(m(z,t)). \tag{1}$$

Deficits accumulate at the natural rate μ. We allow for the existence of exogenous environmental factors that reduce (or increase) the deficit accumulation rate. These factors are captured by the parameter a. The effect of the individual's health investment on the deficit accumulation rate is described by the function $f(m(z,t))$. Finally, we denote by T the individual's longevity, as defined by the identity $D(T,t) = \bar{D}$.

The health investment function takes the following functional form

$$f(m(z,t)) = A\left[\left(\frac{H(t)}{M(t)}\right)^{\epsilon} m(z,t)\right]^{\gamma}, \quad 0 \leq \epsilon, \gamma \leq 1 \tag{2}$$

where $M(t)$ is effective aggregate demand for health care (i.e. $M(t)$ is the sum of $m(z,t)$) and $H(t)$ is the exogenous supply of NHS services.[3] By definition, total consumption of NHS services is constrained by $H(t)$, implying that $H(t) \leq M(t)$ with $H(t)$ equalling the sum of

[3]In practical terms, $H(t)$ can be thought of as the total supply of "bed hours" provided by the NHS, while $M(t)$ is the sum of aggregate "bed hours" (assuming NHS services are used at full capacity) and aggregate waiting time. As we detail in section 3, we calibrate the model to match the steady-state ratio $H(t)/M(t)$ to the observed ratio of the average length of stay in the UK to the sum of the average length of stay and the average wait time in the UK using data obtained from the NHS hospital episode statistics.

$h(z,t)$ across all individuals in the economy.[4] Whenever aggregate effective demand exceeds NHS supply, the average waiting time share $\widehat{\omega}(t)$, which is defined as

$$\widehat{\omega}\left(t\right)=\max\left\{0,1-\frac{H\left(t\right)}{M\left(t\right)}\right\}$$

will be positive and will increase whenever the NHS becomes more congested (i.e. $M(t)$ rises relative to $H(t)$).

Equation (2) implies that the effectiveness of health care is negatively correlated with the level of congestion in the NHS. This assumption is consistent with the fact that for many life threatening conditions, time to treatment is a significant determinant of survival, with survival probabilities declining over time (e.g. Rexius et al. 2004, Sobolev et al. 2006, Doubeni et al. 2018, Hanna et al. 2020). The parameter ϵ governs the returns to timely treatment. When $\epsilon=1$, the effectiveness of health care is directly proportional to the current level of congestion in the NHS. At the other extreme, when $\epsilon=0$, the efficacy of health care is unaffected by NHS congestion.[5]

Before moving on, we briefly digress to consider two important points about how we have chosen to model waiting in our model: (i) While our modeling of $\widehat{\omega}\left(t\right)$ is internally consistent and while it can be calibrated to the data, it constitutes a reduced form of a full model of waiting time dynamics, such as presented e.g. in Siciliani (2008). A more accurate modeling of the waiting time dynamics would add a second and delayed dynamic process in our model, where individuals demand health care at time t but receive it only after some time $t+\omega(z,t)$. We circumvent the complexity involved by conflating a whole treatment episode, as measured by $m(z,t)$, into a single period t, a year say, such that each individual is treated within t but assuming that (unmodelled) waiting reduces the effectiveness of $m(z,t)$.

[4]Indeed, we will show below that $H(t)<M(t)$ must hold in equilibrium. See footnote 10 below.

[5]Note that the effectiveness of each unit of $m(z,t)$ can be expressed as $(H(t)/M(t))^{\epsilon}=(1-\hat{\omega})^{\epsilon}$. For $0<\epsilon<1$, effectiveness is thus decreasing in a convex way with the average waiting time share.

We contend that this is legitimate given that our main concern is with the medical loss of effectiveness and the additional time cost from waiting. (ii) Recalling our broader notion of congestion, we can also understand $\widehat{\omega}(t)$ to reflect a general loss in the effectiveness of any individual effort in accessing the health care system. Such a loss in effectiveness may not only arise from waiting but also from reductions in the quality of care or from extra time costs involved (e.g. the patient's referral to more distant providers with spare capacity).[6]

We assume that individuals receive utility from consumption and disutility from work and waiting for NHS services. The instantaneous utility function for an age z individual at time t is

$$u(c(z,t), m(z,t), l(z,t)) = \frac{c(z,t)^{1-\sigma}}{1-\sigma} - \theta\widehat{\omega}(t)\, m(z,t) - \eta l(z,t), \tag{3}$$

where $c(z,t)$ is current consumption and $1/\sigma$ is the intertemporal elasticity of substitution. The parameter θ is the disutility weight from individual waiting. This disutility may arise from the prolonged pain, suffering, and anxiety incurred by an individual due to unresolved health issues while waiting for NHS services.[7] Labor (i.e. $l(z,t)$) generates disutility as measured by the parameter η. We assume that $l(z,t)$ is a binary variable equaling one when the individual is in the labor force and zero when they exit. Assuming that $R(t)$ is the

[6]Focusing only on elective treatments, there is ample evidence for cardiac treatments that waiting increases pre-treatment mortality (e.g. Rexius et al. 2004, Sobolev et al. 2006) while evidence that waiting also increases in-hospital mortality and other outcomes is more mixed (e.g. Rexius et al. 2005, Sobolev et al. 2008, 2012, Moscelli et al. 2016). Recent evidence suggests that (excessive) waiting is leading to worsening outcomes in cancer treatments (e.g. Doubeni et al. 2018, Hanna et al. 2020). Some (milder) health loss is also associated with waiting for hip and knee replacement surgery (Nikolova et al. 2016). Finally, there is a large body of evidence that hospital capacity strain is associated with higher mortality and poorer treatment outcomes (e.g. Schilling et al. 2010, Eriksson et al. 2016).

[7]Indeed, there is a considerable body of evidence that waiting lowers patients' quality of life and psychological well being (e.g. Sampalis et al. 2001, Oudhoff et al. 2007, Sutherland et al. 2016).

individual's optimal retirement age,[8] we have

$$l(z,t) = \begin{cases} 1 & \text{if } z \leq R(t) \\ 0 & \text{if } z > R(t) \end{cases}. \tag{4}$$

Individuals consume and save out of their asset income and after-tax earnings. Assets are held in the form of physical capital $k(z,t)$ and accumulate according to

$$\dot{k}(z,t) = (1-\tau^k(t))r(t)\,k(z,t) + (1-\tau^l(t))w(t)\,l(z,t) - (1+\tau^c(t))c(z,t) + s(z,t) \tag{5}$$

where $r(t)$ is the real interest rate and $\tau^k(t)$, $\tau^l(t)$, and $\tau^c(t)$ are the capital, labor, and consumption tax rates, respectively. Working individuals (i.e. $l(z,t) = 1$) are paid a wage $w(t)$. We model a simple public pension system that provides a benefit $s(z,t)$ when the individual reaches the statutory eligibility age $Q(t)$ so that

$$s(z,t) = \begin{cases} 0 & \text{if } z \leq Q(t) \\ \kappa(t)w(t) & \text{if } z > Q(t) \end{cases}.$$

where $\kappa(t)$ is the replacement rate.

Individuals are, thus, assumed to maximize lifetime utility

$$V(0,t) = \int_0^{T(t)} e^{-\rho z} u(c(z,t), m(z,t), l(z,t))dz \tag{6}$$

with ρ denoting the rate of time preference, by choosing $c(z,t)$, $m(z,t)$, and $R(t)$ subject to equations (1) and (5). The Hamiltonian function (dropping the age-time indicators z and t

[8]Strictly speaking $R(t)$ amounts to the optimal retirement age of an individual belonging to the birth cohort $t - R(t)$.

for brevity) that characterizes the individual's optimal decision problem is given as

$$\begin{aligned}\mathcal{H} &= e^{-\rho z}\left\{\begin{array}{c} u(c,m,l)+\lambda^D\mu\left(D-a-A\left(\frac{H}{M}\right)^{\gamma\epsilon}m^{\gamma}\right) \\ +\lambda^k[(1-\tau^k)rk+(1-\tau^l)wl-(1+\tau^c)c+s]\end{array}\right\} \\ \text{s.t.} &\quad D(0,t)=D_0,\ D(T,t)=\bar{D},\ k(0,t)=k(T,t)=0,\end{aligned} \tag{7}$$

where λ^D and λ^k are the costate variables on the stock of deficits and capital, respectively. The resulting first-order conditions are[9]

$$c^{-\sigma}=\lambda^k(1+\tau^c) \tag{8}$$

$$-\lambda^D\gamma\mu A\left(\frac{H}{M}\right)^{\gamma\epsilon}m^{\gamma-1}=\theta\widehat{\omega} \tag{9}$$

$$\frac{(1-\tau^l)w}{(1+\tau^c)c(R)^{\sigma}}=\eta \tag{10}$$

$$u(T)=\frac{\theta\widehat{\omega}m(T)^{1-\gamma}}{\gamma A}\left(\frac{H}{M}\right)^{-\gamma\epsilon}\left(\bar{D}-a\right)-\frac{\theta\widehat{\omega}m(T)}{\gamma}+\frac{(1+\tau^c)c(T)-s}{(1+\tau^c)c(T)^{\sigma}} \tag{11}$$

$$-\dot{\lambda}^k/\lambda^k=r-(\delta+\rho) \tag{12}$$

$$-\dot{\lambda}^D/\lambda^D=\mu-\rho. \tag{13}$$

Equation (8) is the standard outcome, equating the marginal utility of consumption to the shadow value of financial wealth, weighted by the after-tax cost of a unit of consumption.

