The Role of Waiting Time in a Prepaid Health Care System: Evidence from the British National Health Service

ROGER D. FELDMAN, PHLD.* and DAVID J. BALLARD†

I. Introduction

This paper estimates a model of demand for the services of general practitioner physicians (GPs) in the British National Health Service (NHS). A wide range of services is available from the NHS to any person free of charge at the time of use. Money prices have been replaced, except for nominal charges for prescriptions and appliances, by nonprice rationing methods such as queuing for hospital admissions and waiting for physicians’ office visits.

Nonprice rationing has been cited as a major difficulty of the NHS [1, p. 219]. While this problem has not surfaced in the United States to the same degree as in England, it is nevertheless instructive to study the British experience. About 40 percent of U.S. health expenditures in 1980 were paid by public sources [2] and, if national health insurance legislation is enacted here, the public share could rise toward the level observed in the nationalized British system.

In addition, some private health insurance plans in the U.S. have features that resemble the NHS, e.g., free care at the time of use. The NHS may thus provide a model for the analysis of these U.S. plans.

We examine the effect of office waiting time on the demand for GPs’ services. Using a utility-maximizing model of physicians’ behavior, we predict that waiting time should be directly related to exogenous demand pressures on the physician’s practice. Longer waiting time, in turn, should reduce both the number of visits per person and the number of patients in the physician’s practice (which the British call the physician’s “listsize”).

We specify and estimate an empirical version of the model with three equations: waiting time, visits per patient, and listsize. A key finding is that the elasticity of demand for GPs’ services with respect to waiting time is between −.48 and −.92, depending on our definition of variables and functional form. This result has important implications for national health insurance for outpatient visits (if such legislation is passed in the U.S.). If insurance were to reduce money prices to zero, demand would become relatively more sensitive to differences in time prices. In turn, as Achen has noted [3], this would permit persons with a lower opportunity cost of time to bid services away from those with a higher opportunity cost of time. National health insurance might, therefore, have important distributional effects, as well as the intended reduction of money price barriers to access.

Even without national health insurance, policy-makers in the U.S. might want to study the British National Health Service. Increasing numbers of Americans are joining Health Maintenance Organizations (HMOs), HMOs, which provide health services for a fixed prepaid fee to a voluntarily-enrolled
utility function to represent the possibility that waiting time is a source of disutility to physicians. Health economists commonly analyze physician’s values as consumption commodities. Evans, for example [6], included discretionary demand-shifting in the utility function, and Simon and Feldman [7] analyzed Evans’ model in detail, under the assumption that discretionary power is a “bad,” which physicians use only if they “must.” Here, we assume that waiting time is a “bad,” i.e., physicians would rather provide the optimal style of treatment without inconvenience to their patients.

Let the demand per patient in the physician’s list be

\[ V = V(W,Y) \]  

a standard demand function where waiting time decreases quantity demanded and a vector of exogenous variables \( V \) increases demand. The number of patients in the physician’s list is

\[ P = P(W,X) \]  

We assume that the partial derivatives are \( P_w < 0 \) and \( P_s > 0 \) (\( X \) is a vector of exogenous variables). By definition, total visits is the product of \( P \) times \( F \) as, in addition to patient visits, the physician has a fixed workload, \( F \) (filling out government forms may be a fixed workload), his or her total workload is

\[ L = V(W,Y) \cdot P(W,X) + F \]

The British GP is paid by a capitation system based on the number of patients on his or her list (average list size is about 2,500), plus a variety of allowances and item-of-service payments for drug dispensing, contraceptive counseling, immunizations, and other special services [8]. In 1971, the annual

population [4], have reduced the money price of outpatient care to zero for a significant number of Americans! Thus, waiting time in an HMO should, as in the NHS, be an important deterrent to demand. We are not aware of any systematic studies of this hypothesis, but impressionistic evidence suggests that “excessive” waiting time in HMOs is a major source of patient dissatisfaction [5].

