A Systems Thinking Approach to Analyze the Waitinglist Phenomenon

Ann van Ackere
London Business School
Sussex Place
Regent's Park
London, NW1 4SA, UK

Abstract

The aim of the National Health Service (NHS) is to provide access to medical care to all. One of the challenges is to balance demand and resources while avoiding unduly long waiting lists, either for treatment (the `traditional' waiting list phenomenon), or to see a practitioner. This note illustrates how Systems Thinking can contribute to the understanding of this complex issue, and to assessing the impact of various policies. Systems thinking differs from traditional approaches in that it looks at the problem as a whole, as a *system*, and focuses on identifying key inter-relationships and feedback loops between different components of the system. We present three simple examples which illustrate the concepts of systems thinking, and indicate how this approach could be useful to analyse various policies.

A Systems Thinking Approach to Analyze the Waitinglist Phenomenon

Introduction

The funding of the NHS is a very sensitive issue, both politically and socially. In the past, it has been common to observe announcements of additional funding in pre-election years. These have led to short term reductions in waiting lists, but a year later, these effects had disappeared, and waiting lists reached new heights. The debate about optimal funding of the NHS is as old as the NHS itself. The complexity of the issue can be better understood when, following Musgrave's classification, the NHS is considered as a public good characterized by rival consumption (if one patient sees a doctor, this time is not available to another patient) and the impossibility of exclusion. It is this latter characteristic (everyone's right to free health care through the NHS) which makes the provision of health care in the UK a public good. Impossibility of exclusion usually occurs for technical or economic reasons. In the case of the NHS, the exclusion problem has a moral origin: no one should be excluded from health care. Consequently, private companies have entered the market, providing a less congested alternative to the NHS.

It has been suggested that rather than providing the service free, patients should bear part of the cost. This would reduce congestion, as potential customers would avoid using the service without good cause. If this is indeed the case, McConnell's study of a congested recreation facility indicates an undesirable side-effect. He shows that under plausible assumptions about the effects of income and price responsiveness, the use of a fee would tend to ration the service towards people who are less price responsive and more averse to congestion than the average user, typically higher income people. Such a result would be contrary to the aim of the NHS. Equally important, a fee could cause low-income individuals to postpone seeking medical advice, or to avoid preventive care (e.g., timely eye tests), leading to higher demand in the long run.

Some authors have argued that congestion is not necessarily inefficient. Kornai and Weibull consider an economy characterized by chronic shortage and queueing. Although this market does not satisfy the conditions of a Walrasian equilibrium, it is in a stationary state. Queueing is used as a rationing device, leading to postponement of purchase or forced substitution. It is worth noting that they deal with a deterministic model because ``This choice reflects our belief that in situations characterized by chronic shortage, the stochastic element is secondary in comparison to the inter-dependencies and feedback mechanisms regulating the system." (p. 376).

Over the years, there has been a continuous debate over *what represents an adequate level of funding for the NHS*. The most vocal point of view claims that the existence of lengthy waiting lists is clear evidence of a lack of resources. But would an increase in resources lead to significantly shorter waiting lists and less congestion costs? Following Kornai and Weibull, we want to argue that if a service is provided free, delay is a natural way to allocate a limited resource. Therefore, waiting lists are endemic to the NHS, and simply throwing in more resources will not solve the problem. At least two arguments support this view.

First, the demand for NHS health care is determined jointly by the resource level of the NHS and the availability of private care. An increase in NHS resources, resulting in less congestion, may cause patients to switch from private care to the NHS. In 1988 public health care expenditure in the UK amounted to 86% of total health care expenditures, and preliminary estimates for 1990 yield a figure of 84.5%. A large switch from private to public health care would have a non-negligible impact. Government policy may also influence patients' choice by making private health care more or less attractive. For instance, Orros examines the implications of tax relief on health insurance for people over 60, both for the insurance industry and for the NHS. Second, demand for health care is not a constant. If health care became more easily accessible (shorter waiting lists, less crowded emergency rooms, less time spent in the GP's waiting rooms, etc) demand might actually increase. For instance, patients would more easily seek care for minor health problems, the average time of a doctor's visit may increase, more elective surgery may take place, people incurring an injury or becoming ill abroad would be more likely to seek immediate repatriation, etc. At this stage we do not consider the implications of increased demand due to the progress of medicine.

We aim to provide a better understanding of the interaction between funding policies, resulting waiting lists and an individual's decision to seek either NHS or private health care. We hope to build a convincing argument that the *inject cash to clear the waiting list* policy is not workable. We also intend to emphasize that decisions regarding NHS funding should not be taken in isolation. Rather, these should be considered jointly with other policy measures which determine the attractiveness of private health care, both at the individual level and at the employer level (is this a perk worth offering?) as this influences demand for NHS care.