[9]Note that the second-order conditions can be verified numerically.

Equation (9) describes the individual's optimal demand for health care. This condition implies that the optimal allocation of time to health (i.e. total time waiting for and receiving NHS services) at age z is set so that the marginal product of health care is equal to the marginal disutility from waiting.[10] Since we are studying congestion (as measured by waiting time) in a public health care system, we have chosen to abstract away from a monetary cost of health care to the family in favor of a time cost. Moreover, given that we are calibrating our model to match UK data, this assumption is consistent with how health care is delivered in the UK, with NHS services provided for free at the point of service and the public sector accounting for approximately 86% of all health care expenditures. Under the NHS's queuing rules, waiting is the only cost imposed on patients. Equation (9) captures these costs in the two ways discussed previously: (i) direct utility loss from waiting and (ii) loss in the effectiveness of health care.

Each worker will engage in the labor market as long as the marginal benefit of working exceeds the disutility of labor, implying that the endogenous retirement age R coincides with the age at which the after-tax earnings, weighted by the marginal utility of consumption, is equal to the marginal disutility from labor (see equation (10)). Note that, unlike Dalgaard and Strulik (2017), neither the individual's health investments, nor their current number of accumulated deficits factor directly into their retirement decision. Nevertheless, as our analysis will demonstrate, in equilibrium health will indirectly impact the retirement decision by influencing λ^k through the individual's consumption and savings decision. Additionally, in order to simplify the modeling of the public pension system and the endogenous choice of R, we treat Q as exogenous so that the public pension does not impact the individual's retirement decision.

[10]It is easy to infer from the first order condition (9) that the equilibrium allocation will necessarily involve some waiting, i.e. $\hat{\omega}(t) > 0$ and, thus, $H(t) < M(t)$, where $H(t)$ is a given capacity and $M(t)$ is the aggregate demand. Suppose by contradiction that $\hat{\omega}(t) = 0$, which in equation (9) implies a zero marginal cost for the utilization of health care. Accordingly, individuals would then increase $m(t)$ and only stop if $\hat{\omega}(t) > 0$ (and sufficiently large). But then, for a given $H(t)$, we must have $H(t) < M(t)$.

Finally, equation (11) determines optimal longevity. This condition is derived by noting that the Hamiltonian function will be equal to zero in the terminal period. Setting $\mathcal{H}(T) = 0$, substituting for $\lambda^k(T)$ and $\lambda^D(t)$ using equations (8) and (9), and rearranging terms yields equation (11) and allows us to solve for T, using the following condition:

$$u(T) = -\lambda^D(T)\mu\left(\bar{D} - a - A\left[\left(\frac{H}{M}\right)^{\epsilon} m(T)\right]^{\gamma}\right) + \lambda^k(T)[(1+\tau^c)c(t) - s].$$

This condition implies that individuals prefer to live up to the point that their instantaneous utility at age T just equals the utility cost of facing a further incremental accumulation of deficits and of the net loss in wealth due to a further year's worth of consumption spending. In other words, by age T, utility no longer compensates for the cost of sustaining further survival (in terms of wealth and deficits).

The Euler equation for consumption is obtained by differentiating equation (8) with respect to z and substituting for the dynamics for λ^k from (12). Rearranging terms yields

$$\frac{\dot{c}}{c} \equiv g_c = \frac{r - (\delta + \rho)}{\sigma}. \tag{14}$$

Similarly, the Euler equation for health investments is derived by differentiating (9) with respect to z and substituting for $\dot{\lambda}^k$ and $\dot{\lambda}^D$ from (12) and (13):

$$\frac{\dot{m}}{m} \equiv g_m = \frac{\rho - \mu}{1 - \gamma}. \tag{15}$$

This equation implies that the lifetime path of m is positively correlated with the rate of time preference, ρ.[11] Note that as $\rho \to 0$, g_m becomes negative, implying that the optimal health investment strategy for perfectly patient individuals is to invest more heavily in their

[11]Note here the difference to the dynamics in Dalgaard and Strulik (2014, 2017), where the interest rate, r, shows up instead of the rate of time preference, ρ. This difference follows, as in our model, the consumption of health care merely takes up time and time has a pure utility value.

health when they are young and relatively healthy, as opposed to deferring the cost of health investment (in terms of waiting) until later in life. For $\rho > \mu$, $g_m > 0$, the individual is sufficiently impatient and their optimal strategy flips, with the individual opting to push the undesirable cost of queuing off until late in life when their health is poorer. Similarly, the growth rate of health investment is inversely affected by the degree of diminishing returns γ. If diminishing returns set in slowly (i.e. γ is close to one) and $\rho > \mu$, then the individual's optimal g_m will be relatively large. Put differently, if the rate of diminishing returns to health investment is small and people are relatively impatient, then the individual can afford to delay investing heavily in their health until late in life so that initial health investment will be relatively low and then will increase rapidly throughout the individual's lifetime. In our calibration, $g_m > 0$, implying that individual demand for NHS services will increase with age (see figure 2 below in section 3).

Finally, the solution to the individual's problem requires solving for $c(0)$, $m(0)$, R, and T using the following system of equations: (i) the intertemporal budget constraint, which is derived from equation (5), (ii) the intertemporal deficit constraint that is obtained by integrating equation (1) forward, (iii) the optimal retirement condition described by equation (10), and (iv) $\mathcal{H}(T) = 0$ from equation (11). We cannot derive an analytical solution for the individual's problem. Instead, we solve for $c(0)$, $m(0)$, R, and T numerically in general equilibrium.

2.2 Production

The economy consists of two sectors: a private final goods sector, in which firms produce under perfect competition; and a public health care sector where total output is decided by the government. Both sectors employ labor and capital from competitive markets. Turning to the final goods sector more specifically, we assume that there is a single representative

firm. The final good $F(t)$ is produced according to a Cobb-Douglas technology

$$F(t) = Z(t) K_F(t)^{\alpha} N_F(t)^{1-\alpha},$$

where $Z(t)$ is total factor productivity, where $K_F(t)$ and $N_F(t)$, respectively, are the capital stock and labor employed in the final goods sector, and where α is the output elasticity with respect to capital.

The profit function for the final goods producer is

$$\pi(t) = F(t) - w(t)N_F(t) - (r(t) + \delta)K_F(t).$$

For simplicity, we assume that there is a single competitive market for homogeneous labor. Consequently, the representative final goods firm and the NHS will pay workers the same market clearing wage rate $w(t)$. Likewise, we assume that capital markets are perfect and that capital can be costlessly moved between sectors. The representative firm and the NHS borrow for investment at the market clearing interest rate $r(t)$; and we further assume them to directly pay for depreciated capital, where δ is the depreciation rate of physical capital. Since we are assuming that the representative final goods firm operates under perfect competition, the real interest rate and the wage rate will be equal to the marginal product of capital (minus δ) and labor respectively:[12]

$$r(t) = \alpha Z(t) \left[\frac{K_F(t)}{N_F(t)}\right]^{\alpha-1} - \delta,$$

$$w(t) = (1-\alpha) Z(t) \left[\frac{K_F(t)}{N_F(t)}\right]^{\alpha}.$$

Let $H(t)$ be the aggregate supply of NHS services. The NHS production function takes

[12]Note that in the long-run the economy will converge to a steady-state with a constant capital per worker ratio in the final goods sector.

the CES functional form[13]

$$H\left(t\right)=B\left(t\right)\left[\beta K_H(t)^{\xi}+(1-\beta)N_H\left(t\right)^{\xi}\right]^{1/\xi},\quad -\infty\leq\xi\leq 1.$$

