Congestion in a Public Health Service: A Macro Approach

May 2022

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 to two groups of individuals, the healthy and the sick, 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.

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 by 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 is 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 that 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 have chosen to employ a representative agent, life-cycle model of health care demand introduced by Dalgaard and Strulik (2014 & 2017). Dalgaard and Strulik (2014 & 2017) advances on the seminal work of Grossman (1972), by incorporating a more realistic and empirically valid biological aging process that motivates

health care demand. While this approach potentially misses some important microeconomic aspects of health care that are captured by models of idiosyncratic health shocks.[1], we contend that the level of complexity of these models limits the intuition that can be derived from the model itself. Further, given the scope of our study, analysis of the results would rely heavily on averaging across different segments of the economy. Therefore, to address the issue of intra-generational heterogeneity in health status that is important in determining the welfare implications of health care policy, we have adapted Dalgaard and Strulik's model to include within cohort heterogeneity in health endowments, allowing us to evaluate how policy preferences may differ both within and across generations. We reserve a more detailed discussion of the limitations of our work for the conclusion.

The model economy is composed of a continuum of finitely-lived individuals born into distinct birth cohorts. Individuals, in turn, are assumed to differ in their health status. 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). Here, unhealthy as opposed to healthy individuals are assumed to start with a higher level of deficits. We assume that the government is the sole provider of health care services and supplies a predetermined level of medical capacity which is financed out of an earnings tax. Health care services are provided for free at the point service and are therefore allocated according to a queuing rule, where congestion (i.e. waiting time) becomes both the cost of obtaining medical services and a negative externality that reduces the effectiveness of health care services. Congestion 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

[1]e.g. French and Jones (2004), De Nardi, French, and Jones (2010), and Fonseca et. al. (2021), and Hosseini et al. (2021).

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 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.[2] 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 improves instantaneous welfare (across all cohorts) even if this comes at the cost of a lower 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

[2]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.

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 a little below age 40 tend to suffer from a reduction in their life-cycle utility, implying that the capacity expansion fails to satisfy Pareto optimality. While the welfare impacts are somewhat more pronounced both positive and negative for the unhealthy, there is no qualitative difference by health status.

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. Again this holds for healthy and unhealthy individuals alike.

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. Moreover, 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.

From a modeling perspective, our approach is following the literature on health deficit accumulation, a concept pioneered by epidemiologists (e.g Mitnitski et al. 2002) and first adapted to economic analysis by Dalgaard and Strulik (2014). In this approach individuals are assumed to accumulate health deficits (or frailty) over their life-cycle, a process that can be delayed but not reversed through health care investments. This approach can be shown to generate realistic correlations between health care spending and age or health status. Recent work has brought forth the importance of heterogeneity in the processes of frailty accumulation (Hosseini et al. 2021). We take account of this and explore its role in respect to the allocations and policies studied by distinguishing healthy and unhealthy individuals.

The remainder 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 Model

2.1 Individuals

The economy is composed of a continuum of identical birth-cohorts, consisting of two groups of individuals; "sick" and "healthy." Let the indicator $i \in \{h, s\}$ denote the health group the individual belongs to, with h representing the "healthy" group and s representing "sick" group. To abstract away from population growth, we assume that the number of individuals born into each group is constant across time.

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 spend some time on a waiting list before becoming eligible to consume NHS services. Let $h_i(z,t)$ be the utilization of NHS services (in terms of time) by an individual aged z and belonging to group i at time t.[3] We distinguish between $h_i(z,t)$ and $m_i(z,t)$, the individual's current demand for health care, which is the sum of $h_i(z,t)$ and $\omega_i(z,t)$ the amount of time the individual spends waiting for NHS services. For simplicity we assume that wait times are uniform across the economy, implying that individual waiting time is directly proportional to $\widehat{\omega}(t)$, the average share of total health care time that individuals spend waiting for NHS services at time t. Thus, we have $\omega_i(z,t) = \widehat{\omega}(t)m_i(z,t)$ and $h_i(z,t) = (1-\widehat{\omega}(t))m_i(z,t)$.

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 $D_i(z,t)$ be the individual's total number of accumulated health deficits at age z and time t. The deficit accumulation function is defined as

$$\dot{D}_i(z,t) = \mu_i[D_i(z,t) - a_i - f(m_i(z,t))]; \quad D_i(0,t-z) = D_i(0). \tag{1}$$

Deficits accumulate at the natural rate μ_i which we allow to differ according to the agent's individual health status i.[4] Likewise, we allow the initial deficit endowment $D_i(0,t-z)$ to differ according to health group. Exogenous environmental factors that reduce (or increase) the deficit accumulation rate are captured by the parameter a_i. The effect of the individual's health investment on the deficit accumulation rate is described by the function $f(m_i(z,t))$. Finally, we assume that an individual will die once their total accumulated deficits exceed the threshold $\bar{D}$. We then denote by T_i the individual's terminal age which corresponds to

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

[4]While we allow for the possibility here that the natural force of aging differs according to health type, in our calibration we assume that $\mu_s = \mu_h = \mu$. Our justification for this assumption is provided in section 3.

the age the individual reaches the deficit threshold (i.e. $D_i(T_i, t) = \bar{D}$).

The health investment function takes the following functional form

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

where $M(t)$ is effective aggregate demand for health care (i.e. the sum of $m_i(z,t)$ over all i and z) and where $H(t)$ is the exogenous supply of NHS services.[5] In the supply-constrained equilibrium of a public health care sector total consumption of NHS services is constrained by $H(t)$, implying that $H(t) \leq M(t)$.[6] Supply $H(t)$ then equals total utilization, i.e. the sum of $h_i(z,t)$ across all individuals in the economy. Whenever aggregate demand exceeds NHS supply, the average waiting time share $\widehat{\omega}(t)$, which is defined as

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

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

[5] 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.

[6] Indeed, we will show below that $H(t) < M(t)$ must hold in equilibrium. See footnote 12 below.

is unaffected by NHS congestion.[7] Substituting $h_i(z,t) = (1-\widehat{\omega}(t))m_i(z,t)$ into equation (2) we obtain the alternative expression

$$f(m_i(z,t)) = A\left(\frac{H(t)}{M(t)}\right)^{\epsilon+\gamma} m_i(z,t)^{\gamma}, \quad 0 \leq \epsilon, \gamma \leq 1. \tag{3}$$

There are two important points with respect to how we have chosen to model waiting in our model that are worth discussing in detail: (i) While our modeling of $\widehat{\omega}(t)$ 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 $t + \omega_i(z,t)$. We circumvent the complexity involved by conflating a whole treatment episode, as measured by $m_i(z,t)$, into a single period t, a year say, such that each individual is treated within t but assuming that (unmodeled) waiting reduces the effectiveness of $m_i(z,t)$. 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).[8]

Finally, before proceeding, we briefly digress to discuss the role of chronological age

[7] Note that the effectiveness of each unit of $m_i(z,t)$ 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.

[8] 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).

in the accumulation of health deficits. For this discussion, we refer the reader to equation (A.1) in the Appendix, which presents the solution for the agent's health deficit stock at any arbitrary age z. Several empirical studies, such as Contoyannis, Jones and Rice (2004), French and Jones (2004), and Hosseini et al. (2021), have found evidence of a significant role for chronological age in determining the life-cycle dynamics of health. While equation (1) does not include a separate age effect, chronological age nevertheless plays a crucial role in determining the lifetime dynamics of the agent's health deficit stock.

Absent medical care investments, the agent's current stock of deficits would be solely a function of exogenous environmental factors a, their initial deficit endowment $D_i(0)$, the natural force of aging μ_i, and the agent's age z. Intuitively, in Dalgaard and Strulik's formula of the aging process, μ_i represents the mean rate at which individuals randomly accumulate medical conditions overtime. Naturally, the longer an individual lives, the more opportunities they have to acquire frailties. In equation (A.1), this is captured by the multiplicative term $e^{\mu_i z}$. Thus, the accumulating nature of health deficits in our model implicitly captures the age effect, while preserving the gerontological view of the aging process as being state dependent.

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

$$u(c_i(z,t), m_i(z,t), l_i(z,t)) = \frac{c_i(z,t)^{1-\sigma}}{1-\sigma} - \theta\frac{\omega_i(z,t)^{\phi}}{\phi} - \eta l_i(z,t), \tag{4}$$

where $c_i(z,t)$ is current consumption and $1/\sigma$ is the inter-temporal elasticity of substitution. The parameter θ is the disutility weight from individual waiting time, while $\phi \geq 1$ governs the curvature of waiting time disutility. 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.[9] Labor (i.e. $l_i(z,t)$) generates disutility as measured by the parameter

[9]Indeed, there is a considerable body of evidence that waiting lowers patients' quality of life and psycho-

η. We assume that $l_i(z,t)$ is a binary variable equaling one when the individual is in the labor force and zero when they exit. Assuming that $R_i(t)$ is an i-type individual's optimal retirement age,[10] we have

$$l_i(z,t) = \begin{cases} 1 & \text{if } z \leq R_i(t) \\ 0 & \text{if } z > R_i(t) \end{cases}. \tag{5}$$

Individuals consume and save out of their asset income and after-tax earnings. Assets are held in the form of physical capital $k_i(z,t)$ and accumulate according to the agent's flow budget constraint

$$\dot{k}_i(z,t) = (1-\tau^k(t))r(t)\,k_i(z,t) + (1-\tau^l(t))w(t)\,l_i(z,t) - (1+\tau^c(t))c_i(z,t) + s(z,t), \tag{6}$$

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. $l_i(z,t) = 1$) 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

$$V_i(0,t) = \int_0^{T_i(t)} e^{-\rho z} u(c_i(z,t), m_i(z,t), l_i(z,t))dz \tag{7}$$

with ρ denoting the rate of time preference, by choosing $c_i(z,t)$, $m_i(z,t)$, and $R_i(t)$ subject to

logical well being (e.g. Sampalis et al. 2001, Oudhoff et al. 2007, Sutherland et al. 2016).