The Systems Thinking Approach

The aim of the NHS is to provide access to medical care to all. One of the challenges is to balance demand and resources while avoiding unduly long waiting lists, either for treatment (the `traditional' waiting list phenomenon), or to see a practitioner. This note illustrates how Systems Thinking can contribute to the understanding of this complex issue, and to assessing the impact of various policies. Systems thinking differs from traditional approaches in that it looks at the problem as a whole, as a *system*, and focuses on identifying key inter-relationships and feedback loops between different components of the system. Causal loop diagrams are a way to sketch these inter-relationships, and are especially useful in brain-storming sessions, where the purpose is to identify the key factors and policy levers, and how they are interrelated. Translating these causal loop diagrams into a stock and flow network allows one to simulate the system, and evaluate the impact of alternative policies. The simulation models differ from traditional simulations in that one attempts to model decision policies and *soft* variables (e.g. perceived waiting times), rather than focus solely on objectively measurable physical elements. Wolstenholme applies this approach successfully to analyze community care.

In the next section we use simple examples to introduce the two basic building blocks of causal loop diagrams: the *balancing loop* and the *reinforcing loop*. Next we translate two of these causal loop diagrams into stock and flow networks. This enables us to carry out simulations and observe the impact of different policies over time. These simulations can be the starting point of a constructive discussion, resulting in a variety of options being evaluated and tested. Our aim is to illustrate how such an approach could provide a constructive input into the debate on the funding of the NHS. Often, sensibly sounding, over-simplified arguments are put forward, and it can be hard to refute these. One classical example is the argument that one should inject a lot of cash in the NHS to get rid of waiting lists *once and for all*. The simple model in example 3 illustrates that such a course of action, taken in isolation, is unlikely to be successful, any improvements being only temporary.

Causal Loop Diagrams

Consider first the inter-relationship between demand for NHS care and the length of waiting lists (Figure 1.A). As demand for NHS care increases, waiting lists get longer, other things being equal. This is indicated by the sign S on the arrow from **demand for NHS care** to **Waiting list**: both entities move in the Same direction. But, as waiting lists get longer, people seek alternative care, i.e., **Fraction to NHS** goes down. This is indicated by the symbol **0**, as the entities move in opposite directions. This in turn leads to a decrease in NHS demand (S as both entities move in the same direction). Going around the loop, an increase in demand for NHS care leads to longer waiting lists, which implies a lower fraction to NHS and results in lower demand for NHS care. This is called a balancing loop, and is indicated by a **B^** in the diagram. We assumed total demand to be given. A more detailed analysis would consider untreated conditions and elective procedures.

A balancing loop represents a system which is striving for an equilibrium position. The pattern over time depends on whether or not there are delays in the system. For instance, in Figure 1.A, the // sign indicates a delay between a change in the length of waiting lists and the resulting change in the **fraction to NHS**. Consider for instance a demand pattern as in Figure 1.B(i). Demand suddenly increases due to an external factor (e.g., an epidemic or a new disease). What is the impact on the waiting list and future demand for NHS care? If there are no delays, waiting lists will evolve as pictured in Figure 1.B(ii): after the initial shock, the waiting list moves gradually to a new equilibrium position. If there are delays, behaviour will be as in figure 1.B(iii): fluctuations arise, as the delays cause the waiting list to overshoot its new equilibrium position.

Next, let us elaborate on this simple model, and include the issue of untreated cases, some of which become serious due to the lack of treatment. This is pictured in Figure 2.A. As the **NHS load** increases, access becomes less convenient (**O**), less convenient access implies more **Untreated cases** (O). But in the long run, more untreated cases implies a larger **Number of serious cases** or emergencies (**S**), which results in a higher load (**S**). This loop is reinforcing in nature: an initial increase in the load results in a further increase later on. This is indicated by the symbol **R^** in the centre of the loop. Figure 2.B shows the typical pattern over time resulting from a reinforcing loop. Note that in the short run, more untreated cases implies a lower load (O). This yields a balancing loop, which will counteract the reinforcing loop.

The third example considers resources, building on the balancing loop of the first example (Figure 3.A). Longer waiting lists create pressure for more resources. Over time this results in increased resources, which reduces the waiting list. This yields a second balancing loop. The resulting behaviour depends on the magnitude of the delays in the two loops.

