Journal of Public Health Medicine

Vol. 22, No. 3, pp. 413–421
Printed in Great Britain

Ambulance economics

A. J. Fischer, P. O'Halloran, P. Littlejohns, A. Kennedy and G. Butson

Abstract

Background Ambulance services produce a large quantity of data, which can yield valuable summary statistics. For strategic planning purposes, an economic framework is proposed, and the following four resource allocation questions are answered, using data from the Surrey Ambulance Service: (1) To satisfy government response time targets, how many additional ambulances will be required, *ceteris paribus*? (2) To minimize average response time (r^*) with given resources, how should ambulances be rostered temporally? (3) Which innovations are worth undertaking? (4) How would an increase in demand affect r^*?

Methods The 'Ambulance Response Curve' – the relation between response time and the number of available but not-in-use ambulances – is used to estimate how much r^* will be reduced by deploying an additional ambulance. Estimating the marginal cost of an ambulance allows us to estimate the opportunity cost of each second of response time, and to compare the cost of three 'innovations' with that of increasing resources. The time savings of adding an extra ambulance at each of the 168 h of the week are examined.

Results In 1997–1998, r^* was 8 min 52 s. An additional ambulance reduces r^* by 8.9 s. Each reduction of 1 s in r^* costs £28 000 per year. Fourteen additional ambulances are required to meet response time targets if the 8.9 s reduction per ambulance is maintained. r^* reduces by 4.6 s when ambulances are shifted from early mornings to Saturday evenings. Activation time reduces by 38 s when crews sit in their ambulances. A 1 min decrease in overall call time decreases r^* by 1.1 s. Answering only 10 per cent of all calls reduces r^* by 63 s. An increase of demand of 10 per cent increases r^* by 7.8 s.

Conclusions Ambulance services will be better able to determine which innovations are worth undertaking. Policy makers will be better placed to determine funding levels to achieve response time targets.

Keywords: ambulance, economics, resource allocation, innovation

Introduction

A fast and reliable ambulance service is taken for granted in modern societies. In Britain, demand for ambulances has been increasing at the rate of about 4 per cent each year throughout most of the last decade, and at 9 per cent, 5 per cent, 7 per cent and 8 per cent over the last four years.[1] Yet very little work has been undertaken on the efficacy of innovations in the ambulance service.[1,2]

Improvements in automated data capture in British ambulance services over the past few years now make it possible to analyse their performance in meeting time targets. A national target for meeting a response time for life-threatening calls of 8 min or less in 75 per cent of cases has been set for the fiscal year 2000–2001. ('Response time' refers to the time elapsed from when the ambulance control room answers an incoming '999' call to when the ambulance arrives at its destination.)

Currently, only one of 37 NHS ambulance trusts (Staffordshire) meets this target. An important question of resource allocation is to determine how many additional ambulances would be required to meet the standard, given that current performance standards are maintained (i.e. neither improved nor made worse) and provided there were no changes in demand for ambulances. This paper develops an economic framework to answer this question.

In this paper, we show how our framework is also capable of answering a number of other questions pertinent to the provision of ambulances:

(1) How should ambulances be rostered (at different times of the day and days of the week) to minimize average response time for a given budget?
(2) Will an innovation with the same cost as an additional ambulance provide more or less benefit than the additional ambulance? That is, is an innovation worth undertaking? We consider three possible innovations, as examples of how the method may be employed. Other innovations may also be evaluated using this framework.
(3) Other things being equal, how would an increase in demand affect response times? We look at the case of a 10 per cent increase in demand.

Health Care Evaluation Unit, Public Health Sciences Department, St George's Hospital Medical School, Cranmer Terrace, London SW17 0RE.

A. J. Fischer, Senior Lecturer in Health Economics

John Lewis Partnership, Oxford Street, London W1A 1EX.

P. O'Halloran, Statistician (and formerly Health Care Evaluation Unit)

National Institute for Clinical Excellence (NICE), 90 Long Acre, London WC2E 9RZ.

P. Littlejohns, Clinical Director (and formerly Director, Health Care Evaluation Unit)

Surrey Ambulance Service, The Horseshoe, Banstead SM7 2AS.

A. Kennedy, Chief Executive Officer

G. Butson, Fleet and Resource Manager

Address correspondence to Dr A. J. Fischer.
E-mail: afischer@sghms.ac.uk

It is important to note that this analysis does not model where ambulances should be located: that is a further exercise.

We have used data from the 1997–1998 records of calls to the Surrey Ambulance Service (SAS) to analyse response times. Over 10 months from 1 April 1997 to 31 January 1998, 75 239 calls were attended. The SAS is classified as an urban service and must meet targets set for urban areas (currently 50 per cent of calls within 8 min and 95 per cent within 14 min). It operates up to 37 ambulances at a time from 19 bases. Usually, up to 37 ambulances are used during the day, dropping to 19 in the early hours of the morning, and averaging 28. When the ambulances at a particular base are all in use, ambulances not in use in adjacent areas will be sent to 'cover' for that area. This may cause a domino effect, of ambulances further afield covering for the ambulance which is providing the primary cover. Ambulances on cover may travel all the way to another base, whose ambulances are already in use, but may also stand in a roadside lay-by, part of the way to the base. Even where no ambulance from a base is in use, an ambulance from that base may be required to stand in a lay-by as part of a strategy of covering the county more evenly, rather than at a base. In this way, it can reach some of the population more quickly than otherwise.