[10] Strictly speaking $R_i(t)$ amounts to the optimal retirement age of a type-i individual belonging to the birth cohort $t - R_i(t)$.

equations (1) and (6). The Hamiltonian function that characterizes the individual's optimal decision problem is given as[11]

$$\begin{aligned}\mathcal{H}_i &= e^{-\rho z}\left\{\begin{array}{l} u(c_i,m_i,l_i)+\lambda_i^D\mu_i\left(D_i-a_i-A\left(\frac{H}{M}\right)^{\epsilon+\gamma}m_i^{\gamma}\right) \\ +\lambda_i^k[(1-\tau^k)rk_i+(1-\tau^l)wl_i-(1+\tau^c)c_i+s] \end{array}\right\} \\ &\text{s.t.} \quad D_i(0,t)=D_i(0),\ D_i(T_i,t)=\bar{D},\ k_i(0,t)=k_i(T_i,t)=0, \end{aligned} \tag{8}$$

where λ_i^D and λ_i^k are the costate variables on the stock of deficits and capital, respectively. The resulting first-order conditions are[12]

$$c_i^{-\sigma}=\lambda_i^k(1+\tau^c) \tag{9}$$

$$-\lambda_i^D\gamma\mu_iA\left(\frac{H}{M}\right)^{\epsilon+\gamma}m_i^{\gamma-1}=\theta\widehat{\omega}^{\phi}m_i^{\phi-1} \tag{10}$$

$$\frac{(1-\tau^l)w}{(1+\tau^c)c_i(R_i)^{\sigma}}=\eta \tag{11}$$

$$u(T_i)=\frac{\theta\widehat{\omega}^{\phi}m_i(T_i)^{\phi-\gamma}}{\gamma A}\left(\frac{H}{M}\right)^{-(\epsilon+\gamma)}\left(\bar{D}-a_i\right)-\frac{\theta\left(\widehat{\omega}m_i(T_i)\right)^{\phi}}{\gamma}+\frac{(1+\tau^c)c_i(T_i)-s}{(1+\tau^c)c_i(T_i)^{\sigma}} \tag{12}$$

$$-\frac{\dot{\lambda}_i^k}{\lambda_i^k}=(1-\tau^k)r-\rho \tag{13}$$

$$-\frac{\dot{\lambda}_i^D}{\lambda_i^D}=\mu_i-\rho. \tag{14}$$

[11]Note that the age and time indicators z and t have been dropped for brevity.
[12]Note that the second-order conditions can be verified numerically.

Equation (9) 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. Equation (10) 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.[13] 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. Thus, under the queuing rule, waiting is the only cost imposed on patients. Equation (10) 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_i 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 (11)). 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_i^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

[13]It is easy to infer from the first order condition (10) 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 (10) implies a zero marginal cost for the utilization of health care. Accordingly, individuals would then increase $m_i(z,t)$ and only stop if $\hat{\omega}(t) > 0$ (and sufficiently large). But then, for a given $H(t)$, we must have $H(t) < M(t)$.

of R_i, we treat Q as exogenous so that the public pension does not impact the individual's retirement decision.

Finally, equation (12) determines optimal longevity. This condition is derived by noting that the Hamiltonian function will be equal to zero in the terminal period. Setting $\mathcal{H}_i(T_i) = 0$, substituting for $\lambda_i^k(T_i)$ and $\lambda_i^D(T_i)$ using equations (9) and (10), and rearranging terms yields equation (12) and allows us to solve for T_i, using the following condition:

$$u(T_i) = -\lambda_i^D(T_i)\mu_i \left(\bar{D} - a_i - A \left(\frac{H}{M} \right)^{\epsilon+\gamma} m_i(T_i)^{\gamma} \right) + \lambda_i^k(T_i)[(1+\tau^c)c_i(T_i) - s].$$

This condition implies that individuals prefer to live up to the point that their instantaneous utility at age T_i 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_i, 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 (9) with respect to z and substituting for the dynamics for λ_i^k from (13). Rearranging terms yields

$$\frac{\dot{c}_i}{c_i} \equiv g_c = \frac{(1-\tau^k)r - \rho}{\sigma}. \tag{15}$$

Similarly, the Euler equation for health investments is derived by differentiating (10) with respect to z and substituting for $\dot{\lambda}_i^k$ and $\dot{\lambda}_i^D$ from (13) and (14):

$$\frac{\dot{m}_i}{m_i} \equiv g_{m,i} = \frac{\rho - \mu_i}{\phi - \gamma}. \tag{16}$$

This equation implies that the lifetime path of m_i is positively correlated with the rate of time preference, ρ.[14] Note that as $\rho \to 0$, $g_{m,i}$ becomes negative, implying that the optimal

[14]Note here the difference to the dynamics in Dalgaard and Strulik (2014, 2017), where the interest rate, r,

health investment strategy for perfectly patient individuals is to invest more heavily in their 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 $\rho > \mu_i$, $g_{m,i} > 0$, 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 $\rho > \mu_i$, then the individual's optimal $g_{m,i}$ 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 investments will be relatively low but then increase rapidly throughout the individual's lifetime. In our calibration, $g_{m,i} > 0$, implying that individual demand for NHS services increases with age (see figure 3 below in section 3). Likewise, the life-cycle growth rate of health care demand will be adversely correlated with ϕ, which determines the curvature of disutility from waiting time. For values of $\phi > 1$ there is convexity in the disutility from waiting. Thus, for higher values of ϕ, the marginal disutility increases rapidly with health care demand, making the agents reluctant to incur more waiting time as they age, despite the increasing deterioration of health that occurs with age.

Given the complexity of this system, we cannot derive an analytical solution for the individual's problem. Instead, we solve for $c_i(0)$, $m_i(0)$, R_i, and T_i numerically in general equilibrium. The solution procedure for obtaining the numerical solution is outlined in the Appendix.

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.

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 health care output is determined 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

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

where $Z(t)$ is total factor productivity, where $K_F(t)$ and $L_F(t)$, 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

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

We assume that workers are homogeneous and are perfect substitutes across sectors, implying that there is a single competitive labor market. 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, net of the rate of depreciation δ, and labor

respectively:[15]

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

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

Let $H(t)$ be the aggregate supply of NHS services. The NHS production function takes the CES functional form[16]

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

The NHS employs two inputs; health care labor, $L_H(t)$, and capital, $K_H(t)$. Total factor productivity in health care is represented by $B(t)$, 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 $p(t)H(t)$. The NHS is required to operate as a non-profit institution and therefore chooses $K_H(t)$ and $L_H(t)$ in order to minimize the cost of producing $\bar{H}(t)$. The zero-profit condition that characterizes the NHS' objective is

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

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

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

i.e. a reimbursement of NHS services according to their average cost. Taking $p(t)$ as given,

[15] 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.

[16] 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. This choice also improves the stability of the model.

cost minimization implies that the NHS chooses capital according to the following rule

$$p(t)\frac{\partial H(t)}{\partial K_H(t)} = r(t) + \delta. \tag{17}$$

Similarly, in minimizing costs, the NHS chooses $L_H(t)$ so that the marginal product of labor in the health care sector is equal to the equilibrium wage rate

$$p(t)\frac{\partial H(t)}{\partial L_H(t)} = w(t). \tag{18}$$

2.3 Aggregation and Market Clearance

Let $N_i(t)$ be the size of a new cohort of health type-i at time t. Noting that the number of surviving cohorts in each health category is equal to the respective life expectancy of that group, the total population in the economy is defined as

$$N(t) = N_h(t)T_h + N_s(t)T_s. \tag{19}$$

Aggregate consumption $C(t)$ is obtained by summing up the individual consumption across cohorts and health groups

$$C(t) = N_h(t)c_h(0,t)\int_0^{T_h} e^{g_c z}dz + N_s(t)c_s(0,t)\int_0^{T_s} e^{g_c z}dz. \tag{20}$$

Likewise, the aggregate labor supply $L(t)$ is the sum of the individual labor supplies across cohorts

$$L(t) = N_h(t)\int_0^{T_h} l_h(z,t)dz + N_s(t)\int_0^{T_s} l_s(z,t)dz = N_h(t)R_h + N_s(t)R_s. \tag{21}$$

Finally, the various aggregate measures of health related time allocations are obtained by summing up individual effective demand for NHS services and NHS consumption.

$$M(t) = N_h(t)m_h(0,t)\int_0^{T_h} e^{g_{m,h}z}dz + N_s(t)m_s(0,t)\int_0^{T_s} e^{g_{m,s}z}dz, \tag{22}$$

$$H(t) = N_h(t)m_h(0,t)\int_0^{T_h} (1-\widehat{\omega}(t))e^{g_{m,h}z}dz + N_s(t)m_s(0)\int_0^{T_s} (1-\widehat{\omega}(t))e^{g_{m,s}z}dz. \tag{23}$$

Aggregate capital accumulation is obtained by summing up the individual flow budget constraints, described by equation (6), across all living cohorts. This summation yields

$$\dot{K}(t) = (1-\tau^k)r(t)K(t) + (1-\tau^l)w(t)L(t) - (1+\tau^c)C(t) + S(t). \tag{24}$$

where $K(t)$ denotes aggregate wealth (physical capital) and $S(t)$ denotes aggregate public pension payouts. It should be noted that we are assuming that $K_Y(t)$ and $K_H(t)$ depreciate at the same rate δ.

