THE UNIVERSITY of York

Discussion Papers in Economics

No. 2007/23

Modelling the Dynamics of a Public Health Care System: Evidence from Time-Series Data

By

Fabrizio Iacone, Steve Martin, Luigi Siciliani, Peter C Smith

Department of Economics and Related Studies
University of York
Heslington
York, YO10 5DD

Modelling the dynamics of a public health care system: evidence from time-series data

Fabrizio Iacone* Steve Martin† Luigi Siciliani‡ Peter C. Smith§

12 June 2007

Abstract

The English National Health Service was established in 1948, and has therefore yielded some long time series data on health system performance. Waiting times for inpatient care have been a persistent cause of policy concern since the creation of the NHS. This paper develops a theoretical model of the dynamic interaction between key indicators of health system performance. It then investigates empirically the relationship between hospital activity, waiting times and population characteristics using aggregate time-series data for the NHS over the period 1952–2005. Structural Vector Auto-Regression suggests that in the long run: a) higher activity is associated with lower waiting times (elasticity = -0.9%); b) a higher proportion of old population is associated with higher waiting times (elasticity = 1.6%). In the short run, higher lagged waiting time leads to higher activity (elasticity = 0.2%). We also find that shocks in waiting times are countered by higher activity, so the effect is only temporary, while shocks in activity have a permanent effect. We conclude that policies to reduce waiting times should focus on initiatives that increase hospital activity.

Keywords: Waiting times, Dynamics, Vector Auto-Regression. *JEL classification*: I11, I18, H42, H52.

*Department of Economics and Related Studies, University of York, Heslington, York, YO10 5DD. E-mail: fi501@york.ac.uk.

†Department of Economics and Related Studies, University of York, Heslington, York, YO10 5DD. E-mail: sdm1@york.ac.uk.

‡*Corresponding author.* Department of Economics and Related Studies, and Centre for Health Economics, University of York, Heslington, York, YO10 5DD; C.E.P.R., 90-98 Goswell Street, London EC1V 7DB. E-mail: ls24@york.ac.uk.

§Centre for Health Economics, and Department of Economics and Related Studies, University of York, Heslington, York, YO10 5DD. E-mail pcs1@york.ac.uk. We thank Andrew Jones, Nigel Rice, Richard Cookson and other participants at Health Economics Seminar at University of York (HEDG) for helpful suggestions and discussions.

1 Introduction

Waiting times and waiting lists for health care have been a persistent phenomenon in the UK National Health Service (NHS) since its inception in 1948. Waiting times have more recently become a major health policy issue also in other OECD countries like Australia, Canada, Denmark, Italy, Finland, the Netherlands, Norway, Portugal, Spain and New Zealand (Siciliani and Hurst, 2005). Average waiting times for common procedures, such as hip and knee replacement, cataract surgery or varicose veins, vary from three to eight months. However, it is only in the UK that waiting lists figures have been recorded for a period as long as fifty years. In this study we exploit for the first time these data in order to analyse interactions between the average waiting times patients wait to receive treatment, the volume of activity provided by hospitals, and population demographic characteristics.

When it was created in 1948, the NHS effectively 'nationalized' a previously somewhat anarchic network of hospitals and other providers that were owned and run by local governments, charities and other not-for-profit institutions. These hospitals were placed under the management of local Hospital Management Committees answerable to the national government. The new NHS sought to offer comprehensive health care to all citizens, free of charge. Most of the funding for the NHS came from national taxation, and individual hospitals were funded through annual budgets from the regional offices of the national ministry.

Primary care physicians (general practitioners) remained outside the salaried NHS, and were instead offered a national contract to care for NHS patients. Almost all general practitioners (GPs) accepted this contract, as the scope for private practice was now limited. The NHS gave GPs an important 'gatekeeping' role. Every citizen had to register with a GP, and (except for emergency treatment) none could gain access to a hospital specialist without a referral from his or her GP.

These arrangements gave rise to three potentially lengthy waits for the patient: the time between the GP referral and the appointment with the specialist; the time waiting for the results of any clinical tests and investigations requested by the specialist; and

the wait for receipt of inpatient treatment once a decision had been reached that such treatment was required.

Although the NHS has been subject to many structural reforms since 1948, the founding principles have remained in place largely unchanged in the ensuing sixty years. The major change of note relevant to this paper was the split between local purchasers and providers introduced in 1991. Geographically defined local purchasers, in the form of district health authorities, were given budgets to care for their populations. NHS providers were split from these authorities and given separated boards of management. They were then expected to compete for business from the health authorities in an 'internal market' of hospital provision. Since 2000 the NHS provider market has been gradually augmented to embrace a range of other not-for-profit and for-profit hospitals (Rivett, 1998).

The NHS inherited a waiting list of over 400,000 patients waiting for hospital inpatient treatment in 1948. This waiting list refers to the third of the patient waits described above: the wait for hospital admission once the need for treatment had been agreed. Ever since 1948, this waiting list - and the associated waiting times for treatment - have been a stubborn feature of the NHS that has often become a focus of intense national political debate and controversy. Furthermore, the NHS has collected annual data on waiting lists in a broadly consistent format throughout its existence. It is therefore possible to track the size of the waiting list over time.

On their own, waiting list data are not directly helpful in indicating the real concern of patients: the length of time they have to wait for treatment. However, it is possible to calculate a robust indicator of the expected wait for treatment by dividing the size of the waiting list on an annual census date by the number of hospital admissions in the year. This gives a measure of the 'time to clear the waiting list'. Whilst not the same as the actual waiting time experienced by patients (this has not been collected over the entire NHS lifetime), it is a good proxy that we use throughout this paper, effectively adjusting the magnitude of the waiting list for current levels of supply.