As HMOs attempt to expand their market share against commercial insurance companies, they might examine how waiting time affects their ability to attract and hold customers. Evidence from our British data suggests that waiting time does not significantly reduce a GP’s list size. If this finding can be generalized to the HMOs’ “listsize” (its enrolled population), then, perhaps, waiting time will not be a significant obstacle to continued HMO growth.

In the next two sections we present a theoretical model of waiting time and estimate an empirical version of the model. Implications of the results will be discussed in the concluding section.

II. The Theoretical Model

We postulate that a representative physician maximizes a utility function,

\[ U = U(I,L,W) \]  

where: \( I \) = physician income with utility a positive function of income; \( L \) = physician workload with utility a negative function of workload; \( W \) = average patient waiting time, with utility a negative function of waiting time.

The first two arguments represent conventional wisdom that utility depends on income and work. Waiting time is included in the

\[ \text{income effect cannot be established. In what follows we assume that } \frac{dW}{dW} \text{ is zero, i.e., waiting time is a "superior good." This means that the physician who gets more nonlabour income increases waiting time to reduce workload, practice size and, by implication, earned income.}

Second, the effect of an increase in fixed practice costs is

\[ \frac{dW}{dF} \left( V(Y,F) + P(Y) + U_2 \right) \]

To make this derivative positive, it is sufficient to assume that \( U_2 > 0 \), which is positive, exceed the absolute value of \( U_{11} \), which is negative. The meaning of this assumption is that physicians who experience exogenous pressure on workload tend to institute nonprice rationing.

These first two derivatives are useful mainly because they appear in

\[ \frac{dW}{dW} \frac{dW}{dF} \]

where \( \Delta \) is the opportunity cost of time. To simplify the derivatives that follow, we assume \( \phi = 1 \).

\[ \text{Waiting time in a prepaid system} \]

\[ \text{The simple correlation coefficient between gross quarterly income and listsize was .91 in our survey sample.} \]
separable, i.e., $P_{i,j} = 0$, $P_{j,i} = 0$. Therefore, to establish the signs of $dW/dX$ and $dW/dY$ we have made the following sufficient assumptions: income effects are positive, and patient behavior functions are separable. Waiting time should then be directly related to exogenous demand pressures on the physician's practice, such as more patients wanting to join the list or more visits demanded by each patient.

The effect of a change in the capitation rate on waiting time is a final derivative of interest.\[ \frac{dW}{dC} = \frac{dW}{dG} \cdot \frac{dG}{dC} \]

The first term is assumed to be positive and the second can be given a positive sign. Therefore, higher capitation rates should increase waiting time, causing adverse effects on patient lists and services per patient.

In the next section we describe the British National Health Survey and a survey, conducted by the second author, of general practitioners in the NHS. This survey will be used to estimate an empirical specification of our model. The empirical model will then test hypotheses about the effects of waiting time on lists and per capita demand.

III. Empirical Model and Estimation

The British National Health Service (NHS) was established July 5, 1948 to "secure improvement in the physical and mental health of the people of England and Wales, and the prevention, diagnosis and treatment of illness" [11]. Responsibility for the day-to-day operation of the NHS rests with 205 Health Care Districts.

In the NHS, doctors practice either as specialists in the hospital system or as general practitioners. The GP consults with patients in his surgery (office), makes domiciliary visits, and in some catchment areas, the GP may have access to beds in community hospitals. The GP operates as an independent contractor: he is free to practice from privately owned premises; he has the right to employ his own staff and to accept private patients in addition to his NHS list; and is free from clinical interference.

The English attempt to achieve geographic redistribution of GPs through two mechanisms: by providing income bonuses for practicing in an area designated as "underdoctored," and by designating certain areas as "overdoctored areas" and restricting new practitioners to these. In 1968, the average number of patients per doctor in underdoctored areas was 2,919; in the average in restricted areas was 1,811 [12].

The patient may choose any GP, about 97 percent of the population is enfranchised to receive general medical services through the NHS [11, p. 3]. A wide range of health services is available to anyone free of charge at the time of use. Other than nominal charges for prescriptions and appliances, the NHS is financed almost entirely out of taxation.

