Agent based modelling and policy design: Patient Choice and Activity based hospital finance in the English NHS

Paul Ormerod, Volterra Consulting
James Raftery, Director, WIHRD University of Southampton

[Acknowledgements to Donald Franklin,Martin Chalkley, Ellie Cooper,Matt Salisbury]

HESG June 2005

e-mail: pormerod@volterra.co.uk

Work in progress – comments welcomed

Summary

Up to 2004, individuals using the NHS for elective surgery are essentially obliged to use their local hospital. From 2005 (and in full by 2008)individuals can choose their hospital, and in which the NHS will fund treatment at any hospital, private or public, offering care at the publicised tariff which meets minimum quality standards.

This agent-based model was developed for the economists in the Department of Health to inform on policy advice as to how this market is likely to work. The model contains decision rules for both consumers and providers. These agents are heterogeneous.

The model has been calibrated to actual data on 13 hospitals currently providing primary hip replacement surgery in the Birmingham and Black Country Strategic Health Authority. The results have been presented to professionals in the NHS both inside and outside this area.

Key findings include a high level of entrants and exits, depending on the 'exit rules', the necessity to frame clear exit rules and to stick to them, the formal linking of both the exit rules and the level of tariff to outcomes, and the need for improved specification of the relationship between signals at specialty level and the behaviour of the hospital.

1. Introduction

Within the NHS, demand has been regulated not by price, but by waiting. Substantial waiting lists exist for most of elective surgery. Consumers within the NHS have had very little choice as to where their operation will be carried out. Essentially, they have been assigned to the waiting list of their local hospital.

Under the proposed new system, consumers will in principle be able to choose the hospital where their operation will be carried out. A national tariff (price) for each particular type of operation has been set (with allowance for local market forces), and the hospital which actually performs the operation will receive this tariff.

The model can be calibrated to a specific examples, both in terms of the initial conditions and the behavioural rules themselves. Section 2 describes the operation of the model and the behavioural rules. Section 3 discusses the properties of the model, and section 4 provides examples of the sensitivity of the properties to different assumptions. In section 5, we ssummarise results of application of the model to hip operations in hospitals in the Birmingham and Black Country SHA.

2. **The operation of the model and its behavioural rules** (aiming to reflect NHS policy levers)

There are M suppliers initially. These are placed on a circle, and the distance between each one is d. The maximum distance between any pair is scaled to be equal to 1, so $d \in [0,1]$.

There are N consumers, where N >> M. These are geographically based, and are initially obliged to use the nearest supplier. (For simplicity, we assume that the distance between each consumer and his/her nearest supplier is 0).

We allow the model to move forward in time on a period by period (week by week) basis. Consumers are allowed to choose the hospital where their operation will be carried out. Once the choice is made, no further switching is allowed.

The number of consumers coming forward each period to register for the operation is fixed for each area at the outset. In other words, in each period over which the model is run, the same number come forward in any given area, although this number is different across the different areas. We thus assume that demand is independent of supply. We could incorporate growth in the overall size of the market over time into the model, but this is not done here. They could therefore be interpreted as showing the market share of each hospital in a growing market. The 'period' in the model is deemed to be a week, though this could be varied.

The number allocated to each area is drawn at random from a uniform distribution on [x, y]. These can be thought of either as individual agents or as representative agents of a larger number who have identical preferences.

We specify the initial capacity of each hospital, set equal to the number of consumers coming forward each period in the relevant area. So initially, every hospital is operating at full capacity and waiting lists are stable.

Quality and distance are both measured over the interval [0,1], and waiting times are scaled into this interval for the purposes of consumer choice. A wait of W_{max} weeks is deemed to be equal to 1, and waits below this level are scaled by (wait time in weeks/ W_{max}). A hospital with a wait time in excess of W_{max} is regarded as unacceptable by all consumers, regardless of its quality. Similarly, distance and time are scaled such that hospitals with q = 0 or d = 1 are regarded as unacceptable.

Both distance and waiting times are perceived perfectly for all hospitals by each consumer (who is given a specific admission date for a specified hospital at the point of choice). The quality of the local hospital is also perceived perfectly. However, the quality of all other hospitals is calculated by the j th consumer as $q_{kj} = q^*_k - \varepsilon_j$, where q^*_k is the true quality, and ε_j is drawn from a random uniform distribution on [0, m]. In other words, consumers are aware that they have imperfect information. They therefore make an allowance for this in their calculation of quality.

We allow M to decline over time. This is because the understanding of how a market system operates will increase over time, and a cohort entering after 7 or 8 years, say, of competition is more likely to be able to judge quality accurately than a cohort which enters during the first year of the new system.

The utility obtained by each consumer at each hospital is calculated. If this is maximised at the local hospital, the consumer goes there. If not, a consumer from a given area chooses the best hospital with probability σ_k, where this parameter is drawn from a uniform distribution on [0,1] at the start of each model solution, and is allocated to all consumers in a given area throughout the course of the solution. The value of σ varies across localities. The parameter σ is a simple way of introducing into the model a complex range of factors relating to the ability and propensity of different social groups both to gather information and to exercise choice based on that information, as well as the loyalty of patients are their advisers to historic referred providers.