Of the 75 000 calls, 59 000 of them were so-called 'emergency' calls, and 16 000 'urgent' calls. The urgent calls comprise mainly inter-hospital transportation of patients and calls by GPs to transport patients to hospital appointments. Emergency calls receive priority over urgent calls, the latter of which are generally answered within 2 h. At the time that the data refer to, emergency calls were not prioritized. However, a prioritization scheme has since been introduced.

Method

Data

The dataset of the 75 000 calls to the Surrey Ambulance Service (SAS) consists of detailed information on every call made, including all relevant times, ambulance call number, name and address of the patient, name of caller, reason for call, whether the ambulance was at its own base, at another base or on standby at the time of the call, and the name of the hospital attended. From this, activation time, response time and all-round trip time can all be calculated.

As well as this, at the conclusion of the call, the crew categorized it as either life threatening (category A) or non life threatening (category B). Our analysis of response times is of the emergency calls, although to find out how many ambulances were being used (and therefore not available for emergency use) we had to take urgent use into account.

When more than one ambulance has been called, the response time has been calculated for only the first ambulance called. This is because sometimes the second ambulance is called only after the first has arrived, and therefore with a considerable lag, the time being measured from the first contact with the ambulance service.

Average and 75th percentile response times

We examine average response times, rather than the response time of the 75th percentile, to be the new industry standard. There are four reasons for this: first, the mean contains more information than the 75th percentile; second, it is arguably more easily interpreted. The other two reasons are technical: response times at the time the data were collected were expressed in whole minutes, so that, for example, both 8 min 0 s and 8 min 59 s were recorded as 8 min. (These times were recorded by hand; since then, they have been automatically recorded to the nearest second.) Thus 75th percentiles were recorded to the nearest minute over a number of observations, whereas mean times over a number of observations could be estimated to the nearest second, with an error of only several seconds. Fourth, and more importantly, in the regression equations that follow, the regression of response time (r) against number of available not-in-use ambulances (n) is carried out for each of the 55 319 observations with a valid response time as data points, and not on the 35 mean response times (r^*) for the 1–35 available ambulances. It is not possible to do this calculation for the 75th percentile on an individual call basis, and the regression would have to be carried out using only 35 data points.

Nevertheless, the answers for the mean response times have been translated into 75th percentile response times using the regression equation

$$75M = a + br^* \quad 2 < n < 33 \tag{1}$$

where $75M$ is the 75th percentile response time and r^* is the mean response time when there are n ambulances available and not in use.

The effect of an additional ambulance: the Ambulance Response Curve

The centrepiece of our analysis is what we call the Ambulance Response Curve (ARC). This shows the relationship between the response time for an individual call (r) and the number of ambulances available and not in use (n) at the time the call was made. For example, let us suppose that 35 ambulances are on duty and 10 of them are being used. Then n has the value of 25 when the next call is taken. *Ceteris paribus*, as n increases, we expect that r will fall.

The data were already in chronological order of receipt of call. Using only the standard statistical packages *Excel* and *SPSS*, at activation time of an ambulance, a new variable was created with the value of (-1) to represent its departure from the not-in-use fleet. When the ambulance returned and was once more available for use, a different new variable was given the value of $(+1)$ to represent its addition to the not-in-use fleet. The return times were then put into a separate file and sorted

into chronological order. The (+) and (−) files were then merged chronologically, and a running tally was kept of the number of ambulances being added to and subtracted from the fleet of not-in-use ambulances. Similarly, a third file containing the numbers of ambulances beginning and ending their shifts (+ for beginning and − for ending) was merged with the file that had already been merged, to obtain the numbers of not-in-use ambulances at all times. In this way, the number of available ambulances (not in use at the time of the call) was associated with the response time of the ambulance, for all calls throughout the year.

We estimated the relationship between r and n by means of linear regression with and without hour-of-day covariates to account for traffic congestion, and then investigated several non-linear forms. Mathematical modelling of distances that equally spaced ambulances on a square grid of uniform population density would travel suggests that, ignoring ambulances returning from previous assignments, r will be related to n in the following fashion:

$$t = r - a - d = bn^{-\frac{1}{2}} \tag{2}$$

where t is the traffic-adjusted travel time, a is the activation time, which is known for each of the 55 000 observations, and d is the coefficient of the time dummy already estimated from the linear regression with hour covariates.

Thus, we undertook a regression of the form

$$t = r - a - d = bn^{\alpha} \tag{3}$$

where α is to be determined from the regression equation.[3]

When n is zero or negative (denoting a queue for the next ambulance), this equation will not hold. For such values of n, and to some extent for small positive values of n, the response time will depend on the rate of return of ambulances currently being used. For the small number of cases that this applies to, a linear ARC is thought to be the most likely. (We investigate the shape of the ARC in a separate paper, currently being written.)