The NHS employs two inputs; health care labor, $N_H\left(t\right)$, and capital, $K_H(t)$. Total factor productivity in health care is represented by $B\left(t\right)$, while β is the capital share in health care output. We assume that the government contracts with the NHS to produce health care services. The NHS produces the mandated level of services $\bar{H}(t)$ and charges the government a fee $p(t)$. The NHS is required to operate as a non-profit institution and therefore chooses $K_H(t)$ and $N_H(t)$ in order to minimize the cost of producing $\bar{H}(t)$. The zero-profit condition that characterizes the NHS' objective is

$$\min_{K_H(t),N_H(t)}\quad p(t)H(t)=(r(t)+\delta)K_H(t)+w(t)N_H(t)\quad \text{s.t. } H(t)=\bar{H}(t),$$

Rearranging the NHS' zero-profit condition to solve for $p(t)$ and substituting $\bar{H}(t)$ for $H(t)$ yields

$$p(t)=\frac{(r(t)+\delta)K_H(t)+w(t)N_H(t)}{\bar{H}(t)},$$

i.e. a reimbursement of NHS services according to their average cost. Taking $p(t)$ as given, cost minimization implies that the NHS chooses capital according to the following rule

$$p(t)\frac{\partial H(t)}{\partial K_H(t)}=r(t)+\delta. \tag{16}$$

Similarly, in minimizing costs, the NHS chooses $N_H(t)$ so that the marginal product of labor

[13]The decision to employ different functional forms for the two sectors was made to improve the model's fit with respect to relative employment between the two sectors and the stability of the model.

in the health care sector is equal to the equilibrium wage rate

$$p(t)\frac{\partial H(t)}{\partial N_H(t)} = w(t). \tag{17}$$

2.3 Aggregation and Market Clearance

Aggregate capital accumulation is obtained by summing up the individual flow budget constraints, described by equation (5), across all living cohorts. This summation yields

$$\dot{K}(t) = (1-\tau^k)r(t)K(t) + (1-\tau^l)w(t)N(t) - (1+\tau^c)C(t) + S(t). \tag{18}$$

where $K(t)$ is aggregate wealth (physical capital), $N(t)$ is aggregate labor supply, $C(t)$ is aggregate consumption, and $S(t)$ are aggregate pension payments. It should be noted that we are assuming that $K_Y(t)$ and $K_H(t)$ depreciate at the same rate δ.

For simplicity, we assume a stationary population and normalize the size of each birth cohort to one. The size of the population is thus given by $\int_0^{T(t)} dz = T(t)$, aggregate labor supply follows as $N(t) = \int_0^{R(t)} dz = R(t)$, while aggregate consumption is computed as

$$C(t) = c(0,t)\int_0^T e^{g_c z} dz,$$

where $c(0,t)$ denotes the consumption of the cohort born at t and where g_c corresponds to consumption growth according to the Euler equation (14).

We restrict the government to a balanced budget rule. This implies that total government spending on public consumption $G(t)$, NHS expenditures $p(t)\bar{H}$, and pensions $S(t)$ must equal total tax revenue

$$\tau^k r(t)K(t) + \tau^l w(t)N(t) + \tau^c C(t) = G(t) + p(t)\bar{H} + S(t).$$

For simplicity, we assume that public consumption is equal to a fixed fraction ν of current final goods output, so that $G(t) = \nu F(t)$. Assuming a constant replacement rate $\kappa(t) = \kappa$, in equilibrium aggregate pension payments will be equal to $S(t) = \kappa w(t)(T(t) - Q(t))$, where $T(t) - Q(t)$ is the total population of individuals that are receiving public pension payments.

We can simplify (18) by using the government's budget constraint and noting that in a competitive equilibrium with a neoclassical production function we have $F(t) = (r(t) + \delta)K_F(t) + w(t)N_F(t)$. Inserting this expression into (18) reduces the aggregate capital accumulation function to

$$\dot{K}(t) = (1 - \nu)F(t) - C(t) - \delta K(t). \tag{19}$$

In a steady-state, it holds that $\dot{K}(t) = 0$. From equation (19), we then obtain the goods market clearing condition

$$(1 - \nu)F(t) = C(t) + \delta K(t)$$

Noting that aggregate GDP $Y(t)$ is equal to

$$Y(t) = F(t) + p(t)H(t)$$

and that labor and capital market clearance requires

$$N(t) = R(t) = N_H(t) + N_F(t)$$

and

$$K(t) = K_H(t) + K_F(t)$$

completes the equilibrium description of our economy.

3 Data and Calibration

The model is calibrated to match UK data over the time frame 2007 to 2016. Calibration of the model requires that we employ both aggregate and individual level data. Our data comes primarily from the UK Office for National Statistics (ONS). ONS provides data on GDP (and its components), NHS expenditures, total compensation, total population and employment, and the national life tables.

Using the national life tables, we compute the average life expectancy at age 20 in the UK, which is 61.06 years for the time frame we consider. Likewise, we rely on OECD estimates of the average effective age at retirement, which for males during this period was 63.99 years.

We calibrate the wage rate to match the average annual labor income per worker, which was £28,359.20 on average between 2007 and 2016. At the individual level, we seek to calibrate the model so that average consumption and income per person in equilibrium are close to the observed level of consumption and GDP per capita from the data, which were £16,373.20 and £26,160.20 respectively. This implies an aggregate consumption share of 62.61%.[14] At the same time, NHS expenditures accounted for 8.75% of total output. All other forms of public consumption and investment accounted for 13.45% of GDP. Aggregate employment is obtained from the Labour Force Survey (LFS), while NHS employment is taken from the public sector employment time series (PSE). Between 2007 and 2016, the NHS accounted for an average of 4.35% of total employment.

In calibrating the average wait time and utilization of NHS services we rely on the Hospital Episode Statistics (HES) database published by the NHS. This database provides detailed information on all admissions to NHS hospitals in England, including aggregate estimates of the average wait time and length of stay per episode at NHS hospitals. We

[14] All output shares are in nominal terms.

use the average length of stay as our measure of the average provision of NHS services per person (H/T in the model). This equates to approximately 1.5 bed days per person and year (equivalent to 0.4% of the individual's time endowment). Average wait time per episode was 52.79 days, implying that the aggregate average waiting time was 14.89 days (4.02% of the time endowment).

Table 1: Deficit Accumulation Parameters

Description	Notation	Value	Source
Health investment elasticity of health care	γ	0.19	Dalgaard and Strulik (2014)
Force of aging	μ	0.043	Mitnitski et al. (2002)
Health deficits at age 20	$D(0)$	0.027	Dalgaard and Strulik (2014)
Maximum health deficits	$D(T)$	0.1005	Dalgaard and Strulik (2014)

The majority of the parameters from the deficit accumulation function (1) are taken from Dalgaard and Strulik (2014), who base their calibration of γ, μ, D_0, and $\overline{D}$ on the work of the gerontologists Arnold Mitnitski and Kenneth Rockwood. The values for these parameters, along with their description and notation, are listed in table 1. The natural rate of aging μ is taken from Mitnitski et al. (2002) and is set at 0.043. Furthermore, Dalgaard and Strulik (2014) rely on Mitnitski et al.'s (2002) analysis to estimate the initial and terminal deficit stocks. Based on this, D_0 and $\overline{D}$ are set at 0.027 and 0.1005. Finally, they set the curvature parameter γ to 0.19 in order to calibrate the lifetime growth path of health care in their model to match the observed 2.1% growth rate from the data.

Table 2: Fixed Parameters (Calibration)

Description	Notation	Value
Rate of time preference	ρ	0.05
Intertemporal elasticity of substitution	$1/\sigma$	1
Disutility from waiting time	θ	2.25
Disutility from labor	η	0.975
Medical effectiveness	A	0.0115
Returns to timely treatment	ϵ	0.5
Environmental parameter	a	0.0163
Aggregate supply of NHS services	$\bar{H}$	0.25
Productivity parameter (NHS)	B	0.0008
Capital share (NHS)	β	0.2
EOS between capital and labor (NHS)	$1/(1-\xi)$	1.163
Productivity parameter (final goods)	Z	1246.75
Capital share (final goods)	α	0.3
Depreciation rate of capital	δ	0.04
Government expenditure share of final goods	ν	0.147
Replacement rate	κ	0.246
Growth rate of consumption	g_c	0.0015
Growth rate of demand for NHS services	g_m	0.0086

The remainder of the parameters are listed in table 2 above. These parameters are set to calibrate the equilibrium to match the sample averages from the data. The rate of time preference (ρ) and the intertemporal elasticity of substitution ($1/\sigma$) are chosen to calibrate private consumption. We set $\rho = 0.05$, close to Dalgaard and Strulik's (2014) choice of 0.06, and we follow Dalgaard and Strulik (2014, 2017) in setting $\sigma = 1$, implying that we are assuming log utility for consumption. This choice is consistent with Chetty (2006), whose meta-analysis of numerous labor supply studies supports the conclusion that the coefficient of relative risk aversion is approximately equal to one.