4 types of consumer are modelled, their preferences calibrated on the basis of stated preference research for the London Choice project. Type A places a much higher weight on quality rather than waiting time or distance. Type B puts much higher weight on waiting time rather than quality or distance. Type C allocates its highest weight to distance. Type D is similar to Type C, except that these place even more weight on distance. One interpretation of the latter group is that they are households for whom transport is difficult and/or costly relative to their household budget.

When each consumer enters the model, he or she is allocated to Type A with probability p_A, to Type B with probability p_B, and to Type C with probability p_C, and to Type D with probability $1 - p_A + p_B + p_C$.

Each week, each hospital treats a number of patients which is equal to its capacity, if the waiting list is larger than capacity. If the waiting list is less, it simply treats this number.

Each hospital receives an identical and fixed amount of money for each consumer treated. For simplicity, this is normalised at 1, so that revenue per week is simply the number of consumers treated.

The general specification of the cost function is the same for all hospitals, and depends upon their level of quality, their effort, their efficiency, the number of consumers they treat, and their capacity. Individual hospitals differ in all these factors, so that their specific cost functions differ. It is important to distinguish between effort and efficiency. 'Effort' indicates how hard hospitals aim to keep their costs down. 'Efficiency' is intended to capture the structural efficiency of a hospital –determining the level of fixed costs. Both concepts indicate measures of how well a hospital turns cost into output, but the efficiency parameter is a fixed random draw, whereas effort is a variable that the hospital has the ability to change.

The cost function is given by:

$$(1 + q)^2*(1/(1 + e))*N_{op}*w_1 + w_2*N_{capac}$$

where q is the quality of the hospital and e is its effort, both of which are scaled to be in [0,1]. N_{op} is the number of operations performed in a week, and N_{capac} is the number of operations it could perform at full capacity. The parameters w_1 and w_2 represent the relative weights placed on the two expressions which together make up the cost functions. In the solutions of the model described below, w_1 is set equal to 0.2, and w_2 is drawn at random for each hospital from the uniform distribution on [0.6, 0.7]. The interpretation of w_2 is that this represents the structural efficiency of the hospital.

The three types of hospital (Traditional, Forward Looking (foundation) Trusts and Indpendent Sector) differ in the amounts of information which they consider in making decisions on quality, effort, and capacity, and in their motivations.

The ***Traditional*** hospitals adapt the least to the new market-oriented environment. Their basic motivation is to minimise effort subject to the constraint to break even. They pay little attention to market conditions. However, they will increase quality if they can do so at minimum effort without incurring costs.

If they make a profit in any given year, in the first instance they simply slacken off, and their effort is reduced. Effort is reduced to the level at which they predict that they will not make a profit but will break even in the next year.

The costs they will incur is specified in the cost function above, but to calculate the break even level of effort, they need to predict N_{op}, the number of operations performed in a week. They also need to provide the level of capacity, N_{capac} . They (and all other types of hospital) aim to operate at full capacity, so they adjust capacity so that it is equal to the expected number of customers. The capacity is based upon the change in demand over the previous year, and is set as follows:

$$Capacity[t+1] = Capacity[t]*\{(N_{op}[t] + WLe[t]) / (Nop[t-1] + WLs[t])\}$$

where:

Capacity[t+1] is the capacity which will be provided in the coming year
Capacity[t] is the existing capacity which has existed during the year just ended
$N_{op}[t]$ is the number of operations performed during the year just ended
$N_{op}[t-1]$ is the number of operations performed during the year previous to the one just ended
WLe = waiting list now i.e. at the end of the year just ended
WLs = waiting list at the start of the year just ended

Using this rule, hospitals adjust their capacity by the percentage change in demand which has taken place. They consider the number of operations they have performed in the year just ended and the size of their waiting list, and compare these with the number of operations performed during the previous year, and the size of their waiting list at the end of that year (which is the same as the start of the year just ended). This rule is used in all periods, except that at the end of the first year of the solution of the model, they use their existing capacity for the value of $N_{op}[t-1]$, which latter does not of course exist.

As mentioned, where a Traditional is still predicting that a profit will be made in the next year effortlessly, quality is increased until break-even is predicted. However, the fact that an increase in quality will, ceteris paribus, increase the number of consumers choosing this hospital is not taken into account, and the simple extrapolation rule for revenue is used, described in the previous paragraph.

If Traditional hospitals make a loss, in the first instance they increase effort to the level at which break-even is predicted for the next year.

If they increase effort to its maximum value of 1 and still predict that a loss will be made, they do at last take some account of market conditions, and examine their waiting lists. Each hospital is allocated a target waiting list as part of the initial conditions of the model. If the waiting list is less than the target, the hospital will reduce its capacity. If the waiting list is above target, the hospital will reduce quality by squeezing more patients in and/or in an attempt to discourage demand.

The motivation of ***Forward-Looking Trusts*** is to maximise their objective function, which is given by

$N_{op}*(1 – e)*q$

In other words, they aspire to increase the number of customers they serve, increase their quality, and reduce effort.

They want to treat as many patients as they possibly can. They make a projection of the number of customers over the next year as described above, and they invest in capacity [i.e. incur costs in the coming year] in order to provide for this number. They then choose q and e to maximise utility, given this capacity and subject to the need to break even financially. (If a loss has been made, this constraint implies that they target a profit in the coming year sufficient to offset the loss).