A signal of the long-standing political concern with NHS waiting has been the large number of initiatives launched by the government to reduce the waiting time for treat-

ment. These include: periodic injections of special finance directed at hospitals with especially long waits; a 1991 'Patient's Charter' that guaranteed all patients treatment within two years (soon reduced to 18 months); experiments with a range of patient choice models, designed to encourage patients to seek out providers who offer short waits; and most recently a set of high-profile hospital targets that have since 2001 brought down the maximum waiting time to six months.

The political concern with waiting lists is matched by a concern with the levels of expenditure of the NHS. The national government sets an annual budget for the NHS in the light of national economic circumstances, pressure for public expenditure in other public services, and demand pressures within the NHS. Waiting lists are an important signal of those pressures, and are also politically important in themselves. A key role of the government is therefore to balance the interests of taxpayers and patients, as expressed in the budget it sets for the NHS.

The optimal balance between waiting and activity might be affected by numerous uncertainties and shocks, such as changes in national income, demand side shocks and changes in health care technologies and productivity. It is possible to propose a variety of causal models of how activity and waiting times interact. For example, activity and budget changes might be affected in the light of shocks to the waiting time. But equally, the duration of the waiting time might in time be expected to respond to changes in activity and budgets.

The long time series of data from the NHS offers a unique resource with which to explore the long run dynamics of waiting times. This paper applies general time series methods such as Structural Vector Auto-Regression (SVAR) to investigate the long-run and short-run dynamics of NHS waiting times. More precisely, using aggregate time-series data for the English NHS over the period 1952–2005, we investigate the relationship between activity, waiting times and the proportion of people aged 65 and over (a measure of demographic pressures). The results suggest that in the long run: a) higher activity is associated with lower waiting times (elasticity = -0.9%); b) a higher proportion of the older population is associated with higher waiting times (elasticity = 1.6%). In the short run, positive shocks in waiting times lead to higher activity (elasticity = 0.2%).

We also find that while shocks in waiting times have temporary effect, shocks in activity have permanent effects.

We first give a brief review of the literature in section 2, and present a simple theoretical framework in section 3. The empirical methods (the structural vector auto-regression approach) are then described in section 4. The scope of the data we can use are constrained by the need for availability over the entire time period. They are described in section 5. The results are provided in section 6. Finally, we draw some conclusions on the scientific and policy implications of this study in section 7.

2 Literature review

Economists have traditionally viewed hospital waiting times as a non-monetary rationing mechanism that reconciles limited supply with less limited demand for surgery (Lindsay and Feigenbaum, 1984; Cullis, Jones and Propper, 2002). Over the past decade several studies have sought to estimate models with waiting time affecting both the demand for and supply of elective care. These studies normally use as a unit of observation either a local geographical area (such as an electoral ward or district health authority) or a health care provider (hospital). They are conducted either as static (one-period) cross-sectional analyses or panel-data analyses (albeit using at most seven years of data). The characteristics of these studies therefore contrast markedly with those of this paper, where the analysis is undertaken at the national level with an annual time series of data stretching back over 50 years to the formation of the NHS.

Previous theoretical studies of the relationship between waits and activity, model the demand side on a utility maximising consumer who faces a choice between either delayed access to treatment in the NHS free of charge, or immediate treatment at a financial cost in the private sector. Supply side models are typically based on a utility maximising hospital manager whose utility depends positively on the achievement of waiting time performance (Lindsay and Feigenbaum, 1984; Martin and Smith, 1999; Gravelle, Smith and Xavier, 2003; Siciliani, 2006).

Depending on data availability, some supply models assume that queue length has

reached an equilibrium, so that observed demand (additions to the waiting list) equals supply (the number of inpatient admissions) (Martin and Smith, 1999; Siciliani, 2005). In this equilibrium model, the theoretical impact of waiting time on supply is positive.

Other studies have access to both demand side data (additions to the waiting list) and supply side data (hospital admission data), and so do not have to assume equilibrium (Gravelle, Smith, and Xavier, 2003; Martin, Rice, Jacobs and Smith, 2003). It is then possible to specify a model in which the hospital manager is concerned about waiting time (or list) performance at the end of the current period. This can be forecasted as a function of the waiting time at the beginning of the period, and additions to and removals from the list in the current period. Increased activity improves the manager's end of period performance but induces greater demand in the following period, making it more difficult to hit future waiting times targets. The impact of waiting time on supply is ambiguous, depending on how managers discount future utility.

Early evidence on NHS demand elasticities with respect to waiting time is provided by Goddard and Tavakoli (1998) who estimate a demand function for NHS treatment using panel data for 15 Scottish Health Board areas over 12 quarterly observations (1990-92) for six specialties. They model the number of additions to the waiting list as a function of the expected waiting time for NHS treatment. They find some dynamic effects, necessitating the inclusion of a lagged explanatory waiting time variable, which they suggest indicates the existence of a partial adjustment from one quarter to the next. Elasticities exhibit the anticipated negative sign for all six specialties, ranging from -0.017 for general surgery to -0.096 for orthopaedic surgery.

Martin and Smith (1999) estimate a similar demand model based on a more extensive dataset. This empirical analysis is based on population data for all the routine surgical specialties for 1991-92 and employes about 5,000 electoral wards as the unit of analysis. The model assumes equilibrium in supply and demand and yields an elasticity of inpatient demand with respect to the waiting time of about -0.21. In a subsequent study, Martin and Smith (2003) apply the same model to a panel of seven years' data. The elasticity of demand with respect to the waiting time for all routine surgery is estimated as -0.23 when the demand model is estimated in first difference form. Analogous models

for individual specialties yield demand elasticities of -0.17 for general surgery, -0.14 for orthopaedics, and -0.17 for ophthalmology.

Studies by Gravelle, Dusheiko and Sutton (2002), Gravelle, Smith and Xavier (2003) and Martin, Jacobs, Rice and Smith (2007) apply the same type of demand model to panel data. They do not need to assume equilibrium between demand and supply, as separate data on admissions from and additions to the list are available. Static and dynamic models can therefore be estimated for a number of specialties, and using a number of estimation methods. Demand elasticities with respect to waiting time are broadly in line with previous studies, in the range -0.1 to -0.2.