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

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

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 $S(t) = \kappa w(t)(N(t) - Q(t))$, where $N(t) - Q(t)$ is the total population of individuals that are receiving public pension payments.

We can simplify (24) 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 (24) reduces the aggregate capital accumulation function to

$$\dot{K}(t) = (1-\nu)F(t) - C(t) - \delta K(t). \tag{25}$$

In a steady-state, it holds that $\dot{K}(t) = 0$. From equation (25), 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

$$L(t) = L_H(t) + L_F(t)$$

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. The primary

source of our data is 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.

The definition of "sick" individuals is critical to our calibration exercise. In deciding which childhood conditions to include in our definition, we looked for conditions where (i) the child survives into adulthood, (ii) having that condition in childhood has an adverse impact on that individual's longevity, and (iii) is a condition where the child is likely to survive into adulthood without major disabilities. The availability of empirical evidence of the disease's prevalence and the subsequent mortality gap was also an important factor in determining which conditions to include as part of our definition. According to these criteria, we base our calibration of the sick group on individuals with asthma, type-1 diabetes (T1DM), and survivors of childhood or adolescent cancer. Relying on survey data of patient-reported asthma from 2010-2012, Mukherjee et. al. (2016) estimate the prevalence of asthma in UK at 9.6%.[17] To estimate the prevalence of T1DM, Stedman et. al. (2020) combine data from the NHS's Hospital Episode Statistics (HES) and the National Diabetes Audit (NDA). They estimate a T1DM prevalence of 0.6%. Finally, Yeh et. al. (2020), utilizing data from the Childhood Cancer Survivor Study, estimate that 0.74% of the adult population in the UK are survivors of childhood or adolescent cancer. Combining these three categories, we assume that the sick group accounts for approximately 11% of the population.

Table 1 presents the aggregate variables we calibrate the model to match along with the corresponding values from the model. 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. Using data from the Copenhagen City Heart Study, O'Byrne et. al. (2019) conclude that the average lost life years (LLY) from asthma alone is 3.3 years (relative to the general population). Utilizing data from the 2015-2016 NDA, Heald et. al. (2020) estimate an

[17]Note that we are using patient reports of having clinician-diagnosed and treated asthma.

Table 1: Aggregates

Variable	Description	Data	Model
w	Average wage rate	£28,359.20	£28,325.00
y	GDP per capita	£26,160.20	£30,255.20
c	Consumption per capita	£16,373.20	£20,169.50
h	NHS expenditures per capita	0.40%	0.41%
$\hat{\omega}m$	Average waiting time	4.02%	4.02%
C/Y	Consumpiton output share	62.61%	66.66%
pH/Y	NHS output share	8.75%	8.26%
S/Y	Public pension output share	6.05%	6.03%
L_H/L	NHS labor share	4.35%	4.87%
R^*	Average retirement age	43.99	43.96
$R - R_s$	Retirement gap (relative to the average)	N/A	0.67
T^*	Average life expectancy at calendar age 20	61.06	60.97
$T - T_s$	Life expectancy gap (relative to the average)	3.93	4.06
τ^l	Tax rate on labor income	25.95%	17.74%
τ^k	Tax rate on capital income	28.70%	28.70%
τ^c	Tax rate on consumption	16.10%	16.10%

* Ages are presented in model age. All agents are assumed to enter at chronological age 20.

average of 7.6 LLY from T1DM. Yeh et. al. (2020) estimate that the average LLY of childhood cancer survivors diagnosed in 1990-1999 is 9.2 years. Using the prevalence figures cited above, we obtain a weighted average LLY from these three conditions of 3.93 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. Several recent studies[18] find evidence of increased absenteeism, limited access to certain occupations due to risk, and poor health associated with the conditions under consideration, all of which suppress earnings. However, given that our scope is to evaluate the welfare implications of various NHS policy targets on individuals that differ according to their health endowments, we have opted to simplify the model by abstracting away from other sources of heterogeneity, such as income or wealth.[19]

[18] e.g. Lindahl Norberg et. al. (2017), Teckle et. al. (2018), Persson et. al. (2018), and Belova et. al. (2020)

[19] As the analysis provided in the next section demonstrates, the preferred policy target is the policy that provides the greatest flexibility for NHS supply to respond to shocks to aggregate health care demand. Therefore, even though the omission of wage heterogeneity likely biases our welfare analysis downward, given that we are evaluating a public health service that is provided free of charge, the inclusion of income heterogeneity would likely further reinforce our findings as the sick group derive greater marginal benefit from consuming health care.

Average consumption and income per person were £16,373.20 and £26,160.20 respectively. This implies an aggregate consumption share of 62.61%.[20] NHS expenditures accounted for 8.75% of total output, while 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 employment share averaged 4.35%. We estimate the average retirement age using OECD estimates of the average effective age at retirement, which for males during this period was 63.99 years old. The model implies an 8 month retirement gap for sick workers.

For calibrating health care demand and the average wait time for NHS services we rely on the 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 use the average length of stay as our measure of the average provision of NHS services per person (H/N in the model). This equates to approximately 1.5 bed days per person, per year (equivalent to 0.4% of the individual's time endowment). Average wait time per episode was 52.79 days, which equates to an aggregate average waiting time of 14.89 days (4.02% of the time endowment).

Table 2: Deficit Accumulation Parameters

Description	Notation	Value
Health investment elasticity of health care	γ	0.65
Force of aging	μ	0.043
Health deficits at age 20 (healthy)	$D_h(0)$	0.027
Health deficits at age 20 (sick)	$D_s(0)$	0.0283
Maximum health deficits	$\overline{D}$	0.1005

The majority of the parameters from the deficit accumulation function 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. In these studies Mitnitski and

[20] All output shares are in nominal terms.

Rockwood utilize frailty indexes to measure human frailty over the course of the life-cycle and show that deficits accumulate in an exponential fashion. Though the data sources for these studies differ,[21] they consistently find that the average individual accumulates deficits at a rate between 3%-4% annually. For the sake of brevity, we direct the reader to Dalgaard and Strulik (2014) for a summary of Mitnitski and Rockwood's work as it relates to the motivation and calibration of the deficit accumulation function.

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 as 0.027 and 0.1005 respectively. 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.

To match the life expectancy gap of 3.93 years, we allow the initial health endowments to differ between the two health groups, while maintaining a constant force of aging across the economy. This choice is consistent with Mitnitski and Rockwood (2016), Dragone and Vanin (2020), and Dalgaard, Hansen, and Strulik (2021), who emphasize the "self-productive" nature of the aging process. That is, that the rate at which individuals age is dependent on their current health state, as empirically supported by Mitnitski et al. (2002), Searle et al. (2008), Shi et al. (2011), Harttgen et al. (2013), Mitnitski et al. (2013), Mitnitski and Rockwood (2016), Abeliansky and Strulik (2018a,b), and Abeliansky, Erel, and Strulik (2020). Thus, while the natural force of aging is constant across health groups, the sick group will age faster than the healthy group precisely because entering the economy in worse health than the healthy makes them more susceptible to acquiring additional deficits. As Figure 2 demonstrates, in our model, the small initial gap in deficits between healthy and sick

[21] Mitnitski et. al. 2002 utilizes a data set of 66,589 Canadians, aged 15 to 79, while Rockwood and Mitnitski combine data of elderly community-dwelling individuals from Australia, Sweden, and the United States

individuals amplifies over time.

[Insert Figure 2 here]

In our calibration, the terminal level of deficits $\overline{D}$ is the same for both health groups. In choosing a value for the initial deficit endowment of healthy individuals, we rely on Dalgaard and Strulik's estimate of 0.027 for the general populace. Then, to match the life expectancy gap of 3.93 years, we assume $D_s(0) = 0.0283$. Based on our choice of $\rho = 0.05$, setting $g_h = 0.02$ requires a value of 0.65 for γ. Given our choice of ρ (see Table 3 below), setting $g_h = 0.02$ implies that $\gamma = 0.65$.

Table 3: 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	2.75
Curvature parameter for waiting time	ϕ	1
Disutility from labor	η	0.975
Medical effectiveness	A	0.115
Returns to timely treatment	ϵ	0.25
Environmental parameter	a	0.0199
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	1250
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	0.02

The remainder of the parameters are listed in Table 3 above. These parameters are chosen to calibrate the equilibrium to match the sample averages from the data. The rate of time preference (ρ) and the inter-temporal 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.

As a benchmark, we set $\phi = 1$, implying that there is constant marginal disutility from waiting. We have tested the sensitivity of our results to different values of ϕ and found that our benchmark results are not substantively altered by allowing convexity in the disutility of waiting time.[22] The disutility parameters θ and η are set at 2.75 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. We set $\bar{H} = 0.25$ so that $\bar{h}$, the average time utilizing NHS services, equals 0.4% of the time endowment.[23] Following Acemoglu and Guerrieri (2008), the capital share in the health care sector β is set at 0.2. We assume a value of 0.25 for ϵ, which determines 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

[22] We can provide the results of our sensitivity analysis upon request.

[23] We note here again that this is equivalent to an average of 1.5 bed days, the average utilization implied by the aggregate HES data.

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 $Q(t)$.

In the steady-state the lifetime growth rate of demand for NHS services is 2% respectively. As Figure 3 shows, both non-medical consumption and consumption of NHS services will rise throughout the agents' lifetimes. The exponential life-cycle growth of health care consumption is consistent with the empirical evidence (e.g. Jung and Tran 2014). Moreover, the exponential profile of medical consumption implies that health care utilization is highly skewed in our model, which is also consistent with the stylized facts (e.g. Manning and Mullahy 2001 and Cantoni and Ronchetti 2006). Within each cohort, sick individuals will utilize a greater quantity of NHS services. Thus, our model captures the fact that preferences for health care, and subsequently for NHS policy, will depend on both age and health status (i.e. the agent's deficit stock).