The new entrants from the ***Independent Sector*** aim to make profit. This is of course a simplification. In reality, there will be IS hospitals which act more like FTs or Traditionals and vice versa. The model is designed to explore differences in behavioural types rather than institutional or ownership regimes.

It must also be stressed that profit maximisation is by no means an unambiguous concept in this model. For example, at the point of potential entry, hospitals already in the market are making and implementing decisions about their quality and capacity over the coming year. The decision by an entrant will obviously depend upon whether it can be presumed to have knowledge of these decisions, or whether it has to rely on knowledge which is in the public domain, namely the quality and capacity of existing hospitals in the previous period. As it happens, it seems more realistic to assume this latter rather than the former, which is what in fact we do.

Over time, for example, new entrants may appear in the market whose overall offer is not known at any given time. Existing hospitals may alter their offers to consumers in the future. And they may the market. It is unrealistic to assume that a new entrant could have information on these factors which would enable a decision to be made which would maximise profit over a time interval of several years into the future. Purely by way of example, in order to calculate whether existing hospitals will exit the market, it is necessary to have complete information on their internal cost structures. The decisions of Independent Sector hospitals are therefore made with reference to the profits which they might make in the immediate year ahead.

At all times, these hospitals set effort equal to its maximum value of 1, which of course helps to minimise costs.

We first of all discuss the rules they use in deciding whether or not to enter in the first place.

A new entrant from the Independent Sector must replace a failed hospital in the model immediately (possibly by franchising). There is an important qualification to this, namely that the entrant must be capable of making a profit, no matter how small, by entering at this location. This rule is paramount. If the calculation suggests that a profit cannot be made, no new entrant takes the place of the failed hospital at that time. This decision is reviewed at the end of every year in the model, in other words after every 52 periods. Market conditions change, and an entry which was not profitable in a particular period might become so in future. However, the possibility exists of locations in which it is never possible to make a profit.

In addition to replacing any hospitals which exit the market at any given time, at the end of each year one new entrant considers whether to enter the market and compete at the same location as an existing one. Again, the decision to enter is based upon whether or not a profit can be made. In the current version of the model, there is no minimum profit specified, either in absolute terms or as a percentage of turnover, for entrants onto a site where an existing hospital has failed. But potential entrants which are considering competing directly at the same location with an existing one are required to expect a minimum profit of 5 per cent of total revenue in the first year before they decide to enter.

The difference in required profit margin of the two types of new entrant reflects the assumption that a replacement entrant will be able to make use of the existing capital and good will of the hospital it replaces. A new entrant competing at a location where a hospital already exists will have to attract new consumers, a riskier prospect, and so will require a higher profit margin.

The entrant, whether replacing a failed hospital or whether competing with an existing hospital, needs to decide its quality and capacity, its effort being assumed to be always equal to 1 in order to minimise costs. However, the assessment of the levels of quality and capacity to set is by no means straightforward.

The potential entrant is assumed to have knowledge of consumer behaviour, and uses this as a basis for forecasting demand. If it is replacing a failed hospital, it estimates the demand for its services at this location. If it is considering whether to compete with an existing hospital, it examines the location at which quality is the lowest which is on offer, and considers whether it is profitable to enter there. To repeat an important point made above, the entrants know that they have information on the quality and capacity of other hospitals only for the period which has just ended. They do not have information on the new levels of quality and capacity which are being set at that point in time, and they know that they do not know this.

The entrant knows the consumer demand functions and the percentages of consumers allocated to the four types of function, although it does not know the precise geographical location of the individual consumers by their demand types. The entrant calculates the number of customers it expects, based on the demand it would receive from all sites, making full use of its knowledge of patients' utility functions. It therefore takes into consideration the potential demand from other locations. It is aware that in each location there is a fixed probability, σ_i, that the consumers will switch from their existing local hospital if another offers better value. It is also aware of the fact that consumers realise they are uncertain about the quality of all hospitals apart from their initial, local one.

In general, the quality of all other hospitals is calculated by the j th consumer as $q_{kj} = q^*_k - \varepsilon_j$, where q^*_k is the true quality, and the quality discount ε_j is drawn from a random uniform distribution on [0, a]. For new entrants which replace existing ones, this same rule applies. This reflects the likelihood that the new entrant will have the option of encouraging consumers to view the change as reflecting only improved management. However, for new entrants which compete at the same location with an existing provider, ε_j is drawn from a random uniform distribution on [0, b], where b > a, on the grounds that new entrants represent a particularly unknown quality. As noted above, a is reduced over time, and so is b, so that at the end of a specified period of time (10 years in the results discussed below) they are both equal to 0.1.

A similar distinction is made for σ, the probability of a consumer switching to a hospital which would provide higher utility. For a new entrant which replaces an existing one, the customers in the area regard it as their local hospital. For new entrants which compete with an existing provider at the same location, consumers in all locations will only switch to it if it offers higher utility with the relevant value of σ_i.

Initially, the new entrant chooses its level of quality. It examines the quality provided by any existing hospital at a location, and the qualities of the two hospitals which are nearest to the location. It decides to set a level of quality which is higher than the maximum level of quality provided by this set of hospitals.

