Estimating the Benefits of Regionalizing Emergency Medical Service Provision

James M. Wilson and Daniel J. Dudek

Local area governments have experienced increasingly stringent budget constraints in recent years. Innovations in service delivery provide one avenue for increasing the effectiveness of resource allocations. This paper explores the potential savings available from regionalizing emergency medical service provision. A mixed integer programming model incorporating peak demand considerations is used to minimize service cost given a desired maximum response time. Changes in the weighted average response time measure the quality degradation required to attain the savings from cooperative provision. The results indicate that the benefits are substantial but that distribution of these gains is a possible barrier to implementation.

Introduction

As the Federal fiscal presence continues to shrink and as taxpayers increasingly seek control over budgets, new strategies for cost savings in public service provision are at a premium. Communities, particularly those in rural areas, have been forced to reassess constantly their revenues and spending. Regionalization is one strategy which has been employed effectively to improve the efficiency of delivery of services such as education, public transport, and waste disposal. A neglected area is the potential economies available from regionalizing emergency medical services (EMS), the topic of this paper.

The Northeast has great potential for savings from public service regionalization. For example, there are 351 communities in Massachusetts and 330 have an EMS unit. Within the state, the average community is approximately 25 square miles with a mean population of roughly 16,000 and a median of 7,000, a demographic pattern characteristic of the entire Northeast region. Thus, EMS units generally are operating over small areas with low populations.

Problem

The present use of EMS resources, constrained by geopolitical boundaries, is unlikely to be optimal. Redundancy in supply is likely so that a spatial reallocation of resources could provide services at lower costs with marginal changes in the quality of the service. This analysis requires an estimation of cost, an evaluation of how inputs affect outputs, and the determination of the optimal spatial distribution of such resources over the region. Quality levels, measured in terms of response time, are varied and the benefits calculated for each of these levels. This procedure generates a set of estimates of cost and quality which delineate the tradeoffs available in EMS provision.

The primary objective of this research is to estimate the benefit of efficiently allocating EMS resources at the regional level. As such, this problem focuses on only one dimension of reducing morbidity and mortality due to emergency medical events. For example, the quality of equipment and labor are also under the control of management, but these dimensions of service planning are excluded from the analysis. Further, prevention is likely a more effective strategy to reduce morbidity and mortality than investing in additional EMS. How-

Staff Research Associate, Massachusetts Development Research Institute and Assistant Professor, Department of Agricultural and Resource Economics, University of Massachusetts, Amherst. The research reported in this paper was supported by the Northeast Regional Center for Rural Development under subcontract no. OSP 2832. This paper has benefited from—the comments of P. Geoffrey Allen and two anonymous reviewers. The authors assume complete responsibility for the contents.

ever, prevention requires a system-wide approach, which is rarely feasible for implementation by a single community.

Theory

Most citizens of a community will never use EMS. However, EMS investment occurs to insure access to a particular quality of EMS. This investment represents insurance against suboptimal care when an emergency occurs. The level of the community budget allocation for this insurance is determined by the risk preferences of the community and its budget constraint.

The spatial distribution of demand for EMS is expected to be related to the distribution of population since density will invariably affect the rate of incidence. Further, emergency events will not occur uniformly over time, but can be expected to peak at particular times, for example, during increased social activity or poor weather conditions. Peak demands determine the probability of a subsequent call arising when an ambulance is already in service. Typically, EMS systems are evaluated on the basis of two performance measures: the speed of service delivery (response time) and the proportion of calls receiving immediate service. The probability of immediate response to an emergency call may be set as a criterion. Maintenance of this criterion during peak demand periods may require additional ambulances or redeployment of the present force.

There are several problems inherent in quality measurement for EMS. Willemain discusses three categories of measures to assess EMS: input, process, and outcome measures. Input measures are most often expressed in terms of resources per capita; for example, ambulances or fulltime physicians per 100,000 population. However, input measures are of limited usefulness for EMS evaluation since they reveal nothing of system performance and offer only hints of system potential. Process measures focus on such characteristics as delays in obtaining emergency care and "appropriateness" of service. These efficiency measures (such as delays or degrees of underresponse or over-response) can be quite useful in monitoring performance. Outcome measures including "lives saved" and changes in mortality or morbidity rates are highly desirable, but these are difficult to estimate since the increased morbidity and mortality associated with a marginal increase in response time is generally unknown.

