Congestion in a Public Health Service: A Macro Approach*

Mark Kelly,†Michael Kuhn‡

March 26, 2020

Abstract

Health care services in the UK (and other public health care systems) are provided, mostly free of charge at the point of service, by the central government via the National Health Service (NHS). NHS services are therefore allocated according to a queuing rule, with total NHS expenditures constrained by a predetermined budget. Since there is no price mechanism for allocating NHS services, an increase in the demand for health care services requires that the government either increase funding to the NHS or allow patient waiting times to increase. In this study, we develop a continuous-time lifecycle model with a realistic aging process in the spirit of Dalgaard-Strulik and publicly-provided health care services. Health care slows the rate of aging and is subject to congestion, which lowers its efficacy. Households optimally determine how much time to devote to consuming NHS services given the current wait time. We calibrate the model to match UK data from 2007-2016 and analyze the steady-state, general equilibrium response of the model to various shocks to the economy. Our analysis suggests that the optimal NHS response to an increase in the demand for health care depends strongly on what is driving the demand increase (i.e. income growth or medical progress).

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

†Baylor University, United States

‡Wittgenstein Centre (Univ. Vienna, IIASA, ÖAW/VID) and Vienna Institute of Demography, Austria.

1 Introduction

Between 1990-2014 the average annual growth rate of the output share of healthcare among OECD nations was approximately 1.34%. Moving forward, the rapid rise in the utilization of government-financed healthcare 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 healthcare 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 healthcare by fixing the prices that suppliers can charge for the services they can provide.

Resource rationing, on the other hand, is slightly less common and is most effective in healthcare systems where the government is a major provider of healthcare services. In this case, the government is attempting to indirectly limit the cost of healthcare by limiting the supply of healthcare. 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). The present study focuses on a rationing regime whereby the policy-maker decides on a given level of supply of health care services (as measured in total hours of treatment available) and allocates it, in case of a demand that exceeds the fixed supply, via a waiting list, such that the effective utilization (hours demanded net of waiting time) equals the supply. By lowering the effectiveness of the demand for health care, and raising its time cost to the individual, waiting thus serves as mechanism to contain demand. By maintaining a larger capacity within the health care system, the policy-maker can to some extent control waiting times. This is well-illustrated for the English National Health Service (NHS), where the development of waiting times in NHS hospitals over the time span 1999-2017 is following a trend that is broadly opposite to the trend in the NHS spending share (see figure 1).

[Insert Figure 1 here]

In a different vein, 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 will be using waiting time data for the calibration of our model, our general notion of congestion is broader and includes other forms of rationing, such as reductions in consultation times or reductions in the quality of care.

Considering the macroeconomic efficiency of the public provision of health care services subject to congestion, three key issues emerge: (i) Which mechanisms determine the allocation of health care and its outcomes through waiting and what are the macroeconomic repercussions which arise from funding and from health-related changes in longevity and labor supply? (ii) What constitutes an efficient level of public supply if (a) this determines waiting times and, thus, the effectiveness of medical care, both directly and indirectly through affecting individual demand; and (b) the timely access to health care bears on 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 in respect to the public provision of health services and, more specifically, what rules – targeting fiscal sustainability, i.e. 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, Fonseca et al. forthcoming, Jones 2016, Böhm et al. 2018, Frankovic and Kuhn 2018, 2019)?

In order to analyze these issues, we develop a theoretical model of a public health care system in which the government is the sole provider of healthcare services. We assume that the government supplies a predetermined level of medical capacity which is financed out of an earnings tax. We further assume that the economy is composed of a continuum of finitely-lived individuals born into distinct birth cohorts. Each individual derives utility from

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

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

Our key results suggest that for our calibration, a 10% expansion of the NHS is improving welfare 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

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

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. Unsurprisingly both productivity growth and medical progress contribute to an increase in welfare. However, the overall impact of these shocks hinges on the policy target that accompanies it. We find that aiming for the fiscal target tends to boost the welfare gain in the presence of productivity growth, whereas aiming for a waiting list target tends to boost the welfare gain in the presence of medical progress.

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

effects. Furthermore, the lack of an explicit modeling of production and factor markets does not allow for the analysis of general equilibrium repercussions, which again we show to be of relevance.