More precisely, because it knows that consumers are aware of their own uncertainty about the quality of any new entrant, it sets a level of quality which is just higher than the maximum level of quality provided by this set of hospitals by at least the expected difference in the perception of quality between the new hospital and its immediate competitors. Although the entrant is competing with all existing hospitals, there is no guarantee that this level of quality is the highest in the system as a whole, merely that it is higher than its immediate geographical neighbours. The entrant makes this choice on the grounds that a reasonably high level of quality is necessary both to establish its reputation and to assist its prospects of survival in the short term. The entrant knows the quality levels of existing hospitals in the previous period, but knows that it does not know the new quality levels which they are now setting for the next period. If it sets a quality which is too low, although it will initially gain revenue because of its lack of a waiting list, this advantage will tend to disappear and it will be driven out of business because of its low quality.

Given this level of quality, the entrant then decides the level of capacity at which it will make most profit, in the light of its knowledge of consumer demand. In practice, the simple use of these demand functions would lead hospitals to enter with capacities which are very large with respect to the total size of the market. The entrants modify this naïve calculation of capacity on two grounds. First, as noted above, they know that they only have knowledge of the quality and capacity of existing hospitals for the previous period, and not for the period which is about to begin. Second, the entrant by definition initially has no waiting list. It therefore tends to be able to attract large numbers of consumers. However, a consequence of this is that the waiting lists of existing hospitals falls, and they become more attractive to consumers. Further, not only do they take decisions at the end of the next year based upon what happens to them in the year the new entrant enters, but other new entrants may also decide to enter at that time.

The entrants calculate what their capacity would be if they, too, had the average waiting list which obtains elsewhere at the time of their potential entry. This enables a more realistic assessment of capacity to be made.

There are frequent attempts to enter the market by Independent Sector providers who compete at the same location as an existing hospital. We can see this from the cost function:

$$(1 + q)^2*(1/(1 + e))*N_{op}*w_1 + w_2*N_{capac}$$

The new entrant forms a view on N_{op} as described above. It sets capacity to provide for this expected level, so in terms of the costs it expects to incur in the first year, N_{op} and N_{capac} can be set equal. The Independent Sector hospitals set effort equal to 1, so expected costs are:

$$N_{op}*[\ (1 + q)^2*0.5*w_1 + w_2]$$

The tariff is normalised to be equal to 1, so expected revenue is N_{op}. A new entrant competing at the same location as an existing provider must expect at least a 5 per cent profit rate, so costs must be at most 95 per cent of expected revenue. An entry will therefore happen whenever the term in the square brackets is less than 0.95.

In these simulations, w_1 is set equal to 0.2, and w_2 is chosen at random from a uniform distribution on [0.6, 0.7]. The potential entrant examines the quality of the hospital in the location where it is considering entry, plus its two immediate neighbours. It sets its quality to be above the maximum level of any of these, plus the expected value of the difference between η and ε, plus a small margin set at 0.05, where η and ε are the amount by which consumers discount the quality of a new entrant and an existing hospital, respectively.

Initially, the average value of η is 0.2, and the average value of ε is 0.1, so if a Forward-Looking Trust hospital is in the comparator set, the entrant will set quality to be at least 0.65, given that the initial quality of the Forward-Looking Trusts is 0.5. So the value of the first term in the square brackets at the start of any solution of the model is 0.272. For an entry to take place at the outset of any solution, the assigned value of w_2 must therefore be less than (0.95-0.272) = 0.678. This will be the case 78 per cent of the time. There is a small adjustment to this probability, given that occasionally three Traditional Hospitals will be placed next to each other in the initial random location. So a competitive entry will take place at the outset in the clear majority of solutions.

We can see quite readily that as the quality of existing providers rises, the probability of new, competing entrants (as opposed to those entering to replace exiting hospitals) will decline. It is not clear how to solve this analytically, given that the probability of entry depends not on the average quality across the system, but on the maximum quality provided by a neighbour of the location where quality is the minimum. But the principle that competitive entry becomes less likely as overall quality increases seems clear.

Of course, when the market is switched on in practice in 2008, there will be at least one Independent Sector provider already in the market. If the existence of at least one such provider is required initially, under the existing rules of the model the easiest way to accommodate this is to discard solutions in which an Independent Sector hospital does not enter *ab initio*. As noted above, in most solutions one will enter.

Once they have entered, the Independent Sector hospitals continue to try to make as much profit as possible in the year ahead. The first action which they take at the end of each year is to set quality as low as possible, subject to it being higher than the maximum of their geographically nearest competitors, and subject to the profit constraint. If they have made a loss, their forecasts for the next year lead them to cut capacity.

Finally, a rule specifies when a hospital is deemed to have exited the market. Profits and losses are cumulated from one period to the next. In this version of the model, a hospital is deemed to exit the market for the particular operation if it incurs a cumulative loss of 8 per cent of total cost in any given year.

3. Results with the base version of the model

We start the model with 10 hospitals, 5 of which are allocated at random to behave as Traditionals, and the remaining 5 to behave as Forward-Looking Trusts.

The number of consumers coming forward each week in each area is drawn from a uniform distribution on [10,100]. The average number of consumers is therefore 550. As noted above, each one of these can be thought of as being representative of whatever number of actual consumers is realistic.