The disutility parameters θ and η are set at 2.25 and 0.975 respectively. These values were chosen in order to match the steady-state average waiting time and retirement age in the model to the observed per capita wait time and the average age at retirement in the data. The aggregate supply of NHS services (H) directly affects both the average wait time and

the average utilization of NHS services in the economy. Since we assume that the size of new birth cohorts is constant and normalized to one, the terminal age T also represents the total population of the economy. And, given that we match T to the observed life expectancy at age 20 (i.e. 61.06) and $\bar{h} = H/T = 0.4\%$ by definition, we set $\bar{H} = 0.25$. Following Acemoglu and Guerrieri (2008), the capital share in the health care sector β is set at 0.2. We assume a value of 0.5 for ϵ, which captures the returns to timely treatment. Taking $\bar{H}$, β, and ϵ as given, we choose B and ξ to calibrate the relative supply of NHS workers and the health care output share (pH/Y) to the data.

Following the literature, we set the capital share in the final goods sector $\alpha = 0.3$. Final goods productivity Z is chosen to calibrate the real wage rate and GDP per capita. We fix the depreciation rate of capital δ at 0.04, a standard rate in the literature. The government expenditure to final goods output share ν equals 0.147 in order to match the observed public consumption output ratio of 13.5%.

We set κ, the public pension replacement rate to 0.246 to match the observed public pension-output ratio we obtained from the OECD. According to the OECD, in 2014 (the only year the data is available) the replacement rate for workers earning at the average was 0.216, while workers earning 50% of the average wage had a replacement rate of 0.433. Thus, our calibrated replacement rate fits within this interval and is close to the replacement rate of average wage earners. We use the current statutory retirement age of 65 for males for $Q(t)$.

Finally, in the steady-state the lifetime growth rates of consumption and demand for NHS services are 0.15% and 0.86% respectively. Thus, as figure 2 shows, both c and m will grow slightly throughout the representative individual's lifetime. Since $g_m > 0$, the cost of providing NHS services to an individual will rise throughout the individual's lifetime.

[Insert Figure 2 here]

The calibrated benchmark steady-state is compared with the sample averages from the data in table 3. As table 3 demonstrates, the model closely matches the data for most variables of interest. The average utilization of NHS services (h) and wait time ($\hat{\omega}m$) in the model equal their counterparts in the data. Likewise, the NHS employment and expenditure shares are close matches with the data. We successfully calibrate the wage rate to the data, but the model over-predicts consumption and GDP per capita. The steady-state retirement and terminal ages are equivalent to the sample averages.

Table 3: Model vs. Data

	Data	Model
y	£26,160.20	£30,125.00
c	£16,373.20	£20,079.50
w	£28,359.20	£28,220.30
h	0.40%	0.41%
$\hat{\omega}m$	4.02%	4.04%
C/Y	62.61%	66.65%
pH/Y	8.75%	8.27%
S/Y	6.05%	6.06%
N_H/N	4.35%	4.88%
R	43.99	44.01
T	61.06	61.07
τ^l	25.95%	17.80%
τ^k	28.70%	28.70%
τ^c	16.10%	16.10%

* GDP per capita.

Following Dalgaard and Strulik (2014, 2017) we define the value of life at age z and time t as

$$VOL(z,t) = \frac{\int_z^T e^{-\rho(\hat{z}-z)} u(c(\hat{z}, t+\hat{z}-z), m(\hat{z}, t+\hat{z}-z), l(\hat{z}, t+\hat{z}-z), D(\hat{z}, t+\hat{z}-z), R(t+\hat{z}-z)) d\hat{z}}{u_c(c(z,t), m(z,t), l(z,t), D(z,t), R(t))}.$$

For a newborn cohort, i.e. for $z = 0$, we then obtain a value in the order of £3.25 Million from our model. This is well within the bounds of the estimates based on a range of international studies, as summarized in Viscusi and Aldy (2003). Figure 3 plots the value of life against the individual's age, demonstrating a continuous decline in $VOL(z,t)$ with age z, ultimately

approaching zero at the terminal age $z = T$.

[Insert Figure 3 here]

4 Numerical Experiments

Based on the benchmark calibration, we conduct three numerical experiments:

1. 10% increase to the supply of NHS services
2. 10% increase in the total factor productivity of final goods production
3. 10% increase in medical effectiveness in curbing deficit accumulation

The first experiment captures the impact of an expansion of NHS capacity, as is continuously and widely debated in the UK (e.g. O'Dowd 2016). The second and third experiments embrace the impact of productivity growth and medical progress as two of the well-known drivers of health care expenditures and life-time expansion (e.g. Hall and Jones 2007, Jones 2016, Kelly 2017, Böhm et al. 2018, Fonseca et al. 2020, Frankovic et al. 2020a,b). All experiments are based on a comparison of the underlying steady states.

For experiments 2 and 3, we consider three possible policy scenarios:

(i) Maintain a constant supply of NHS services, i.e. the benchmark

(ii) Maintain a constant medical expenditure to GDP ratio, i.e. $pH/GDP = \overline{pH/Y}$

(iii) Maintain a constant mean waiting time share, i.e. $\widehat{\omega} = \overline{\omega}$, which is tantamount to a constant waiting time per treatment.[15]

[15]Strictly speaking the target in our model is set as a fixed share of waiting in total health care time. Assume that treatments have a standardized (average) duration of $x \leq h(z,t)$, such that $h(z,t)/x$ gives the number of treatments, and note that waiting time is given by $\widehat{\omega} m(z,t) = \frac{\widehat{\omega}}{1-\widehat{\omega}} h(z,t)$. Waiting time per treatment is then given by $\frac{\widehat{\omega} m(z,t)}{h(z,t)/x} = \frac{\widehat{\omega} x}{1-\widehat{\omega}}$, implying that targeting $\widehat{\omega}$ is equivalent to targeting a constant waiting time per treatment.

Note that, in principle, there are three targets for the policymaker: NHS capacity, the NHS expenditure to GDP ratio, and average waiting time. In scenario (i), the policymaker is assumed not to respond to income growth and/or medical progress, by holding the NHS capacity constant and leaving the expenditure to output ratio and the waiting time free to adjust. In scenario (ii), the policymaker is assumed to target a fixed medical expenditure to GDP ratio. In this case, capacity will be adjusted in a way that the fiscal target of a constant expenditure share is met, with waiting time once again emerging endogenously. In scenario (iii), the policymaker is assumed to target a fixed waiting time per treatment by way of appropriate adjustments to the NHS capacity. In this case, it is the expenditure to GDP ratio which is left free.

We now proceed to discussing the outcomes of the main experiments 1-3 in turn. Tables 4-6 report the outcomes as percentage changes for a range of key variables: per capita income y; per capita consumption c; the wage rate w; the income tax rate τ^l;[16] per capita demand for health care (measured in gross time devoted to waiting for and consuming NHS services) m; the the share of gross health care time an individual spends waiting $\hat{\omega}$; the retirement age R; the consumption share in output C/Y; the NHS to output ratio pH/Y; the public pension output share, S/Y; the NHS labor share, N_H/N; longevity T; initial life-cycle utility $V(0,t)$ and instantaneous aggregate (cross-cohort) welfare $\Omega(t)$, where $V(0,t)$ is described by equation (6) and where $\Omega(t) = \int_0^T u(c(z,t), m(z,t), l(z,t))dz$. Note that in the following exposition, we drop the time argument t.

[16]Note that we assume that τ^k and τ^c are exogenous.

Table 4: 10% Increase in $\bar{H}$

y	c	w	τ^l	m	$\hat{\omega}$	R
-0.71	-1.93	-0.03	8.84	1.70	-0.79	-0.75

C/Y	pH/Y	S/Y	N_H/N	T	$V(0)$	Ω
-1.24	10.46	1.57	10.90	0.32	-0.16	0.17

$\Delta VOL(0)$: -2.25%

Experiment 1, as summarized in table 4, shows that a 10% increase in NHS supply $\bar{H}$ reduces the share of health care time spent waiting by 0.79%, motivating a 1.7% increase in m, the average time devoted to health. The greater effectiveness of health care induced by the decline in NHS congestion combined with the increase in the per capita demand of health care causes longevity to increase by 0.32% (approximately 71 additional days). Notably, the addition of old cohorts (as measured by the increase in T) with high demand for health care exacerbates the utilization of NHS services at the extensive margin. Overall, this leads to a sizeable increase by 10.46% in the health share. This is funded via an 8.84% increase of the income tax. In response, workers retire sooner, with the equilibrium retirement age declining by 0.75% (equivalent to 120 days sooner). Declining labor supply, combined with the rise in longevity, reduces income and consumption per capita by 0.71% and 1.93% respectively.