A second paper that is relatively closely in line with our study, if only as it is also based on a calibration for the English NHS, is Böhm et al. (2018). This study features a general equilibrium model with overlapping generations of individuals who are subject to deficit accumulation and examines how the rationing of public health care bears on welfare if public health care investments induce medical innovations. While our model lacks much of the dynamics present in Böhm et al. (2018) it provides a more thorough modeling of waiting/congestion, which is not an issue in Böhm et al. (2018). Other papers featuring a public health care sector are Kuhn and Prettner (2016) who study the role of publicly provided health care in the presence of R&D-driven economic growth and Grossmann and Strulik (2019) who employ a model of deficit accumulation to study the role of social security reform in Germany.

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

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

2 The Model

2.1 Individuals

We assume that NHS services are free to the agent at the point of delivery. Consequently, the NHS will have to rely on queuing to allocate its services, implying that agents will have to spend some time on a wait list before they become eligible to consume healthcare services. For simplicity, we assume that wait times are uniform across the economy, so that each agent's current wait time $\omega(z,t)$ is equal to the product of $\widehat{\omega}(t)$, the (average) propensity to wait at time t (to be defined below), and $m(z,t)$, the agent's total time devoted to health care,[2] which includes both the current amount of time that the agent devotes to consuming NHS services and the time they spend on a NHS wait list.

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 agent's total number of accumulated deficits reaches some maximum survivable threshold $\bar{D}$. Let $D(z,t)$ be the agent's total number of accumulated health deficits at age z and time t.[3] The deficit accumulation function is defined as

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

Deficits accumulate at the natural rate μ. We allow for the existence of exogenous environmental factors that reduce (or increase) the deficit accumulation rate. These factors are captured by the parameter a. The effect of the agents' health investment on the deficit accumulation rate is described by the function $f(m(z,t))$. Finally, we denote by T the

[2]We also refer to $m(z,t)$ as the individual's demand for health care throughout this study.
[3]Note the implied relationship to the birth year $t_0 = t - z$.

individual's longevity, as defined by the identity $D(T,t) = \bar{D}$.

The health investment function takes the following functional form

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

where $M(t)$ is effective aggregate demand for health care (i.e. $M(t)$ is the sum of $m(z,t)$) and $H(t)$ is the exogenous supply of NHS services.[4] By definition, total consumption of NHS services is constrained by $H(t)$, implying that $H(t) \leq M(t)$.[5] Whenever aggregate effective demand exceeds NHS supply, the average propensity to wait $\widehat{\omega}(t)$, 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 intended to be consistent with the fact that for many life threatening conditions time to treatment is a significant determinant of survival, with survival probabilities declining sharply over time. The parameter ϵ governs the returns to timely treatment. When $\epsilon = 1$, the effectiveness of health care is directly proportional to the current level of congestion in the NHS. At the other extreme, when $\epsilon = 0$, the efficacy of health care is unaffected by NHS congestion.[6]

Before moving on, we briefly digress to consider two important points about how we

[4] $H(t)$ can be thought of as the total supply of "bed hours" provided by the NHS, while $M(t)$ is the sum of $H(t)$ (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.

[5] Indeed, we will show below that $H(t) < M(t)$ must hold in equilibrium. See footnote 9 bleow.

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

have chosen to model waiting in our model: (i) While our modeling of $\widehat{\omega}(t)$ is internally consistent and while it can be calibrated to the data, it may not be so easy to reconcile this with a full model of waiting time dynamics, such as presented e.g. in Siciliani (2008). A more accurate modeling of the waiting time dynamics would add a second and delayed dynamic process in our model, where individuals demand health care at time t but receive it only after some time $t+\omega(z,t)$. We circumvent the complexity involved by conflating the whole time allocation process into a single period, a year say, where in a congested health care system, the individual may simply require more time to access a given level of health care. 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).

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

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

where $c(z,t)$ is current consumption and $l(z,t)$ is a binary variable equaling one when the individual is in the labor force and zero when they exit. Assuming that $R(t)$ is the individual's optimal retirement age,[7] $l(z,t)$ will be equal to

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

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

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

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

where capital pays a return equal to the real interest rate $r\,(t)$ and depreciates at the rate δ, and where working individuals are paid a wage $w\,(t)$, which is taxed at the rate $\tau\,(t)$.

