The Optimal Provision of Ambulances: a Modelling Approach

Alastair Fischer, St George's, University of London.

Graham Nichol, Harborview Medical Center, University of Washington.

Paper submitted to HESG, York, July 2006

alastair.fischer@nice.org.uk

The Optimal Provision of Ambulances: a Modelling Approach

Alastair Fischer and Graham Nichol

Introduction

Very little work has been carried out on the optimal provision of emergency services. In part, that is because there is no recognised group of patients in society (other than society itself) that the service supports. If someone knew that they were going to need an ambulance at a particular time in the future, it would be a planned event rather than an emergency. So there are few if any "emergency service support groups" to lobby for better provision. Yet each year, about 15% of all deaths in industrialised countries are caused by sudden cardiac arrest (SCA). Any other cause of death which gives rise to mortality on such a scale has seen huge research funding and efforts and funding devoted to finding the cause and successful treatment of the condition. While a good deal is now known about the cause of sudden cardiac episodes, efforts to apply successful treatment have not received the funding that could be expected of the cause of such a large number of deaths.

What we do know is that the sooner that people who suffer SCA are treated, the higher the probability that they will survive; the survival drops off at about 20% for each minute of delay to defibrillation.(1) There is also a general belief that survivors do not have long to live, whereas the average length of survival is of the order of 6 years, and somewhat longer for those who arrest at a younger age (2).

That is why ambulance response time is so important: each minute, indeed each second, that can be shaved from response time – or more pertinently, time from collapse to defibrillation – has potentially great benefit to those who suffer an episode of SCA.

Speed of response is crucial for survival or avoidance of severe disabilities for many other emergencies, such as for victims of road and other accidents; stroke; choking; electrocution; burns; collapse from diabetes, epilepsy, asthma and anaphylactic shock; drug-related conditions.

So the story is unsung. There is no noisy lobbying and no-one to point out the potential benefits of speedier response to emergencies (except, it seems, for countering terrorist attacks).

But there would seem to be another very good reason for this deficiency in our knowledge and research base: there has not been a proper framework for the evaluation and a lack of available data to fill in the gaps in our knowledge. Although work has been done on the evaluation of emergency service response times, it has either been "black box" or the estimates of marginal costs and benefits have been at a point, and therefore any extrapolations have essentially been linear. Raftery (3) Nichol et al (4) examined the relative merits of adding single responder units or fully-equipped ambulances, but the study was not designed to find out how many additional such units could be added to an EMS fleet for each successive addition to remain cost effective. Costs per QALY were obtained for each process, but these would pertain only to the margin, and could not tell us how long we should keep adding units to the fleet. Fischer et al (5) carried out a somewhat different exercise which also compared the cost effectiveness of a new practice with that of an existing practice. While this study looked beyond marginal changes in resources, it was carried out from the perspective of an emergency service (in terms of cost per second of a reduction in response time). It did not compare the cost effectiveness of an emergency service with that of other parts of a health service or use cost per QALY as a unit of measurement.

This paper takes some first tentative steps towards establishing a highly nonlinear framework for the evaluation of the number of ambulances that should serve a region. It must be regarded as work in progress, because among other things, many of the magnitudes have had to be guessed; additionally, the modelling is primitive, showing the deficiencies of the first author.

Costs, Benefits and Perspective

Perhaps 5% or 10% of calls to an ambulance service are for life-threatening conditions. Typically, about 25% or 30% of all calls are triaged as potentially life-threatening. Since an ambulance service answers all calls, however, the "production" of ambulance services is jointly to answer all calls. The question arises as to the proportion of costs of an ambulance service that should be attributed to life-threatening calls. (In the case of joint production, there is no theorem to determine the correct proportion.) If life-threatening calls did not exist, then an ambulance service probably would not be supplied from public funds. To that extent, the whole cost of an ambulance service should be attributed to life-threatening calls.