Process measures, specifically response time and the probability of immediate service or the peak demand criterion, frequently have been used to assess the quality of EMS. The value of using these measures is usually defended by simply pointing out that by definition an emergency needs as fast a response as possible. However, these performance measures are only valid if it is assumed that treatment is appropriate once the EMS unit arrives. As alluded to earlier, such measures as "appropriateness" and "outcome" are difficult to evaluate. These difficulties leave response time and peak demand as the least expensive and most tractable of the quality measures of EMS. Undoubtedly this is why they appear so often in the literature.

Methodology

For the decisionmaker planning EMS, cost and quality are constraints. The location of ambulance sites and the number of ambulances deployed are the fundamental choice variables. This general problem of locating service facilities subject to a time constraint is commonplace. However, it is useful to review briefly methodological developments in this area in order to elucidate essential features of the locational analysis for EMS service sites. These include total distance travelled, facility numbers, maximum time constraints, nonhomogenous sites and peak demand.

The "p-median" model was formulated to locate "p" homogenous facilities so as to minimize the total distance travelled by users (Hakimi). Facilities are not distinguished by either size or specialization and their number is exogenously given. The neglect of fixed cost differences among sites and the absence of a maximum travel time limit led to Toregas and Revelle' s development of the set-covering model. Set-covering minimizes the number of facilities required to cover all demand sites subject to a maximum time or distance constraint. The term set-covering derives from the mapping of demand sites to supply points within a prespecified distance standard. It is a fundamental requirement of the model that at least one site must be chosen from the eligible locations in each set. Thus all demand points are co/vered. However, the level of demand is

not considered when the minimum number of sets for a region is chosen, nor is the peak demand problem.

None of the preceding methods specifically accounts for the spatial distribution of demand. The maximal-covering problem developed by Revelle and Church readmits the importance of the spatial distribution of population in determining system cost. Thus the number of exogenously given facilities are located so as to cover the maximum population within the stated distance constraint. This formulation can show tradeoffs between percent coverage and the number of supply points and predict the percent of population covered given a number of sites. However, cost is not explicitly calculated and peak demand is also ignored. Toregas, *et al,* noted that for the "p-median" problem, a limit on the distance that users must travel to a facility can be imposed by a set-cover. This modified "p-median" problem, termed the weighted distance (WD) model, implicitly considers variable cost by trying to minimize total distance traveled. It does not consider the fixed cost differences between sites and facility numbers are not endogenously determined. Further, since peak demand considerations are not incorporated, the trade-offs between centralizing and decentralizing ambulance sites is not explored.

Daberkow and King used a mixed integer-continuous variable formulation that builds on the WD model by incorporating fixed and variable costs in the objective function and minimizing cost over the array of set-covers. This cost weighted (CW) model considers reponse time with the number of sites determined endogenously. The model does not consider the effect of peak demands however. Without a constraint on the allocation of demand to supply sites, suppliers could be unable to satisfy the immediate dispatch requirement embodied in the peak demand criterion. Thus, solutions are achieved at a lower cost than if adjustments for peak demands were considered.

The model developed for the present study includes facility site capacity constraints which derive from the discontinuities in EMS cost associated with the lumpiness of ambulance inputs. The novelty embedded in this approach is the recognition of excess capacity when individual community boundaries define service areas. This conceptualization allows the possibility of efficiency gains from spatial reallocation by eliminating redundant facilities and instituting interlocal cooperative supply. Excess capacity is calculated as the amount of additional demand that a specific supplying site could absorb above the resident community demand without any degradation in local service quality.

While excess capacity is expected to exist under the current institutional arrangements, this concept can be extended to any configuration of EMS resources under alternative performance standards. Since excess supply is a function of the mean service time (derived from the response time constraint and speed assumption) and the performance standard as well as the resident demand, results from Bell and Allen can be used to estimate any additional capacity.

To deal with the problem of peaks in the demand for EMS per unit time, Bell and Allen developed a method using queueing theory to calculate the probability of a call receiving immediate dispatch. The flow of emergency calls can be viewed as a queueing system in the sense that arrivals are the calls for emergency service, service commences with dispatch of the ambulance and finishes when the patient is delivered to the hospital. The specific assumptions of this model are that calls arrive randomly at a rate of L per hour, the number of arrivals in any time interval is described by a Poisson distribution and is independent of the number of arrivals in any non-overlapping interval.