Individuals are, thus, assumed to maximize lifetime utility

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

with ρ denoting the rate of time preference, by choosing $c(z,t)$, $m(z,t)$, and $R\,(t)$ subject to equations (1) and (5).

The Hamiltonian function (dropping the age-time indicator z and t for brevity) that characterizes the agent's optimal decision problem is given as

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

where λ^D and λ^k are the costate variables on the stock of deficits and capital, respectively. The resulting first-order conditions are[8]

$$c^{-\sigma} = \lambda^k \tag{8}$$

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

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

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

$$\begin{aligned} & u\left(T\right) - \lambda^D \mu \left(\bar{D} - a - A\left(\frac{H}{M}\right)^{\gamma(1+\epsilon)} m\left(T\right)^{\gamma}\right) - \lambda^k c\left(T\right) \\ = \; & u\left(T\right) + \frac{\theta \widehat{\omega} m\left(T\right)^{1-\gamma}}{\gamma A}\left(\frac{H}{M}\right)^{-\gamma(1+\epsilon)}\left(\bar{D} - a\right) + \frac{\theta \widehat{\omega} m\left(T\right)}{\gamma} - c\left(T\right)^{1-\sigma} = 0 \end{aligned} \tag{11}$$

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

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

Apart from the two features (i) that individuals face a direct utility loss rather than a monetary loss from investing in health and (ii) that the effectiveness of health care depends on the average propensity for waiting, our solution mirrors the one developed in Dalgaard and Strulik (2014, 2017). Equation (8) is the standard outcome, equating the marginal utility of consumption to the marginal cost of consumption. Equation (9) describes the agent's optimal health investment decision. This condition implies that the optimal allocation of time to health at age z is set so that the marginal product of health care is equal to the marginal disutility from waiting.[9] Agents will engage in the labor market as long as the

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

marginal benefit of working exceeds the disutility of labor, implying that the endogenous retirement age R coincides with the age at which the after-tax earnings, weighted by the marginal utility of consumption, is equal to the marginal disutility from labor (see equation (10)). Finally, equation (11) determines optimal longevity.

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

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

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

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

This equation implies that the lifetime path of m is positively correlated with the rate of time preference, ρ.[10] Note that as $\rho \to 0$, g_m becomes negative, implying that the optimal health investment strategy for perfectly patient agents 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 $\rho > \mu$, $g_m > 0$, the agent is sufficiently impatient and their optimal strategy flips, with the agent opting to push the undesirable cost of queuing off until late in life when their health is poorer. Similarly, the growth rate of health investment is inversely affected by the degree of diminishing returns γ. If diminishing returns set in slowly (i.e. γ is close to one) and $\rho > \mu$, then the agents optimal g_m will be relatively large. Put differently, if the rate of diminishing returns to health investment is small and agents are relatively impatient, then the agent can afford to delay investing

[10]Note here the difference to the dynamics in Dalgaard and Strulik (2014, 2016), where the interest rate, r, shows up instead of the rate of time preference, ρ. This difference follows, as in our model, the consumption of health care merely takes up time and time has a pure utility value.

heavily in their health until late in life so that initial health investment will be relatively low and then will increase rapidly throughout the agent's lifetime.

2.2 Production

The economy consists of two sectors: a private final goods sector, in which firms produce under perfect competition; and a public health care sector where total output is decided by the government. Both sectors employ labor and capital from competitive markets. Turning to the final goods sector more specifically, we assume that there is a single representative firm. The final good $Y(t)$ is produced according to a Cobb-Douglas technology

$$Y(t) = Z(t) K_Y(t)^{\alpha} N_Y(t)^{1-\alpha},$$

where $Z(t)$ is total factor productivity, $K_Y(t)$ is the aggregate capital stock, $N_Y(t)$ is the aggregate labor supply in the final goods sector, and where α is the output elasticity with respect to capital. The representative firm operates under perfect competition, implying that the real interest rate[11] and the wage rate are equal to the marginal product of capital and labor respectively:

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

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

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

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

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

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

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

For simplicity, we assume that labor and capital are homogeneous across sectors, so that NHS workers and capital are paid the equilibrium wage rate and interest rate from the final goods sector. Rearranging the NHS' zero-profit condition to solve for $p(t)$ and substituting $\bar{H}(t)$ for $H(t)$ yields

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

Taking $p(t)$ as given, cost minimization by the NHS implies that the NHS will choose capital according to the following rule