The alternative way of considering this is to say that an ambulance service must be judged as a single entity, and the totality of all costs and benefits should be estimated. The first way of considering the problem considers the totality of costs. The alternative way works out the benefits (in some generic fashion, such as QALYs) for all calls. Since the QALY gains for non-life-threatening calls will be small relative to those of life-threatening calls, we shall for the moment ignore them, thus making the two ways of estimating costs and benefits equivalent.

The question of future medical costs (sometimes many years into the future) of someone who would otherwise have died if an ambulance had not arrived when it did must also be considered. This aspect is fraught with difficulties. From the perspective of an ambulance service, this cost should not be included. From the perspective of the NHS or similar third-party payer that pays for both the ambulance service and future medical costs, it probably should be included.

From a practical viewpoint, the estimation of such costs is difficult and is therefore usually ignored. From a theoretical viewpoint, the future costs related to the condition that put the patient into the ambulance should (from an efficiency viewpoint) be included (6), but there is considerable debate about whether to include the costs of non-related future conditions. (The problem is in comparing two people, both of whom would have died were it not for the timely intervention of the ambulance, one of whom has no other future condition, and the other of whom has a costly non-related future condition: might you wish to save the first person but not the second? (7) The question of whether to count the future costs of non-related treatment, and to some degree of related treatment, is thus overwhelmingly an ethical one. If all life-threatening ambulance calls can be covered by the Rule of Rescue, then future costs should almost certainly be excluded.) In what follows, the analysis will firstly include, and secondly exclude, future costs.

From a societal perspective, some allowance for future benefits to society of the person whose life has been extended should also be made.

On the benefits side, we know approximately how many people with SCA are treated by an ambulance service (by being given defibrillation therapy), approximately what proportion survive for a given response time and approximately their average future life expectancy. The benefits of using defibrillators out of hospital are easy to state but uncertain in their extent: they are the sum of the years of additional life gained by survivors discounted by disability and the effluxion of time. The years of extra life can be estimated retrospectively by long-term follow-up of the survivors from about 1970 onwards. However, as therapy for people with heart conditions improves, it is likely that survival from SCA will change. On one hand, it will increase to the extent that follow-up post-survival from SCA will improve. On the other hand, it will decrease to the extent that the kind of person having an SCA episode may be older and most probably sicker than those of past years, because of improvements in therapy for people with heart conditions.

While this is somewhat cautionary, however, the life expectancy of survivors, even if it changes by 20% up or down, will be relatively well estimated. In comparison, estimation of survival *differences* between earlier and later use of defibrillators is fraught with potentially enormous errors.

Furthermore, we know almost nothing about the effect of an ambulance service on the aggregate of all other life-threatening conditions. We shall therefore carry out the analysis firstly assuming no benefits at all for anyone other than SCA patients, and secondly assuming that the total benefits of all patients other than SCA patients are equal to those of SCA patients. It is not clear whether the second assumption is an upper bound. For example, for a young accident victim who survives, the number of expected years of added life will be high, but the decrease in survival probability per added minute of response time may not on average be as high as 20%. And it is not clear how much benefit can be attributed to those whose condition is not life-threatening.

The model

We shall outline a means of estimating the optimal size of an ambulance service. This outline needs to be adapted to local circumstances prior to use to guide decision making in an ambulance service.

The model used is therefore a simple one and is based on UK prices in pounds sterling (£). Since costs will vary from one country to another in ways that are not accounted for by either the official exchange rate or a purchasing power parity exchange rate, the model should use country-specific costs if is to be used outside the UK.

The objective is to express, in terms of response time, both the total costs and the total benefits of increasing the size of the emergency service within a given geographical area and with given technology and procedures. From there, we find the range of response times for which an emergency service should be provided, and estimate the response time that maximizes the total net benefits (= total benefits – total cost) of the service.

Estimation of total cost

First, we find an expression for annual total cost of an emergency service in terms of response time, as follows.

We define T to be the time from collapse to defibrillation.

T comprises a lag of "a" minutes from collapse to calling the ambulance, plus b minutes of activation time from receipt of call to dispatch of responder, plus t minutes of travel time, plus c minutes from arrival to administration of the first shock.