Non-medical consumption is relatively flat throughout the life-cycle, growing by 0.15% annually for both health groups. However, within each cohort we observe two important differences in the behavior of healthy and sick individuals. First, since sick individuals have shorter lifespans, they have less of an incentive to save and will therefore consume slightly more while alive than their healthy counterparts within their birth cohort. Second, since healthy individuals survive longer (an additional 4.7 years) they will consume more, both on average and in total, than sick individuals.

[Insert Figure 3 here]

Finally, our primary measure of individual welfare is $V_i(z, t+z)$, the individual's re-

maining lifetime value function. As Figure 4 demonstrates, $V_i(z, t+z)$ closely maps to $VOL_i(z, t+z)$, the agent's value of life. Following Dalgaard and Strulik (2014, 2017) we define the value of life at age z and time t as

$$VOL_i(z,t) = \frac{\int_z^{T_i} e^{-\rho(\hat{z}-z)} u(c_i(\hat{z}, t+\hat{z}-z), m_i(\hat{z}, t+\hat{z}-z), l_i(\hat{z}, t+\hat{z}-z), D_i(\hat{z}, t+\hat{z}-z), R_i(t+\hat{z}-z)) d\hat{z}}{u_c(c_i(z,t), m_i(z,t), l_i(z,t), D_i(z,t), R_i(t))}.$$

For a newborn cohort, i.e. for $z = 0$, we then obtain a value in the order of £3.25 Million for healthy individuals in the model. This is well within the bounds of the estimates based on a range of international studies, as summarized in Viscusi and Aldy (2003). We observe a continuous decline in $V_i(z, t+z)$ and $VOL_i(z, t+z)$ with age z, ultimately approaching zero at the terminal age $z = T$. Furthermore, at all ages, the mortality gap adversely impacts the remaining lifetime utility of sick individuals relative to their healthy counterparts.

[Insert Figure 4 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 and Charlesworth et. al. 2021). 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 "status quo target" policy

(ii) Maintain a constant medical expenditure to GDP ratio, i.e. the "fiscal target" policy

(iii) Maintain a constant mean waiting time share (which is tantamount to a constant waiting time per treatment), i.e. the "waiting time target" policy[24]

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 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.

For the purpose of evaluating the welfare implications of each scenario, we rely on two measures of welfare. The first is Ω_i, which is the sum of the instantaneous utility of health group i. The second is the value function $V_i(z, t+z)$, which is the remaining life-cycle

[24]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 $x \leq h_i(z,t)$, such that $h_i(z,t)/x$ 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.

utility at age z of an individual belonging to health group i who was born at time t. The former allows us to evaluate the aggregate welfare implications, while the latter allows us to consider how the welfare consequences of each policy for a type-i individual evolve with age.[25] Throughout we provide plots of $V_i(z, t+z)$, which shows how the individual's policy preferences develop over their life-cycle. We now proceed to discuss the outcomes of the main experiments in turn.

4.1 10% Increase in NHS Capacity

Table 4 displays the response (in percent change form) of the representative individuals' average lifetime consumption (c_i), effective medical demand (m_i), age at retirement (R_i), life expectancy (T_i), and initial welfare ($V_i(0)$), as well as the instantaneous aggregate welfare of each group (Ω_i). Table 5 presents the percent change of several key macroeconomic variables including the wage rate, the output shares of aggregate capital ($\frac{K}{GDP}$), consumption ($\frac{C}{GDP}$), NHS expenditures ($\frac{pH}{GDP}$), and social security ($\frac{S}{GDP}$), the NHS labor share ($\frac{L_H}{L}$), the labor tax rate, and the average waiting time ($\widehat{\omega}$).

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

	c_i	m_i	R_i	T_i	$V_i(0)$	Ω_i
Healthy	-1.96	6.71	-0.54	0.60	-0.18	0.39
Sick	-1.97	6.75	-0.48	0.55	-0.17	0.34

Table 5: 10% Increase in $\bar{H}$: Aggregate

w	$\frac{K}{GDP}$	$\frac{C}{GDP}$	$\frac{pH}{GDP}$	$\frac{S}{GDP}$	$\frac{L_H}{L}$	τ^l	$\hat{\omega}$
-0.02	0.74	-1.21	10.22	2.44	10.64	9.11	-0.25

In equilibrium, increasing NHS capacity by 10% reduces the average waiting time by

[25]Note that since we assume that individuals are homogeneous at birth within their respective health group, in a stationary equilibrium $V_i(z, t+z)$ represents both the individual life-cycle profile of a type-i individual's value function as well as a cross-cohort profile of the remaining life-cycle utility flows of the type-i health group at current time t. Thus, we can state that $V_i(z, t+z) = V_i(z, t) = V_i(z)$ in equilibrium.

0.25%. The reduction in waiting time increases the efficacy of NHS services, raising T_h and T_s by 0.6% and 0.55%, approximately 135 and 115 additional days respectively. These two factors combine to raise the average demand for NHS services of both groups by a little more than 6.7%. Furthermore, as panels (a) and (b) of Figure 5 demonstrate, utilization of NHS services rises as well. The additional NHS outlays are financed by a 9.11% increase in the income tax rate. Thus, while NHS services remain free at the point of service, households ultimately pay for the additional healthcare services in the form of higher taxes. In response, households reduce their average consumption by just under 2% (see panels (c) and (d) of Figure 5). In addition, the increase to the income tax rate reduces the return to supplying labor, leading to a 0.54% and 0.48% decline in the retirement ages R_h and R_s despite the increase in longevity.

By assumption, NHS expenditures rise by 10%, diverting resources away from final goods production, as the NHS labor share will rise by 10.64% and the NHS capital share increases by 9.36%. Final goods output declines by 1.1%. These effects translate to a 0.18 decline in GDP. Aggregate capital rises by 0.74% relative to GDP. Aggregate consumption falls by 1.21% relative to GDP, while the NHS expenditure share increases by 10.22%. The rise in life expectancy increases social security outlays, resulting in a 2.44% rise of the social security share.

In the aggregate both health groups benefit from this policy, suggesting that the NHS was previously underfunded. Interestingly, the healthy group benefits more than the sick group, with Ω_h increasing by an additional 0.05 percentage points relative to Ω_s, as healthy individuals enjoy slightly greater gains in lifespan and leisure time. However, it is important to note that the benefits of this policy are not evenly distributed throughout the economy. At the point of entering the economy, young individuals from both the healthy and sick groups are made worse off relative to the benchmark economy. Panels (e) and (f) of Figure 5 plot the lifetime path of the representative agents' value function. From these graphs, we can see that

young individuals are initially made worse-off by this policy. However, with increasing age, the values individuals assign to the benchmark as opposed to the scenario with a capacity increase converge and eventually cross. Thus, sick and healthy agents, respectively, tend to prefer the capacity increase from age 37.04 and 38.18 onward (note that these ages are represented by the vertical red lines).[26]

4.2 10% Increase in general productivity Z

Table 6: 10% Increase in Z: Individual

	c_i	m_i	R_i	T_i	$V_i(0)$	Ω_i
1. Fixed $\bar{H}$						
Healthy	16.24	1.18	0.68	-0.02	1.66	1.56
Sick	16.26	1.19	0.65	-0.02	1.66	1.57
2. Fixed pH/Y						
Healthy	14.40	7.31	0.30	0.51	1.52	1.93
Sick	14.40	7.35	0.31	0.48	1.53	1.89
3. Fixed $\hat{\omega}$						
Healthy	15.29	4.36	0.49	0.25	1.59	1.75
Sick	15.30	4.38	0.48	0.23	1.59	1.74

Table 7: 10% Increase in Z: Aggregate

	w	$\frac{K}{GDP}$	$\frac{C}{GDP}$	$\frac{pH}{GDP}$	$\frac{S}{GDP}$	$\frac{L_H}{L}$	τ^l	$\hat{\omega}$
1. Fixed $\bar{H}$	14.61	-0.47	0.96	-8.33	-0.52	-9.78	-6.57	0.12
2. Fixed pH/Y	14.59	0.15	-0.03	0.00	1.58	-1.27	0.89	-0.11
3. Fixed $\hat{\omega}$	14.60	-0.15	0.45	-4.05	0.56	-5.42	-2.74	0.00

Experiment 2, as summarized in Tables 6 and 7, examines a 10% increase in factor productivity in final goods production, Z, which is equivalent to an economic growth impulse. The latter results in a sizable aggregate cross-cohort welfare gain of over 1.5% for both health groups in the status quo setting, wherein NHS capacity is held constant. The rise in productivity has a large, positive impact on the wage rate (14.6% increase). Households

[26]Note that since we assume that agents enter at age 20, these ages correspond to $z = 17.04$ and $z = 18.18$ in the model.

further benefit from a 6.57% decline of the income tax rate that is realised from the increase in the tax base while NHS spenidng remains constant. Workers respond positively to these effects, increasing their labor supply, leading to an increase of R_h and R_s by 0.68% and 0.65%. As household income rises, agents are able to afford more consumption, and average annual consumption increases by 16.25%.