The capacity of each hospital is set at the same number as the number of consumers coming forward each week, so that each hospital is initially operating at full capacity. The same length of waiting list is allocated to each hospital, at 19 weeks.

The initial quality of the Traditionals is set at 0.225 and of the Forward-Looking Trusts at 0.5.

The consumer demand functions are as follows:

8 per cent are Type A, and their utility is given by

$$U_A = \text{quality}^{0.8} * (1-\text{wait})^{0.1} * (1-\text{distance})^{0.1}$$

23 per cent are Type B, and their utility is given by

$$U_B = \text{quality}^{0.1} * (1-\text{wait})^{0.8} * (1-\text{distance})^{0.1}$$

45 per cent are Type C, and their utility is given by

$$U_C = \text{quality}^{0.2} * (1-\text{wait})^{0.2} * (1-\text{distance})^{0.6}$$

24 per cent are Type D, and their utility is given by

$$U_D = \text{quality}^{0.2} * (1-\text{wait})^{0.2} * (1-2*\text{distance})^{0.6}$$

where in the above equations all variables are scaled in [0,1] . As noted above, there is a maximum wait time which consumers will consider (W_{max}), which here is set at is 39 weeks.

The weights on distance in the Types C and D functions and the relatively large sizes of these two groups have an important influence on the overall properties of the model. To anticipate somewhat, they make it easier for the initial group of Traditional and Forward-Looking Trusts to survive. Utility functions in which the average weight on distance across all consumers is considerably less often give solutions to the model in which, over a period of years, the original providers are all driven out of business. Provision in such solutions is concentrated upon a relatively small number of Independent Sector hospitals, and there are locations in which it is not profitable for a hospital to exist. The size of such hospitals tends to be large, and the improvements to quality and the falls in average waiting times tend to be greater than is the case with the above set of utility functions.

All consumers coming forward each week in a given area are allocated a value of σ, the propensity to switch from the local provider if a higher level of utility could be obtained elsewhere. The values of σ are fixed throughout all the periods of each individual solution.

In addition, the quality of the local hospital is also perceived perfectly. However, the quality of all other Traditional and Forward-Looking Trust hospitals is calculated by the j th consumer as $q_{kj} = q^*_k - \varepsilon_j$, where q^*_k is the true quality, and ε_j is drawn from a random uniform distribution on [0, 0.2]. The value of ε_j declines linearly over time, so that it is 0.1 after 10 years. For Independent Sector hospitals which enter and compete at a location with an existing hospital, their discount value, η_j is drawn from a random uniform distribution on [0, 0.4]. The value of η_j declines linearly over time, being equal to 0.1 after 10 years.

At the outset of the model, one Independent Sector hospital considers whether or not to enter at a location where a hospital exists.

The model advances in a series of steps, or periods. In each period, the consumers coming forward for the operation choose their hospital. The model progresses for 52 weeks, when all hospitals in the market either exit or alter their offers in accordance with the above rules. At the same point in time, an Independent Sector hospital considers whether or not to enter at a location where a hospital has just exited the market. And another Independent Sector hospital considers whether or not to enter at a location where a hospital continues to exist.

We run the model for 10 successive batches of 52 periods. In other words, for 520 weeks or 10 years.

A hospital is deemed to exit the market for the particular operation if it incurs a cumulative loss of 8 per cent of total revenue in any given year.

We describe the results obtained with 500 separate solutions of the model. This is sufficient to describe the range of results which the model generates. The model is programmed in Matlab, and a single solution on a standard PC takes around 1 minute to solve for 10 hospitals over a period of 520 weeks.

In general, we find that:

- average quality improves
- average waiting times fall
- consumer utility increases
- capacity utilisation falls

In each case, averages are the appropriate weighted average by number of consumers.

Figure 1 summarises the key findings for consumers in the simulations. This plots results from the 500 solutions at the end of year 10 in each of the solutions.