A number of features of the underlying adjustments following an increase in NHS capacity are worth noting. First, surprisingly perhaps, the expansion of NHS capacity by 10% actually raises per capita waiting time $\hat{\omega}m$ despite the 0.79% decline in the share of health care time devoted to waiting. This is because of the 1.7% increase in the per capita demand for health care m that is driven both by the increase in individual demand at the intensive margin and by the addition of old cohorts with high levels of demand at the extensive margin. Second, the expansion of the NHS leads to a more than proportional increase in the health care output and employment ratios. The former reflects the reduction in final goods production, while the latter results from the fact that the capital share is

lower in the health care sector than in the final goods sector. Third, the reallocation of resources away from consumption towards health care is reflected in a 2.25% decline in the value of life $VOL(0)$. As Frankovic et al. (2020a) note, the value of life can be interpreted as the marginal rate of substitution between longevity and consumption. Given the ensuing increase in longevity following the capacity increase, this necessarily implies a reduction in its value.

Aggregate welfare, as measured by Ω as a cohort cross-section, increases by a modest 0.17% despite the decline in per capita income and consumption. However, the welfare gain is not uniform across age groups, with the level of life-cycle utility of a newborn cohort $V(0)$ declining by some 0.16%.[17] This implies that older generations benefit from NHS expansion by more than younger generations who are relatively healthier, bear the burden of the tax increase, and are heavily discounting their future utility flows from early retirement and additional lifespan. Evaluating the remaining life-cycle utility $V(z)$ for each age z we find that individuals are indifferent between the two policies around age 40.674 (i.e. $z = 20.674$) with older individuals preferring the increased NHS supply and younger individuals preferring the lower supply. Thus, we conclude by noting that an expansion of a public health care sector within a stationary economy fails Pareto optimality and has the potential to create intergenerational conflict. Against, this backdrop, we now explore the impact of different policy responses to general productivity growth and medical progress.

[17]Note that the life-cycle utility of a newborn cohort $V(0)$ differs from the value of life at birth $VOL(0)$, which should be interpeted as a measure of the willingness to pay for survival (or the marginal rate of substitution between longevity and consumption – see e.g. Frankovic et al. 2020a). Thus, although in our analysis $V(0)$ and $VOL(0)$ decline in tandem, it is easily conceivable that a policy raises life-cycle utility $V(0)$ by increasing survival despite a (modest) reduction in consumption but at the same time lowers the value of (further) survival $VOL(0)$. At a mechanical level, a further distinction between $V(0)$ and $VOL(0)$ arises for the separable functional form we employ for instantaneous utility. Here $VOL(0)$ places a heavier emphasis on consumption (as opposed to utility loss from labor and waiting) due to the weighting with the inverse of the marginal utility from consumption.

Table 5: 10% Increase in Z

	y	c	w	τ^l	m	$\hat{\omega}$	R
1. Fixed $\bar{H}$	15.12	16.23	14.61	-6.49	1.56	0.16	0.74
2. Fixed pH/Y	14.49	14.46	14.59	0.59	3.09	-0.53	0.17
3. Fixed $\hat{\omega}$	14.97	15.83	14.61	-4.87	1.92	0.00	0.61

	C/Y	pH/Y	S/Y	N_H/N	T	$V(0)$	Ω
1. Fixed $\bar{H}$	0.97	-8.38	-0.32	-9.83	0.04	1.67	1.62
2. Fixed pH/Y	-0.02	0.00	1.00	-1.27	0.32	1.55	1.80
3. Fixed $\hat{\omega}$	0.74	-6.47	-0.01	-7.88	0.11	1.64	1.67

$\Delta VOL(0)$: 1. 18.32%; 2. 16.23%; 3. 17.85%

Experiment 2, as summarized in table 5, shows that a 10% increase in factor productivity in final goods production, Z, which is tantamount to an economic growth impulse, leads to a sizable aggregate cross-cohort welfare gain of 1.62% for households in the benchmark setting in which NHS capacity is held constant. The economic growth impulse is magnified in the benchmark setting by an expansion of labor supply due to the reduction in the NHS tax. The latter is feasible due to the expansion of the tax base. Notably, the welfare gain arises predominantly from an increase in per capita output and consumption in excess of 15% and 16%, respectively. Since capacity is held constant, average consumption of NHS services does not change, leaving life expectancy unaltered. This is all the more striking as the increase in the value of life by more than 18% indicates a strong willingness to pay for health care. However, for a capacity constrained NHS this willingness to pay does not boost demand by much, as individuals anticipate that an elevated level of overall demand will only serve to raise waiting times, thereby curbing the effectiveness of health care.

This situation is not much improved when the health policy is changed to holding average waiting times constant in scenario 3. Indeed, even in the benchmark, average waiting time does not increase by much following the economic growth impulse due to what might be considered a "voluntary" restraint of demand in a capacity constrained health care system. Therefore, counteracting the modest increase by a slight expansion of health care capacity

does not result in a significant change in welfare. The more appropriate policy, from an aggregate welfare perspective, in this setting amounts to maintaining a constant medical expenditure to income ratio in scenario 2. The additional tax receipts in the presence of income growth allow for a sizable expansion of NHS supply by 8.84% which triggers simultaneously a reduction in the average waiting time by some 0.53% and a boost in the per capita demand for NHS services by 3.09%. Overall, this translates into a longevity gain by 0.32% (the equivalent of approximately 71 additional days) and a modest increase in labor supply (due to the postponement of retirement). Despite the smaller increase in labor supply as compared to the benchmark, the expansion of the health care sector will be partially "self-financing," with the income tax rate rising only slightly by 0.59%. In total, per capita income and consumption grow by only 0.62 and 1.77 percentage points less than in the benchmark. Overall, the sizable increase in longevity increases the aggregate gain in cross-cohort welfare by 0.18 percentage points from 1.62% to 1.8%.

This finding is consistent with earlier findings by Hall and Jones (2007), Kuhn and Prettner (2016), Jones (2016), Böhm et al. (2018), Frankovic and Kuhn (2018), Fonseca et al. (2020) and Frankovic et al. (2020b) who all show that within growing economies the willingness to pay for life-expanding health care, as measured by the value of life, is so high as to warrant an expansion of the health car sector even at the expense of economic performance or potential consumption growth. Indeed, the strong increase in the value of life at birth even within the scenario in which the health expenditure share is kept constant indicates that a policy that would not only aim at maintaining the NHS spending share but rather at raising it (to some extent) may yield an even larger increase in aggregate welfare.

A more diverse picture emerges, when moving from a cross-cohort welfare perspective to the individual cohort. From the perspective of a newborn individual, the welfare gain following the productivity shock is greatest when NHS supply is held constant, with $V(0)$ increasing by 1.67%. Conversely, welfare gains for the youngest individuals are lowest under

a policy that fixes the medical expenditure output ratio as target, which is the policy that maximizes aggregate welfare. Reflecting the age-dependent preferences about a capacity increase, the preferred policy response to a productivity shock is again depends on age. At age 38.974 individuals switch from preferring a fixed NHS supply to a policy that fixes waiting time as a target. From age 39.366, individuals begin to prefer the medical expenditure share policy relative to a policy with fixed supply. Thus, while all policy responses allow all individuals to participate in the welfare gains from productivity growth, which thus amounts to a Pareto improvement per se, there is disagreement (by age) on what would be the best policy. Notably, the variance in the distribution of welfare gains across cohorts and policy measures is limited, implying a relatively even distribution of the benefits from productivity growth regardless of the particular policy response.