The removal of financial barriers to access should increase patient waiting times and other forms of nonprice rationing. Evidence

A catchment area is a geographical subdivision of a hospital's catchment area based on socioeconomic, demographic, and socio-economic characteristics. Physicians' delight in non-institutional treatment may be necessary for a regular solution to the utility-maximization problem. For further discussion of this issue, see Shao and Feldman [7; especially footnote 3].

from Sweden and Quebec indicates that waiting time for an appointment increased tremendously following the introduction of NHI [13]. There has been no systematic study of the effect of waiting time on demand in the British NHS, despite the allegation that national health insurance is responsible for long delays in obtaining surgical procedures.[1]

Data for this study were collected by the second author in the West Dorset Health Care District, Dorset, England, from January to April, 1973. In addition to the data collection mentioned in the paper, the author was fortunate to observe the daily work of GPs over a six-week period. The field work included participation in over 300 surgery consultations and 100 domiciliary visits.

There were 86 GPs in West Dorset and the NHS maintains a confidential Personal Record A Form with information on the GPs' income and listsizes for each. Through a mailing from the Family Practitioners Committee offices in Ferndown, Dorset, the author asked for access to individual Personal Record A Forms for the first quarter of 1977. Following a second request by mail, a total of 38 GPs agreed to release this form.

The author also mailed a survey to the 86 West Dorset GPs, enclosing stamped return envelopes. After a follow-up letter to the non-respondents, 51 GPs participated in the survey designed to elicit information relating to practice workload. The office waiting time variable was taken from this survey.

The waiting lists in 1968 for plastic surgery and hernioplasty/adnexitomies were, respectively, 63 percent and 49 percent of the total annual hospital discharges and deaths for these services. The overall surgical waiting list, as a percent of deaths and discharges, was 20 percent [13].

Located centrally on the southern coast of England, West Dorset, 469 square miles in area with a population of 160,000, is one of the 205 Health Care Districts in the NHS.

Thirty-four GPs responded to the survey as well as granting Personal Record A Form disclosure. Those 34 GPs were compared with the other members of the GP population (those who either did not return the survey or did not reveal their income records) based on the FPC Index numbers. Three GPs with significantly lower-than-average listsizes were eliminated from the sample, after which standard t-tests revealed no mean differences between respondents and other members of the GP population on a number of common variables. The remaining 31 GPs constitute the sample used in this study.

The general stochastic specification of our model has three equations:

\[ W = W(F, F, X, Y, C, \bar{c}_i) \]
\[ P = P(W, \bar{X}, \bar{e}_i) \]
\[ V = V(W, \bar{Y}, \bar{e}_i) \]

where $c_i$, $\bar{c}_i$, $\bar{e}_i$ are random error terms, assumed to be independently normally distributed. The signs above variables in equation (13) are predicted by the comparative statics analysis of Section II, and $Y$ are assumed to have positive effects on listsizes and visits per person; and the sign of $W$ is listsizes and visits equations implies that higher time prices will cause some patients to switch doctors and others to reduce their use of services.

In the absence of any explicit utility function, there is no unique form of equation (13). We will estimate both linear and logarithmic functions. Likewise, we will estimate both linear and logarithmic versions of (14) and (15)
Both specifications satisfy our assumption that (14) and (15) are separable functions. In our model, waiting time affects \( P \) and \( W \), but not vice versa; however, the model cannot be estimated recursively by ordinary least squares regression because \( W \) depends on a random error term, \( e_2 \). This occurs because the utility-maximizing value of \( W \) depends on all exogenous variables in the model, including \( e_1 \) and \( e_2 \). To correct for simultaneous equations bias, we estimate the model by two-stage least squares, using an instrumental variable for \( W \) in equations (14) and (15). 