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

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

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

2.3 Aggregation and Market Clearance

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

$$\dot{K}(t) = (r(t) - \delta)K(t) + (1 - \tau(t))w(t)N(t) - C(t). \tag{18}$$

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

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

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

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

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

$$\tau(t)\,w(t)\,N(t) = G(t) + p(t)\bar{H}.$$

For simplicity, we assume that public consumption is equal to a fixed fraction ν of current output, so that $G(t) = \nu Y(t)$. Noting that in competitive equilibrium with a neoclassical production function we have $Y(t) = r(t)K_Y(t) + w(t)N_Y(t)$, substituting for $G(t)$ and $K_H(t)$ in the government's budget constraint and inserting the resulting expression into (18)

allows us to rewrite the aggregate capital accumulation function to

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

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

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

Noting that labor and capital market clearance requires that

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

and

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

completes the equilibrium description of our economy.

3 Data and Calibration

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

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

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

In calibrating the average wait time and utilization of NHS services we rely on the Hospital Episode Statistics (HES) database published by the NHS. This database provides detailed information on all admissions to NHS hospitals in England, including aggregate estimates of the average wait time and length of stay per episode at NHS hospitals. We use the average length of stay as our measure of the average provision of NHS services per person (H/T in the model). This equates to approximately 1.5 bed days per person (equivalent to 0.4% of the agent's time endowment). Average wait time per episode was 52.79 days, implying that the aggregate average waiting time was 14.89 days (4.02% of the time endowment).

Table 1: Deficit Accumulation Parameters

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

The majority of the parameters from the deficit accumulation function (1) are taken

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

Table 2: Fixed Parameters (Calibration)

Description	Notation	Value
Rate of time preference	ρ	0.05
Intertemporal elasticity of substitution	$1/\sigma$	$1.0\bar{1}$
Disutility from waiting time	θ	26.7
Disutility from labor	η	1.227
Medical effectiveness	A	0.0115
Returns to timely treatment	ϵ	0.5
Environmental parameter	a	0.0163
Aggregate supply of NHS services	$\bar{H}$	0.25
Productivity parameter (NHS)	B	0.000725
Capital share (NHS)	β	0.2
EOS between capital and labor (NHS)	$\frac{1}{1-\xi}$	1.163
Productivity parameter (final goods)	Z	2080
Capital share (final goods)	α	0.25
Depreciation rate of capital	δ	0.04
Output share of government expenditures	ν	0.135

The remainder of the parameters are listed in table 2 above. These parameters are set to calibrate the equilibrium to match the sample averages from the data. The rate of time preference (ρ) and the intertemporal elasticity of substitution ($1/\sigma$) are chosen to calibrate private consumption. We set $\rho = 0.05$, close to Dalgaard and Strulik's (2014) choice of 0.06. Our choice of $\sigma = 0.99$ implies an intertemporal elasticity of substitution (IES) of 1.01. This is consistent with Chetty (2006), whose meta-analysis of numerous labor supply studies

supports the conclusion that the coefficient of relative risk aversion is approximately equal to one.

The disutility parameters θ and η are set at 26.7 and 1.19 respectively. These values were chosen in order to match the steady-state average waiting time and retirement age in the model to the observed per capita wait time and the average age at retirement in the data. The aggregate supply of NHS services (H) directly affects both the average wait time and the average utilization of NHS services in the economy. Since we assume that the size of new birth cohorts is constant and normalized to one, the terminal age T also represents the total population of the economy. And, given that we match T to the observed life expectancy at age 20 (i.e. 61.06) and $\bar{h} = H/T = 0.4\%$ by definition, we assume that $H = 0.25$. Following Acemoglu and Guerrieri (2008), the capital share in the health care sector β is set at 0.2. We assume a value of 0.5 for ϵ. Taking H, β, and ϵ as given, we choose B and ξ to calibrate the relative supply of NHS workers and the healthcare output share (pH/Y) to the data.

Following the literature, we set the capital share in the final goods sector $\alpha = 0.25$. 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. Finally, the government expenditure output share ν equals 0.135 and is used to aid in the calibration of the consumption, healthcare, and investment output shares.