On the *supply* side, there are fewer estimates of the responsiveness of inpatient supply to waiting times. Using an equilibrium model, Martin and Smith (1999) report an elasticity of supply with respect to waiting time for all routine surgical specialties of 2.93. A later dynamic specification yields an elasticity of 5.29 (Martin and Smith, 2003). These estimates are rather large and subsequent studies in which the assumption of equilibrium is dropped, yield markedly lower elasticities, in the range 0.07 to 0.18 (Martin, Rice, Jacobs and Smith, 2007).

Where separate supply and demand data are available, Gravelle, Smith and Xavier (2003) estimate an elasticity of supply of 0.083 with respect to mean waiting time, a similar figure to the one reported by Martin, Jacobs, Rice and Smith (2003). In a study of a single hospital in Scotland over the period 1997-2001, Windmeijer, Gravelle and Hoonhout (2005) report a slightly higher positive elasticity of overnight inpatient admissions with respect to waiting times (0.40) and a more modest response for day cases (with an elasticity of 0.13).

3 Theoretical framework

The dynamic behaviour of waiting times can be modelled in a variety of ways. In this section we develop a very simple but quite general model to help motivate the subsequent empirical work. Define z_t, w_t and s_t respectively as activity, waiting time

and the proportion of older people at time t. The demand for care at time t is

$$D(w_t, s_t) + u_t^d$$

where u_t^d denotes a shock on the demand. We assume that: a) demand is decreasing in waiting time ($D_w < 0$): the higher the wait, the higher is the number of patients who opt for the private sector or give up the treatment (Lindsay and Feigenbaum, 1984; Martin and Smith, 1999);[1] b) demand is increasing in the proportion of older people ($D_s > 0$): we use the proportion of older people s as a proxy for medical needs in the population. Technological development is also likely to increase demand, but it is likely to be captured by the same variable s as this is a trending variable.

Waiting times act as a non-monetary price which helps to bring the demand for and the supply of heath care in equilibrium. We assume that the market for health care does not clear instantaneously, so that an excess demand in one period increases waiting time in the following period, while an excess supply reduces the waiting time. The speed of adjustment of waiting times is denoted with θ. Analytically,

$$w_{t+1} - w_t = \theta \left(D(w_t, s_t) + u_t^d - z_t \right) \tag{1}$$

We assume that the dynamics of the proportion of elderly people, our proxy of medical needs in the population, is given by:

$$s_{t+1} = s_t + c^s + u_t^s \tag{2}$$

where c^s is a positive constant and u_t^s is a shock.

Finally, we assume that activity evolves over time according to:

$$z_{t+1} = z_t + c^z + \gamma \left(w_t - w_{t-1} \right) + u_t^z \tag{3}$$

[1] We provide a reduced-form specification of the demand function. A more general specification which models the choice of individual patients in terms of different alternatives (public treatment, private treatment and no treatment) is possible but would make the presentation more complex without providing any additional insights.

where c^z is a positive constant, u_t^z is a supply shock and γ denotes the responsiveness of activity to past variations in waiting times. This formulation implies that when policy makers observe an increase in waiting time in one period, they are more willing to fund increases in supply the subsequent period. Alternatively, providers might be willing to work harder when waiting times increase either because of altruism or because of financial or non-financial incentives attached to waiting-time targets. Supply shocks might be caused by changes in technology or efficiency in organization, following reforms in payment schemes for healthcare providers (for example a switch from fixed budgets to activity-based funding).

In the long-run equilibrium, waiting times do not vary over time ($dw_t = dw_{t-1} = 0$) so that demand for and supply of treatment have reached equilibrium ($D(w_t, s_t) = z_t$).

4 Methods

Our data naturally constitute a vector of time series, so we analyse them using the popular Structural Vector Auto-Regression (S-VAR) approach (Hamilton, 1994). To allow for the nonstationarity generated by potential unit roots, we distinguish between long and short-term dynamics. For a generic vector of n variables $y_t = [y_{1t}, y_{2t}, ..., y_{nt}]'$, we assume a representation

$$y_t = \sum_{j=1}^{p} \Phi_j y_{t-j} + u_t \tag{4}$$

where u_t is an independent, identically distributed sequence, with $E(u_t) = 0$, $E(u_t u_t') = \Omega$ (full rank), for a finite p (notice that Φ_j is a matrix with dimension $n \times p$). This is a generalisation to a vector of a standard AR(p) model, and it is known as VAR(p). We assume that the elements of y_t are subject to at most one unit root.[2] The model in (4) does not include any deterministic component, like an intercept or a time trend, for example. As such, y_t can be considered as a model for the deviations from a deterministic term, usually a mean or a trend. If the data are not stationary, then the Φ_j in (4) implicitly differentiates the data, so the introduction of the constant in (4) generates a

[2] In this case, an initial condition like $y_t = 0$ for all $t \leq 0$ is also specified.

linear trend in the levels y_t, and a linear trend in (4) generates a quadratic trend in the levels y_t.