Stock and Flow Networks

Causal loop diagrams help in understanding the issue, identifying the relevant parties and key decisions, and developing a common language. The next step consists of translating the casual loop diagrams into stock and flow networks. These networks consist of the following building blocks: *stocks*, represented by a box; *flows* and *regulators*, which regulate the inflows and outflows of stocks; and *converters*, represented by a circle. The stocks, flows and regulators represent the *plumbing* of the network, while converters represent the *information* network.

We first consider a simplified version of example 1, assuming no delays are present. In this example (Figure 1.C) we have one stock, representing patients on a **Waiting list**. The level of this stock (number of patients on the waiting list) is affected by one inflow and one outflow. The inflow is labelled **New NHS patients** and represents the rate at which patients join the waiting list, expressed for instance as patients per year. Similarly the outflow, labelled **Patients treated** represents the rate at which patients are treated, and therefore leave the waiting list. We could include a second outflow, representing patients who leave the waiting list without receiving treatment. The inflow of patients depends on two elements: the total number of new patients (**Total new patients**) and the fraction of this total seeking NHS care (**Fraction to NHS**). The **Fraction to NHS** in turn depends on the **Waiting list**. In Figure 1.D the bold line indicates the balancing loop we discussed earlier.

An interesting issue is: who decides on whether or not a patient selects the NHS? Is it the patient, his GP, or both? In other words, is the relationship between the length of the waiting list and the fraction of patients selecting NHS dependent on the patients attitude or the GP's decision? This information should come from discussions with individuals involved in these decisions.

At this stage we are close to having a simulatable model. We only need to formalise the dependencies between the various converters, and provide an initial value for our stock. We consider a scenario where initially the annual number of new patients equals 100, of whom 80% select NHS care when the waiting list has 40 people on it, i.e., an expected waiting time of 6 months if patients are treated at a rate of 80 per year. We set the initial value of the waiting list at 40. The equation for **New NHS**

patients is straight forward: **Total new patients** multiplied by **Fraction to NHS**. The relationship between **Waiting list** and **Fraction to NHS** is trickier. Most people would agree that the longer the waiting list, the smaller the fraction of customers selecting NHS. A more precise relationship could be derived from historical data, or from the expertise of individuals dealing with this issue.

Figure 1.E sketches one possible relationship. As stated earlier on, we assume that when there are 40 people on the waiting list, 80% of new patients select NHS. We also assume that, if there is no waiting list, everyone selects NHS, and if the waiting list is long (100 or more people), 40% select NHS. These three points are indicated on the graph. We then draw a smooth line linking these three points. The graphical function is a useful discussion tool: it allows a team to discuss different assumptions without having to resort to complicated mathematical expressions. Figure 1.F summarises the equations. Our assumptions lead to a steady state situation where each year 80 new patients join the waiting list, 80 are treated, and the waiting list remains at 40. This is illustrated in Figure 1.G, up to year 4. We next modify the total number of new patients to see the impact of a sudden increase in demand. We model a step increase of 20% in year 4. The results are illustrated in Figure 1.G, years 4 to 16. Initially the number of **New NHS patients** increases in line with the total number of patients. This causes the waiting list to increase. Consequently the fraction of customers selecting NHS is reduced, leading to a lower number of new patients. The system moves smoothly to a new equilibrium with a somewhat longer waiting list, and two thirds of the patients (80 out of 120) selecting NHS.

Next, consider a scenario where total demand remains at 100 per year throughout the simulation, but the rate at which patients are treated increases from 80 to 90 patients per year in year 4. The result is shown in Figure 1.H. Again we observe a smooth transition to a new equilibrium: the increase in the number of patients treated causes a gradual decline in the waiting list, which triggers a gradual increase in the number of new NHS patients. In reality, people do not react instantly to changes such as a shorter waiting list. It takes time before they realize something is happening. Let us consider the impact of such delays on the behaviour of the system. Specifically, assume it takes some time before potential NHS patients become aware of the changes in the length of the waiting list. This is modelled in Figure 1.I. We add a converter, labelled **Time to perceive change in waiting list**, and define the **Perceived waiting list** to depend on the actual **Waiting list** and this delay. Specifically, we assume the **Perceived waiting list** to be an exponentially smoothed average of the actual waiting list, over the past year. This is easily done by using a build-in function called SMOOTH1. The fraction of customers selecting the NHS is assumed to depend on the perceived waiting list in the same way it depended on the actual waiting list in the previous model.