Thus $T = a + b + t + c$. (1)

Response time $R = b + t$ (i.e. activation time + travel time) (2)

Travel time t depends on the distance travelled by a responder, and we assume direct proportionality between time and distance travelled. The distance travelled will depend on the number of responders and whether they are optimally positioned.

In the simplest case, assume that the population is uniformly distributed on an infinite plane and that the ambulances (which are uniformly spaced) travel on a rectangular grid. Think of the ambulances at the centre of each square on a chess board. The average distance travelled will be half the distance to the top or bottom of the square, plus half of the distance to the left or right boundary. The total distance travelled on average will therefore be the distance from the centre of the square to the top or bottom. When there are four times as many ambulances, the average distance will halve; when there are nine times as many ambulances, the average distance travelled will be 1/3 of the original. Thus, for n on-call ambulances, the average distance will be 1/(the square root of n) compared with the distance for one ambulance. It is likely that as the number of responders increases and average distances fall, average speeds will also fall marginally, and the centralized dispatcher of vehicle placement will place them less optimally, but in the base case of this model, those factors have been ignored.

Thus $T = a + b + d.n^{-1/2} + c$ (3)

If the average speed of responders falls and they become less optimally placed as n increases, this may be modelled by changing the exponent on n in the direction of zero. The exponent on n was estimated by Fischer to be –0.37 for the Surrey Ambulance Service, but the magnitude of this exponent estimated using a different but plausible functional form was –0.43. A reasonable alternative magnitude for sensitivity analysis would therefore be –0.4.

The formulation of T as an inverse function of a power of the number of available ambulances, and a plausible value for that power, is central to this exposition.

Since $R = b + t = b + d.n^{-1/2}$, by rearranging, we obtain

$n = (d/(R - b))^2$ (4)

We assume that costs are proportional to the number of ambulance units on duty, which we call n + n*, where n are on-call (and waiting) and n* are in use.

The total cost of equipping, staffing and operating an ambulance for a year in England is about £300,000. Thus the total cost (TC) of running (n + n*) ambulances is about £300,000 (n + n*).

Substituting from (4) for n yields $TC = 300{,}000((d/(R - b))^2 + n^*)$ (5)

Note that this analysis has been carried out from an emergency service perspective, because the cost of hospitalisation and future illness costs for survivors have not been included in (5). From a societal perspective, we must add to (5) an average hospital cost H and a cost of future illness of F for each of the N.S survivors, where N is the number of patients annually carried to hospital with SCA and where S is the probability of survival for an individual.

We assume that the survival curve T minutes after collapse, derived from the OPALS study (1), is given by

$$S = \exp(-0.23T) \quad (6)$$

Since T = R + a + c, then

$$S = \exp(-0.23(R + a + c)) \quad (7)$$

If N people with SCA are carried by the emergency service in a year, the expected number of survivors will be $NS = N.\exp(-0.23(R + a + c))$. (8)

Thus the total cost of current and future illnesses of survivors is thus

$$(H + F)N.\exp(-0.23(R + a + c)). \quad (9)$$

Thus from equations (5) and (9), total societal cost becomes

$$TC(soc) = 300{,}000((d/(R - b))^2 + n^*) + (H + F)N.\exp(-0.23(R + a + c)) \quad (10)$$

<u>Estimation of total benefits</u>

First, we find the total benefits (TB) for SCA patients in terms of R, from the perspective of the emergency service.

We assume that the NS survivors on average live Y = 6 years at a utility value of 0.7 after an SCA episode, and that the value of a life year saved (at utility = 0.7) to the health service/third-party payer is £21,000, equivalent to £30,000 per QALY. Clearly, the threshold cost per QALY varies from country to country: for the USA, a figure often used is $100,000 (about £60,000); for low-income countries, the figure would be substantially lower than £30,000.