The elements on the diagonals of Φ_j describe the dependence of each variable on its own past, while those off the diagonal describe the interaction with the past of the other variables. To disentangle and distinguish the different contributions, the dynamics of the VAR are expressed as a function of the original shocks u_t. A stationary y_t can be expressed in general as

$$y_t = \sum_{j=0}^{\infty} \Psi_j u_{t-j}. \tag{5}$$

Let $\Psi_{(kl)j}$ be the element in the k-th row, l-th column of Ψ_j: a plot of $\Psi_{(kl)j}$ against the lags (j) is known as *Impulse Response Function* (IRF). The matrices Ψ_j describe the effect of past shocks.[3]

Another useful tool is the *Forecast Error Variance Decomposition* (FEVD), which describes how much of the variation of each variable in y_t is generated by shocks of any variable: formally, indicating by $\widehat{y}_{t+d|t}$ the forecast of y_{t+d} made at time t, the FEVD is[4]

$$V\left(\widehat{y}_{t+d|t}\right) = \sum_{j=0}^{d-1} \Psi_j \Omega \Psi_j'. \tag{6}$$

If the multivariate process y_t is stationary, the shocks have only temporary effects, and the process reverts to zero (or, more in general, to the mean) over time. If y_t is not stationary, the shocks may have permanent effects as well. Even under nonstationarity, however, there may nevertheless be some non-trivial vectors a of dimension $n \times 1$ such that $a'y_t$ is stationary: in that case the deviations from the linear relation $a'y_t$ are only

[3] When Ω is a diagonal matrix, $\Psi_{(kl)j}$ is the effect of a shock to the l^{th} variable on the k^{th} variable j periods ahead when the interactions embedded in the model are taken into account. However, when Ω is not diagonal, shocks that affect y_{lt} influence y_{kt} as well, so a further hypothesis is necessary to attribute the simultaneous movement to u_{lt} or u_{gt}: this may be done by introducing the independent vectors $\varepsilon_t = \Omega^{-1/2} u_t$, where $\Omega^{-1/2}$ is the Choleski decomposition of $\Omega = \Omega^{1/2}\Omega^{1/2\prime}$, so that $E(\varepsilon_t) = 0$, $E(\varepsilon_t \varepsilon_t') = I$ (identity matrix). The elements of ε_t then identify shocks of each variable. Transforming (5) as $y_t = \sum_{j=0}^{\infty} \Upsilon_j \varepsilon_{t-j}$, where $\Upsilon_j = \Psi_j \Omega^{1/2}$, a plot of the matrices Υ_j against time gives the Orthogonalised Impulse Response Function. Notice that, contrary to the matrices Ψ_j, the matrices Υ_j depend on the ordering of the elements of u_t.

[4] As for the IRF, the decomposition is only in terms of u_t, so the individual shocks may potentially be correlated with each other: in order to attribute the shocks to one source only an identification assumption is necessary.

temporary, and $a'y_t$ emerges as a stable relation in the long run. The variables in y_t that have non-zero coefficient in $a'y_t$ are then co-integrated, and a is called a *cointegrating vector*. For an $n \times 1$ vector y_t, there may be h (such that $h < n$) non-trivial, linearly independent vectors a: these may be indexed as $a_1, ..., a_h$, and may be stacked in a $h \times n$ cointegrating matrix $A = [a_1, ..., a_h]$. h is then the rank of A and it is also known as the *cointegrating rank*.

Finding the h stable long-run relationships is of interest for their potential interpretation in terms of economic theory. But there are also statistical reasons to motivate the cointegration analysis: when y_t is not stationary, the estimate of the VAR in (4), of the IRF and the FEVD, are still consistent, but less efficient, unless integration and cointegration are properly taken into account.

Inference in a potentially cointegrated VAR is often done rewriting (4) as a Vector Error Correction Mechanism (VECM): under cointegration, any VAR(p) may indeed be expressed as

$$\Delta y_t = \sum_{j=1}^{p-1} \alpha_j \Delta y_{t-j} + BAy_{t-1} + u_t \tag{7}$$

for a $n \times h$ matrix B. In this representation, long and short-run dynamics are modeled separately, and the matrix B is the link between the two, because it expresses the effect of a deviation from the long term equilibrium Ay_{t-1} on the short term dynamics Δy_t. The matrices α_j then express the short-term interactions among the variables of interest.

When the cointegration rank h is known, simultaneous estimation of α_j, B and A, and inference in (7), can be obtained following Johansen (1991). When the cointegration rank is not known, it must be estimated in advance: Johansen (1991) shows that it is possible to test the null hypothesis that the cointegration rank is actually h against the alternative that it is $h + 1$ (maximum eigenvalue test) or against the alternative that it is n (trace test) (when $h = n$, the data are actually stationary). These tests however require at least some preliminary knowledge of a potential h: when this is not available, the tests are usually applied sequentially, starting from $h = 0$ and increasing the cointegration rank that is being tested as long as the null hypothesis is rejected.

5 Data

All of the data employed have been extracted from the Compendium of Health Statistics 2005-06 published by the Office of Health Economics (Yuen, 2005). Unless otherwise stated, the data cover the period 1952 to 2003 and relate to England and Wales.

Two indicators of the need for health care have been constructed: the all-age resident population (which increases from 44 million in 1952 to 51.8 million in 2003) and the percentage of the resident population aged 65 years and over (which rises from 11% in 1952 to 16% in 2003). Figure 1 also shows how the number of individuals aged 65 years and over (measured in millions) rises from 5.8 millions in 1952 to 10.6 in 2003.

As a measure of NHS activity we employ the number of discharges and deaths (for 1952 to 1986) and then the number of finished consultant episodes adjusted for multiple episodes within a single spell of care (from 1987 to 2003) both divided by the all-age resident population. This generates a measure of hospital activity per 1000 population which varies between 74.4 discharges in 1952 and 261.1 discharges in 2003. Figure 1 shows how NHS activity (in 1'000s) rises from 3,414 thousands in 1952 to 11,658 thousands in 2003.