We consider the same equilibrium scenario up to year 3, followed by a 20% increase in total demand in year 4. Figure 1.J shows the simulation results. We observe the same initial increase of the flow of patients to the NHS, and of the waiting list. But the move to a new equilibrium is far from smooth. The waiting list builds up as patients initially fail to perceive its increases. When they do, many look for alternative treatment, which results in a small decrease in the waiting list. This in turn attracts more patients. Over time, fluctuations level out as the system reaches a new equilibrium. Note that in this simple model, the resulting equilibrium is the same with and without delays. In more complex models this would not be the case, as the fluctuation would trigger reactions in other parts of the system. Figure 1.K shows the impact of increasing the rate at which patients are treated from 80 to 90 in year 4. Again we observe fluctuations before the system reaches a new equilibrium.

Next, let us consider the stock and flow network for example 3 (Figure 3.B). Note that we have divided the model into two sectors, one concerning the waiting list, and one dealing with resources. The connectors crossing the boundaries (e.g., from **Level of resourcing** to **Rate of treatment**) indicates that these sectors are interdependent. The waiting list sector has one stock (**Waiting list**) whose level is regulated by the **Inflow of patients** and the **Patients treated**. The resourcing sector also has one stock (**Level of Resourcing**). Changes in the level of resourcing depend on the pressure experienced by decision-makers due to the perceived length of the waiting list. The bold line traces the balancing loop in the top part of Figure 3.A, the dotted line traces the bottom one. An increase in the **Waiting list** results, after some delay, in increased **Pressure for resources**, which in turn leads to increased resources. This results in a larger number of **Patients treated**, which reduces the waiting list. How exactly increasing pressure leads to additional resources would be an interesting topic for further exploration. We have postulated a simple graphical relationship. Input should come from individuals involved in these issues, and whatever assumptions are made should be subjected to careful sensitivity analysis.

In the second loop, an increase in the **Waiting list** decreases the **Fraction of demand to NHS**, which results in a smaller **Inflow of patients** and thus a shorter **Waiting list**. Note that the **Waiting list** is part of both loops. This implies that the consequences of a change in **Demand** will spill over into the resource sector. Similarly, a change in the **Rate of treatment** will impact the **Inflow of Patients**. This will be illustrated in the simulations discussed below.

This example illustrates that translating a circle diagram into a stock and flow network often involves the inclusion of additional concepts. It requires exploring the assumptions implied by the circle diagram in more detail. For instance, how exactly do resources depend on the pressure created by the existence of long waiting lists?

Three graphical converters are used in the model. They are indicated by a "~"-sign in the converter. As discussed earlier, this is a convenient way to incorporate qualitative data in a simulation model. The base, steady state scenario assumes a demand of 100 people per quarter, of whom 80% select NHS. The initial value for the waiting list is 160 patients, and the available resources allow 80 patients to be treated each quarter. This implies an average waiting time of 2 quarters. There is no pressure for additional resources. The equations are summarized in Figure 3.C. The graphical converters are shown in Figure 3.D, 3.E and 3.F.

In the first simulation, the system remains in steady state for the first 16 quarters. At this point in time, demand is increased by 20% (Figure 3.G). The increase in the **Inflow of patients** causes longer waiting lists, which in turn triggers an increase in the **Level of resourcing**. The increase in the **Waiting list** also causes a decrease in the inflow of patients. This, together with the increased resources reduces the length of the waiting list. Again, delays cause fluctuations before the system settles down to a new equilibrium. In the second simulation (Figures 3.H and 3.I), demand remains constant throughout the 64 quarters. We model a situation where additional resources (a 25% increase, labelled **Temporary resources**) are provided for a 1 year period, starting in quarter 16. Note that these additional resources are the amount required to treat all the patients on the waitinglist. The increase in resources initially causes a considerable decrease in the **Waiting list**. Consequently, more patients choose NHS, leading to a larger **Inflow of patients**. This, together with the resource level being reduced to its initial level after one year, reverses the trend in the waiting list, which actually overshoots its initial value. This leads to pressure for additional resources. The system fluctuates further before returning to its initial steady state. This very simple simulation seems to indicate that increasing resources temporarily to eliminate the waiting list is unlikely to work, unless it is accompanied by measures preventing recurrence.

Concluding Remarks

The purpose of this note was to introduce the concepts of systems thinking, and illustrate how this approach could be useful in increasing the understanding of the complex policy issues surrounding demand for NHS care and its funding. We first discussed causal loop diagrams, and presented three simple examples. We then developed a stock and flow network for two of these examples, and simulated two scenarios, one relating to increasing demand, and one relating to increased resources. We also discussed how delays (e.g. between an increase in the length of the waiting list and when people perceive this increase) can cause the system to fluctuate over time. Consequently, the effects of a specific course of action may not be immediately apparent, as short term and long term consequences can differ. For instance, referring to Figures 3.H and 3.I, a temporary increase in resources does indeed have a dramatic impact on the length of the waiting list in the short run, but the long run picture does look very differently.