Visits per capita are measured in two ways: first, by the physician’s weekly surgery (office) visits per 1,000 patients on his or her list; second, by the weighted average of surgery and domiciliary visits (housecalls), where the weight for housecalls is the sample average time per housecall divided by the sample average time per office visit. The second measure recognizes that most GPs in our sample make housecalls, and the time required for a housecall generally exceeds that for an office visit. 

A list of variables, with means and standard deviations, is given in Table 1. Variables selected as shift factors in the demand function for visits are DRUGS, GPBEBDS, RURAL, F6ST074, and FOVER74. The dummy variable DRUGS denotes whether the GP dispenses drugs to home patients. If the GP does dispense drugs, there will be a greater demand for office visits. The variable GPBEBDS represents the average number of general practitioners beds available to each GP in the catchment area. More GPBEBDS may increase or decrease demand for ambulatory visits, depending on whether hospital and ambulatory visits are complements or substitutes. 

*Rural practice units*: In addition to capitalization payments are awarded to GPs who must travel long distances to visit patients. RURAL measures the number of these units per patient on the physician’s list. Given the access problem to the GP’s surgery, RURAL should be negatively related to demand per patient. 

Patients over 65 years of age demand more health services than other groups. The oldest, those 75 and older, are restricted in their mobility. We include the fraction of patients between 65 and 74 (F6ST074) and the fraction 75 and older (FOVER74) separately in the demand function. We predict that the coefficient of F6ST074 is positive, but the coefficient of FOVER74 may be negative. 

The equation for the GP’s listsize includes, in addition to waiting time, GPERCAP, RURAL, AGE, and YEARSWD. We predict that more patients will want to join the list if there are fewer alternative GPs per 1,000 people in the catchment area. 

GP’s listsize in rural areas tends to be smaller than in urban areas. This may be due to the additional travel time required for rural GPs to make home visits. We control for this effect by including RURAL in the listsize equation. 

The GP’s willingness to accept patients depends on unknown parameters of his or her utility function such as tastes for income, workload, and waiting time. Lacking specific hypotheses about tastes, we include two proxy variables in the listsize function: the GP’s age and the number of years he or she has practiced in West Dorset. If the relative marginal utility of leisure increases with age, older GPs will tend to have smaller listsize. 

The waiting time equation, number (13), is a reduced form containing all the variables that affect patients and physicians. We predict that waiting time is directly related to the “\( x \) and “\( y \) variables from equations (14) and (15). Waiting time should also be directly related to capitalization payments per patient, but, since capitalization payments per patient are constant in a cross section analysis, their effect is absorbed into the constant term of equation (13). 

We do not have measures of fixed time costs and nonlabor income. Omission of these variables may bias the coefficients of other variables in equation (13), but the direction of bias is unknown. 

Estimates of the demand equation for visits are shown in Table 2. Waiting time is negatively associated with the quantity demanded per patient in all specifications. The elasticity of demand in the logarithmic specification is 

\[ -4.83 \] (visits per GP), or 

\[ -7.82 \] (weighted visits); the elasticity of demand in the linear specification, at the mean quantity and waiting time, is 

\[ -5.2 \] (visits per GP) or 

\[ -9.2 \] (weighted visits). Demand increases if the GP dispenses drugs during home visits; the coefficient of DRUGS is larger in equations which include weighted home visits. Patients aged 65 to 74 use more visits, but, after age 75, increasing age is a barrier to access to ambulatory care. Ruralness of the practice and the number of available hospital beds per GP are statistically insignificant in all specifications of the demand equation.
TABLE 2  Demand Equations for the Services of GPs in the British National Health Service