Table 3: Model vs. Data

	Data	Model
gdp	£26,160.20	£28,291.90
c	£16,373.20	£18,797.60
w	£28,359.20	£28,277.50
h	0.40%	0.41%
ωm	4.02%	4.02%
C/GDP	62.61%	66.44%
pH/GDP	8.75%	8.35%
N_H/N	4.35%	4.51%
R	43.99	43.98
T	61.06	61.06

The calibrated benchmark steady-state is compared with the sample averages from the data. As table 3 demonstrates, the model closely matches the data for most variables of interest. The average utilization of NHS services (h) and wait time (ωm) in the model equal their counterparts in the data. Likewise, the NHS employment and expenditure shares are close matches with the data. We successfully calibrate the wage rate to the data, but the model over-predicts consumption and GDP per capita. The steady-state retirement and terminal ages are equivalent to the sample averages. Lastly, we obtain a value of life[12] in the order of £3.7 Million from our model. This is well within the bounds of the estimates based on a range of international studies, as summarized in Viscusi and Aldy (2003).

4 Numerical Experiments

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

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

The first experiment captures the impact of an expansion of NHS capacity, as is continuously and widely debated in the UK (e.g. O'Dowd 2016). The second and third experiments embrace the impact of productivity growth and medical progress as two of the well-known drivers of health care expenditures and life-time expansion (e.g. Hall and Jones 2007; Fonseca et al. forthcoming; Böhm et al. 2018; Frankovic and Kuhn 2018, 2019). All experiments are based on a comparison of the underlying steady states.

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

[12] $VOL = \int_0^T [e^{-\rho z} u(c(z,t), m(z,t), l(z,t), D(z,t), R(t))] dz / u_c(c(0,t), m(0,t), l(0,t), D(0,t), R(t))$.

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

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

(iii) Maintain a constant mean wait time, i.e. $\widehat{\omega} = \overline{\omega}$.

Note that, in principle, there are three targets for the policy, maker: NHS capacity, the NHS expenditure to output ratio, and average waiting time. In scenario (i), the policy-maker is assumed not to respond to income growth and/or medical progress, by holding the NHS capacity constant and leaving the expenditure to output ratio and the waiting time free to adjust. In scenario (ii), the policy maker is assumed to target a fixed medical expenditure to output 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 policy maker is assumed to target a fixed waiting time by way of appropriate adjustments to the NHS capacity. In this case, it is the expenditure to output ratio which is left free.

We now proceed to discussing the outcomes of the main experiments 1-3 in turn. Tables 4-6 report the outcomes as percentage changes for a range of key variables: per capita income y; per capita consumption c; the wage rate w; the tax rate τ; NHS supply H; per capita demand for health care (measured in gross time) m; average propensity for waiting (the share of gross health care time, an individual spends waiting) $\hat{\omega}$; the consumption share in output C/Y; the NHS to output ratio pH/Y; the NHS labor share, N_h/N; retirement age R; longevity T; life-cycle utility $V(0,t)$ and instantaneous aggregate welfare $\Omega(t)$, where $V(0,t)$ is the agent's value function as described by equation (6) and $\Omega(t) = \int_0^T u(c(z,t), m(z,t), l(z,t))dz$.

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

gdp	c	w	τ	m	$\hat{\omega}$	p
-0.38	-1.68	0.00	3.95	1.83	-0.78	0.00

C/Y	pH/Y	N_H/N	R	T	V	Ω
-1.31	10.06	10.53	-0.49	0.32	0.02	0.30

ΔVOL: -1.59%

Experiment 1, as summarized in table 4, shows that a 10% increase in NHS supply $\bar{H}$ increases welfare, as measured by V and Ω, by a modest amount, despite raising the tax rate by 3.95% and imposing a drag on per capita income and per capita consumption. Our result mirrors earlier findings by Kuhn and Prettner (2016), Jones (2016), Böhm et al. (2018), Frankovic and Kuhn (2018), and Fonseca et al. (forthcoming) who all show that within "rich" (or growing) economies the willingness to pay for life-expanding health care, as measured by the value of life, is so high as to warrant an expansion of the health car sector even at the expense of economic performance. In the context of this study, our finding indicates that NHS capacity is under-supplied from a welfare perspective. A number of features of the underlying adjustments following an increase in NHS capacity are worth noting. First, surprisingly perhaps, the expansion of NHS capacity by 10% actually raises the time share devoted to waiting, despite the 0.78% decline in the average wait time. This is because it triggers an increase in the demand for health care both at the intensive margin, as measured by m, and at the extensive margin, as measured by the expansion of the population at highest ages due to the increase in longevity T.