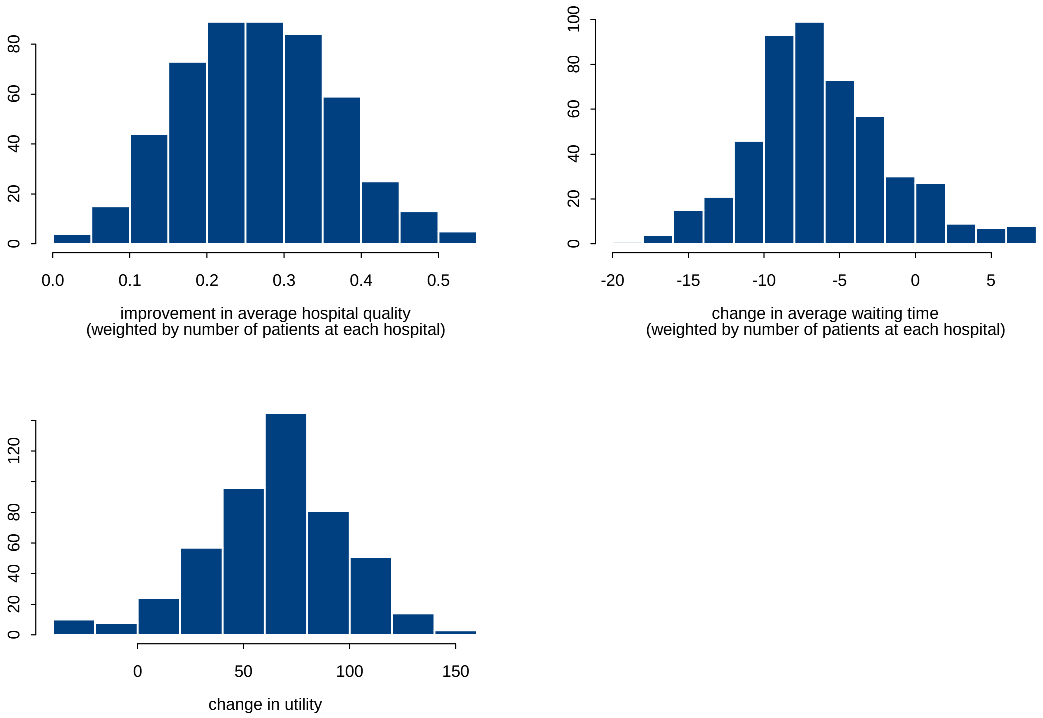

Figure 1 *Changes in average quality, average waiting times and total consumer utility. Summary of results from 500 separate solutions of the base model, at year 10*

In order to put the results into perspective, we need to recall the scales over which the various variables are measured. At the outset of each solution, the average weighted quality of hospital is 0.3625. The 5 Traditional hospitals have quality of 0.225 and the Forward-Looking Trusts 0.5. The number of consumers allocated to each one is chosen at random, so the actual weighted initial quality will vary from solution to solution, but on average it will be 0.3625. Quality is scaled within the interval [0,1], so the increases in quality which are observed are substantial with respect to the initial level, the average being 0.26.

In the consumer demand functions, waiting times are also scaled in [0,1], but here we show them in terms of actual weeks. The initial level is 19 weeks for each hospital. 90 percent of solutions show average waiting times to fall. The average (weighted) fall of almost 6 weeks is again non-trivial compared to the initial 19 weeks' wait.

The change in total utility is obtained as follows. At the outset, the level of utility for each consumer is obtained by calculating it from the appropriate utility function using the values of the local hospital, which the consumer until then had been obliged to visit, and is stored in a vector. At the end of 10 years, the same calculation is performed for the consumers whose operations have just been carried out and is stored in another vector. Of course, these are different individuals from the ones at the start of the model, but they have identical utility functions. To gauge the improvement in utility, we simply subtract the two vectors and sum the differences.

We can examine in a little more detail the outcomes on quality, waiting times and utility. Figure 2 plots the increase in quality and the final number of hospitals, in other words the number of hospitals at the end of year 10.

Final Number of Hospitals and Change in Quality

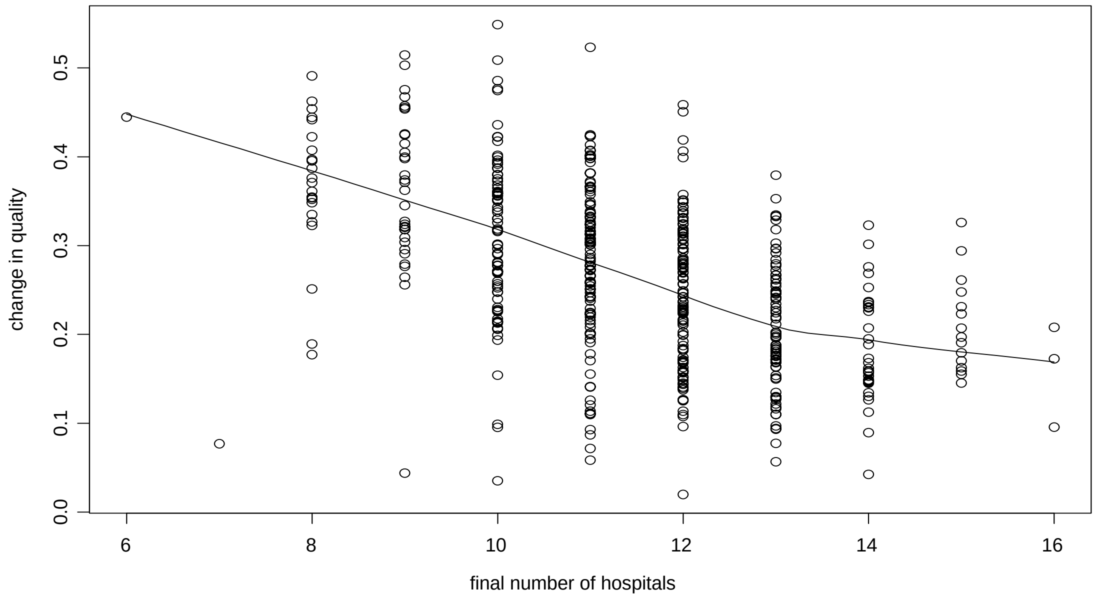

Figure 2 *Increase in weighted average quality and number of hospitals in existence at the end of year 10. Solid line is a (mildly) non-linear least squares fit*

A simple linear regression shows that a further determinant is the final number of Independent Sector hospitals:

Increase in quality: OLS regression

```
Coefficients:
                  Value Std. Error  t value Pr(>|t|)
   (Intercept)   0.4538   0.0230    19.7225   0.0000
final.num.hosp  -0.0307   0.0018   -16.9107   0.0000
  remaining.IS   0.0308   0.0019    15.9512   0.0000

Residual standard error: 0.06806 on 497 degrees of freedom
Multiple R-Squared: 0.5128
```

where final.num.hosp is the number of hospitals at the end of year 10 and remaining.IS is the number of Independent Sector hospitals. The number of the other two types is not statistically significant.