This approach has been used successfully with management teams to tackle complicated policy issues. In several instances, brain-storming sessions resulting in a causal loop diagram have provided considerable insights into the problem, and it was felt that the step to a simulation model was unnecessary. In other cases, the team has pursued the development of a simulation model to enable experimenting with a variety of policies.

References

Kornai, J., and J.W. Weibull, The Normal State of the Market in a Shortage Economy: A Queue Model, Scandinavian Journal of Economics, Vol 80, 1987, p.375--398.

McConnell, K.E., Heterogeneous Preferences for Congestion, Journal of Environmental Economics and Management, Vol 15, 1988, p. 251-258.

Musgrave, R.A., P.B. Musgrave, Public Finance in Theory and Practice, 4th ed, McGraw-Hill International student edition, 1984, 824p.

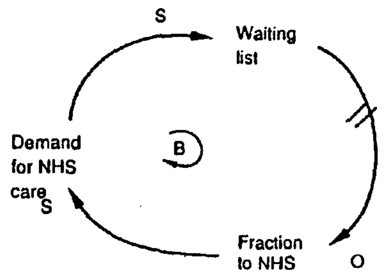
Orros, G.C., Private Health Insurance -- A new Era for the NHS, Health Services Management, June 1989, p. 118-120.

Statistical Abstract of the United States, 1991 and 1992.

van Ackere, A., Competing Against a Public Service, Working paper, 1993.

Wolstenholme, E.F., A Case Study in Community Care Using Systems Thinking, Journal of the Operational Research Society, Vol 44(9), p.925-934.