Second, the expansion of the NHS leads to a more than proportional increase in the health care output and employment ratios. The former reflects the reduction in final goods production, while the latter results from the fact that the capital share is lower in the health care sector than in the final goods sector. Third, due to the sizable increase to the tax rate, labor supply, as measured by the retirement age, declines by nearly 0.5%. Fourth, the

expansion of the health care sector has only a minor positive impact on life-cycle utility (0.02%), but a slightly larger positive impact on aggregate welfare (0.25%). This implies that older generations benefit from NHS expansion by more than younger generations who are relatively healthier, bear the burden of the tax increase, and are heavily discounting their future utility flows from early retirement and additional lifespan.

Table 5: 10% Increase in Z

	y	c	w	τ	m	$\hat{\omega}$	p
1. Fixed $\bar{H}$	13.93	15.07	13.53	-2.98	0.14	0.01	5.06
2. Fixed pH/Y	13.62	13.60	13.53	0.06	1.69	-0.65	5.05
3. Fixed $\hat{\omega}$	13.93	15.03	13.53	-2.92	0.17	0.00	5.06

	C/Y	pH/Y	N_H/N	R	T	V	Ω
1. Fixed $\bar{H}$	0.99	-7.80	-9.20	0.64	0.00	0.14	0.14
2. Fixed pH/Y	-0.02	0.00	-1.20	0.29	0.28	0.16	0.40
3. Fixed $\hat{\omega}$	0.97	-7.62	-9.03	0.63	0.01	0.14	0.15

ΔVOL: 1. 14.90%; 2. 13.52%; 3. 14.87%

Experiment 2, as summarized in table 5, shows that a 10% increase in factor productivity in final goods production, Z, which is tantamount to an economic growth impulse, leads to a sizable welfare gain of 0.14% for households in the benchmark setting in which NHS capacity is held constant. The economic growth impulse is magnified in the benchmark setting by an expansion of labor supply due to the reduction in the NHS tax. The latter is feasible due to the expansion of the tax base. Notably, the welfare gain arises predominantly from an increase in per capita final goods production and consumption of nearly 14% and in excess of 15%, respectively. Since capcity is held constant, average consumption of NHS services does not change, leaving life expectancy unaltered. This is all the more striking as the increase in the value of life by nearly 15% indicates a strong willingness to pay for health care. However, for a capacity constrained NHS this willingness to pay does not boost demand by much, as individuals anticipate that an elevated level of overall demand will only serve to raise waiting times, thereby curbing the effectiveness of health care.

This situation is not much improved upon were the health policy to be changed to holding waiting times constant in scenario 3. Indeed, even in the benchmark, waiting time does not increase by much following the economic growth impulse due to what might be considered a "voluntary" restraint of demand in a capacity constrained health care system. Therefore, counteracting the modest increase by a slight expansion of health care capacity does not result in a significant change in welfare. The more appropriate policy in this setting amounts to maintaining a constant medical expenditure to income ratio in scenario 2, which is equivalent to holding the NHS tax rate constant. The additional tax receipts in the presence of income growth allow for a sizable expansion of NHS supply by 8.15% which triggers simultaneously a reduction in the average waiting time by some 0.65% and a boost in individual demand for health care by 1.69%. Overall, this translates into a longevity gain by 0.28% (the equivalent of approximately 62 additional days) and a modest increase in labor supply (due to the postponement of retirement). Despite the smaller increase in the labor supply as compared to the benchmark, the expansion of the health care sector will be mostly "self-financing." As a result, the income tax rate will be mostly unaffected, implying that income and consumption grow by only 0.68 and 1.53 percentage points less than in the benchmark. Overall, the sizable increase in longevity increases the aggregate welfare gain by 0.22 percentage points from 0.13% to 0.35%, nearly three times more than in the benchmark. This outcome is consistent with the finding of Hall and Jones (2007) that it is optimal for the health care spending share to increase with income growth and suggests that a policy that would not only aim at maintaining the NHS spending share but rather at raising it (to some extent) would yield an even larger increase in welfare.