The Independent Sector hospitals do of course maintain a reasonably high level of quality for reputational purposes to assist their long-term prospects of making profits. However, their entry is not essential for changes in quality to take place, simply allowing choice forces hospitals to improve their quality, regardless new entrants. In the absence of new entrants, choice will reward

Traditional and Forward-Looking Trust hospitals with higher quality allowing them to further invest in quality. However entry is a significant driver for improvements in quality.

The main determinant across the solutions as a whole of both the change in waiting times and the changes in utility is in fact the number of Independent Sector entrants at the end of year 10 rather than the final number of hospitals as such.

The performance of hospitals across the 500 solutions is set out in Figure 3.

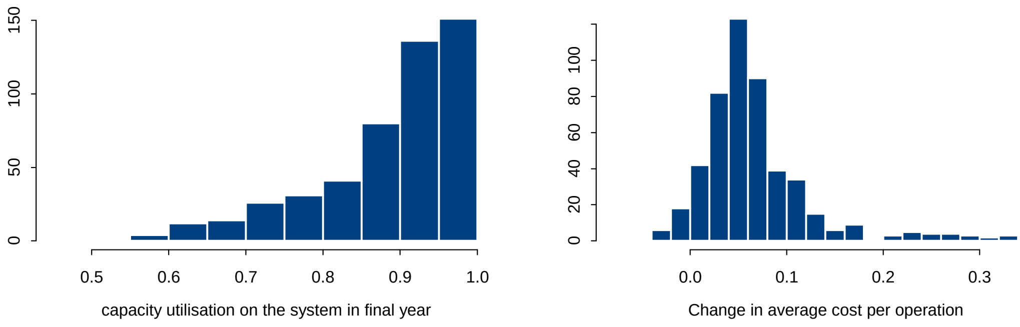

Figure 3 *Capacity utilisation in year 10 and the change in average cost per operation*

Capacity utilisation is initially, by design, equal to 1 (or 100 per cent). In most solutions of the model, a high capacity utilisation continues to obtain. The average increase in capacity across the 500 solutions is not that high. Initially, it averages 560 per week, and at the end of year 10 the average is 624.

In general, the cost per operation increases. This arises in the first instance from the increases in quality and capacity. There are two separate terms in the cost function, one of which describes what we can think of as the variable costs, factors such as the number of operations actually performed, the level of quality and the level of effort. The other describes the costs associated with the level of capacity provided. Total costs, as we have seen, are a weighted sum of the two factors. Increases in both quality and capacity, *ceteris paribus*, increase the costs of an operation.

Providers are able to influence the first part of the cost function by the level of effort which they put in. However, there is no mechanism in the rules for the providers to increase the efficiency with which they absorb the costs associated with their capacity. The weight which is used on the level of capacity is fixed at the outset for each provider whenever it enters the market. It is drawn at random from a uniform distribution on [0.6, 0.7], so the providers differ in the efficiency with which they can absorb costs. But there is no rule which enables them to increases efficiency over time. Further, the range from which the weight is drawn remains fixed over time, so the entrants in the later years are just as likely to draw an 'inefficient' value close to 0.7 as are the initial providers at the outset of the solution.

In a centrally planned system, this assumption may very well be realistic. In a market-oriented one, it is more likely that efficiency gains would be made in all aspects of provision.

However, the absence of such a mechanism in the model avoids the potential charge that any desirable results which emerge are obtained by positing an effect which may or may not occur in practice.

Finally, Figure 4 shows what happens to the number of providers in the various sectors.