Consistent with the findings of Hall and Jones (2007), we find that rising income levels will raise effective demand for health care. This occurs even though NHS capacity is held constant and health care services are provided as a public good. Given that average demand for health care is rising for a fixed supply of NHS services, the average waiting time increases slightly by 0.12%. This has a modest, negative impact on health care efficacy, resulting in a negligible decline in life expectancy. We note that in the absence of adjustments to health care capacity, income growth triggers a treadmill effect. While individuals increase their demand for health care, this does not translate into any sizable benefit but merely works to offset the loss in the effectiveness of care.

Nevertheless, the large increase in consumption made possible by the increase in income results in a substantial rise in aggregate welfare. Moreover, the welfare gain occurs at all ages, with the initial value function for each health group increasing by around 1.66%.

In the aggregate, NHS expenditures increase by 5.5% following the large boost to the wage rate. The demand for NHS labor drops by 9.17% and the government invests more in NHS capital in order to maintain the status quo capacity. Nevertheless, GDP outpaces NHS expenditures, causing the NHS share to fall by 8.33%. Similarly, since life expectancy is unaffected under this policy, social security expenditures will be essentially unaltered. The aggregate consumption share increases by just under 1%, while the aggregate capital share falls by approximately 0.5%.

As Figure 6 and Tables 6 and 7 demonstrate, the response to a productivity shock varies significantly depending on which policy target the government pursues. Figure 6 displays

the lifetime paths of NHS utilization (panels (a) and (b)), consumption (panels (c) and (d)), and the agents' value function (panels (e) and (f)). The original benchmark paths are represented by the black, dashed-dotted line. The status-quo paths are represented by the orange solid line. The blue dashed lines represent the fiscal target, while the green line with larger dashes represent the waiting time target.

Both the fiscal and waiting time targets provide some flexibility for NHS capacity to adjust. So, while consumption of NHS services is held constant under the status quo policy, it is allowed to rise under the other two policies we consider. In this experiment, targeting a constant average NHS expenditure ratio allows for the greatest flexibility in adjusting NHS capacity and therefore results in the greatest increase in the consumption of NHS services of the three cases we consider. Nevertheless, the large increase in NHS capacity (approximately 9%) required to maintain a constant NHS output share is substantial enough to reduce average waiting time. The combined effects of increased consumption of NHS services and improved health care efficacy from diminished waiting time leads to modest gains in life expectancy (approximately 99 additional days for the sick and 116 days for the healthy).

On its own, expanding NHS capacity in this scenario does not impact the income tax rate. However, the moderate impact on longevity that follows increases the social security share by 1.58%, causing the income tax rate to increase by 0.89%. Consequently, as panels (c) and (d) of Figure 6 show, the increase in average consumption is considerably smaller (nearly 2 percentage points below the status quo policy) under this policy. Similarly, the rise in the income tax rate reduces the returns to supplying labor time slightly, muting the impact on the labor supply (via delayed retirement). Both health groups experience the greatest welfare gains under the fiscal target regime in this experiment, making it the preferred policy in the aggregate. Note however, that the preference ranking of the policies depends on the agents' age. Young individuals of each health group initially prefer the status quo policy, which allows them the largest reduction in their tax burden, and benefit the least

from the fiscal target, which leads to a modest increase in taxation in spite of productivity growth while young agents do not benefit yet much from the improved provision of health care However, as panels (e) and (f) demonstrate, the preference ranking changes as the agents' marginal benefit to health care increases with age relative to their marginal utility for consumption.

4.3 10% Increase in the state of medical technology A

Table 8: 10% Increase in A: Individual

	c_i	m_i	R_i	T_i	$V_i(0)$	Ω_i
1. Fixed $\bar{H}$						
Healthy	-0.10	6.75	0.53	0.69	0.02	0.60
Sick	-0.10	6.80	0.54	0.63	0.02	0.53
2. Fixed pH/Y						
Healthy	-0.20	7.14	0.50	0.72	0.01	0.62
Sick	-0.20	7.19	0.51	0.67	0.02	0.56
3. Fixed $\hat{\omega}$						
Healthy	-6.64	28.99	-1.38	2.77	-0.58	1.95
Sick	-6.69	29.19	-1.18	2.55	-0.57	1.70

Table 9: 10% Increase in A: Aggregate

	w	$\frac{K}{GDP}$	$\frac{C}{GDP}$	$\frac{pH}{GDP}$	$\frac{S}{GDP}$	$\frac{L_H}{L}$	τ^l	$\hat{\omega}$
1. Fixed $\bar{H}$	0.01	-0.01	0.06	-0.53	2.07	-0.56	0.65	0.71
2. Fixed pH/Y	0.01	0.03	0.00	0.00	2.21	0.00	1.13	0.70
3. Fixed $\hat{\omega}$	-0.06	2.39	-3.89	32.88	10.63	34.54	31.01	0.00

Experiment 3, summarized in Tables 8 and 9 and Figure 7, shows that a 10% increase in A, the effectiveness of medical care in curbing deficits, results in modest cross-cohort welfare gains under the status quo and fiscal target policies. In this experiment, the improvements to the medical technology predictably increase the demand for health care and improve life expectancy. The fixed supply of NHS services under the status quo policy causes the average waiting time to rise by 0.71% in response to the increase in demand for health care. Life

expectancy increases by 0.63% for sick agents and 0.69% for healthy agents. Workers will remain engaged in the labor market longer due in part to the rise in longevity. The increase in longevity also impacts social security payouts, requiring the income tax rate to increase by 0.65% to fund these additional expenditures. Average annual consumption of both groups falls in response by 0.1%, while aggregate consumption actually outpaces GDP gains by 0.06% due to the increase in lifespans. The NHS expenditure share falls by 0.53%. Given the minimal effect of this shock on income, maintaining a constant NHS expenditure share under the fiscal target regime implies only a small increase to NHS supply (i.e. 0.53%). Consequently, the response to the rise in medical effectiveness under this policy closely maps to the status quo regime.

Panels (a) and (b) of Figure 7, which display the life-cycle trajectories of effective utilization, $h_i(z, t+z)$, shows the dramatic difference between the health care demand under the waiting time regime and the alternative policies. Health care demand under all three policies rises in response to the improvements in health care technology. However, the constraints placed on NHS capacity under the status quo and fiscal target regimes are such that the growth in aggregate demand for NHS services outpaces that of NHS supply. As a result, congestion rises under these two policies to the point that $h_i(z, t+z)$ actually falls at all ages relative to the individuals' pre-shock consumption of NHS services. The rise in congestion notwithstanding, the improvements in medical technology still translate to substantial gains in longevity.

In contrast, maintaining a constant average waiting time requires that NHS supply rise in tandem with the aggregate demand for health care $M(t)$. Referencing Table 8, average demand for medical care increases by approximately 29% for both groups, which translates to a 32.5% increase in NHS capacity in order to maintain waiting times. While, this diverts significant resources away from final goods production (as evidenced by the 34.54% increase of the NHS labor share), the increased capacity provides agents with the opportunity to

consume considerably more NHS services than they did prior to the shock. This, combined with the increased efficacy of health care, contributes to significant life expectancy gains ranging from 1.45 years for sick agents to 1.7 years for healthy ones.

Maintaining constant waiting times is costly. In addition to the substantial increase in NHS capacity, the positive effect on longevity causes social security spending to increase by 10.63% relative to GDP. Funding this requires that the income tax rate increase by 31%. Such a large increase to the income tax rate discourages labor, encouraging sick workers to retire about six months sooner and healthy workers to retire over 7 months sooner. To compensate for the lost income from early retirement and taxation, households cut back on consumption. Average annual consumption declines by just under 6.7% (see panels (c) and (d) of Figure 7) and the aggregate consumption share falling by 3.89%.

The cross-cohort welfare gain from the waiting time target regime exceeds the respective gain in the status quo and fiscal target regimes by a factor of more than 3 for both health groups. Here again, the preference ranking depends on age. Young individuals, who benefit the least from the additional NHS expenditures, are made worse-off. Referencing panels (e) and (f), around age 37.4 for the sick group and 38.6 for healthy group, the fixed waiting time target becomes the preferred policy. This is due to the decline in the marginal utility of consumption that occurs as consumption increases with age, an effect that is present in all scenarios. However, the second and dominating driver in the context of medical progress lies in the complementarity between the effectiveness of medical technology, A, and health care utilization h in reducing deficits. As utilization increases with age, this implies that individuals tend to benefit disproportionately from health care technology in old age. For young individuals these benefits accrue only far into the future and are, therefore, discounted, whereas for older individuals the benefits from medical treatment are more imminent. Notably, the complementarity between A, the individual demand for health care m and the loss in treatment effectiveness $(1-\widehat{\omega})^{\epsilon+\gamma}$ due to waiting also explains the disproportion-

ate increase in life expectancy and welfare when a waiting time target is pursued in the context of medical progress, as opposed to the gain from imposing a fiscal target in the presence of conventional productivity growth: In the presence of this complementarity, the effective utilization of health care, and thus a containment of waiting, becomes a prerequisite for the benefits from better medical technology to materialize. In contrast, the gains from productivity growth materialize regardless of how effective the utilization of health care is.

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 consume health care in order to curb the accumulation of health deficits over the life cycle 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. Here, we allow for differences across individuals in the initial level of health deficits and consequently

their accumulation.

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 contemporary cross-cohort welfare gains. 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. The results do not differ qualitatively for healthy and unhealthy individuals.

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 is the policy that provides the greatest flexibility for NHS capacity to respond the increase in demand for health care. We also identify an important role for complementarity between medical technology and its effective implementation, which requires a limit in congestion. Medical progress strongly leveraged to translate into sizable gains in longevity (only) if it comes with a containment of waiting time.

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.