Table 6: 10% Increase in A

	y	c	w	τ	m	$\hat{\omega}$	p
1. Fixed $\bar{H}$	0.13	0.58	-0.02	-1.30	9.24	1.16	0.04
2. Fixed pH/Y	0.01	0.03	-0.02	0.00	9.89	0.92	0.03
3. Fixed $\hat{\omega}$	-0.48	-2.13	-0.02	5.06	12.33	0.00	0.03

	C/Y	pH/Y	N_H/N	R	T	V	Ω
1. Fixed $\bar{H}$	0.45	-3.34	-3.46	3.66	3.35	0.40	3.25
2. Fixed pH/Y	0.01	0.00	0.01	3.53	3.49	0.41	3.38
3. Fixed $\hat{\omega}$	-1.66	12.90	13.52	2.95	3.98	0.44	3.84

ΔVOL: 1. 1.05%; 2. 0.53%; 3. -1.51%

Experiment 3, as summarized in table 6, shows that a 10% increase in medical effectiveness in curbing deficits, A, leads to a sizable welfare gain of 0.37% (for individuals) and 2.99% (aggregate) in the benchmark setting in which NHS capacity is held constant. In this case, the welfare gain predominantly flows from the expansion of longevity by 3.09% (equivalent to nearly 2 additional years of life). The ensuing increase in the retirement age and, thus, in labor supply by 3.55% contributes to modest growth in income by 0.32% and consumption by 0.63%. More notably, the hike to medical effectiveness triggers a substantial increase in the demand for health care by 8.54%, which for a constant capacity, boosts the average wait time by 1.08%. The resulting loss in medical effectiveness from increased wait times suggests that a significant part of medical progress is neutralized through the increase in congestion.

Interestingly, in this case an expansion of NHS capacity by 3.47% that would secure a constant medical expenditure to output ratio adds only 0.1 percentage points to the aggregate welfare gain. This outcome is likely driven by the smaller increase in the average wait time under this policy which increases the longevity gain by a further 0.1 percentage points. In the presence of medical progress a policy that aims at containing waiting times turns out to be much more effective and yields an additional welfare gain of 0.45 percentage points relative to the benchmark. This is due in large part to the additional 0.47 percentage point

increase to the longevity gains following the 11.11% boost to the average utilization of health care. This policy requires a large-scale expansion of NHS supply (16.81%), financed through a tax increase by some 4.78%. This tax increase reduces the labor supply growth by 1.39 percentage points relative to the benchmark. Consequently, average income and consumption will actually fall by 0.85% and 1.96% respectively.

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

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

- Whether the government should pursue a fiscal target (i.e. maintain the NHS expenditure to output share, and a constant tax) as opposed to a waiting time target depends on the dominant type of technical progress - general productivity growth as opposed to medical progress. If productivity growth dominates and would, for a constant NHS supply, lead to a strong reduction in the health share and only a modest increase in waiting, it is more effective to focus on the fiscal target, i.e. maintain the health share.

This is because the implied reduction in waiting yields an additional boost to the effectiveness and demand for medical care. If, in contrast, medical progress leads to a sharp increase in waiting times but only a modest decline in the health share, it is more effective to focus on the waiting time target. This is because the commitment to lower waiting time tends to boost the demand of health care that allows individuals to fully participate in the gains from medical progress.

- In all scenarios, it turns out, that the macreoeconomic feedback from changes in longevity on the retirement age and, thus, on labor supply constitutes a quantitatively strong general equilibrium effect, which tends to compensate for a large part the negative general equilibrium effects that arise from the tax increases that are warranted by an expansion of the public health service. This result is subject to the caveat that with our abstraction from a public pension system we have modeled the strongest possible case for savings and labor supply to increase with longevity, which leaves us with the more general insight that the macroeconomic repercussions in such a model yield strong quantitative effects. We conclude by noting that while the introduction of a public pension system may, thus, qualify some of our quantitative results, we do not envisage it to have a bearing on the qualitative insights summarized above.

5 Conclusions

We have considered the macroeconomic effects of congestion within a public health care system. In its purist 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 policy-maker, but it builds up or diminishes based on individual decisions on the utilization of health care. Thus, the policy-maker only has limited control through the choice of health care capacity.