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Linear Equation</th>
<th>Logarithmic Eqn.</th>
<th>Linear Equation</th>
<th>Logarithmic Eqn.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficient</td>
<td>T-Value</td>
<td>Coefficient</td>
<td>T-Value</td>
</tr>
<tr>
<td>VISTI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WAITING TIME (endogenous)</td>
<td>-2.861 1.54</td>
<td>-0.493 1.85</td>
<td>-10.600 2.26</td>
<td>-0.582 3.41</td>
</tr>
<tr>
<td>DRUGS</td>
<td>27.376 1.36</td>
<td>-2.258 1.26</td>
<td>113.520 2.08</td>
<td>0.655 1.54</td>
</tr>
<tr>
<td>GPBREADS</td>
<td>1.547 0.67</td>
<td>0.046 0.49</td>
<td>1.020 1.00</td>
<td>-0.00132 0.02</td>
</tr>
<tr>
<td>RURAL</td>
<td>1.339 -0.20</td>
<td>0.159 0.09</td>
<td>2.570 -0.14</td>
<td>0.144 0.08</td>
</tr>
<tr>
<td>FOYESTO74</td>
<td>645.773 1.25</td>
<td>-4.400 1.82</td>
<td>766.172 0.81</td>
<td>0.500 0.17</td>
</tr>
<tr>
<td>FOYESTO73</td>
<td>-557.030 1.80</td>
<td>-4.239 2.11</td>
<td>-582.236 0.74</td>
<td>-0.240 0.93</td>
</tr>
<tr>
<td>CONSENT</td>
<td>121.576 4.33</td>
<td>0.250 0.60</td>
<td>0.275 0.34</td>
<td>0.715 0.50</td>
</tr>
</tbody>
</table>

catchment area and the ruralness of the area have negative signs, as expected, but their coefficients are also insignificant. The reduced form equation for waiting time has an adjusted $R^2$ of .29 in the logarithmic specification. Several variables are statistically significant at $a = 0.1$ or less: DRUGS ($a = .088$); GPBREADS ($a = .009$); and GPPEPREADS ($a = .002$). The same variables are significant in the linear waiting time equation. The signs of these variables all agree with our comparative statics predictions.

IV. Discussion and Implications

The most important result from the demand equation is that the elasticity of demand with respect to waiting time is between -48 and -.92. This range is entirely more negative than estimates with U.S. data of the money price elasticity of demand [16]. However, a study by Jan Acton [17] of the effect of travel time on the demand by users of New York's "taxi" outpatient departments found that the elasticity of demand with respect to travel time was between -6 and -1.0. Acton showed that, if we let $P$ equal the total price of medical services, then

$$\eta_P = \frac{P}{\Pi} \text{ and } \eta_W = \frac{W}{\Pi}$$

where $P$ = money price per unit of service $W$ = time input per unit of service $\Pi = W + P$ $\eta_P$ = money price elasticity $\eta_W$ = time price elasticity $\eta = \text{total elasticity of demand}$

In most U.S. data $\Pi$ is less than one, so the absolute value of the money price elasticity will be less than $\eta$. Since Acton observed a situation where $\Pi/W = 1$, his estimate would be larger in absolute value than the others, i.e., he estimated $|\eta_P| > |\eta| > |\eta_W|$. The same should be true for our data since patients in the NHS face a zero money price. And, in fact, the range of our estimates is approximately equal to Acton's.

The significance of this result for national health insurance is twofold. First, demand for physicians' services is more elastic than commonly supposed, so removal of all barriers to access to physicians' services would substantially increase demand. Second, if the money price were reduced and the time price were not, the time price elasticity would become larger in absolute value than the money price elasticity. Demand would become relatively more sensitive to time prices. This would permit persons with a lower opportunity cost of time to bid services away from those with a higher opportunity cost of time. Persons unable to travel to physicians' offices, for example, would face substantial barriers of access to medical care. And providers might attempt to ration services by waiting. Our study suggests that this has occurred in England.

Our results from the demand and waiting time equations also are relevant for the current debate among health economists whether physicians can shift the demand curve for their services. Evidence for demand-shifting is derived from the positive elasticity of physicians per capita in a demand function for patient visits per capita. Fuchs and Kramer [18], using two-stage least squares regression analysis, estimated this elasticity as 4. Sloan and Feldman [7] pointed out that this positive association may be due to lower waiting times, which would increase demand in physician-dense areas. But they conceded that, with U.S. data, individual time elasticities (waiting time, etc.) could explain less than one-tenth of the composite effect of physicians per capita on demand per capita.