The slope of the ARC tells us the amount by which response time would improve if there were one additional ambulance at the time the call were made. This is a measure of the marginal benefit of additional resources. What is understood by this is rather subtle, as it is utilizing the *ceteris paribus* ('all other things equal') condition of economics and scientific method. This condition would allow an additional ambulance to be used in no better or worse a manner than existing ambulances. It would not be put in an out-of-the-way location where it was rarely if ever used. It would, however, need to be used as a single ambulance, day and night, as that would allot the same marginal resource (one additional ambulance) to all time periods. However, it is likely that an ambulance service would allocate the additional hours in a more optimal fashion than this. For example, if the additional ambulance hours were allocated in the same way as the existing hours, then there would be about 1.3 additional ambulances allocated to day shifts and 0.7 to night shifts. This would give an average of one additional ambulance overall. However, apart from being much more complicated, it is also unnecessary to consider these things, as we can consider modelling the exchange of ambulances between time-slots to effect a better allocation of resources as a separate exercise. We do this in the subsection after next.

Marginal cost

To determine the marginal cost of running an ambulance continuously for a year, we found the average number of crew per ambulance, and the proportions that were paramedics and technicians.

Ambulances are leased, and the annual leasing cost includes maintenance and equipment. Fuel costs are only those of idling time, as demand is assumed not to change when there is an additional ambulance. Small additional costs of uniforms, administrative consumables and additional insurance were used to round up the total figure to the nearest £10 000 per year.

Optimally allocating ambulances by time-of-day and day-of-week

For the Surrey Ambulance Service, let us suppose that we compare ambulance usage between 4 a.m. and 5 a.m. Tuesdays with that between 11 p.m. and midnight on Saturday nights. There were 143 calls in the data period in the Tuesday timeslot, and 535 calls in the Saturday slot. On Tuesday morning, the average value of n was 15.9 whereas for Saturday night it was 11.1. From the ARC [see Fig. 1 and equation (8), below] we estimate that, for $n = 15.9$, the marginal response time savings of an ambulance are 9.4 s, whereas for $n = 11.1$, the time savings are 14.4 s. The marginal response time savings, aggregated over all calls, for each period are

Tuesday 4 a.m. to 5 a.m.: $143 \times 9.4 = 1344$ s
Saturday 11 p.m. to 12 p.m.: $535 \times 14.4 = 7734$ s.

A marginal ambulance will save more response time in aggregate late on Saturday evening than in the early hours of Tuesday. The formula used to calculate the aggregated marginal response time savings is given by $c.\Delta t$, where c is the number of calls in the time-period and Δt the slope of the ARC for the average number of available not-in-use ambulances at that time.

We can use this method to show how to allocate ambulances at different points in time. If we subtract an ambulance from Tuesday morning, Δt will rise above 9.4 s, as a result of the curvature of the ARC. If we add that ambulance to Saturday evening, the Saturday Δt will fall below 14.4 s. As we continue to shift ambulances from Tuesday to Saturday, Tuesday aggregate time savings will increase and the Saturday aggregate will decrease until they are equilibrated. Any further shifting of ambulances beyond this point will result in aggregate time losses, so the process is self-limiting.

This process has been undertaken for each hour of the day and day of the week. Where $c.\Delta t$ is below average, resources