To calculate the percentage of calls that receive immediate response, the mean number of calls per unit time, L, and the number of "services" per unit time, M, are needed. The statistic R ~ L/M can be considered a measure of traffic intensity. The probability of **x" customers in service is calculated from the Poisson density function as:

$$(1) \qquad P(n - x) - \{R"\exp(-R)\}/n!$$

If an EMS facility has "n" ambulances, a request for service will receive immediate attention if and only if there are "n — 1" or fewer customers already in service when the request is made. Thus, the statistic of interest is the probability that there are "n — 1" or fewer customers in service. This cumulative probability can be estimated from:

$$(2) \quad P(n =s n - 1) = \sum_{j=0}^{2} \{R^J\exp(-R)\}/j!$$

This probability of immediate dispatch (legis-

latively identified as 0.95) plus the response time limits embodied in the set-covering provide the quality constraints for the EMS location model developed in this study. The set-cover generated from a specific response time requirement defines a feasible service area and the total demand to be served. The maximum service time, M*, is used to calculate the initial number of ambulances needed to satisfy demand at the supply site. M* is computed as the time required to respond to an emergency at the most distant location in the set-cover, plus the time for on-site treatment of the victim and the time for transport to the hospital. Clearly, this service time is a conservative estimate since it is derived from the furthest demand point distance in the set-cover allowed by the response time constraint.

The mean number of calls, L, for a set-cover is estimated by uniformly distributing the annual emergency demand generated by the supply site community. L and M* are used in equation (2) to solve for the number of ambulances, n, which guarantee that at least 95% of the calls will receive immediate dispatch (the minimum performance standard). From this initial calculation, demand is added iteratively until the proportion of calls immediately served within the particular response time assumption degrades to the 95% minimum. That demand level is the maximum amount of demand the site can absorb without requiring an additional ambulance and it is this figure which is used for the capacity constraint. The complete model is specified as:

$$\text{Minimize: } Z = \sum_{l=1}^{i} \sum_{i=i}^{j} C_u X_{tj} + Y \sum_{i=1}^{i} f_t y_t$$

Subject to: $\sum_{J \in NJ} {}^{D}i\ {}^{x}u = Dj \quad j = 1,2, \ldots, n$

$$\sum_{je Ni} \ldots$$

where:

TI is the upper demand capacity threshold for site i
Cu is the cost of transport from i to j
fj is fixed cost determined by the method of Bell and Allen
Dj — demand at location j
I = denotes the set of facility sites
J — denotes the set of demand points
N_t = the set of facilities "i" serving "kj"
NJ = the set of demand points "j" served by "i"
Xu = fraction of location j population or demand
yt = facility site variable: 0 if no site; 1 if sited
qi = the elements of N_t

Choosing the above measures to represent EMS performance is consistent with the aim of exploring regionalization. Setting a response time constraint (RTC) restricts the maximum size of the service area while the peak demand criterion restricts deployment to guarantee immediate service for a given percentage of calls. These measures conform to the quality standards mandated by the EMSS Act of 1973 which require a maximum 30 minute response time for rural areas and a maximum 10 minute response time for urban areas. Further, 95% of all requests must be accommodated within these standards.

Empirical Studies of EMS Cost

A number of studies of the cost of EMS have been conducted in the past. In a comprehensive study, Dunlop and Associates estimated that approximately 85% of total costs are fixed with the remainder occurring on a per call basis. Labor costs, which account for about 60% of total cost, are the single most expensive item. In an unpublished study of private EMS costs in Massachusetts, Zebrowski surveyed 29 ambulance services and 7 transfer services. Per trip costs were estimated for all categories of service provision. For 1980, he estimated that it cost $49.23 to service an emergency call and $13.13 for a transfer call. This study used a synthetic economic-engineering approach to generate consistent cost estimates over the set of communities and between the base and regionalized case. A number of simplifying assumptions about the services in the area were invoked in order to make the base case and regionalization case comparable. The synthesized costs differ from the actual expenditures of towns in the base case. However, the difficulties in capturing community preferences for level of service, differences in cost structures and the existence of cooperative arrangements among services, such as with fire or police units, would make any other method unreliable.