Finally, let us revert to the issue of heterogeneity in health care needs as driven by processes of idiosyncratic health shocks. As documented by Hosseini et al. (2021), the cumulation of such shocks can generate fat-tailed distributions of health care utilization, with much of the utilization concentrated on a small share of the population and a large share consuming little or no health care. An exploration of the implications of such highly uneven

patterns of health care utilization for waiting and the allocation of health care resources is important. A similar argument applies to an analysis of correlated health shocks, as experienced e.g. during a pandemic. While we need to relegate such analyses to future work, we note that our macroeconomic model of a public and capacity constrained health care sector can serve as a useful starting point.

6 References

Abeliansky, Ana Lucia, Devin Erel, and Holger Strulik (2018a). "How we fall apart: Similarities of human aging in 10 European countries." *Demography 55*, 341-359.

Abeliansky, Ana Lucia, Devin Erel, and Holger Strulik (2018b). "Hungry children age faster." *Economics and Human Biology 29*, 211-222.

Abeliansky, Ana Lucia, Devin Erel, and Holger Strulik (2020). "Aging in the USA: similarities and disparities across time and space." *Scientific Reports 10(1)*, 1-12.

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.

Belova, Anna, Neal Fann, Jacqueline Haskell, Bryan Hubbell, and Tulika Narayan (2020). "Estimating lifetime cost of illness. An application to asthma." *Annals of the American Thoracic Society 17(12)*, 1558-1569.

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*.

Cantoni, Eva and Elvezio Ronchetti (2006). "A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures." *Journal of Health Economics 25(2)*, 198-213.

Charlesworth, Anita et. al. (2021). "What is the right level of spending needed for health and care in the UK?" *The Lancet*, 2012-2022.

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.

Contoyannis, Paul Andrew Jones and Nigel Rice (2004). "The dynamics of health in the British Household Panel Survey." *Journal of Applied Econometrics 19(4)*, 473-503.

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.

Dalgaard, Carl-Johan, Casper Worm Hansen, and Holger Strulik (2021). "Fetal origins—A life cycle model of health and aging from conception to death." *Health Economics*

De Nardi, Mariacristina, Eric French, and John B. Jones (2010). "Why Do the Elderly Save? The Role of Medical Expenses." *Journal of Political Economy 118(1)*, 39-75.

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.

Dragone, Davide, and Paolo Vanin (2020). "Substitution effects in intertemporal problems." *Working Paper DSE 1147, University of Bologna.*

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 (2021). "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.

French, Eric and John B. Jones (2004). "On the Distribution and Dynamics of Health Care Costs." *Journal of Applied Econometrics 19(6)*, 705-721.

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

Grossman, Michael (1972). "On the Concept of Health Capital and the Demand for Health." *The Journal of Political Economy 80(2)*, 223-255.

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.*

Harttgen, Kenneth, Paul Kowal, Holger Strulik, Somnath Chatterji, and Sebastian Vollmer (2013). "Patterns of frailty in older adults: Comparing results from higher and lower income countries using the survey of health, ageing and retirement in Europe (SHARE) and the study on global AGEing and adult health (SAGE)." *PloS one 8(10)*, e75847.

Heald, Adrian H., Mike Stedman, Mark Davies, Mark Livingston, Ramadan Alshames, Mark Lunt, Gerry Rayman, and Roger Gadsby (2020). "Estimating life years lost to diabetes:

outcomes from analysis of National Diabetes Audit and Office of National Statistics data." *Cardiovascular Endocrinology & Metabolism 9(4)*, 183.

Hosseini, Roozbeh, Karen A. Kopecky, and Kai Zhao (2021). "The evolution of health over the life cycle." *Review of Economic Dynamics*, in press.

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

Jung, Juergen and Chung Tran (2014). "Medical consumption over the life-cycle." *Empirical Economics 47(3)*, 927-957.

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.

Lindahl Norber, Annika, Scott M. Montgomery, Matteo Bottai, Mats Heyman, and Emma I. Hoveń (2016). "Short-term and long-term effects of childhood cancer on income from employment and employment status: A national cohort study in Sweden." *Cancer 123(7)*, 1238-1248.

Manning, Willard G. and John Mullahy (2001). "Estimating log models: to transform or not to transform?" *Journal of Health Economics 20(4)*, 461-494.

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.

Mitnitski, Arnold B., Xiaowei Song, and Kenneth Rockwood (2013). "Assessing biological aging: the origin of deficit accumulation." *Biogerontology 14(6)*, 709-717.

Mitnitski, Arnold B. and Kenneth Rockwood (2016). "The rate of aging: The rate of deficit accumulation does not change over the adult life span." *Biogerontology 17(1)*, 199-204.

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.

Mukherjee, Mome, et. al. (2016). "The epidemiology, healthcare and societal burden and costs of asthma in the UK and its member nations: analyses of standalone and linked national databases." *BMC Medicine 14(1)*, 1-15.

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'Byrne, Paul, Leonardo M. Fabbri, Ian D. Pavord, Alberto Papi, Stefano Petruzzelli, and Peter Lange (2019). "Asthma progression and mortality: the role of inhaled corticosteroids." *European Respiratory Journal 54(1)*, 1-14.

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.

Persson, Sofie, Gisela Dahlquist, Ulf-G. Gerdtham, and Katarina Steen Carlsson (2018). "Why childhood-onset type 1 diabetes impacts labour market outcomes: a mediation analysis." *Diabetologia 61(2)*, 342-353.

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.

Rockwood, Kenneth and Arnold (2007). "Frailty in relation to the accumulation of deficits." *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences 62(7)*, 722-727.

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.

Searle, Samuel D., Arnold Mitnitski, Evelyne A. Gahbauer, Thomas M. Gill, and Kenneth Rockwood (2008). "A standard procedure for creating a frailty index." *BMC Geriatrics 8(1)*, 1-10.

Shi, Jing, Xiaowei Song, Pulin Yu, Zhe Tang, Arnold Mitnitski, Xianghua Fang, and Kenneth Rockwood (2011). "Analysis of frailty and survival from late middle age in the Beijing Longitudinal Study of Aging." *BMC Geriatrics 11(1)*, 1-8.

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.

Stedman, Mike et. al. (2020). "Cost of hospital treatment of type 1 diabetes (T1DM) and type 2 diabetes (T2DM) compared to the non-diabetes population: a detailed economic evaluation." *BMJ open 10(5)*, e033231.

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.

Teckle, Paulos, Stuart Peacock, Mary L. McBride, Colene Bentley, Karen Goddard, and Paul Rogers (2018). "Long-term effects of cancer on earnings of childhood, adolescent and young adult cancer survivors–a population-based study from British Columbia, Canada." *BMC health services research 18(1)*, 1-10.

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.

Yeh, Jennifer M. et. al. (2020). "Life expectancy of adult survivors of childhood cancer over 3 decades." *JAMA Oncology 6(3)*, 350-357.

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

Tables and Figures

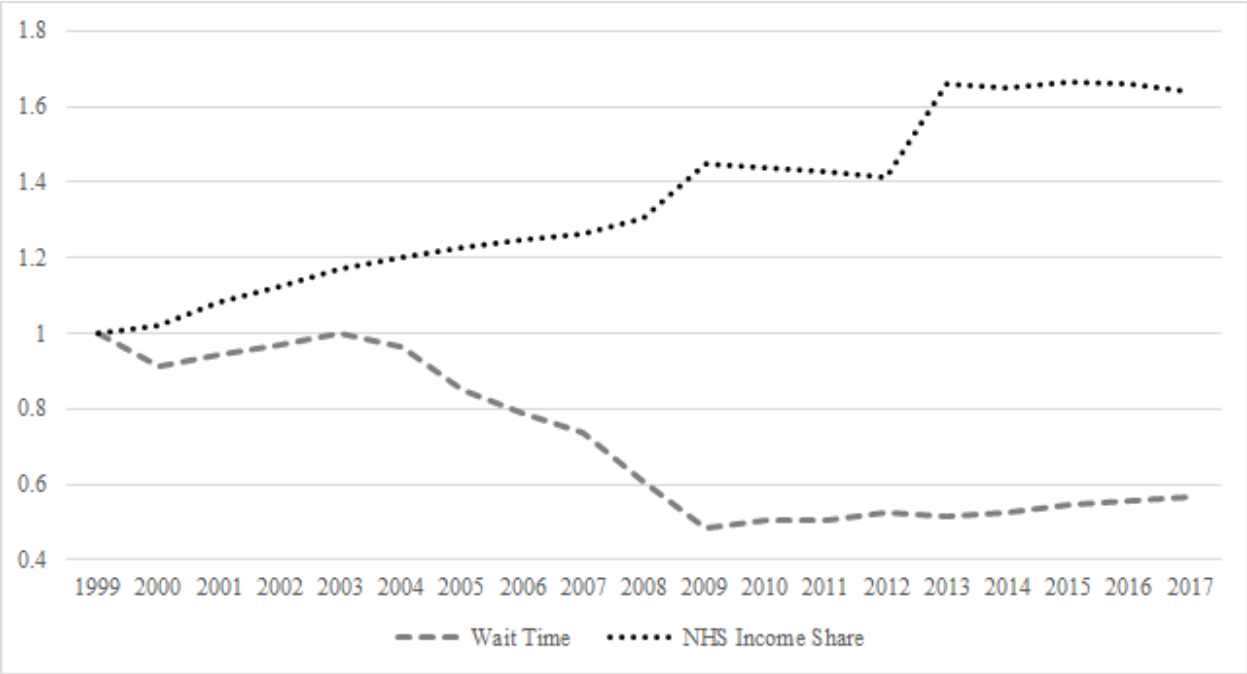


Figure 1: NHS waiting times and NHS Income Share 1999-2017. Source: UK Office of National Statistics (ONS).