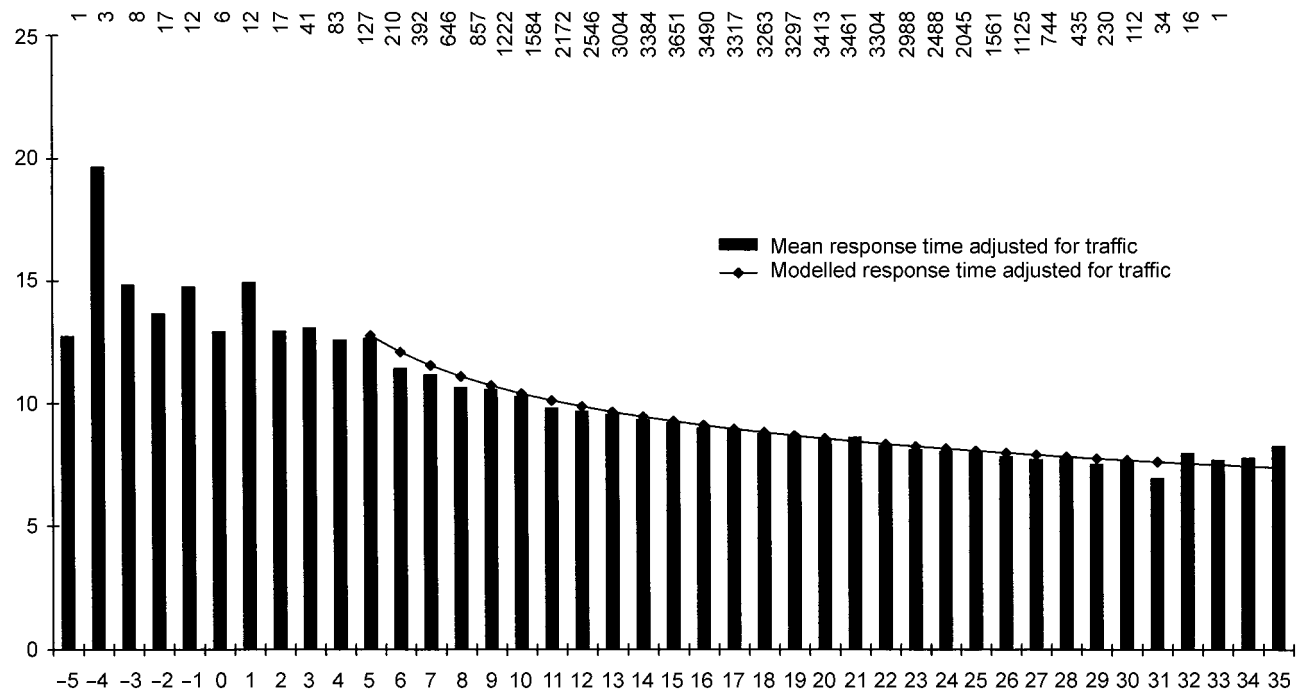

Figure 1 Predicted mean response time adjusted for traffic (filled diamonds and curve) and actual mean response time adjusted for traffic (filled bars) by number of available ambulances.

should be subtracted, and, conversely, they should be added when $c.\Delta t$ is above average.

Innovations

The first innovation we wished to test was to see whether there were time savings from sitting the crew in the ambulance. We estimated from the data the activation time for crews at base and in lay-bys. This gives us a measure of gains or losses in r that may be made from sitting the crew in the ambulance (lay-bys) or not (home base). We estimated separately the value of r^* for ambulances going to home-base zones from home base, to home-base zones from lay-bys, and to non-home-base zones from home-base and from lay-bys.

The second innovation was to examine the effect of having a faster job turnaround time.

The following equation describes how the average time (T) between one emergency call and the next is composed:

$$T = j + u + s \tag{4}$$

where j is the average overall time for an emergency call, u is the average time spent on urgent calls between emergency calls, and s is slack time. A decrease in j of 1 min translates into an increase of 1 min in s. It is the proportionate increase in s that matters, as s, the so-called 'slack', is the true time resource of an ambulance service: it is slack time that determines the value of n.

(It has been suggested by a reviewer that this analysis is too simplistic, and that the duration of the interval between the calls 'depends entirely on the shapes of the response and interval distributions'. We have looked at this proposition by disaggregating the data, and have not found that it makes any difference to our results.)

The third innovation was to examine what response times would have been if the SAS were only to answer the calls its crew said were life threatening. Other calls would be left forever unanswered! (Admittedly, this example is unrealistic, but it provides an upper limit on the time savings from triage.) We did this by finding the value that n would have been if only the A-classified calls had been answered. Let us call this number n_A. We then subtracted $[r^*(n) - r^*(n_A)]$ from the value of $r(n)$ for each of the A calls, where $r^*(n)$ is the average value of the response time when there are n ambulances available and not in use, and similarly for $r^*(n_A)$. [For example, let us suppose $n = 21$ for a particular A call, and the response time for that call was 11 min. We suppose the average response times for all calls when $n = 21$ was 9 min. If only A calls were being answered, we suppose that $n_A = 29$, for which r^* is assumed to be 8 min. There is a time saving of $(9 - 8 =)$ 1 min, by having 29 ambulances instead of 21 when the call is made. So we amend the time of the call from 11 to 10 min. We do this for all the A calls. If the savings are proportional to the length of the trip and not fixed, this will understate the savings in this example, where the trip is longer than average. But this will be offset completely by trips that are shorter than average.]

Results

Average and 75th percentile response times

There was a strong relationship between average and 75th percentile response times, as given in the equation

$$75M = 0.136 + 1.163r^* \quad 2 < n < 33 \quad R^2 = 0.981. \quad (5)$$

(0.285) (0.030) (SDs in parentheses)

When $r^* = 9$ min, the error on $75M$ will therefore be 3.2 s.

r^* for all calls was 8 min 52 s, or 8.87 min, with a standard deviation of only 0.99 s. A 75th percentile response time of 10 min 26 s (10.44 min) is implied from equation (5).

Estimating the Ambulance Response Curve

To picture the Ambulance Response Curve, it is instructive to examine the relationship between the mean response time (r^*) for each n (number of available ambulances not in use). This is given in Fig. 1. (The relationship between r and n is a scatter of over 50 000 points and is not informative.) Negative values of n represent the size of the queue at that time. There were very few observations with negative n, so the standard error for the mean response time for each negative value of n is fairly high. However, for $n > 2$, the number of observations for each n rises from over 100 to several thousand when n is between 20 and 30. (The number of observations on which the mean response time is based for each n is given at the top of Fig. 1.)

In part because of the large number of observations making up r^* for each n, the graph shows a remarkably steady decline in r^* as n increases, from $n = 2$ to $n = 34$.