Demand Estimation

Given the central role of population in the spatial optimization of supply sites, the estimation of EMS demand is critical. Community statistics concerning EMS call frequency are very rare. In the absence of detailed call data for the individual communities within the study area, it was necessary to develop a model to estimate the number of emergency calls occurring within each town. These annual emergency call estimates are then uniformly disaggregated over time parametric ally varied to simulate various levels of peak demand for each community. One would expect medical emergencies to occur differentially among age groups within a population. Consequently, the independent variables chosen for this study were five age groups of 0-14 years (AG1), 15-44 (AG2), 45-64 (AG3), 65-75 (AG4), and those over 75 years of age (AG5). As indicated in the theory section, population density is expected to have a great effect on the incidence of emergency calls per unit area. Thus, the distinction between rural and suburban demand for EMS was investigated.

The only available data for emergency calls over a set of communities was provided by the House Post Audit and Oversight Bureau of the Massachusetts Legislature. In 1980, surveys were sent to the EMS units serving all 351 communities. The survey requested a report of total calls, emergency calls and transfer calls. The survey elicited 106 responses. Of these, 82 reported their emergency calls and 33 reported their transfer calls. The 43 communities with fewer than 14,000 persons were denned as nonurban on the basis of conversations with officials from the Center for Massachusetts Data and the Center for Vermont Data. Since there are a large number of alternative definitions to demarcate the rural-urban dichotomy, the ultimate choice is somewhat arbitrary. Using an emergency call demand model specified as:

(4) ' $E = a_0 +$

1-1

a Fisher test was performed to determine whether the communities in each of these groups derived from a common population relative to the way that the age distribution of population affects the level of emergency calls. The null hypothesis that emergency call generation rates is the same generally was rejected via the F test statistic at the 1% level of

Table 1. Emergency and Transfer Call Sample Characteristics

Variable	Variable Name	Mean	Standard Deviation
Communities < 14 ,000			
Emergency Calls	E	357.58	270.72
Age Group 1 (0-14 years)	AG1	1704.26	899.72
Age Group 2 (15-44 years)	AG2	3382.60	1866.37
Age Group 3 (45-64 years)	AG3	1566.44	911.29
Age Group 4 (65-74 years)	AG4	528.44	355.32
Age Group 5 (75 years and older)	AG5	346.58	246.68
All Communities			
Transfer Calls	T	203.30	213.34
Total Population	POP	17,124.94	18,998.86

significance. Consequently, the nonurban and urban data were not pooled. Summary statistics describing the responding sample of communities are presented in Table 1.

Some EMS providers also service transfer calls, i.e. non-emergency transportation requests. Consequently, it is important to estimate total call demand in order to assess any locational or cost differences from serving transfers as well as emergencies. The simpler transfer demand model was specified as:

(5) $T - Q = {}_{O} + /3POP + e$

The results of the emergency and transfer call demand estimation are shown in Table 2. The adjusted R^2 for the emergency call model was 0.57, well within the range of experience described by other researchers (Daberkow

Table 2. Emergency and Transfer Call Estimation Results

Variable	Estimated Coefficient	Standard Error
Emergency Call Model		
aO	61.0225	61.0060
AG1	0.0525	0.1218
AG2	0.0027	0.0598
AG3	-0.1360*	0.0776
AG4	0.6983**	0.2076
AG5	0.1215	0.2815
Transfer Call Model		
aO	98.5649*	42.8697
POP	6.1161**	1.6902

and Stevenson). Multicollinearity between the independent variables was evident in the emergency call model as some correlations were over 0.80. However, due to the relatively high R^2, and the predictive use intended, no attempt was made to remedy the problem. The transfer demand results yielded a relatively low R^2 of 0.27. However, the regression and the explanatory variables were highly significant. Furthermore, transfer demand is included only to allow an analysis of facility siting sensitivity to demand changes.

Both the emergency and transfer call models were used to predict EMS requirements for the nine town study region described in Table 3. None of the towns within the study area were among the respondents to the State's EMS survey. However, the demographics of the sample and the study region are quite similar. The mean population of the 43 community sample is 7,528 while that for the 9 towns comprising the test case is 7,704. Population forecasts developed by the Department of Public Health were used to produce projections of emergency calls and transfers for 1985 and 1990. The relative accuracy of these forecasts can be evaluated by comparison with those from the Massachusetts survey data. A rate of 70 emergency calls per thousand persons was projected for the study region while 52 per thousand was the mean call rate observed within the sample. However, the standard deviation of that per thousand call rate was slightly over 36. These projected emergency calls were presumed to be sufficiently accurate for use as the EMS demand estimates in the objective function of the model as specified in equation (3).