Table 6: 10% Increase in A

	y	c	w	τ^l	m	$\hat{\omega}$	R
1. Fixed $\bar{H}$	-0.75	-0.47	0.06	3.12	9.18	1.15	2.61
2. Fixed pH/Y	-0.93	-0.95	0.06	5.32	9.63	0.97	2.43
3. Fixed $\hat{\omega}$	-1.95	-3.62	0.03	17.52	11.97	0.00	1.37

	C/Y	pH/Y	S/Y	N_H/N	T	$V(0)$	Ω
1. Fixed $\bar{H}$	0.28	-2.57	9.98	-2.69	3.35	0.32	3.25
2. Fixed pH/Y	-0.02	0.00	10.44	-0.02	3.45	0.29	3.31
3. Fixed $\hat{\omega}$	-1.70	14.17	12.91	14.77	3.96	0.06	3.62

$\Delta VOL(0)$: 1. 0.07%; 2. -0.49%; 3. -3.59%

Experiment 3, as summarized in table 6, shows that a 10% increase in medical effectiveness in curbing deficits, A, leads to a sizable welfare gain of 3.25% across all cohorts in the benchmark setting in which NHS capacity is held constant. In this case, the welfare gain predominantly flows from the expansion of longevity by 3.35% (equivalent to over 2 additional years of life). The retirement age and, thus, the aggregate labor supply, increases by 2.61%. However, the large increase in life expectancy causes income and consumption per capita to decline, albeit moderately. More notably, the improvement of medical effectiveness

triggers a 9.18% increase in m, which for a constant capacity, boosts the average health care time share allocated to waiting by 1.15%. The resulting loss in medical effectiveness suggests that a nontrivial portion of medical progress is neutralized through the increase in congestion.

In order to secure a constant medical expenditure to output ratio, NHS capacity must increase by 2.57%.[18] Interestingly, despite this significant increase in the supply of available health care, this policy adds only 0.06 percentage points to the aggregate welfare gain. This is likely driven by the fact that much of the capacity increase is absorbed by a further boost in demand that neutralizes any mitigating impact on waiting times. Thus, under a policy that is aimed at a constant health care expenditure share, longevity increases by only a further 0.1 percentage points. In the presence of medical progress a policy that aims at containing waiting times turns out to be much more effective and yields an additional aggregate welfare gain of 0.37 percentage points relative to the benchmark. This is due in large part to the additional 0.61 percentage point increase to the longevity gains following a 11.97% boost to m. This policy requires a large-scale expansion of NHS supply (16.41%), financed through the proceeds from a 17.52% increase in the income tax rate. Despite this dramatic increase to the income tax rate, the labor supply grows by 1.37 percentage points, reflecting the need for individuals to expand their labor supply for the purpose of financing the additional life years. Nevertheless, average income and consumption fall by 1.95% and 3.62% respectively.

While all cohorts benefit from medical progress, notably, in contrast to experiment 2, there is now a strong age-gradient in the extent to which they do with the newborn cohort experiencing only a 0.32% increase to their life-cycle utility in the scenario with a constant NHS capaity.[19] Nevertheless, the increase in medical effectiveness turns out to be a

[18]This finding may appear to be at odds with the fact that per capita income y is declining. Note, however, that for a given size of the health care sector, the increase in retirement age and, thus, labor supply leads to an increase in aggregate output, Y, and its share in the GDP. The decline in per capita income is thus reflecting the increase in the old-age dependency ratio and the need for consumers to reallocate consumption to the life-years gained.

[19]We should caution that this result is to some extent driven by the assumption of a comparative static

Pareto improvement regardless of the policy response. Similar to experiment 2, however, the preference ranking for each policy in experiment 3 varies strongly with age. In contrast to the aggregate cross-cohort welfare assessment the youngest cohort stands to gain most, again in the presence of a constant capacity and, indeed, least under the waiting time target. This is again suggestive of older and retired cohorts being the ones with the highest propensity to benefit from the expansionary policy implied by a waiting time target. Studying the age-gradient in $V(z)$ we find that individuals tend to prefer a fixed NHS expenditure share as a policy target beyond age 39.8 ($z = 19.819$) and a waiting time target from age 40.6 ($z = 20.62$).

Drawing on our findings from all three experiments we can summarize the following key insights.

1. The presence of congestion, and perhaps more importantly, its anticipation, plays a crucial role in determining the macroeconomic and welfare effects of productivity growth and medical progress as two key drivers of economic development. While waiting time per se curbs the effectiveness of medical care and, thus, individuals' incentives to invest into it, the anticipation of increases in waiting time puts an additional break on health investments, whereas an anticipated reduction in waiting times boosts the demand for health care (as we have seen in Experiment 2). But then a policy that places a commitment to a waiting time target in the expectation of a strong increase in demand due to medical progress (as is true in Experiment 3) provides additional leverage. In either case, it is worth noting that the preferred policy, from an aggregate cross-cohort welfare perspective, is the one that allows the greatest flexibility for NHS capacity to respond the increase in demand for health care (i.e. is the policy that has the greatest increase in $\bar{H}$ of the three that we consider).

variation in medical effectiveness rather than its ongoing growth. Considering ongoing growth in medical effectiveness, Frankovic and Kuhn (2018) and Frankovic et al. (2020b) show that young cohorts tend to benefit more from an increase in the growth rate through its cumulative impact.

2. Based on our calibration of the UK economy and the NHS, a 10% capacity expansion turns out to trigger an increase in welfare across all cohorts. However, it does not imply a Pareto improvement, as young individuals experience a drop in their life-cycle utility. The concentration in welfare gains among the older population also affects the preferred policy to productivity growth and medical progress, where the youngest individuals would tend to prefer a stationary capacity. The disparities between the individual and aggregate welfare implications of the three policy regimes we consider underscore the importance of accounting for the variance in the value of timely health care according to age and current health status.

3. For simplicity, we have chosen not to include the individual's health deficit state in the instantaneous utility function, nor do we consider other potentially important determinants of individual welfare related to NHS efficiency such as the health status of family members, the positive correlation between health and labor productivity, and individual preferences for equity in the NHS. Thus, our results likely undervalue both the individual and social value of efficiency gains in the NHS.

5 Conclusions

We have considered the macroeconomic effects of congestion within a public health care system. In its purest form the consumption of health care is entirely free of charge, with the time cost to individuals placing the sole limit on the demand for health care. In such a setting, a certain extent of rationing is typically considered as helpful in so far as waiting imposes a time price on the consumption of health care. Wherever waiting is present in the context of possibly severe diseases for the diagnosis and/or treatment of which time is essential there is an additional welfare cost of waiting. Furthermore, the waiting list can typically not be directly controlled by the policymaker, but it builds up or diminishes based

on individual decisions on the utilization of health care. Thus, the policymaker only has limited control through the choice of health care capacity.

Building on an overlapping generations economy in which individuals can consume health care in order to curb the accumulation of health deficits a la Dalgaard and Strulik (2014) and thereby affect their longevity, we study how specific health policy rules shape the individual allocation across health care and consumption, the resulting waiting times and health outcomes (i.e. longevity), as well as the macroeconomic repercussions as transmitted through changes in the cross-sectoral allocation and labor supply.

Calibrating our model to reflect the English NHS and economy over the time frame 2007-2016, we first study an increase in NHS capacity and find that although it tends to lower per capita income, the resulting gain in longevity is more than compensating the reduction in consumption and generates in contemporary cross-cohort welfare. However, such a policy is not Pareto improving, as it reduces the life-cycle utility of the youngest cohorts who face higher taxation over their working lives and discount strongly the benefits from better access to health care in their old age.

We then study policy responses to productivity growth and medical progress and find that their welfare impact varies depending on the type of technological progress. Focusing on target-based policy rules we find that a fiscal target of maintaining the health expenditure share in GDP tends to boost the cross-cohort welfare impact of productivity growth, whereas a waiting list target tends to boost the cross-cohort welfare impact of medical progress. Indeed, the superiority of the respective rules arises from their particular impact on the expectation of consumers with respect to the development of the waiting list and their consequent demand for health care. Accordingly, we find that the preferred policy in response to technology shocks that increases the demand for health care will always be the policy that provides the greatest flexibility for NHS capacity to respond the increase in demand for health care.

However, the uneven distribution across cohorts of the net welfare gains from NHS capacity expansion also leads to a divergence (by age) of the preferred policies. While older individuals tend to prefer in response to both productivity growth and medical progress the most expansionary policy, younger individuals tend to prefer the status quo, i.e. a constant NHS capacity. Thus, again none of the policies satisfy the criterion of a Pareto improvement. This underscores the complexity of setting NHS policy, which must balance individual preferences by age for NHS services against the aggregate implications of each policy under consideration.