For the relationship between the 55 000 observations linking r to n, however, we first list the equation for the simple linear regression. It should be noted that in this and all other regression equations, the times are all quoted in minutes.

$$r = 10.74 - 0.100n \quad R^2 = 0.020. \quad (6)$$

(0.06) (0.003) (SDs in parentheses)

The coefficient of n shows that there is a decrease in response time (on average) of 0.100 min, or 6 s, for each additional ambulance. When time dummies are put into the equation, however, the coefficient of n changes sharply from −0.100 (6 s) to −0.161 (9.7 s). The full equation, where h02 refers to the hour between 1 a.m. and 2 a.m., etc., is given by

$$\begin{aligned} r^* = {} & 11.50 - 0.161n + 0.19\text{h02} - 0.01\text{h03} - 0.01\text{h04} \\ & -0.10\text{h05} - 0.24\text{h06} - 0.01\text{h07} - 0.20\text{h08} + 0.81\text{h09} \\ & +0.41\text{h10} + 0.26\text{h11} + 0.46\text{h12} + 0.68\text{h13} + 0.37\text{h14} \\ & +0.55\text{h15} + 0.69\text{h16} + 1.00\text{h17} + 1.20\text{h18} + 1.35\text{h19} \\ & +0.84\text{h20} - 0.18\text{h21} - 0.46\text{h22} - 0.41\text{h23} + 0.13\text{h24} \\ R^2 = {} & 0.029. \end{aligned} \quad (7)$$

(0.10) (0.0046) (0.11–0.15 for the hour variables)

(SDs in parentheses).

The reason that the n-coefficient value changes so much is that, in essence, there are two patterns of ambulance provision. The value of n during the day hours was almost always above 19, but at night hours was never above 19. This is shown in Fig. 2, where the regression line of response times for different n during night hours is shown by the curve with triangles (the hour chosen to represent these times being 6 a.m.), whereas for day hours (the representative hour being 6 p.m.) it is shown by the curve with diamonds. The regression line without the dummy variables is shown in black, and is the line whose slope is 6 s. It may be thought of as the line that fits the scatter of points of both curves together, but it clearly underestimates the true slope of either the daytime or night-time ARC. This is an important point, because it shows that each additional ambulance provides more benefit than the simple linear regression would have us believe.

As the ARC is clearly convex to the origin, however, a linear curve is not the most appropriate, and we proceeded by estimating equation (3):

$$t = r - a - d = 19.028n^{-0.376}. \quad (8)$$

Confidence limits for the value of the exponent are (−0.364, −0.388). For $n > 4$, the curve is an extremely good fit for the data, as is shown in Fig. 1. As the slope of equation (8) is changing, we could not find the average time saving of running an additional ambulance directly from a coefficient of equation (8). To achieve this, we calculated $\hat{t}(n) - \hat{t}(n+1)$ for all 55 000 observations, and averaged them. [$\hat{t}\,(n)$ is the value of $t(n)$ estimated from equation (8).] This gave us 8.90 s. The importance of this figure is that it is what we believe to be the best estimate of the marginal benefit of an ambulance. As we cannot readily determine the error on this estimate, we use the percentage standard error (of 3 per cent) on the n coefficient from equation (7) as a proxy. Equation (7) is not such a good fit as equation (8), but is more easily interpretable. However, the error estimated in this way will be an overestimate.

For $n < 5$, we estimate

$$r = 14.013 - 0.397n. \quad (9)$$

(0.583) (0.181)

At this low level of n, an additional ambulance reduces response time by 0.4 min = 24 s [confidence interval (CI) 2–45]. However, this level of n is associated with only 0.4 per cent of emergency calls.

Marginal cost

Cost of crew, including on-costs, in 1999 £ per year:	
paramedics	146 378
technicians	78 322
Leasing of ambulance	21 000
Other annual costs (to round up total)	4300
Total	250 000

The last amount (£4300) includes an amount of about £1000 (5 per cent of £21 000) to cover for ambulance down-time – it takes about 1.05 ambulances to provide one working