In the present study different results are obtained. We find that waiting time is highly sensitive to physicians per capita (GPPEPREADS), with an elasticity, $\eta_P$, of -2.02 (linear specification evaluated at the means) or -2.676 (logarithmic specification).14 In the notation of equation (16), the effect of more GPs on demand is $(W/\Pi)\eta_P$. Sloan and

15It may be argued that $\eta_P + \eta_W$ is in the NHS because patients must "spend" time waiting to schedule an appointment and travel to the physician's office. Waiting time for an appointment is not significant---most GPs have private or "walk-in" bases during surgery hours. To the extent that travel time is important, our results will underestimate the total elasticity of demand.

TABLE 3  Listwise and Waiting Time Equations for GPs in the British National Health Service

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Linear Equation</th>
<th>Logarithmic Eqn.</th>
<th>Linear Equation</th>
<th>Logarithmic Eqn.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficient</td>
<td>T-Value</td>
<td>Coefficient</td>
<td>T-Value</td>
</tr>
<tr>
<td>VISTI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WAITING TIME (endogenous)</td>
<td>37.876 1.42</td>
<td>-2.12 1.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DRUGS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPBREADS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RURAL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FOYESTO74</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FOYESTO73</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONSENT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td></td>
<td></td>
<td></td>
<td>0.245</td>
</tr>
</tbody>
</table>

14Sloan and Lurie [19], using OLS regression analysis of U.S. data, found much smaller elasticities---on the order of 1. But their estimate may be too low if U.S. physicians are attracted to areas with low waiting times. Both areas have higher-income populations and other features that attract physicians, according to Sloan and Lurie's regressions.
Feldman estimated that \( W/H \) was approximately .53 in the U.S. in 1969. Therefore, depending on which values of \( a \) and \( b \) are used, a composite elasticity between .55 and 1.1077 is obtained. In other words, the entire effect of physicians on demand observed by Fuchs and Kramer might be explained by "standard" theoretical arguments without recourse to supplier-created demand.

Variables representing an elderly patient population (FSY074 and FOVE74) both have positive effects on waiting time. These results, although not statistically significant, indicate that the elderly can better afford to spend time waiting in physicians' offices than the nonelderly. A similar effect has been observed in the U.S. for people with Medicare (government health insurance for the elderly) [19]. Taken together, these findings support our earlier suggestion that national health insurance might have important distributive effects among people with different values of time.

The coefficient of GPBIARDS in the waiting time equation indicates that office waiting time is shorter when GPs have greater access to hospital beds. This suggests that a "triple" process is taking place, where GPs tend to hospitalize more difficult cases when they can. Taken together, the conditions most conducive to short waiting time are an adequate supply of GPs who have access to hospital beds.

The important "noresort" from the listwise equation is that waiting time does not reduce a physician's listsize. The immediate implication for a British GP is that waiting time does not threaten his or her income. This result may not generalize to the U.S. Many impressionistic observers have noted that the British tolerate long waiting lines. However, in the U.S., Health Maintenance Organizations (HMOs) act as if the analogy holds. HMOs (like the National Health Service) provide health care for a fixed prepaid fee and, like the NHS, physicians in an HMO are often compensated by capitation payments. Hospitalization rates for HMO members are about 30 percent below those of comparable populations with conventional health insurance [4], enabling HMOs to offer more comprehensive benefits for competitive premiums. The price of "free" care in HMOs appears to be longer waiting times and patient dissatisfaction.

Tesler and Mechanic have noted the tension between the benefits of HMOs and patient dissatisfaction [20]. They observe that "each form of practice organization offers distinctive advantages and potential problems... The challenge we face is how to develop practice contexts and physician incentives that maximize the advantages of prepaid practice organization but that are also responsive to consumer perceptions and expectations." If the British experience is a guide, waiting time will ration demand without driving away too many customers.

References