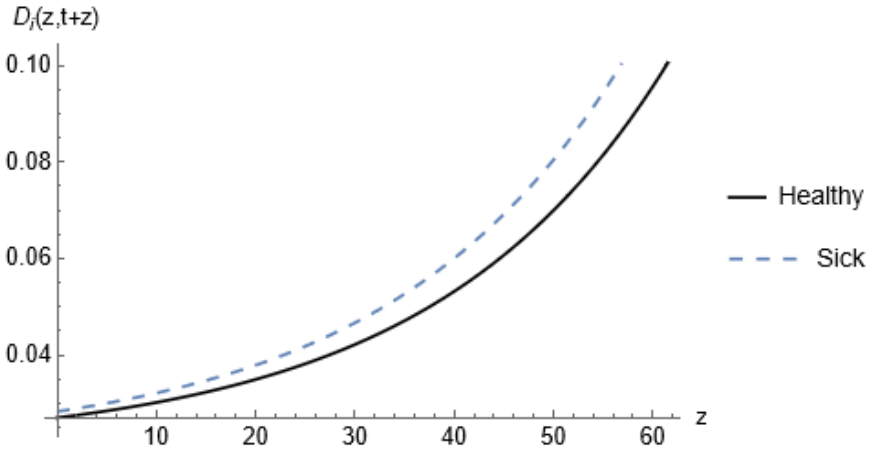
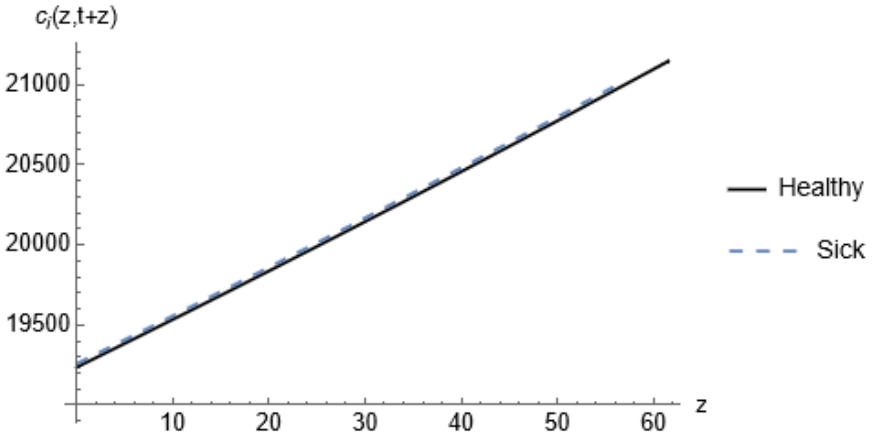
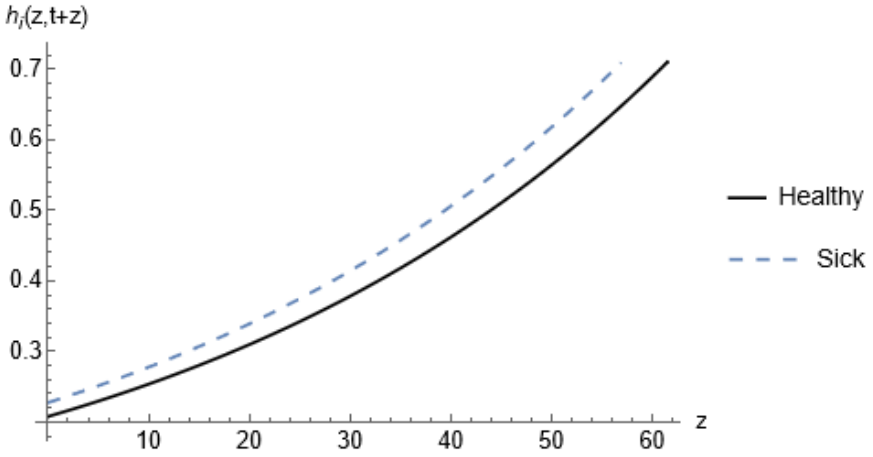