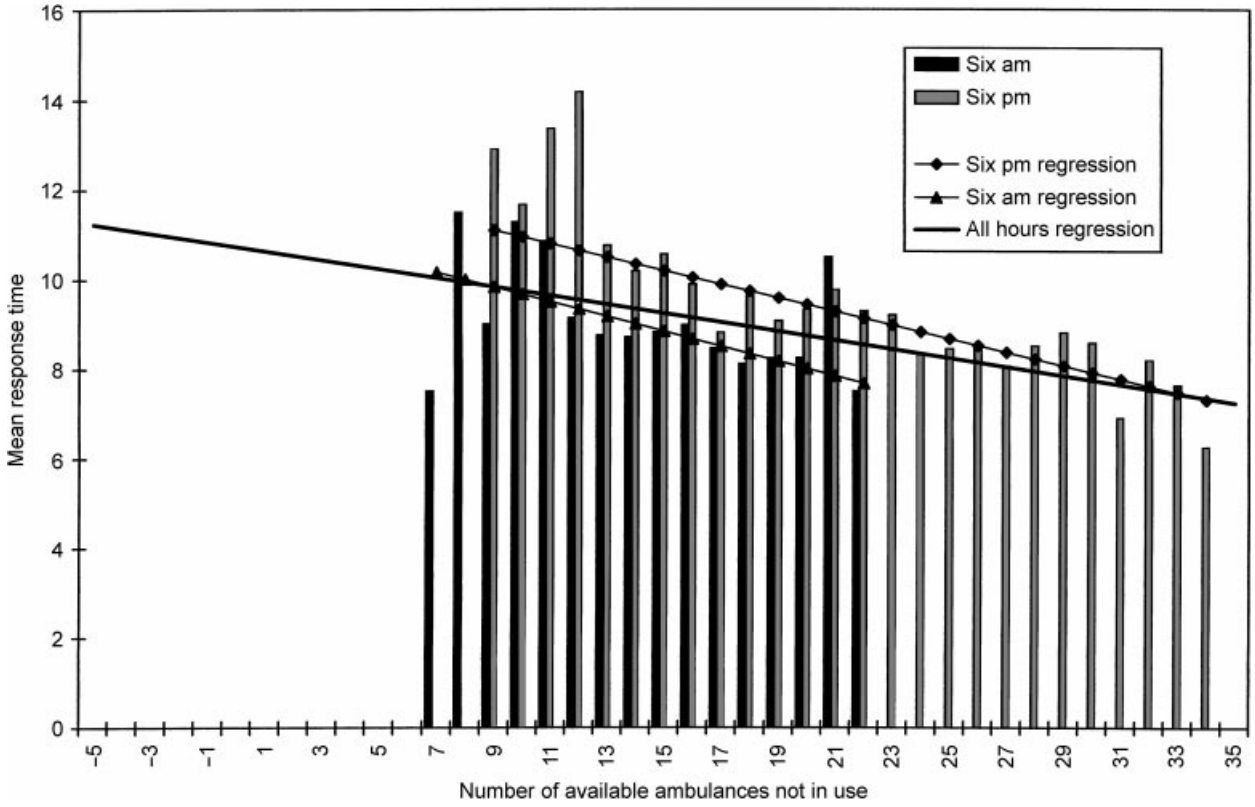

Figure 2 Daytime and night-time ambulance response curves. Black bars, 6 a.m.; grey bars, 6 p.m.; filled diamond and curve, 6 p.m. regression; filled triangle and curve, 6 a.m. regression; curve without symbols, all hours regression.

ambulance. [For the purposes of finding the error of the cost of a second of response time, we shall assume that the uncertainty of the total annual cost (which we shall treat as if it were a standard error) is 3 per cent, or £7500.]

The price, or opportunity cost, of a second of response time

We are now in a position to estimate the cost of improving response time by one second. This is, in effect, the price of a marginal second of response time. As we know by how much response time needs to improve to meet the government target, we can then estimate the cost of doing so, if it is achieved by increasing resources, *ceteris paribus*. From the above estimates, the price of a second of response time is estimated to be £28 000 (uncertainty range[4] £25 600–£30 400). It should be noted that the standard errors on the numerator and denominator are independent and both equal to 3 per cent, so the standard error on the £28 000 will be approximately 4.2 per cent.

This now also acts as an opportunity cost when considering an innovation in ambulance services, because it is the cost of the best alternative to the innovation.

We answer the first question posed (How many additional ambulances would be required to meet the government target, *ceteris paribus*?). From equation (5), to reduce the 75th percentile response to 8 min requires us to reduce the average response to 6.78 min. Thus, the 1997–1998 value of $r^* = 8.87$ min would have to be reduced by 2.07 min. (The standard error on the 2.07 min consists of components from estimating the 6.78 min and 8.87 min. It is estimated as 3.4 s, or 2.7 per cent.) As an additional ambulance saves 8.9 s of response time, this implies 14.0 additional ambulances (CI 12.9–15.1). (This assumes the errors on numerator and denominator are independent. As they will be negatively correlated, there will be some slight overestimation of error.) The annual cost is estimated to be £3.5 million (uncertainty range £3.16–£3.84). However, as the value of n increases, the saving of response time per additional ambulance declines. To account for this, we must extrapolate our model beyond the limits of our data, so the answer is only likely to be approximate. Nevertheless, by this method, we estimate that the number of additional ambulances required would be 30, at an annual cost of £7.5 million.

Optimally allocating ambulances by time-of-day and day-of-week

Table 1 shows the change to the shift patterns which equilibrate $c.\Delta t$ for each hour of the day and day of the week, given the constraint that the only change possible in shifts is to split the 12 h shifts into two 6 h shifts. We estimate that these changes would only reduce average response times by 4.6 s, the equivalent of about 0.5 additional ambulances, or a saving of about £130 000

Table 1 Suggested changes to the current ambulance deployment schedule

	Changes to shift pattern						
Hour	**Monday**	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**	**Sunday**
0.00	2	1	1	1	1	4	7
1.00	2	1	1	1	1	4	7
2.00	−3	−3	−3	−3	−3	0	4
3.00	−3	−3	−3	−3	−3	−2	0
4.00	−3	−3	−3	−3	−3	−2	−2
5.00	−3	−3	−3	−3	−3	−2	−2
6.00	−1	−2	−2	−2	−2	−2	−2
7.00	−1	−2	−2	−2	−2	−2	−2
8.00	0	0	0	0	0	−2	−2
9.00	0	0	0	0	0	0	0
10.00	0	0	0	0	0	0	0
11.00	0	0	0	0	0	0	0
12.00	0	0	0	0	0	0	0
13.00	0	0	0	0	0	0	0
14.00	0	0	0	0	0	0	0
15.00	0	0	0	0	0	0	0
16.00	0	0	0	0	0	0	0
17.00	0	0	0	0	0	0	0
18.00	0	0	0	0	0	0	0
19.00	0	0	0	0	0	0	0
20.00	1	1	1	1	4	3	2
21.00	1	1	1	1	4	5	2
22.00	1	1	1	1	4	7	2
23.00	1	1	1	1	4	7	2

a year if achieved by increasing resources. The before and after shift-change values of $c.\Delta t$ are given in Table 2.