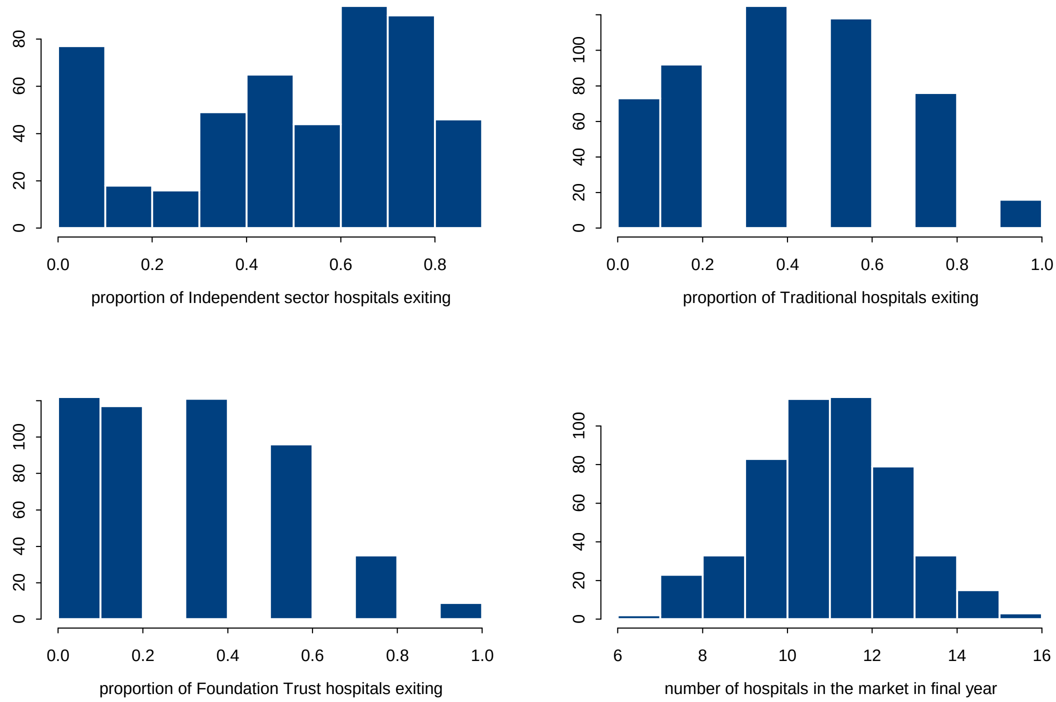

Figure 4 *Proportions of Hospitals Exiting and Number of Hospitals in the market after 10 years*

There is no individual solution in which all the original 10 providers are driven out of the market. On average, 2.2 out of the original 5 Traditional hospitals exit by the end of year 10, and 1.7 of the original 5 Forward-Looking Trusts. The exit rate for the new Independent Sector entrants is high, the average being 50 percent across the 500 solutions, but this is entirely typical of what happens when markets are opened up to competition. The incumbents have a survival advantage because of consumer familiarity with them, and new entrants have relatively high death rates. Land line telecommunications and the domestic supply of energy are examples of such phenomena.

In the course of developing the model, a certain amount of experimentation with rules took place, which enables us to offer some very brief qualitative reflections on how different results might emerge. As noted above, the high weights on distance in the utility functions of Types C and D consumers are important. These reinforce the fitness for survival of the initial set of hospitals, giving as they do a strong advantage to the local provider even when choice is introduced. The less the average weight which consumers place on distance, it appears that the more likely it becomes that the original providers will be driven out of business. The quality and utility gains are higher, but there tends to be fewer final providers with high capacity, and some locations are unable to support a profitable hospital. Similar results emerge whenever Independent Sector entrants take particularly optimistic views of the number of customers they are likely to obtain, so that they tend to choose high levels of capacity on entry.

4. Alternative scenarios

In terms of a more formal analysis of the sensitivity of results in this particular model, we now consider three scenarios in which, in turn:

- the exit rule is made much more lax. Specifically, a hospital now exits only if its losses exceed 20 per cent of turnover in any given period
- the tariff per patient is increased by 10 per cent from 1 to 1.1
- the probability of consumers switching to the utility maximising provider [σ_j] is drawn from [0.5, 1] rather than [0, 1] as in the base solution

The results of these three variants are summarised in Table 1, in each case taking the average of 500 separate solutions of the model and comparing them with the base case described above.

Table 1

	Change in average waiting time, weeks	Increase in quality	Increase in utility %
Base case	-5.92	0.265	17.1
Lax exit rule	-5.08	0.210	16.3
Higher tariff	-6.36	0.466	28.0
Higher propensity of umers to switch	-8.45	0.349	18.4

The reasons for the differences in average outcomes from those of the base case are essentially as follows:

Lax exit rule

This makes it easier for Traditional and Forward-Looking Trust hospitals to survive

Higher tariff

This makes it easier for Traditional and Forward-Looking Trust hospitals to survive. but it also makes it easier for Independent Sector hospitals to enter. In particular, the requirement on entrants competing at the same location with an existing provider to make 5 per cent profit, is more easily satisfied

Higher propensity to switch

This makes entry easier for the Independent Sector. Note, however, that the increase in consumer utility is not much above that of the base case, despite the fact that average wait times are considerably less and the increase in quality definitely higher. But consumers of Types C and D, who attach high weights to distance, suffer in solutions in which locations appear in which no hospital can survive.

5. Results from application to SHA

The model was applied to hip replacement in one SHA covering 13 hospitals, both NHS and private. The results indicated a range of problems to do with three different levels: procedure, specialty and hospital level. Data availability varied by level (unit cost by procedure, waiting time by specialty, quality by hospital) posing aggregation problems. More seriously, the lack of well established relationships between levels, particularly between specialty and hospital, not only raised questions about the behaviour of hospitals as ‘firms’, but also prompted issues to do with cross subsidisation. To what extent could a hospital cross subsidise a loss making specialty, and might this be deemed anti-competitive?. The scope for such cross subsidisation might well vary between large general (NHS) and specialist (often private) hospitals. This in turn raised questions about appropriate regulation of the emerging market.

6. Conclusions

Key findings include:

- a high level of entrants and exits are likely, the level depending on the exit rules
- the necessity for clear exit rules that are maintained if the market is to develop
- the formal linking of exit rules, level of tariff to patient outcomes
- further research is needed on how hospitals as organisations respond to signals at specialty level, which may be different across specialties,
- issues of cross subsidisation may arise between NHS hospitals (which typically have both elective and emergency cases) and private providers (typically elective only). This in turn prompts questions about the role of the regulator.

7436 words