Figure 1.A



igure 1.B

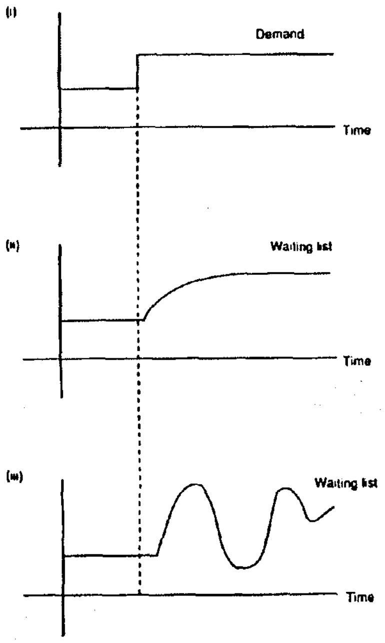


Figure 1.C

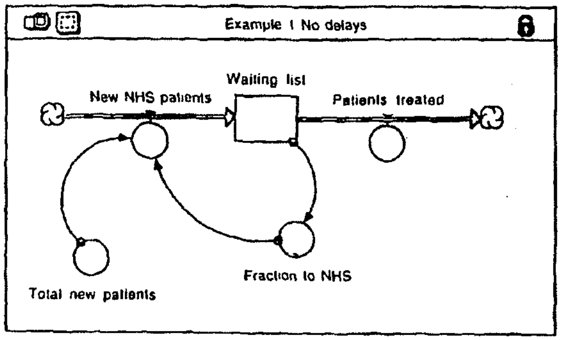


Figure 1.D

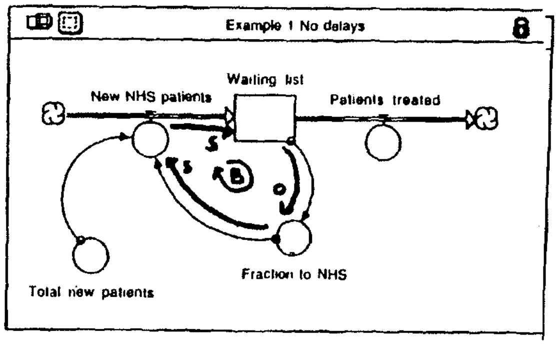


Figure 1.E

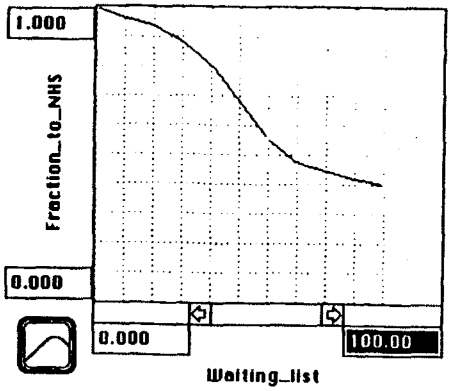


Figure 1.F

```
Example 1 No delays
□ Waiting_list(t) = Waiting_list(t - dt) + (New_NHS_patients - Patients_treated) * dt
  INIT Waiting_list = 40 {patients 80 NHS patients per year, waiting on average 6 months}

  INFLOWS:
     New_NHS_patients = Total_new_patients*Fraction_to_NHS
  OUTFLOWS:
     Patients_treated = 80 + 0*step(10,4) {patients per year}
○ Total_new_patients = 100 + 0*step(20,4) {new patients per year}
⊘ Fraction_to_NHS = GRAPH(Waiting_list)
  (0.00, 1.00), (10.0, 0.965), (20.0, 0.935), (30.0, 0.885), (40.0, 0.8), (50.0, 0.675), (60.0,
  0.55), (70.0, 0.475), (80.0, 0.445), (90.0, 0.42), (100, 0.4)
```