Innovation 1: sitting in an ambulance

Table 3 shows the activation times and travel times for ambulances in four different categories.

For those ambulances travelling to a home zone, the time taken to activate was 0.59 min (35 s) less for those occasions when the crew was sitting in the ambulance. The equivalent time difference for ambulances travelling to a non-home zone was 0.67 min (40 s). The average reduction was 38 s. A 10 percentage point increase in the number of crews sitting in the ambulance would therefore improve response time by 3.8 s, at an opportunity cost of £106 000 per year. This benefit from the innovation can now be compared with the costs of achieving it, including any additional salaries required as a trade-off for less congenial working conditions.

Innovation 2: faster job turnaround

We estimate that $j = 45$, $u = 15$ and $s = 130$ (see equation (4)).

If j could be reduced to 44, then s would increase to 131, or by 0.77 per cent. An increase of this amount, when the average value of $n = 18$ ambulances, amounts to 0.14 ambulances, an improvement of 1.1 s in average response time or an annual opportunity cost saving of £31 000. The question now must be asked: is it worth while to attempt to achieve a 1 min reduction in overall trip time if that results in only a 1.1 s reduction in r^*? (That is, will it cost more than £31 000 a year to achieve?)

Innovation 3: triage – answer only category A calls

The ambulance crews categorized only 9.5 per cent of calls as life threatening. When we ran the system to estimate r^* if only these calls were ever answered, the time reduced by 63 s. By interpolation, if 30 per cent of calls were to be triaged as either life threatening or potentially so, and thus were all given an A categorization, then the reduction would have been only 49 s. This work may be checked by noting that there were 8.8 ambulances on duty on average in Surrey. If only 9.5 per cent of these calls were answered, then 8.0 of these 8.8 ambulances would not be used. As each unused ambulance reduces response time by 8.9 s, the reduction in r^* would be $8 \times 8.9 = 71$ s (compared with 63 s previously estimated). (Of course, if the remaining calls were to be answered at some time as well, the time-reduction on the A calls would be substantially less. Current work is engaged in looking at the effect of dedicating a certain number of ambulances for category A calls, and the effects of doing so on average response times for A, for B and for all calls. This research will therefore be more realistic than the case so far considered.)

Demand

If an increase in demand does not alter the variance of the number of ambulances in use, a 10 per cent increase in

Table 2 Total savings for an additional ambulance under current schedule and under revised schedule by hour of day and day of week

Hour	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Under current schedule							
1	−60.06	−56.80	−55.23	−69.16	−52.25	−127.79	−165.23
2	−35.18	−43.50	−40.30	−58.60	−37.88	−81.32	−101.04
3	−31.31	−30.17	−28.09	−41.06	−29.52	−55.59	−78.55
4	−18.16	−22.44	−23.22	−28.63	−24.30	−33.87	−56.03
5	−22.00	−22.47	−20.58	−24.26	−17.84	−26.04	−29.73
6	−23.03	−23.10	−25.03	−23.99	−23.51	−25.02	−29.25
7	−39.12	−38.03	−35.03	−30.61	−38.16	−31.85	−35.76
8	−47.60	−45.91	−31.99	−36.93	−35.49	−34.32	−38.98
9	−40.47	−40.55	−36.37	−40.31	−42.48	−32.83	−31.15
10	−49.37	−47.18	−52.91	−47.75	−53.74	−42.06	−38.87
11	−55.75	−49.12	−49.02	−50.52	−45.77	−51.99	−49.15
12	−45.94	−45.06	−40.93	−41.90	−52.51	−52.68	−50.85
13	−44.61	−39.03	−43.13	−42.58	−48.08	−54.40	−50.52
14	−52.40	−46.33	−50.34	−49.84	−60.18	−55.53	−44.49
15	−52.48	−44.73	−45.30	−47.10	−50.92	−48.53	−43.28
16	−48.31	−36.62	−45.01	−37.78	−46.97	−47.45	−39.01
17	−40.75	−35.49	−40.31	−35.73	−42.17	−50.47	−40.85
18	−36.25	−38.56	−35.83	−37.01	−36.95	−41.51	−41.48
19	−34.86	−36.02	−31.54	−30.19	−37.67	−39.07	−43.40
20	−35.67	−37.27	−31.94	−33.91	−39.36	−38.67	−45.44
21	−71.94	−67.76	−63.92	−64.39	−62.17	−72.47	−79.94
22	−73.45	−71.94	−71.07	−74.05	−83.73	−95.24	−77.01
23	−62.18	−58.58	−66.22	−66.88	−101.92	−99.94	−74.66
24	−67.21	−62.67	−72.78	−61.97	−112.70	−134.66	−74.54
Under revised schedule							
1	−48.49	−51.10	−49.67	−61.33	−47.61	−82.34	−75.94
2	−29.17	−39.37	−36.57	−52.39	−34.70	−54.11	−51.38
3	−43.01	−41.60	−38.60	−59.17	−39.84	−55.59	−51.36
4	−24.35	−30.36	−31.34	−39.41	−32.29	−41.62	−56.03
5	−29.41	−30.09	−27.56	−32.51	−23.45	−31.43	−36.53
6	−31.02	−31.19	−33.70	−32.36	−31.01	−30.03	−35.87
7	−43.15	−46.62	−42.67	−37.25	−46.10	−38.53	−43.87
8	−51.93	−54.80	−37.62	−43.68	−41.80	−40.84	−47.01
9	−40.47	−40.55	−36.37	−40.31	−42.48	−37.46	−35.59
10	−49.37	−47.18	−52.91	−47.75	−53.74	−42.06	−38.87
11	−55.75	−49.12	−49.02	−50.52	−45.77	−51.99	−49.15
12	−45.94	−45.06	−40.93	−41.90	−52.51	−52.68	−50.85
13	−44.61	−39.03	−43.13	−42.58	−48.08	−54.40	−50.52
14	−52.40	−46.33	−50.34	−49.84	−60.18	−55.53	−44.49
15	−52.48	−44.73	−45.30	−47.10	−50.92	−48.53	−43.28
16	−48.31	−36.62	−45.01	−37.78	−46.97	−47.45	−39.01
17	−40.75	−35.49	−40.31	−35.73	−42.17	−50.47	−40.85
18	−36.25	−38.56	−35.83	−37.01	−36.95	−41.51	−41.48
19	−34.86	−36.02	−31.54	−30.19	−37.67	−39.07	−43.40
20	−35.67	−37.27	−31.94	−33.91	−39.36	−38.67	−45.44
21	−65.33	−61.57	−58.40	−58.81	−44.35	−55.60	−65.96
22	−66.19	−65.03	−64.06	−66.75	−58.79	−60.68	−63.05
23	−56.53	−53.40	−60.16	−60.72	−69.96	−54.60	−61.34
24	−60.72	−56.83	−65.73	−56.19	−76.31	−68.84	−60.89