Figure 2: Health Deficits

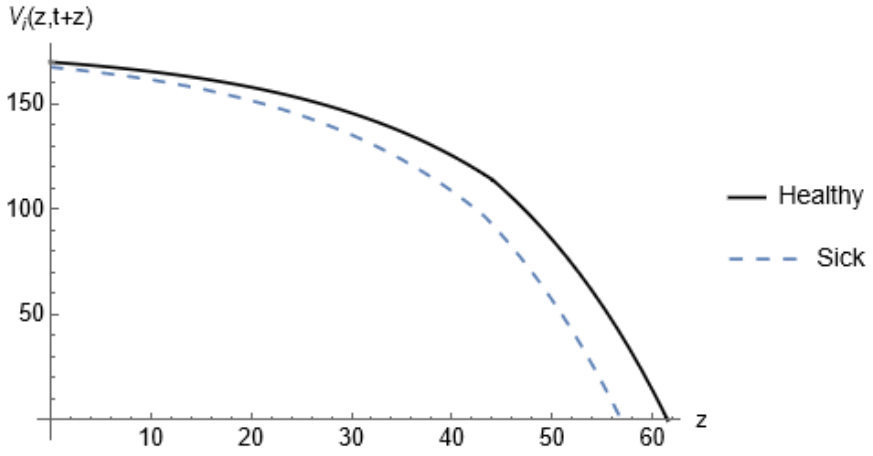


(a) Consumption

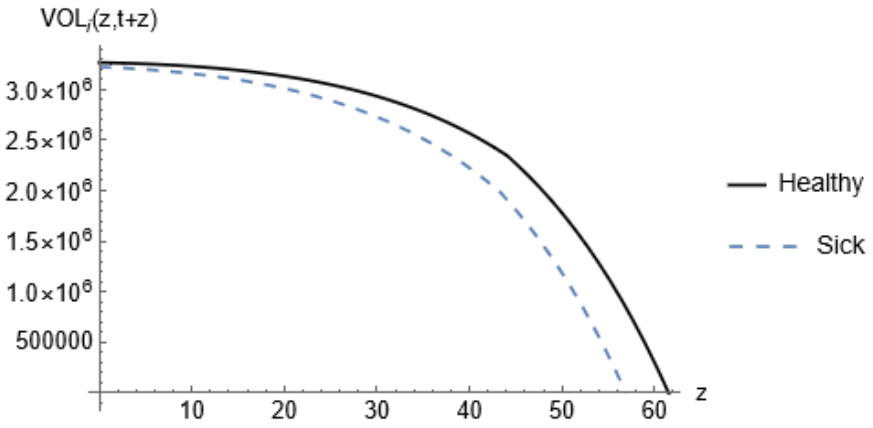


(b) Consumption of NHS Services

Figure 3: Lifetime Path of Consumption and Consumption of NHS Services

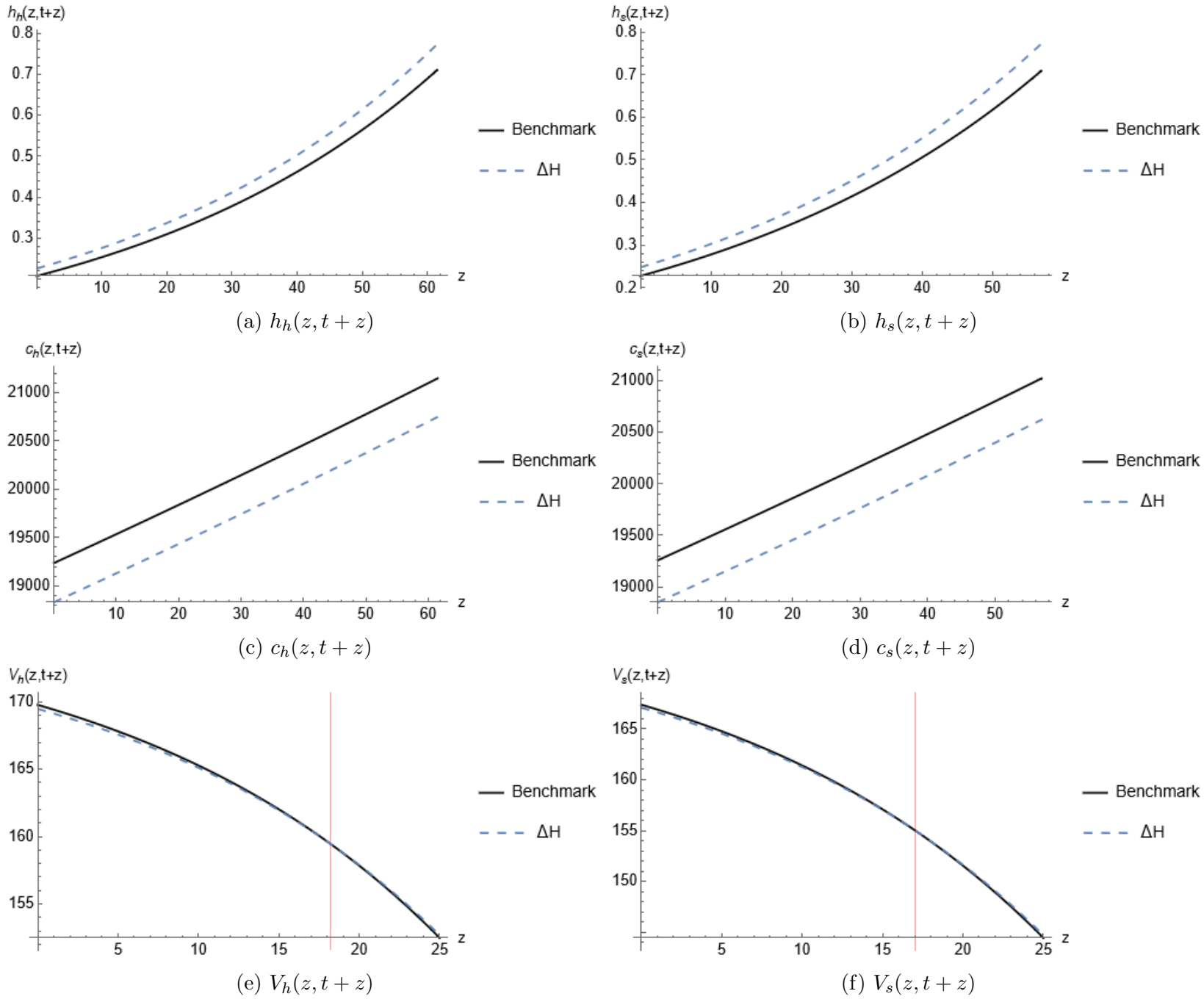


(a) Value Function



(b) Value-of-Life

Figure 4: Lifetime Path of the $V_i(z, t+z)$ and $VOL_i(z, t+z)$



(a) $h_h(z, t+z)$

(b) $h_s(z, t+z)$

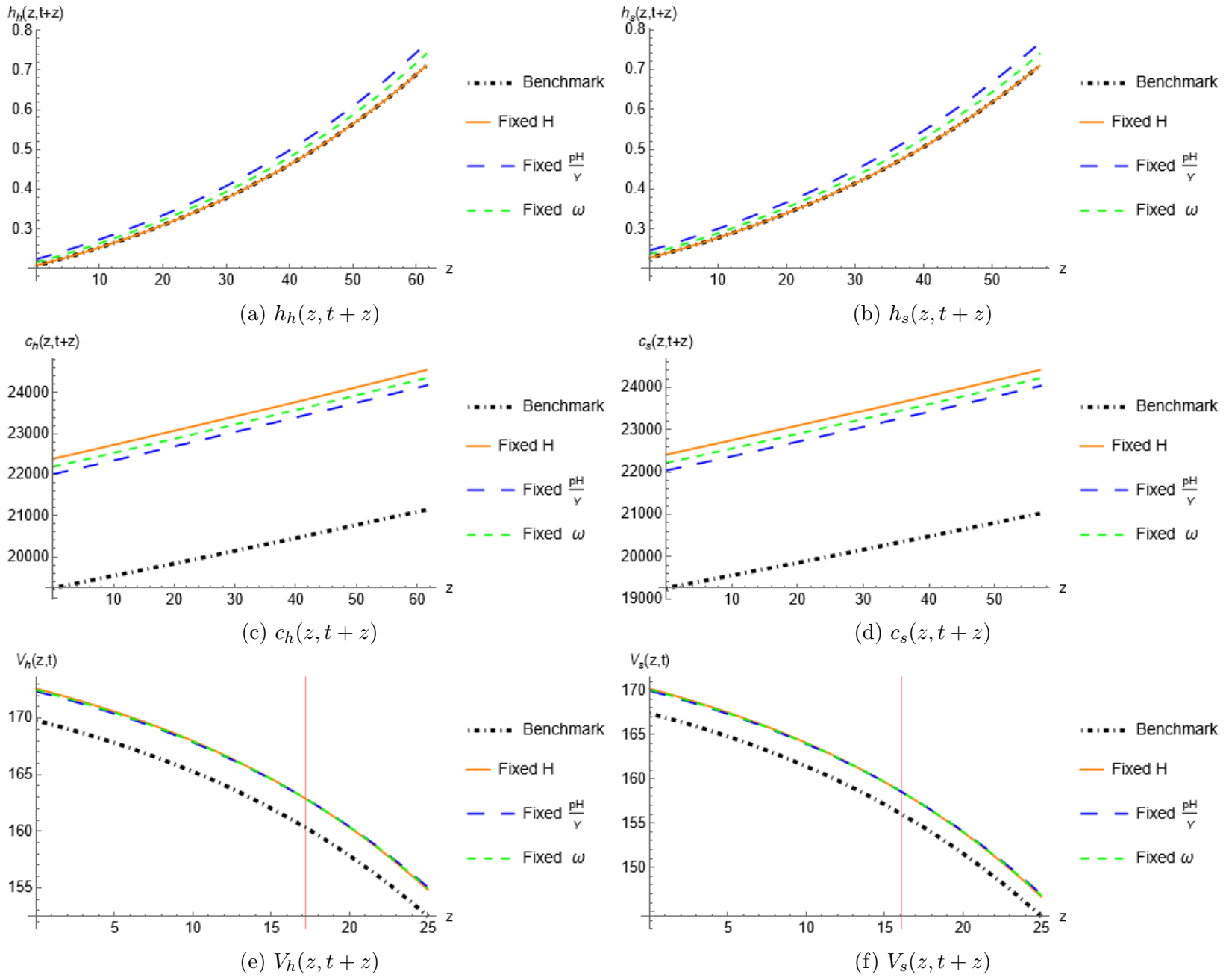
(c) $c_h(z, t+z)$

(d) $c_s(z, t+z)$

(e) $V_h(z, t+z)$

(f) $V_s(z, t+z)$

Figure 5: 10% Increase of $\bar{H}$

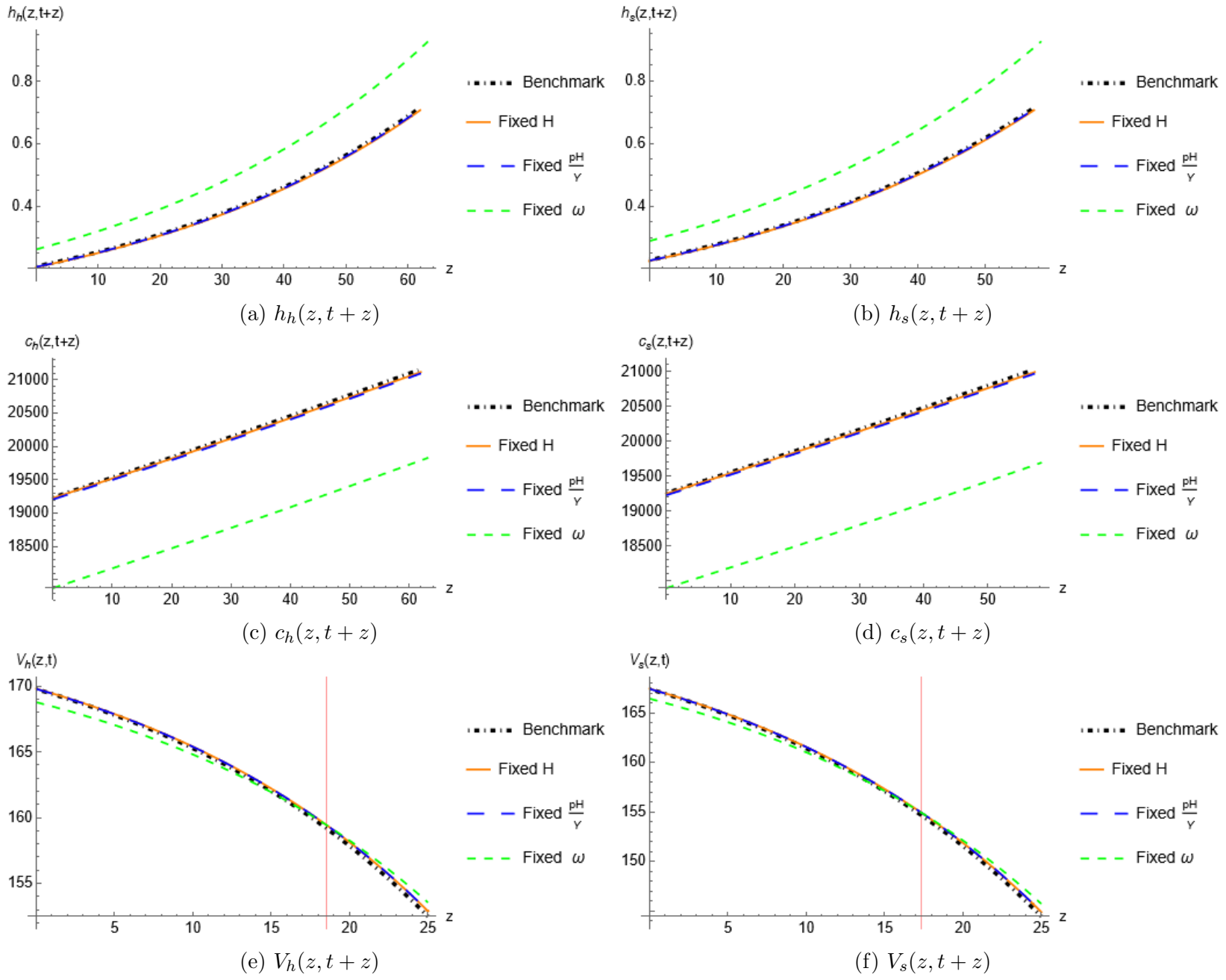


(a) $h_h(z, t+z)$ (b) $h_s(z, t+z)$

(c) $c_h(z, t+z)$ (d) $c_s(z, t+z)$

(e) $V_h(z, t+z)$ (f) $V_s(z, t+z)$

Figure 6: 10% Increase of Z



(a) $h_h(z, t+z)$ (b) $h_s(z, t+z)$

(c) $c_h(z, t+z)$ (d) $c_s(z, t+z)$

(e) $V_h(z, t+z)$ (f) $V_s(z, t+z)$

Figure 7: 10% Increase of A