Thus, using equation (8),

$$TB = 21{,}000.N.Y.\exp(-0.23(R + a + c)) = 126{,}000.N.\exp(-0.23(R + a + c)) \quad (11)$$

From a societal perspective, some survivors will be able to return to work. Compared with being on a state pension, society benefits by the additional production that is measured by their earned income. For the sake of this exercise (as we have no data to inform this scenario), we assume that 10% of survivors work at an annual salary of £20,000 for an additional 5 years, giving an additional benefit of an average of £10,000 per survivor. Thus total societal benefit is given by TB(soc) = 136,000.N.exp(-0.23(R + a + c) (12)

Estimation of total net benefit

A ***net*** benefit is defined as the difference between benefit and cost. If the net benefit is positive, the activity will be worth undertaking; if it is negative, then some combination of other activities will be worth doing instead.

Thus the total net benefit from the perspective of the emergency service, using equations (5) and (11), is given by TNB = TB – TC (13)

Societal net benefit (SNB) is given by equations (10) and (12) as

SNB = TB(soc) – TC(soc) (14)

The objects of the exercise are to establish, first, the range of values of R for which TNB > 0, and second, the value of R which maximizes TNB, and then to repeat the exercise using SNB in place of TNB.

We assume that

- a + c = 4.3 minutes (the sum of the time lag between collapse and contact with the emergency service, plus the lag between arrival of the emergency service and the first shock). The magnitude of the first lag, a, will probably never be known accurately, as it relies on people's notoriously-poor recollections, but it is probably approaching 2 minutes. The magnitude of the second lag, c, appears to be at least 2 minutes (5). The figure of 4.3 minutes makes the data consistent with the OPALS predicted survival curve.
- b = 1.5 minutes (control room activation time, from receipt of call to the departure of a responder). The literature suggests that this magnitude is between 1 and 2 minutes.
- N = 1,100 SCA patients carried in any one year
- n = 20 (average number of ambulances on call (and not in use)
- n* = 8 (average number of ambulances in use)

N, n, n* and d are derived from Fischer's study of the Surrey Ambulance Service (5). The average response time was 8 minutes 52 seconds at the time that the Service was studied. Note that as N, n, n* and d are all interrelated, it would not be appropriate to change one of these variables without changing the others. A different set of these four variables will pertain to each emergency service.

- d = 33 (assuming average response time is 8.9 minutes when n = 20, and using equation 4))
- F = £40,000 of future illness costs for each survivor
- H = £1,700 of hospitalization costs

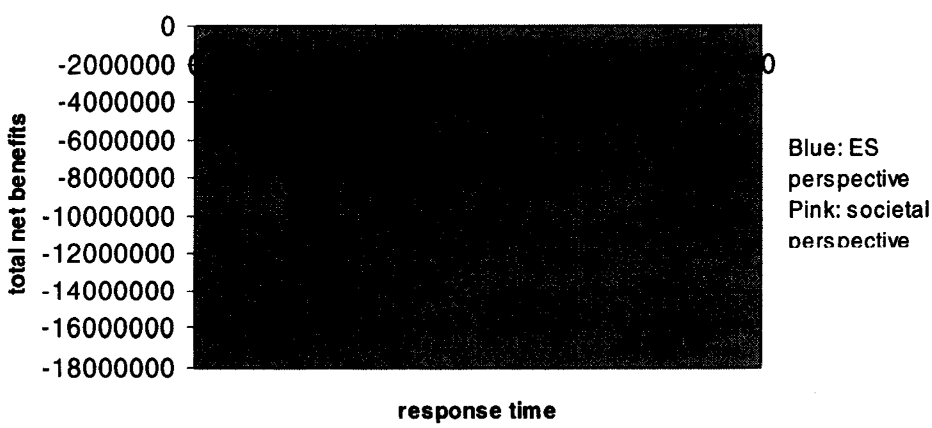
The values for F and H come from Nichol's (1998) study of emergency services in the USA (8). They have been converted to UK£ but have not been adjusted for inflation, because it is likely that the UK costs would have been a little lower in the first instance.