demand would increase the average number of ambulances in use from 8.8 to 9.68. This would increase average response time by $0.88 \times 8.9\,\text{s} = 7.8\,\text{s}$. A 10 per cent increase in demand thus increases average response time by 1.5 per cent, an elasticity of 0.15. If the variance of the use distribution also increases as demand increases, however, the increase in average response times would be somewhat larger.

Table 3 Activation, travel and response times by category of journey

Category	Activation time	Travel time	Response time
1	2.44	5.33	7.79
2	1.85	5.94	7.79
3	2.61	8.88	11.49
4	1.94	8.28	10.22

Categories: 1, at home base, going to a home zone; 2, at lay-by, going to a home zone; 3, at home base, going to a non-home zone; 4, at lay-by, going to a non-home zone.

Discussion

The model of ambulance usage developed here should help authorities decide how to allocate resources. It is able to give fairly accurate estimates of the cost of reducing average response times by providing additional ambulances. In providing a framework for comparing innovations against spending additional money to lower response times, it can lead to a way of measuring the likely success of an innovation.

It must be noted that the ARC (Ambulance Response Curve) estimated in this paper, and which is the centrepiece of our analysis, is specific to Surrey. Other ambulance services would need to estimate their own ARC for themselves, as each ARC will differ because of geography, demography, scale of operation, etc.

In the current context, i.e. without innovation of any kind, it is also implied that response time targets are unlikely to be met without the spending of substantial additional resources.

In this paper, we have not tried to model changing the nature of the service, as we have been looking at changes in the *status quo* situation. It is an open question as to whether radical changes can meet response time targets without an increase in resources, although the orders of magnitude presented here suggest that they could fall somewhat short. The question should, however, be amenable to the framework devised here.

Other questions that can be examined within the framework we have devised are

(1) How will the introduction of triaging affect response times?
(2) Will changing from manual to automated timing and recording times to the nearest second rather than minute affect average response times?
(3) Are some ambulance controllers better than others in allocating ambulances?

Besides this, the estimates arrived at in this process may be used to better inform location simulation models.

Acknowledgements

Core funding for the Health Care Evaluation Unit, which pays for Dr Fischer's salary, has been provided by the NHS R&D initiative (formerly South Thames Region and now South East and London Regions). Thanks are due to Giao Tran, Joanne Lord and Janet Peacock, for helpful advice; to Linda Almeida and Tracey Jaisingh, for cost data; and to numerous others at Surrey Ambulance Service, for helpful discussions.

References

1 Audit Commission. *A life in the fast lane: value for money in emergency ambulance services*. London: Belmont Press, 1998.

2 Chapman R. *Review of ambulance performance standards: final report of steering group*. London: NHS Executive, 1996.

3 Rawlings JO. *Applied regression analysis*. Cole Statistics Probability Series. Pacific Grove, CA: Wadsworth and Brooks, 1988: 388–408.

4 Lord J, Asanti M. Estimating uncertainty ranges for costs by the bootstrap procedure combined with probability sensitivity analysis. *Hlth Econ* 1999; **8**(4): 323–334.

Accepted on 3 December 1999