Waiting time has been calculated as the number of patients on the waiting list (overnight and day cases) as at the annual census date divided by either the number of discharges and deaths (1952 -1986) or the number of finished consultant episodes adjusted for multiple episodes within a single spell of care (from 1987 to 2003). This provides an indicator of the "time to clear the waiting list", albeit only a proxy measure, as waiting list admissions comprise less than one half of all hospital admissions. Although the waiting list has increased since 1952 – rising from just under 500,000 to a peak of 1.25 million in 1997 – hospital capacity has increased at a faster rate so that the time to clear the list (the list divided by annual activity) has declined from 1.7 months

in 1952 to one month in 2003 (see Figure 1).[5]

[Figure 1 here]

Three indicators of the volume of NHS inputs are also employed: first, the average daily number of available NHS hospital beds across all specialties, which declines from 467,000 in 1951 to 198,000 in 2003; second, the number of medical doctors employed in NHS hospitals, which increases from 13,639 in 1951 to 73,761 in 2003; and third, the number of nursing and midwifery staff, which rises from 162,000 in 1951 to over 416,000 in 2003. All three variables were standardised by the size of the resident population so that in the analysis they track changes in physical inputs per 100,000 population. Average inpatient length of stay in NHS hospitals across all specialties has declined from 44 days in 1952 to 5 days in 2003.

Total NHS expenditure (available only at UK level) includes NHS charges paid by patients for prescription medicines and dental charges, and has been adjusted to constant (1949) prices using the GDP deflator. National income is measured as GDP at constant (1949) market prices.

6 Results

We have tried a number of specifications which included NHS expenditure, staff (doctors and nurses), beds and GDP, but we found that these variables did not have a significant role in explaining variations of waiting times over time. Also, they did not always lead to stable models. We therefore excluded them from the final specification. A more detailed discussion of the specification can be found at the end of this section.

In our preferred specification presented below, we analyse the dynamics and the

[5] Note that the use of day case surgery has become increasingly common for many procedures previously requiring inpatient treatment, and waits for day case treatment have been included in the NHS waiting list data since 1987. We estimated the relatively small number of patients awaiting day case treatment before 1987. This estimate was based on the number of day case admissions as a proportion of ordinary (overnight) admissions. The estimated number of day case patients awaiting admission was added to the official figure for inpatients awaiting overnight admission to derive a total awaiting admission figure (Martin, Jacobs, Rice and Smith, 2003).

interactions of waiting times w, hospital's activity z (measured as discharges per capita) and a demographic variable s (percentage of the population older than 65 years), over the years between 1952 and 2003.

All the variables are in logarithms, in order to reduce the risk of heteroskedasticity. Since our data are collected on a yearly basis, we have 51 observations for each time series: with such a small sample, the reliability of the estimations and tests, that are based on asymptotic theory, may sometimes be only approximative. We however feel that this problem is mitigated by the fact that the data refer to a long period of over fifty years: since the system has evolved a lot over such a long span of time, it should make it easier to detect the presence of long-term characteristics.

All tests assume a critical value of 5%. We consider a VAR augmented with a constant: since we suspect that each element of $y_t = [s_t, z_t, w_t]'$ is subject to a unit root, this corresponds to having a linear trend in the levels of the data.

We define the order of lags by estimating the VAR by OLS and then test if the last lag is not significant with a likelihood ratio test: this procedure selects three lags ($p = 3$). We also check that the residuals of the VAR(3) are not subject to heteroskedasticity and do not exhibit structural instability in the equations.

We then test for cointegration. The test statistic and its limit distribution depend on the nature of the deterministic component: given that the data exhibit a linear trend, we assume a model with a deterministic trend. Since we have no preliminary information on the potential number of cointegrating relations, we estimate the cointegration rank iterating the cointegration test starting from $h = 0$: summary statistics for the maximum eigenvalue and the trace tests are in Table 1; both tests indicate $h = 1$.

In this case, both A' and B are 3×1 vectors (recall that B expresses the effects of deviations from the long-term equilibrium Ay_{t-1}). Notice that vector A (the weights that characterize the long-term equilibrium) is only identified up to a scaling parameter, because if $A'y_t$ is stationary, then for any non-trivial scalar k, $kA'y_t$ is also stationary. Therefore, we normalise the estimated cointegrating relation for the waiting times. This is also in accordance with the estimate of vector B, which describes how the cointegrating errors affect the short-run dynamics. In our case, deviations from the long-run

equilibrium do not seem to affect the short-run dynamics of the old populations and the discharges: the coefficients that correspond to the weights of the population and of the discharge short-run dynamics, are jointly not significant (summary statistics of this test, including the restricted estimate of vector B are in Table 2).

The estimated cointegrating equation is, under this assumption,

$$\widehat{w}_t = 12.03 - \underset{(0.14)}{0.89} z_t + \underset{(0.30)}{1.64} s_t \tag{8}$$

Standard errors are in parenthesis. In the long run, the waiting time w is lower when the volume of discharges z is higher and is higher when the fraction of old population s is higher. Since the model is in logarithms, the coefficients in (8) can be interpreted as long-run elasticity. Therefore, a 1% increase in activity is associated in the long term with 0.9% reduction in waiting times. Equation (8) also suggests that a 1% increase in the older population is associated with 1.6% increase in waiting times: a larger proportion of older people is likely to increase need and demand, driving up waiting times.

Our theoretical model in section 3 suggests that in the long-run equilibrium we have $w_{t+1} - w_t = 0$ so that $D(w_t, s_t) = z_t$. After differentiation we obtain $\partial w = \frac{1}{D_w} \partial z$, which combined with the second coefficient in (8) provides $D_w = \frac{1}{-0.89} = -1.12$. This suggests a long-run demand elasticity of just above 1.1: an increase in waiting time of 1% implies a reduction in demand of 1.1%.

Similarly, $\partial w = \frac{D_s}{-D_w} \partial s$ which implies that the elasticity of demand with respect to the proportion of elderly people is $D_s = 1.64 * 1.12 = 1.83$: an increase in the proportion of older people of 1% implies an increase in demand of 1.8%.

Notice that although activity may respond to variations in waiting times in the short run (see equation (3)), this is not the case in the long run. Since by assumption in the long run $w_{t+1} - w_t = 0$, then the supply equation reduces to $z_{t+1} = z_t + c^z + u_t^z$.