Figure 1.G

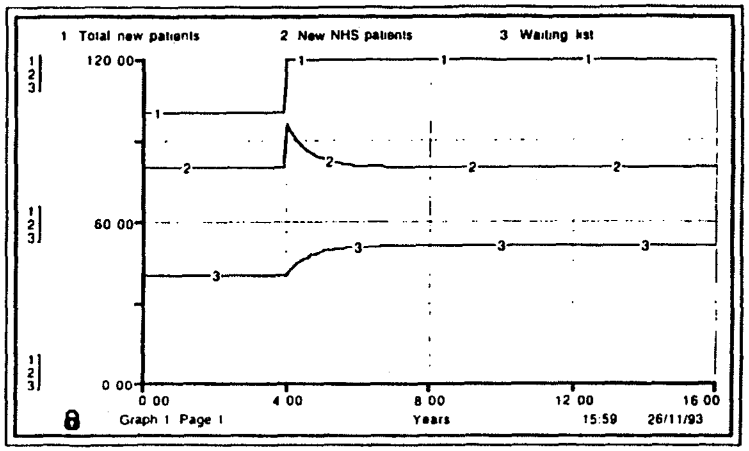


Figure 1 H

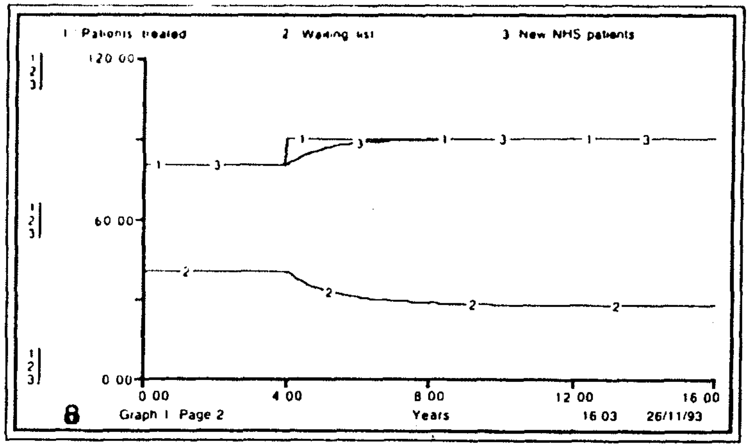


Figure 1 I

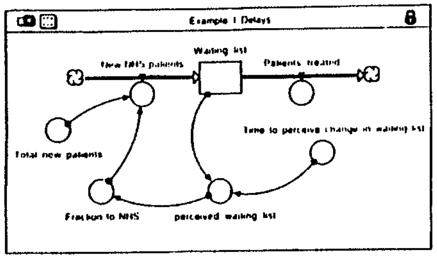


Figure 1 J

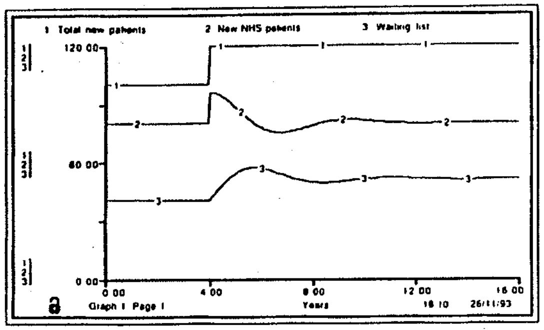


Figure 1.K

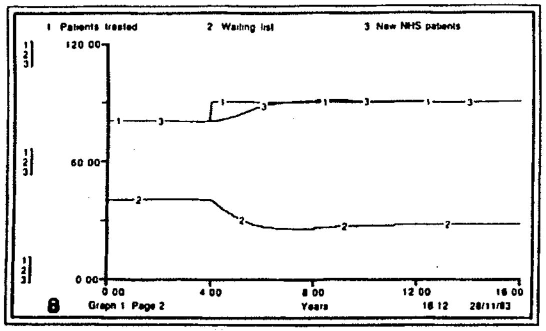


Figure 2.A

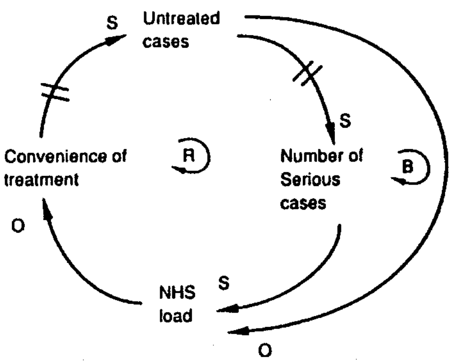


Figure 2.B

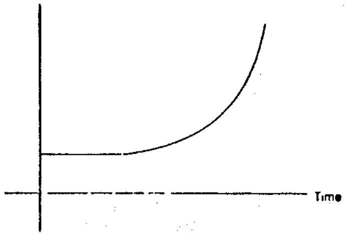


Figure 3.A

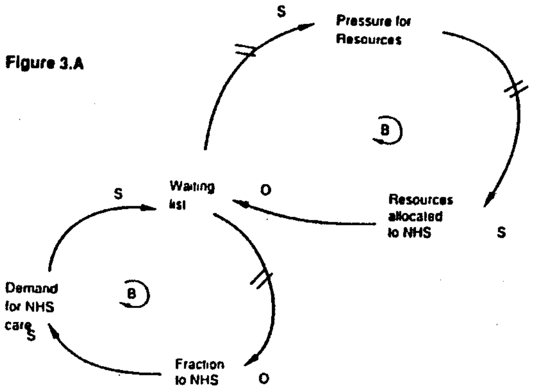


Figure 3 B

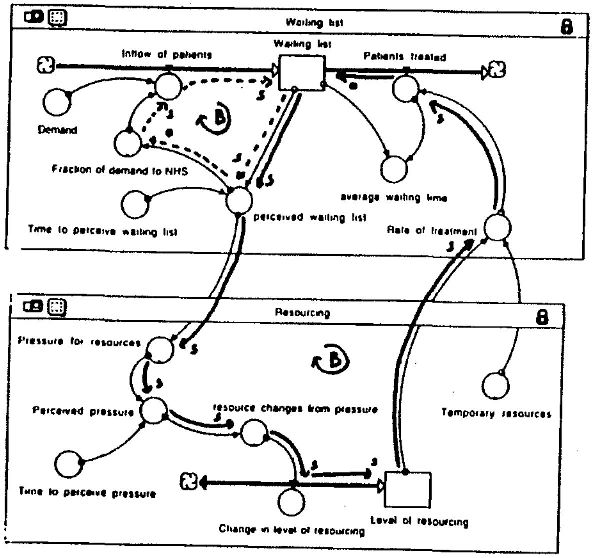


Figure 3.C

Resourcing

```
□ Level_of_resourcing(t) = Level_of_resourcing(t - dt) + (Change_in_level_of_resourcing) * dt
  INIT Level_of_resourcing = 80 {resources allow to treat 80 peopel per quarter}

  INFLOWS
     Change_in_level_of_resourcing = resource_changes_from_pressure
○ Perceived pressure = SMTH1(Pressure_for_resources,Time_to_perceive_pressure)
○ Time_to_perceive_pressure = 2 {quarters: There is a 6 month delay between the time when pressure
  for resources changes and when resources are changed}
⊘ Pressure_for_resources = GRAPH(perceived_waiting_list)
  (0.00, -1.00), (40.0, -0.5), (80.0, -0.25), (120, -0.1), (160, 0.00), (200, 0.1), (240,
  0.18), (280, 0.31), (320, 0.46), (360, 0.68), (400, 1.00)
⊘ resource_changes_from_pressure = GRAPH(Perceived pressure)
  (-1.00, -10.0), (-0.8, -5.40), (-0.6, -2.80), (-0.4, -1.80), (-0.2, -1.00), (-6.66e-17,
  0.00), (0.2, 2.20), (0.4, 5.40), (0.6, 12.4), (0.8, 17.2), (1, 20.0)
⊘ Temporary resources = GRAPH(TIME)
  (0.00, 0.00), (4.00, 0.00), (8.00, 0.00), (12.0, 0.00), (16.0, 20.0), (20.0, 0.00), (24.0,
  0.00), (28.0, 0.00), (32.0, 0.00), (36.0, 0.00), (40.0, 0.00), (44.0, 0.00), (48.0, 0.00),
  (52.0, 0.00), (56.0, 0.00), (60.0, 0.00), (64.0, 0.00)
```