Table 3. Study Region Characteristics

		Projected 1985 Values		
Community	Area (square miles)	Population	Emergency Calls	Transfer Calls
Berlin	13.0	2,424	193	113
Bolton	19.7	3,365	170	120
Boylston	19.5	3,924	331	123
Clinton	7.0	14,210	1,270	186
Lancaster	28.0	6,834	430	141
Northborough	18.5	11,125	526	167
Sterling	30.5	5,965	336	135
West Boylston	13.5	6,590	995	190
Westborough	21.5	14,903	590	136
Totals	171.2	69,340	4,841	1,311

Results

The methodology developed in this study was tested on nine communities in Central Massachusetts. These towns were chosen because each had its own EMS unit, they were not urban, and the hospitals were distributed uniformly over the region. The present locations of EMS facilities, one within each community, is referred to as the base case. The general procedure was to estimate the cost of this base case, compare it to a simulated regionalized case, and calculate the difference as the benefits from cooperation.

The multi-purpose optimization system (MPOS) developed by Northwestern University was used to solve the mixed-integer optimization problem. Forty-two model runs were made using response times that ranged from 5 to 30 minutes at five minute increments and average speeds of travel varying from 10 to 40 miles per hour in 5 mile per hour increments. The optimization model was run for emergency calls alone and for combined emergency and transfer calls for two time periods: 1985 and 1990. For each of these scenarios a base case was also estimated.

The results from the optimization model for the average speed assumption of 25 miles per hour (MPH) for the year 1985 are presented in Table 4. This speed level is assumed to be representative of average conditions within the region. The results for other speeds are detailed in Wilson. Savings are defined as reductions in the total cost for EMS provision in the region due to regionalization. These savings or avoided costs provide the estimates of

Table 4. Regional EMS Costs under Varying Response Time Constraints

	Million $			Minutes		
Response Time Constraint (minutes)	Base Case	Regional Case	Saving	Base Response Time	Regional Response Time	Change in Response Time
Emergency Calls Only — 1985						
10	2.121	1.742	0.379	6.78	6.91	-0.13
15-20	2.121	0.864	1.257	6.78	8.80	-2.02
25-30	2.121	0.565	1.556	6.78	12.09	-5.31
Emergency and Transfer Calls— 1985						
10	2.139	1.759	0.380	6.78	7.11	-0.34
15-20	2.139	0.910	1.229	6.78	9.00	-2.22
25-30	2.139	0.596	1.543	6.78	12.09	-5.31

benefits. The regional time and base time columns present the demand weighted average response times in minutes.

Focusing initially on service provision for emergency calls only, the greatest savings, $1.556 million, are achieved under the 25 and 30 minute response time constraints (RTC). Since these are the most liberal response time constraints, the most centralized allocation of EMS resources results and the greatest savings are attained. One facility housing three ambulances and their crews located in Clinton (site *#4,* see panel C Figure 1) is able to service the entire region's emergency demand (4,840 annual calls). Achieved response times in the region are substantially lower than the stated response time constraints since those weighted average calculations reflect the spatial distribution of demand organized into least cost service areas. This point illustrates the important role of demand weighting in locational analysis. However, these savings through consolidation of supply sites have resulted in an overall 5.31 minute increase in regional response time. Changes in response times for the individual communities are presented in Table 5. Those not experiencing any change are the supplying sites identified in the particular least cost model solution.

As the RTC is decreased (i.e. tightened) from the 25 minute level, the size of the set-covers contract and the maximum service time decreases. As a result, the amount of excess supply for any service site increases since the possible number of "services" per unit time has risen due to the smaller distances. Recall that under the peak demand criterion, the probability that "n — 1" customers (where n is