As we briefly discussed in the introduction, in this study we have chosen to model waiting as a form of congestion. This renders apparent another aspect of waiting, namely that it is associated with an externality: When individuals plan their utilization of the health care service they take the wait time as given and do not recognize that by contributing toward waiting time (by joining a waiting list, for instance) they are imposing a negative externality on others. This externality comes in the form of an increased time price of health care services as well as the reduced effectiveness of health care due to the strong correlation between timely delivery and patient outcomes. Thus, estimating the size of the externality and deriving a solution to the social planner's should be the focus of future work. Another set of issues worthy of further exploration relates to the inequality in the access to health care according to income and education which, strikingly, has been documented also for public health care systems with free provision of care (e.g. Bago d'Uva and Jones 2009, Vallejo-Torres and Morris 2013, Fiva et al. 2014). Frankovic and Kuhn (2019) show how unequal access coupled with medical progress have likely magnified the life expectancy gap in the US. It remains to be shown whether a public health care system is prone to guarantee more equitable outcomes.

6 References

Acemoglu, Daron and Veronica Guerrieri (2008). "Capital deepening and nonbalanced economic growth." *Journal of Political Economy 116(3)*, 467-498.

Bago d'Uva, Teresa and Andrew M. Jones (2009). "Health care utilisation in Europe: new evidence from the ECHP." *Journal of Health Economics 28*, 265-279.

Böhm, Sebastian, Volker Grossmann and Holger Strulik (2018). "R&D-driven medical progress, health care costs, and the future of human longevity." *CESifo Working Paper 6897*.

Chetty, Raj (2006). "A New Method of Estimating Risk Aversion." *American Economic Review 96*, 1821-1834.

Conesa, Juan C., Daniela Costa, Parisa Kamali, Timothy J. Kehoe, Vegard M. Nygard, Gajendran Raveendranathan and Akshar Saxena (2018). "Macroeconomic effects of Medicare." *Journal of the Economics of Ageing 11*, 27-40.

Dalgaard, Carl-Johan and Holger Strulik (2014). "Optimal Aging and Death: Understanding the Preston Curve." *Journal of the European Economic Association 12*, 672-701.

Dalgaard, Carl-Johan and Holger Strulik (2017). "The genesis of the golden age: accounting for the rise in health and leisure." *Review of Economic Dynamics 24*, 132-151.

Doubeni, Chyke A., Nicole B. Gabler, Cosette M. Wheeler, Anne Marie McCarthy, Philip E. Castle, Ethan A. Halm, Mitchell D. Schnall et al. (2018). "Timely follow-up of positive cancer screening results: A systematic review and recommendations from the PROSPR Consortium." *CA: A Cancer Journal for Clinicians 68(3)*, 199-216.

Eriksson, Carl O., Ryan C. Stoner, Karen B. Eden, Craig D. Newgard, and Jeanne-Marie Guise (2017). "The association between hospital capacity strain and inpatient outcomes in highly developed countries: a systematic review." *Journal of General Internal Medicine 32(6)*, 686-696.

Fiva, Jon H., Torbjorn Haegeland, Marte Ronning and Astri Syse (2014). "Access to treatment and educational inequalities in cancer survival." *Journal of Health Economics 36*, 98-111.

Fonseca, Raquel, Pierre-Carl Michaud, Titus J. Galama and Arie Kapteyn (2020). "Accounting for the Rise of Health Spending and Longevity." *Journal of the European Economic Association*, 1-44.

Frankovic, Ivan and Michael Kuhn (2018). "Health insurance, endogenous medical progress, and health expenditure growth." *TU Vienna Econ Working Paper 01/2018*.

Frankovic, Ivan and Michael Kuhn (2019). "Access to health care, medical progress and the emergence of the longevity gap: A general equilibrium analysis" *Journal of the Economics of Ageing 14*, 100188.

Frankovic, Ivan, Michael Kuhn, and Stefan Wrzaczek (2020a). "On the anatomy of medical progress within an overlapping generations economy." *De Economist 168*, 215-257.

Frankovic, Ivan, Michael Kuhn, and Stefan Wrzaczek (2020b). "Medical innovation and its diffusion: Implications for economic performance and welfare." *Journal of Macroeconomics 66*, 103262.

Gaudette, Étienne (2014). " Health care demand and impact of policies in a congested public system." *CESR-Schaeffer Working Paper No: 2014-005.*

Grossmann, Volker and Holger Strulik (2019). "Optimal social insurance and health inequality." *German Economic Review 20(4)*, e913-e948.

Hall, Robert E. and Charles I. Jones (2007). "The Value of Life and the Rise in Health Spending." *Quarterly Journal of Economics 122*, 39-72.

Hanna, Timothy P., Will D. King, Stephane Thibodeau, Matthew Jalink, Gregory A. Paulin, Elizabeth Harvey-Jones, Dylan E. O'Sullivan, Christopher M. Booth, Richard Sullivan, and Ajay Aggarwal (2020). "Mortality due to cancer treatment delay: systematic review and meta-analysis." *BMJ 371.*

Jones, Charles I. (2016). "Life and growth." *Journal of Political Economy 124*, 539-578.

Jung, Juergen and Chung Tran (2016). "Market inefficiency, insurance mandate and welfare: U.S. health care reform 2010." *Review of Economic Dynamics 20*, 132-159.

Kelly, Mark C. (2017). "Health capital accumulation, health insurance, and aggregate outcomes: a neoclassical approach." *Journal of Macroeconomics 52*, 1-22.

Kelly, Mark C. (2020). "Medicare for all or medicare for none? A macroeconomic analysis of healthcare reform." *Journal of Macroeconomics 63*, 103170.

Kuhn, Michael and Klaus Prettner (2016). "Growth and welfare effects of health care in knowledge based economies." *Journal of Health Economics 46*, 100-119.

Mitnitski, Arnold B., Alexander J. Mogilner, Chris MacKnight and Kenneth Rockwood (2002). "The accumulation of deficits with age and possible invariants of aging." *Scientific World 2*, 1816-1822.

Moscelli, Giuseppe, Luigi Siciliani, and Valentina Tonei (2016). "Do waiting times affect health outcomes? Evidence from coronary bypass." *Social Science and Medicine 161*, 151-159.

Nikolova, Silviya, Mark Harrison, and Matt Sutton (2016). "The impact of waiting time on health gains from surgery: Evidence from a national patient-reported outcome dataset." *Health Economics 25(8)*, 955-968.

O'Dowd, Adrian (2016). "NHS reports record waiting times in busiest year ever." *British Medical Journal 353*, i2724.

Oudhoff, J. P., D. R. M. Timmermans, D. L. Knol, A. B. Bijnen, and G. Van der Wal (2007). "Waiting for elective general surgery: impact on health related quality of life and psychosocial consequences." *BMC Public Health 7(1)*, 1-10.

Rexius, Helena, Gunnar Brandrup-Wognsen, Anders Odén, and Anders Jeppsson (2004). "Mortality on the waiting list for coronary artery bypass grafting: incidence and risk factors." *The Annals of Thoracic Surgery 77(3)*, 769-774.

Rexius, Helena, Gunnar Brandrup-Wognsen, Anders Odén, and Anders Jeppsson (2005). "Waiting time and mortality after elective coronary artery bypass grafting." *The Annals of Thoracic Surgery 79(2)*, 538-543.

Sampalis, John, Stella Boukas, Moishe Liberman, Tracey Reid, and Gilles Dupuis (2001). "Impact of waiting time on the quality of life of patients awaiting coronary artery bypass grafting." *Canadian Medical Association Journal 165(4)*, 429-433.

Schilling, Peter L., Darrell A. Campbell, Michael J. Englesbe, and Matthew M. Davis (2010). "A comparison of in-hospital mortality risk conferred by high hospital occupancy, differences in nurse staffing levels, weekend admission, and seasonal influenza." *Medical Care*, 224-232.

Siciliani, Luigi (2008). "A note on the dynamic interaction between waiting time and waiting lists." *Health Economics 17*, 639-647.

Siciliani, Luigi and Tor Iversen (2012). "Waiting times and waiting lists." in A.M. Jones (ed.), *The Elgar companion to health economics*.

Siciliani, Luigi, Valerie Moran and Michael Borowitz (2014). "Measuring and comparing health care waiting times in OECD countries." *Health Policy 118*, 292-404.

Sobolev, Boris G., Adrian R. Levy, Lisa Kuramoto, Robert Hayden, and J. Mark FitzGerald (2006). "Do longer delays for coronary artery bypass surgery contribute to preoperative mortality in less urgent patients?." *Medical Care*, 680-686.

Sobolev, Boris G., Guy Fradet, Robert Hayden, Lisa Kuramoto, Adrian R. Levy, and Mark J. FitzGerald (2008). "Delay in admission for elective coronary-artery bypass grafting is associated with increased in-hospital mortality." *BMC Health Services Research 8(1)*.

Sobolev, Boris G., Guy Fradet, Lisa Kuramoto, and Basia Rogula (2012). "An observational study to evaluate 2 target times for elective coronary bypass surgery." *Medical Care*, 611-619.