The short-term dynamics can be analysed through the orthogonalised Impulse Response Function (IRF, see Figure 2) and the relative importance of the shocks in the

orthogonalised Forecast Error Variance Decomposition (FEVD, see Figure 3).

[Figure 2 here]

In Figure 2, s, z and w refer respectively to the proportion of older people, discharges and waiting times. Every diagram plots the response of each variable against itself and the other two variables for a time lag that goes from 1 to 10 years.

Given that we have three variables, we need three restrictions to identify the orthogonalised IRF. We assume that the proportion of the elderly s is not affected by either discharges z and waiting times w at least at the same time (first and second restriction), which is plausible: while we expect the proportion of older people to affect discharges and waiting times, the proportion of older people should be exogenous. Graphically, these identification restrictions are reflected in the second and third figure in the first row of Figure 2 where the estimated coefficient at $t = 1$ is equal to zero. We also assume that discharges do not react to waiting times simultaneously but only with a lag (third restriction, third figure in the second row of Figure 2 at $t = 1$). This may seem a more arbitrary restriction as we also may expect waiting times to affect discharges simultaneously. However, if we impose the alternative restriction that discharges have no simultaneous effects on waiting times, we still find that waiting time has no simultaneous effect on discharges. In contrast if we impose the restriction that waiting times have no simultaneous effect on discharges, discharges have a negative simultaneous effect on waiting times (see second figure in third row of Figure 2 at $t = 1$). Therefore, our chosen restriction appears empirically to be more appropriate.

The long-run relation of the cointegrating equation (8) can be observed by looking at the three diagrams in the last row of Figure 2. *Response of w to s*: the waiting time increases in the long run when the old population increases; and in *Response of w to z*: the waiting time decreases in the long run, when the discharges increase.

Another relevant result is that a shock that increases the waiting time is countered in the short term by an increase in activity (see *Response of z to w*), and that the waiting-time shock is quickly absorbed (see *Response of w to w*), after which the increase of

discharges reverts to zero. At $t = 2$ the elasticity of discharges with respect to waiting time is 0.2.[6] Once a waiting-time shock occurs, the providers react by raising activity to counter that shock so that the waiting time reverts to the mean within two or three years. As soon as the waiting-time shock is reabsorbed, activity also goes back to the original level.

While waiting-time shocks have only a temporary effect (see again *Response of w to w*), the effect of changes of discharges is permanent, as suggested by the *Response of z to z*: a shock on activity (for example due to a technological innovation or to a policy) has a permanent effect on activity so that activity settles to the new level (changes in policy that affect activity may indeed be permanent too).

Activity responds positively to the size of the older population in the long run, as suggested by the *Response of z to s*. Finally, the proportion of elderly population responds only to shocks of itself (which is consistent with our identification strategy) and these shocks have a permanent effect: when the proportion of the elderly increases it is unlikely to reduce afterwards.

All these results are also supported by the tests summarised in Table 3. The first column suggests that the short-term dynamics of the elderly population does not depend on activity and waiting time (test statistics are respectively 1.08 and 0.56 against a critical value of 5.99). The second column suggests that the short-term dynamics of activity does not depend on the dynamics of the elderly population but the positive effect of waiting time is statistically significant (test statistics are respectively 1.90 and 8.11). The third column suggests that the short-term dynamics of waiting time does not depend on the past short-term dynamics of the elderly population and activity (test statistics are respectively 1.84 and 3.88). The reaction of waiting times to the proportion of the elderly and activity is either instantaneous (as it is assumed for the orthogonalisation of the IRF and the FEVD) or as a form of adjustment to the long-run relationship.

[6] This is obtained as $\frac{\partial z_{t+1}/\partial u_t^d}{\partial w_{t+1}/\partial u_t^d}$.

[Figure 3 here]

The relative importance of the shocks can be analysed through the orthogonalised FEVD (see Figure 3). With the same identification of the IRF for the contemporaneous shocks, more than 50% of the variation of waiting time can be attributed in the long run to changes in the proportion of old population and to discharges (more than 20% and of 30% respectively). The reverse effect, from waiting time to discharges, is rather small.

The above analysis focuses on three variables only. This is mainly due to sample size which obliges us to keep the number of variables limited. In order to consider alternative specifications we also estimated and simulated a variety of non-nested models. Preliminary investigations in these models was based only on robust, possibly inefficient procedures. The most interesting result is that the effect of NHS expenditure on waiting times seems at most weak, and that NHS expenditure does not seem to react to waiting times. Indeed, funding constraints may have played a more important role in determining the amount of NHS expenditure, because NHS expenditure seems to be more convincingly linked to variations in GDP. The numbers of doctors, nurses and beds have no material effect on waiting times, nor do they react to it. The inclusion of length of stay in the basic model suggests that lower length of stay decreases waiting times in the long run (possibly due to the higher efficiency of healthcare providers) and that in the short run higher waiting times may reduce length of stay (providers work harder when waiting times are higher). However, the inclusion of this variable comes at the cost of inducing instability in the estimated model (in which case all the estimates would be altogether inconsistent), so we excluded length of stay from our final specification.

7 Conclusions

This study has examined the short-run and long-run relationships between certain important policy variables within the UK health system over a fifty year period. The relatively small number of observations means that we have had to be parsimonious in

the model specification. We found that the three most salient variables in modelling the dynamics of the NHS were waiting times, hospital activity and the demographic profile. Other variables like total NHS expenditure, the supply of hospital beds and medical and nursing staff did not seem to affect the dynamics of waiting times for treatment and were therefore excluded from the final specification.

Our favoured model includes a clearly exogenous demographic variable (proportion of elderly people) and the two endogenous variables, inpatient activity and waiting time. The role of the demographic variable is straightforward: other things equal, it increases waiting time in the long run (elasticity is 1.6%).