Thus $SNB = [149{,}600{,}000.\exp(-0.23(R + 4.3))] - [300{,}000((1095/(R - 1.5))^2 + 8) + 41{,}700.\exp(-0.23(R + 4.3))]$

$= 107{,}900{,}000.\exp(-0.23(R + 4.3)) - 2{,}400{,}000 - 328{,}500{,}000/(R - 1.5)^2$

We show the graphs of total costs and total benefits and the total net benefits from both a societal perspective and an emergency service perspective. The difference between the two results is that the emergency service perspective stops at the hospital door, while the societal perspective includes proximate hospital costs and all future health costs of survivors as well as all indirect costs.

Figure 1: Total net benefits by response time if the only benefits are for SCA



Results

According to the assumptions of this model, if the only benefits from operating an emergency service accrue to SCA survivors, and to nobody with other conditions, the service should not be funded, as the net benefits are negative for all response times. This is shown in Figure 1, where the net benefits from an emergency service perspective are maximized at an average response time of 9.2 minutes and a net benefit of -£1.7m per year. From a societal perspective, the net benefit curve is flat from an average response time of 10 minutes or more, at a value of -£2.8m per year.

If the benefits of operating an emergency service for other conditions are proportional to those of cardiac arrest for all average response times, and we ascribe a high enough level of benefit to other conditions, then the net benefits of an emergency service become positive for a range of response times. We illustrate this by assuming that all other uses of an emergency service (at each average response time) produce the same net benefits as are gained by people with SCA.

Figure 2: Total costs and benefits by response time when benefits for other treatments equal the benefits for SCA

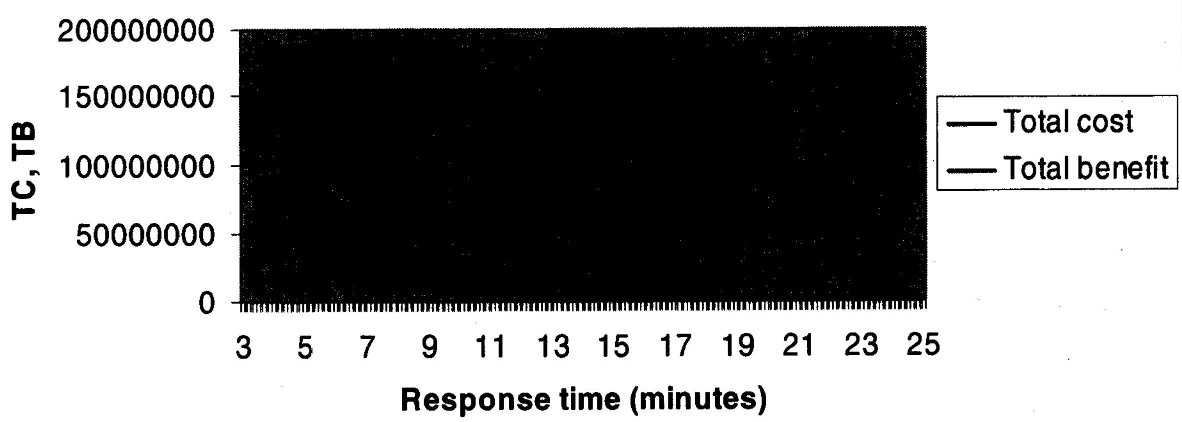
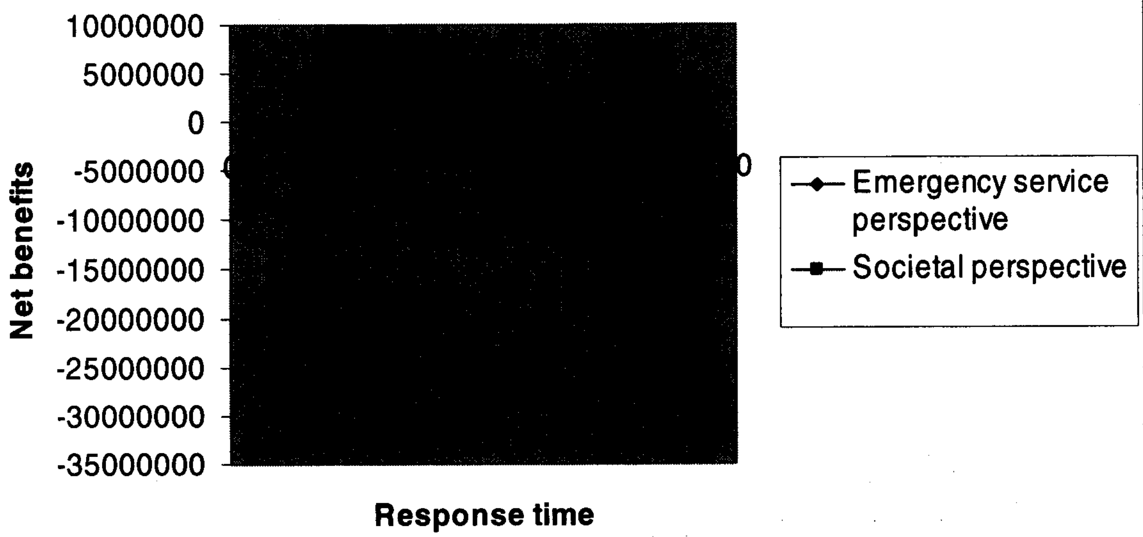