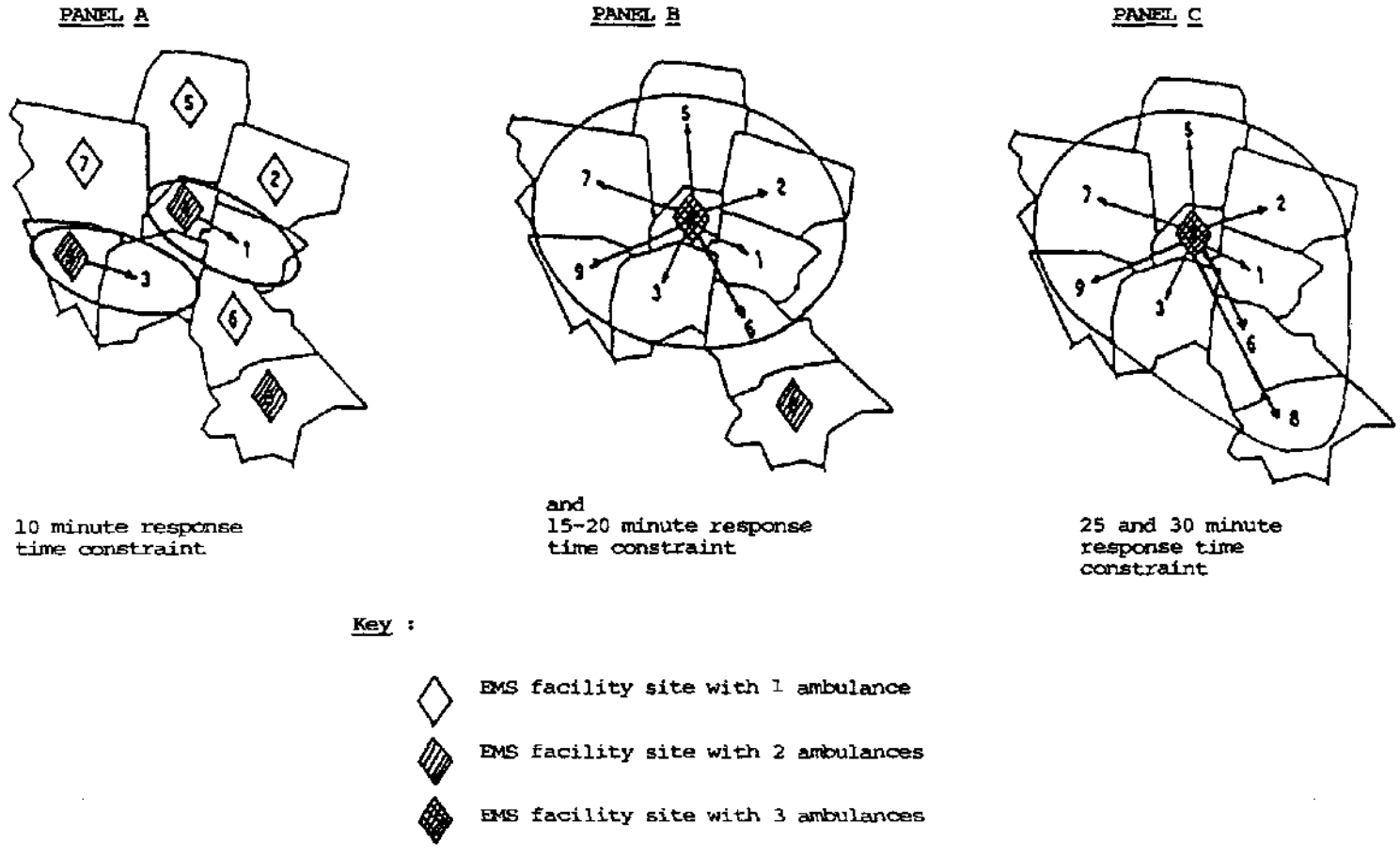

Figure 1. The Spatial Distribution of EMS Resources Under 1985 Emergency Only Demand Conditions

Table 5. Changes in Community Response Time

Community	Base Response Time	Minutes: Response Time Constraint 10 Increases in Response Time	15-20 Increases in Response Time	25-30 Increases in Response Time
1 Berlin	5.40	2.47	2.47	2.47
2 Bolton	6.74	0.00	3.53	3.53
3 Boylston	5.40	1.80	4.87	4.87
4 Clinton	2.71	0.00	0.00	0.00
5 Lancaster	8.64	0.00	3.50	3.50
6 Northborough	8.11	0.00	6.53	6.53
7 Sterling	8.64	0.00	1.63	1.63
8 West Boylston	8.64	0.00	0.00	16.08
9 Westborough	5.40	0.00	7.32	7.32

the size of the ambulance fleet) are in service must be no more than 0.95. In terms of equation (2), $dP(\ )/dR > 0$. Since the maximum possible service area is larger under the less stringent RTC (i.e. larger), the available excess capacity at any site is reduced and the tendency toward centralization is ameliorated by the immediate dispatch requirement. However, for the study area evaluated, the service capacity of a well located multi-unit facility is so large that response time within this small region becomes the binding constraint. Thus, regional system costs increase uniformly as the RTC is tightened.