The most interesting policy findings relate to the interaction between activity and waiting time. In the long run, higher activity is associated with lower waiting times (elasticity is -0.9%), while in the short run, positive shocks in waiting times lead to higher activity (elasticity is 0.2%). We also find that while shocks in waiting times are likely to be temporary, shocks in activity are permanent. Furthermore, we find that alternative specifications of the model, in which NHS expenditure is substituted for activity, do not exhibit such effects, suggesting that specific initiatives to increase NHS activity are likely to be more successful in reducing waiting times than general injections of extra expenditure.

Compared to the existing literature, our implied demand elasticity (-1.12) is higher in absolute values than suggested by existing cross-sectional or panel-data studies, which find an elasticity between -0.1 and -0.2. Therefore, a further policy conclusion is that the reduction in waiting times arising from increases in supply may in the long run be smaller than expected from previous studies. One possible explanation for this result is that over time doctors might change their referral patterns. For example, when more resources are made available to the NHS, doctors might relax the severity threshold for referring patients for treatment, thereby muting the contribution of extra NHS resources to reductions in waiting times.

8 References

Cullis, P., Jones, J.G., Propper, C., 2000, Waiting and medical treatment: analyses and policies, Chapter 28 in A.J. Culyer, Newhouse, J.P. (eds), *North Holland Handbook on Health Economics*, Amsterdam: Elsevier.

Goddard, J.A., Tavakoli, M., 1998, Referral rates and waiting lists: some empirical evidence, *Journal of Health Economics,* 7: 545-549.

Gravelle, H., Dusheiko, M., Sutton, M., 2002, The demand for elective surgery in a public system: time and money prices in the UK National Health Service, *Journal of Health Economics,* 21: 423-49.

Gravelle, H., Smith, P.C., Xavier, A., 2003, Performance signals in the public sector: the case of health care, *Oxford Economic Papers*, 55: 81-103.

Hamilton, J., 1994, *Time series analysis*, Princeton University Press.

Johansen, S., 1991, Estimation and hypothesis testing of co-integrating vectors in Gaussian vector autoregressive models, *Econometrica,* 59: 1551-1580.

Lindsay, C.M., Feigenbaum, B., 1984, Rationing by waiting lists, *American Economic Review*, 74(3): 404-417.

Martin, S., Smith, P.C., 1999, Rationing by waiting lists: an empirical investigation, *Journal of Public Economics*, 71: 141-164.

Martin, S., Smith, P.C., 2003, Using panel methods to model waiting times for National Health Service surgery, *Journal of the Royal Statistical Society*, 166: Part 2, 1-19.

Martin, S., Jacobs, R., Rice, N., and Smith, P.C., 2003, The UK evidence on waiting for health care, Mimeo, CHE, University of York.

Martin, S., Rice, N., Jacobs, R., Smith, P. C., 2007, The Market for Elective Surgery: Joint estimation of supply and demand, *Journal of Health Economics*, 26(2): 263-285.

Rivett, G., 1998, From cradle to grave: fifty years of the NHS, King's Fund, London.

Siciliani, L., 2005, Does more choice reduce waiting times? *Health Economics*, 14(1): 17-23.

Siciliani, L., 2006, A dynamic model of supply in the presence of waiting times and waiting lists, *Journal of Health Economics*, 25(5): 891-907.

Siciliani, L., Hurst, J., 2005, Tackling excessive waiting times for elective surgery: a comparison of policies in twelve OECD countries, *Health Policy*, 72(2): 201-215.

Windmeijer, F., Gravelle, H., and Hoonhout, P. 2005, Waiting lists, waiting times, and admissions: an empirical analysis at hospital and general practice level, *Health Economics*, 14: 971-985.

Yuen, P., 2005, Compendium of Health Statistics 2005-06, Office of Health Economics, London.

Tables

Table 1: summary statistics for the maximum eigenvalue and for the trace tests

Hypothesized h	$\widehat{\lambda}$	$\sum\widehat{\omega}$	5% *c.v.*	$\widehat{\omega}$	5% *c.v.*
$h=0$	0.419734	37.67	29.68	27.21	20.97
$h=1$	0.140614	10.46	15.41	7.58	14.07
$h=2$	0.056027	2.88	3.76	2.88	3.76

Note: $\widehat{\lambda}$, estimated eigenvalue; $\widehat{\omega}$, Max-eigenvalue statistic;
$\sum\widehat{\omega}$, Trace statistic; 5% *c.v.* = 5% critical value
Trace test indicates 1 cointegrating equation
Max-eigenvalue test indicates 1 cointegrating equation.

Table 2: estimated restricted SVAR model

$$\Delta y_t = \widehat{c} + \sum_{j=1}^{2}\widehat{\alpha}_j\Delta y_{t-j} + \widehat{B}\widehat{A}y_{t-1} + \widehat{u}_t,\ y_t = [s_t, z_t, w_t]'$$

$\widehat{c}$	$\widehat{B}$	$\widehat{A}$
0.008661 (0.00381)	0^*	−1.637722 (0.29787)
0.012898 (0.00633)	0^*	0.892054 (0.14364)
−0.027191 (0.01845)	−0.440155 (0.08541)	$1^{\#}$

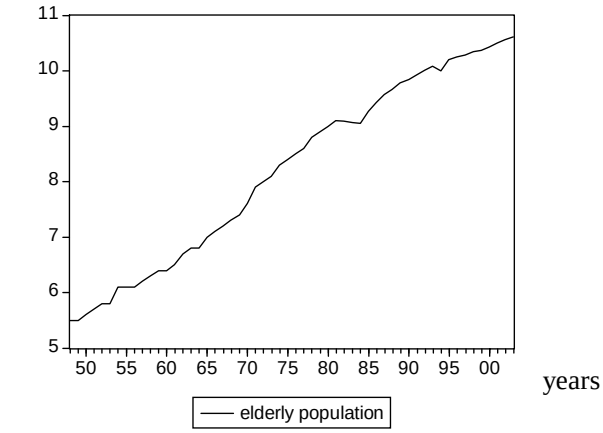
Note: standard errors in parenthesis;
* indicates the restrictions imposed; $^{\#}$ indicates the normalisation imposed.
LR test statistic: 3.93 (5% critical value, 5.99).