Figure 2 shows how both total costs and total benefits change as R changes. As more resources are made available to the service (that is, as R falls) total benefits at first rise faster than total costs. However, as response times fall sufficiently, total costs begin to rise faster than total benefits. The curves cross twice, and at these points, TC equals TB, and net benefits are zero. Between the two crossing points (R = 4.7 minutes and R = 13.4 minutes), an emergency service is justified. The greatest net benefit of £7.6m per year occurs when R = 6.5 minutes, as is shown in Figure 3. From a societal perspective, the two crossing points are at 5.8 minutes and 11.1 minutes, the optimal R is at 7.4 minutes and the maximum benefits are estimated to be £2.2m per year.

Figure 3: Net benefits by response time if benefits for other treatments equal benefits for SCA



The relationship between 75% response time and average response time was estimated by Fischer (5) to be "75Resp" = 0.136 + 1.163R, so R= 6.5 (the optimum from the emergency service perspective) corresponds to "75Resp" = 7 minutes 42 seconds. That is, according to this model, the optimal target for emergency services in urban areas in Britain is approximately correct. From the optimum response time, we calculate that optimum value of n will be

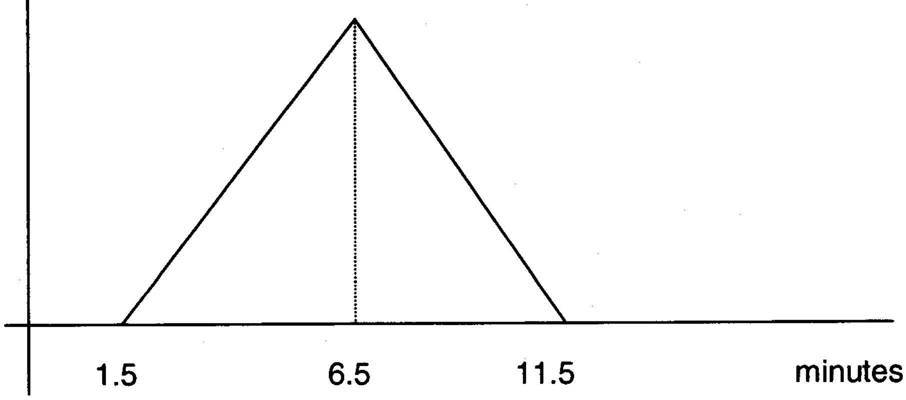
$n = (33/(6.5 - 1.5))^2 = 44$ compared with actual 20, and opt ($n + n^*$) will be 44 + 8 = 52 instead of 20 + 8 = 28, a considerable increase. From a societal perspective, opt R = 7.4 minutes, and "75Resp" = 8.7 minutes, requiring n = 31 ambulances, or $n + n^* = 39$ rather than the actual 28. At the time of Fischer's (5) study, the actual average R for the Surrey Ambulance Service on which the numbers were based was 8.9 minutes. Institutional and other improvements have improved this performance since then.