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

Calibrating our model to reflect the English NHS and economy over the time frame 2007-2016, we first study an increase in NHS capacity and find that although it tends to lower per capita income, the resulting gain in longevity is more than compensating the reduction in consumption and generates a welfare gain. Strikingly, although the capacity increase reduces the average wait time, average total time waiting increases due to a more than proportional increase in the demand for health care both at the intensive and extensive margins. 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 welfare impact of productivity growth, whereas a waiting list target tends to boost the welfare impact of medical progress. Indeed, the superiority of the respective rules arises from their particular impact on the expectation of consumers with respect to the development of the waiting list and their consequent demand for health care. Accordingly, we find that the preferred policy in response to technology shocks that increases the demand for health care will always be the policy that provides the greatest flexibility for NHS capacity to respond the increase in demand for health care.

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.

6 References

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

Bloom, David E., David Canning, Richard K. Mansfield and Michael Moore (2007). "Demographic change, social security systems, and savings." *Journal of Monetary Economics 54*, 92-114.

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

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

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

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

Dalgaard, Carl-John 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.

Fonseca, Raquel, Pierre-Carl Michaud, Titus J. Galama and Arie Kapteyn (forthcoming). "On the rise of health spending and longevity." *Journal of the European Economic Association*.

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, in press*.

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

Grossmann, Volker and Holger Strulik (2019). "Optimal social insurance and health inequality." *German Economic Review, in press*.

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.

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

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

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

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

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

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

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

Schneider, Maik T., Christian P. Traeger and Ralph Winkler (2012). "Trading off generations: equity, discounting, and climate change." *European Economic Review 56*, 1621-1644.

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.

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(1)*, 5-76.

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

7 Appendix

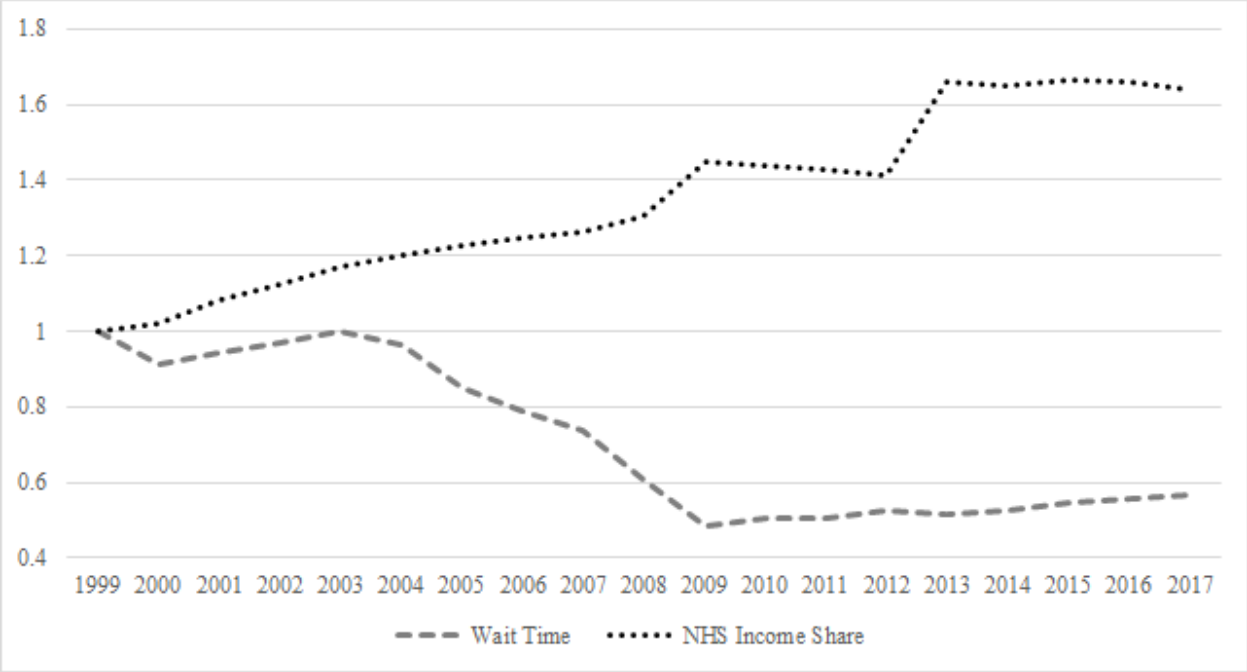

Figure 1: NHS waiting times and NHS Income Share 1999-2017. Source: UK Office of National Statistics (ONS).