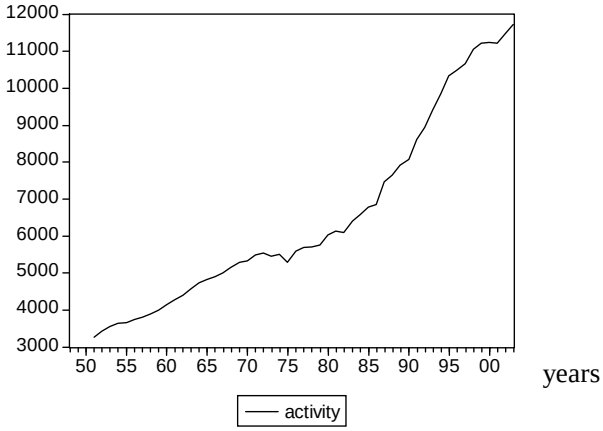
Table 3: tests on the short term dynamics (pairwise Granger Causality tests)

Test Statistic			
	Dependent variable		
Excluded	ds_t	dz_t	dw_t
ds_t		1.90	1.84
dz_t	1.08		3.88
dw_t	0.56	8.11	

Note: 5% critical value, 5.99.

Figure 1. Dynamics of elderly population (millions of individuals aged 65 or older), NHS activity (thousands of patients treated) and waiting time (months)





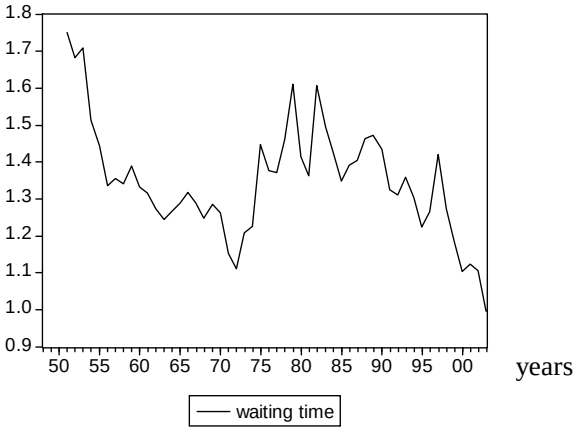
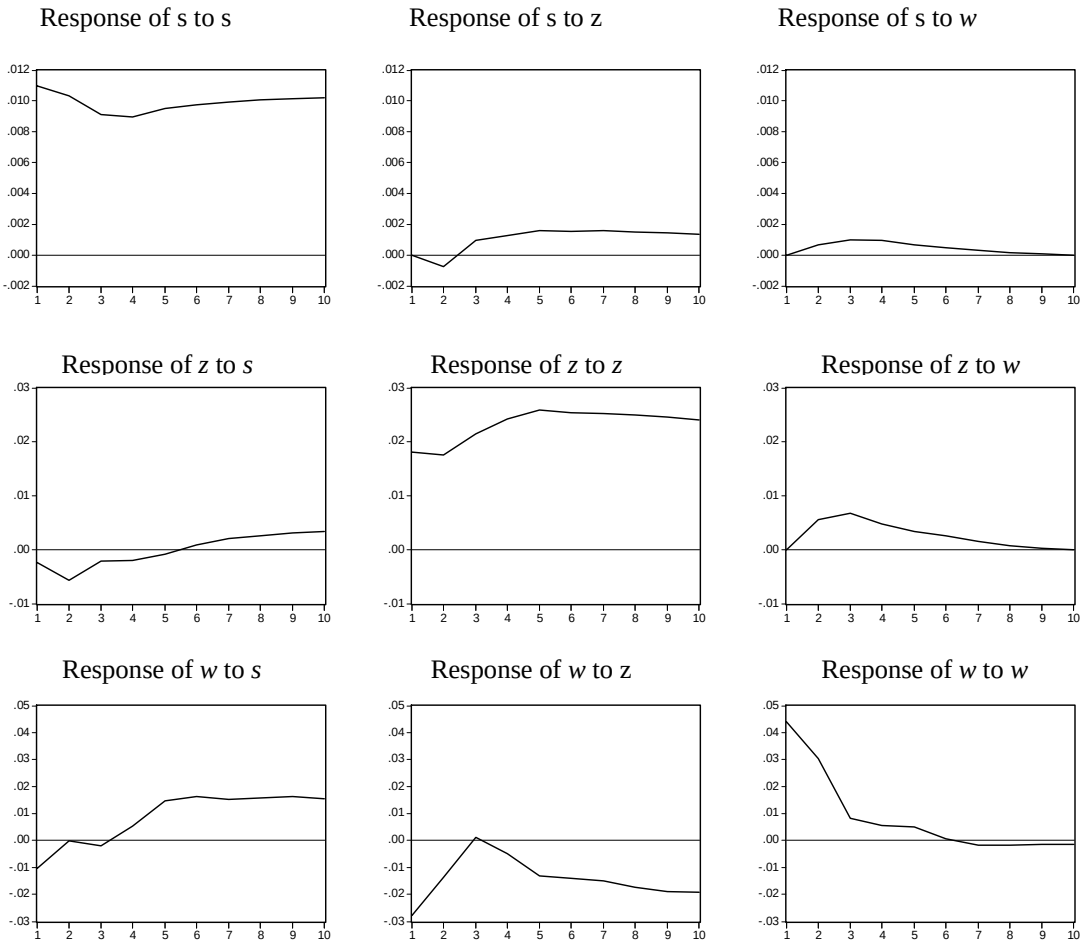


Figure 2. IRF when cointegration and restrictions are imposed



Note: s = proportion of elderly people; z = activity; w = waiting time.

Figure 3. FEVD when cointegration and restrictions are imposed