Sensitivity Analysis

The assumption that all calls will have the same average response time is a simplification. To be carried out more realistically, response time should have a distribution and the whole exercise should be done either as a Monte Carlo simulation or iteratively. Since the TB curve is convex to the origin, the benefit of having the response time for person A of (say) 6 minutes and for person B of (say) 8 minutes will exceed that of a response time of 7 minutes for both A and B. That is because, for a convex survival function, the probability of survival at 6 minutes + the probability of survival at 8 minutes will exceed twice the probability of survival at 7 minutes. We have simulated this by the optimal response time will be slightly lower than the one calculated in this exercise.

We have tested the effect of this by assuming that the optimum value of R is 6.5, of which 1.5 minutes is a constant and 5 minutes is the average travel time. The distribution of response times in the theoretical case where each ambulance is at the middle of a square grid of equal size is given by the "double-triangular" distribution illustrated in Figure 4. The benefits obtained from assuming this distribution of response times are 52% higher than those given in the simple case of assuming the same average response time for all calls for a given number of ambulances. This aspect has not been fully explored, but it is likely to decrease the optimum value of R by about half a minute, with a corresponding decrease in the optimal 75th percentile response to a little over 7 minutes.

Figure 4: Assumed distribution of response times for an ambulance with an average response time of 6.5 minutes



Using an exponent of -0.4 instead of -0.5 on n in the equation for T does not make a great deal of difference to the results, with the optimal value of R rising to 7 minutes.

The results would also indicate that running an emergency service in a low-income country is probably not a good use of scarce health care resources, because the willingness and/or ability to pay for the service is not likely to be sufficient to allow net benefits to become positive for any size of the ambulance fleet. In a country that is sufficiently wealthy to afford emergency services, the results are likely to depend critically on the extent of benefits from treating and transporting people with conditions other than cardiac arrest. The results would also appear to indicate that having an ambulance travel too far is not a good use of resources, and that therefore, if an ambulance service is provided in large conurbations, the minimum size of fleet is unlikely to be n = 1 or any other small value of n. Once the decision to provide a service is made, then a number of vehicles should be purchased, leased or rented.

Discussion

In the context of the appropriate provision of resources for emergency services, the most appropriate approach is to consider the costs of expanding or contracting the emergency service provider as a whole, and to compare these costs with the estimated benefits of doing so.

The appropriate level of funding of emergency services, particularly emergency services, has probably never been properly estimated. The fact that in this paper, a crude illustration of how to go about working out the cost effectiveness of an emergency service is apparently as good an estimate as has ever been produced, shows the rudimentary state of the art in this area. From the tentative results obtained, it would seem that in many countries, emergency services have been something of a Cinderella service in health care.

Turning to the model in more detail, changing the values of each parameter in turn could be used to explore the appropriateness of the model in different circumstances. For example, changing the willingness-to-pay variable (assumed to be £30,000 per QALY in the above) would allow a more appropriate analysis for higher-income countries (USA) and lower-income countries. Changing N would act as a surrogate for lower population density. Changing the different assumed time lags would give an indication of their importance and whether it would be worthwhile reducing them. Estimating the benefits for carriage of people with life-threatening conditions other than SCA is very important, as the size of the total benefit of an emergency service determines whether the service should be provided, and if provided, the level of its provision.

The model could also be extended to include single responders as well as fully equipped ambulances. The optimal mix of these two types of responder could be estimated from such a model.