Waiting list

```
□ Waiting list(t) = Waiting list(t - dt) + (Inflow of patients - Patients treated) * dt
  INIT Waiting list = 160 {Assuming that under 'normal conditions' 80% of demand is dealt with by NHS,
  the waiting list equals 6 monts }
  INFLOWS
     Inflow of patients = Demand*Fraction of demand to NHS
  OUTFLOWS
     Patients treated = Rate of treatment
○ average waiting time = Waiting list/Patients treated
○ Demand = 100+ 0*step (20 16) {demand per quarter}
○ perceived waiting list = SMTH1(Waiting list,Time to perceive_waiting list)
○ Rate of treatment = Level of resourcing+ 0*Temporary resources { we assume that resources are
  expressed as the number of patients that can be treated per quarter}
○ Time to perceive waiting list = 2 {quarters it takes about 6month for people to react to changes in
  waiting list}
⊘ Fraction of demand to NHS = GRAPH(perceived waiting list)
  (0.00, 1.00) (40.0, 0.98) (80.0, 0.96), (120, 0.905), (160, 0.8), (200, 0.71), (240,
  0.635), (280, 0.555), (320, 0.535) (360, 0.52), (400, 0.5)
```

Figure 3.D

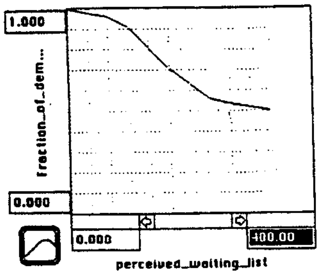


Figure 3.E

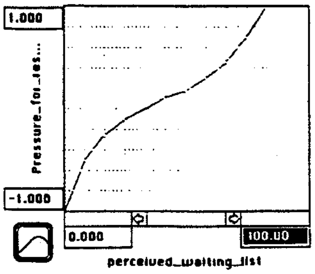
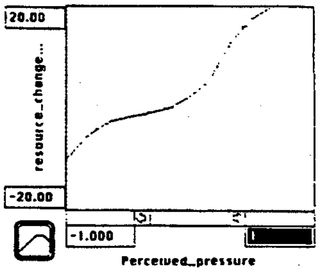


Figure 3.F



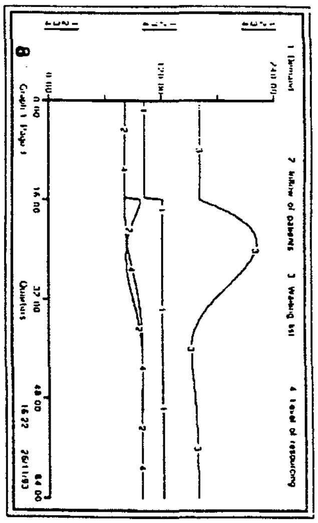


Figure 3.G

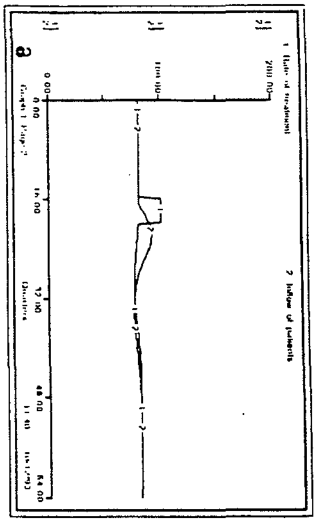


Figure 3.H

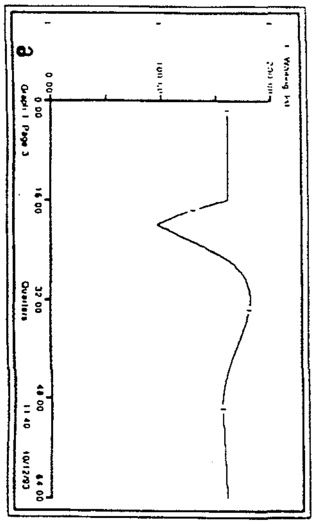


Figure 3.I