Sutherland, Jason Murray, R. Trafford Crump, Angie Chan, Guiping Liu, Elizabeth Yue, and Matthew Bair (2016). "Health of patients on the waiting list: Opportunity to improve health in Canada?." *Health Policy 120(7)*, 749-757.

Vallejo-Torres, Laura and Stephen Morris (2013). "Income-Related Inequity In Healthcare Utilisation Among Individuals With Cardiovascular Disease In England—Accounting For Vertical Inequity." *Health Economics 22*, 533-553.

Viscusi, W. Kip and Joseph E. Aldy (2003). "The Value of a Statistical Life: A Critical Review of Market Estimates Throughout the World." *Journal of Risk and Uncertainty 27*, 5-76.

Zhao, Kai (2014). "Social security and the rise in health spending." *Journal of Monetary Economics 64*, 21-37.

7 Appendix

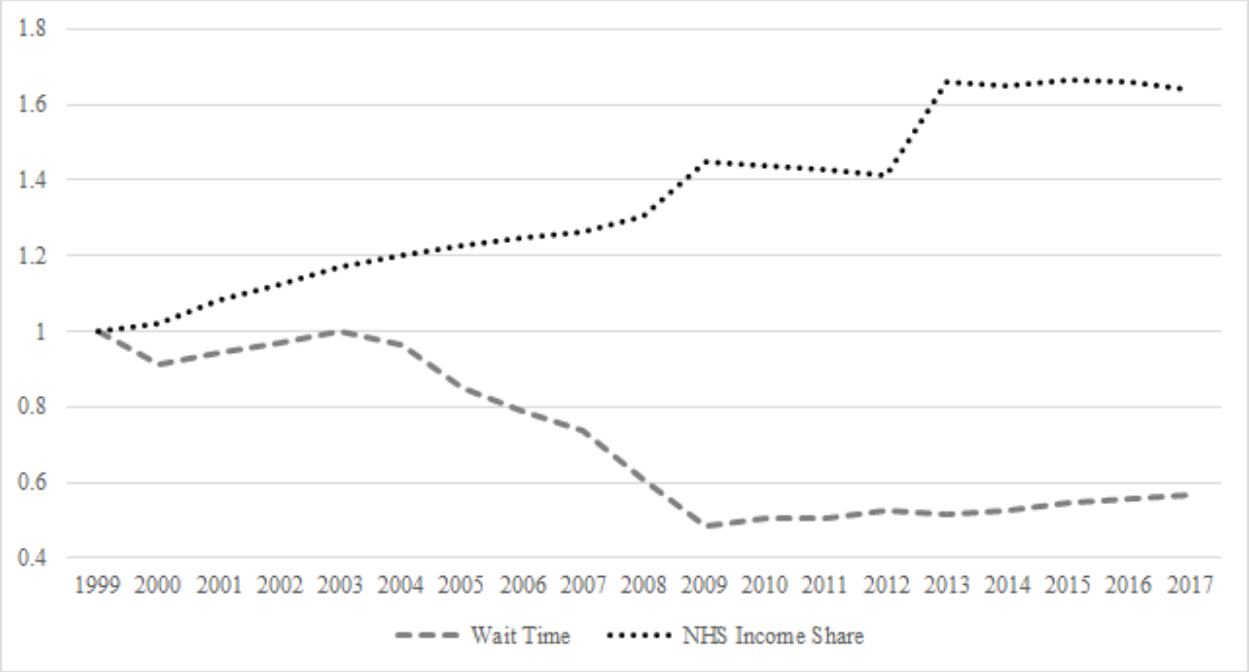
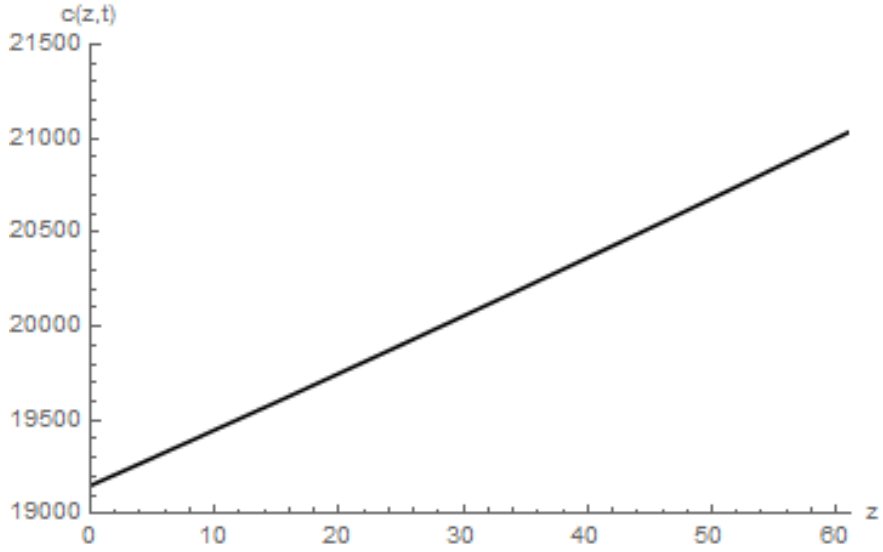
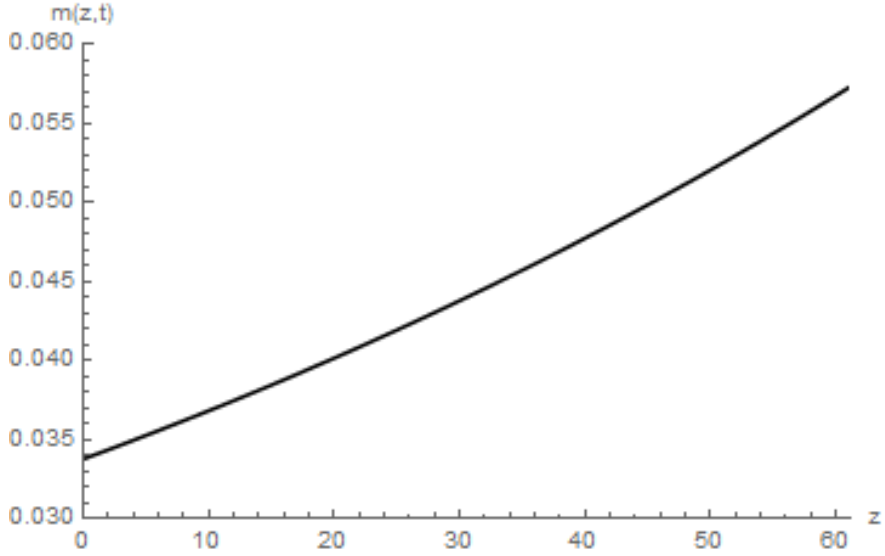


Figure 1: NHS waiting times and NHS Income Share 1999-2017. Source: UK Office of National Statistics (ONS).



(a) Consumption



(b) Demand for NHS Services

Figure 2: Lifetime Path of Consumption and Demand for NHS Services

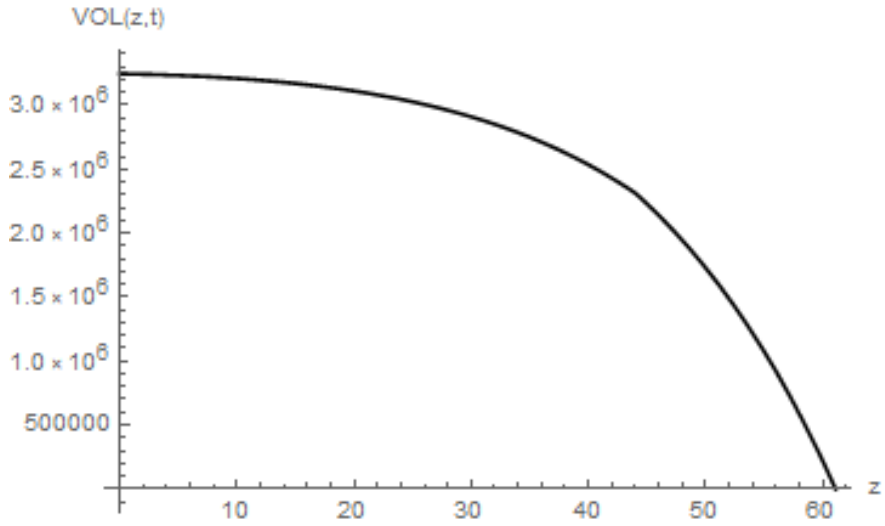


Figure 3: Lifetime Path of the Value of Life