In most situations, decreasing returns mean that the provision of a resource will be limited, because eventually, the costs of a further increase of provision will outweigh its benefits. However, for emergency services, we have shown that benefits may increase faster than costs at certain levels of provision, even though the eventual result will be a limit in provision. According to the lower curve in Figure 3, the costs of providing a service with an average response time of over about 12 minutes would generally outweigh the benefits of the service (which is why sufficiently remote areas are not served by emergency services). But as the response time dips below 12 minutes, the benefit (per minute of decrease in response time) increases at an increasing rate, as can be seen from the slope of the curve. However, the cost of resources that will allow a decrease of a minute of response time will increase as response time decreases. That is because, if there is a four-fold increase in the number of ambulances, the travel time will not on average be ¼ of the previous time, but ½. So at first, the cost effectiveness of ambulances will actually improve as more resources are added, and in countries where this is so, there is added reason for an increase in those resources.

This analysis has several strengths. First, it provides a framework for decision makers to assess the marginal cost of improving response time by addition of emergency response vehicles to an existing ambulance system whereas most previous analyses have considered the incremental cost of improving the skill level of providers in the field. Second, the framework can be adapted to local circumstances by substituting local effectiveness values. Although the overall survival observed in the OPALS study, the basis of this economic analysis, was low, an ambulance system staffed by highly-trained, experienced providers who treat a high volume of calls and perform a high volume of procedures each year, with independent medical oversight and quality assurance can achieve survival of 15 to 20%.(Unpublished data, Seattle Medic One program, August 28, 2005) In Seattle, WA, survival decreased by 3% with every 1-minute delay in CPR and by 4% with every 1-minute delay in defibrillation. Third the framework can be adapted to local circumstances by comparison to local thresholds for willingness to pay. The latter is important since a society's willingness to expand a service can depend on the availability of competing alternatives and adequate resources.

This analysis also has weaknesses. First, it provides a point estimate of the incremental cost of improving response time. Since it relies on several assumptions, there is likely to be large uncertainty in this estimate. Future extensions of this model should use Monte Carlo simulation to evaluate this uncertainty. Second, the model assumes that technology and dispatch protocols are constant through time. However, response time can be improved by changes in both technology and protocols, or location of vehicles closer to where calls are likely to occur. The costs of these alternate interventions were not considered. Third, this analysis is likely to underestimate cost-effectiveness by assuming no benefit to use of emergency vehicles for other conditions in one case, and by artificially assuming that cardiac benefits are half of all benefits in the other case. Fourth, the use of OPALS data underestimates survival by assuming zero survival once a VF rhythm has ceased. Despite these limitations, we believe that this analysis represents an important advance over previous analyses of the economics of emergency service interventions, and should be validated in a variety of settings.

References

1. De Maio VJ, Stiell IG, Wells GA, Spaite DW Optimal defibrillation response intervals for maximum out-of-hospital cardiac arrest survival rates. *Ann Emerg Med. 2003 Aug; 42, 2: 242-250*
2. Engdahl J, Herlitz J. Localization of out-of-hospital cardiac arrest in Goteborg 1994-2002 and implications for public access defibrillation. *Resuscitation. 2005 Feb;64(2):171-5.*
3. Raftery J et al
4. Nichol G, Laupacis A, Stiell I, O'Rourke K, Anis A, Bolley H, et al. A cost-effectiveness analysis of potential improvements to emergency medical services for victims of out-of-hospital cardiac arrest. *Ann Emerg Med. 1996;27(6):711-20.*
5. Fischer AJ, O'Halloran P, Littlejohns P, Kennedy A and Butson G Ambulance Economics *Journal of Public Health Medicine 2000; 22, 3: 413-21.*
30.
6. Meltzer D Accounting for future costs in medical cost-effectiveness analysis *J Health Econs 16* (1997) 33-64
7. Richardson J (personal communication)
8. Nichol G, Hallstrom A, Ornato JP, Riegel B, Stiell IG, Valenzuela T, et al. Potential Cost-Effectiveness of Public Access Defibrillation in the United States. *Circulation. 1998;97(13):1315-20.*