It is intuitively obvious that for a given region, increasing the number of supplying sites will reduce response time.[1] In the cases shown in Table 4 and illustrated in Figure 1, the set-covers are such that a central site (#4) can cover itself and some combination of other demand points depending upon the RTC. This configuration is due to the shape and demographics of the study region. Clinton, geographically central in the region, also has the highest population and a hospital. Its relatively large population indicates that excess capacity is quite likely because the town will require at least 2 ambulances for all RTCs considered. Also, because a hospital is in the town, the variable costs are low relative to other sites in the region which adds further impetus to its choice as a major supply point.

At the 10 minute RTC, the size of feasible service areas is very limited even within this small region. Consequently, 10 ambulances distributed at 7 sites are still required to service EMS demand. At the 15 and 20 minute RTCs, the set-covers were such that the maximum mean service time, M*, implied a service capacity for Clinton, the central site, sufficient to absorb demand in excess of that for the entire region.[2] Only the binding 15 minute RTC prevented Clinton from serving the entire region. At this solution, the regional ambulance fleet drops to 5 units at 2 sites.

Benefit-Risk Tradeoff

For the EMS decisionmaker it is important to know the effects of the regional solution on cost and observed response time in the region. As expected, Table 4 reveals that the greater the total cost of a system, the lower the average response time in the region since there are more sites. Consequently, an important relationship is the change in response time relative to a change in cost between a regional and base system since response time is a proxy for risk. In the absence of a mapping of response time increases into changes in morbidity and mortality, the ratio of savings to increase in response time describes the benefit-risk tradeoff. The results for the emergency only service analysis show approximately a $428,000 annual saving for every minute increase in the weighted average regional response time. Translated into EMS resource equivalents, that sum would purchase 6 new ambulances for the region or pay 18 EMTs annual salaries. If the money were divided evenly among the nine towns, each would receive approximately $47,500 or roughly one-fifth of the average yearly community expenditure for EMS.

This very positive regional picture belies the experiences of individual communities. For example, under the 10 minute RTC, only two communities cooperate in service provision. As indicated by Figure 1 (panel A), Berlin (#1) would be served by Clinton (#4) and Boylston (#3) by Westborough (#9). Only Berlin and Boylston would experience an increase in response time, however, as all other communities would continue to supply EMS

[1] This result holds irrespective of the number of ambulances in the total regional fleet. In the extreme, increasing the number of supply sites would simply mean decentralizing some ambulance units.

[2] A less conservative approach would have used the average service time in the set rather than the maximum to calculate excess capacity. Since only approximately 30% of all emergency calls result in patient transport to hospitals, the use of the maximum potential service time is conservative.

(Table 5). Consequently, the savings per minute increase in response from the siteless communities view should be calculated as $185,000. Similarly, under the 15 and 20 minute RTCs, Clinton serves all but one outlying town. The weighted average response time increase for those communities experiencing service degradation is 4.85 minutes and the benefit per minute is about $259,000. Under the 25 and 30 minute RTCs, the demand weighted average increase for those communities experiencing changes is 7.98 minutes which implies a $195,000 saving per minute increase.

The disaggregated estimates of per minute savings are lower than those presented for the region as a whole, but these are still substantial benefits. This disaggregated view of the results is intended to highlight the expected difficulties of distributing the savings from any cooperative efforts. This research does not address the difficult questions of compensation for quality changes. Regionalization does not result necessarily in an unambiguous improvement for all communities. Some communities (those without supply sites after regionalization) must weigh the increase in response time against the potential cost savings. In any actual implementation, sites may be situated to ameliorate inter-local response time differences.

Transfer Demand

Many EMS units provide both emergency and transfer ambulance service. The model in this study was initially optimized using only emergency calls. By adding transfers, demand increases and influences both variable costs (more trips) and fixed costs (if the additional demand requires that the system add an additional ambulance or site). For the EMS manager the issue is cost; is it worth it to include transfer calls? For the 25 mph case, the changes in costs from the emergency only case to the emergency plus transfer case was examined (Table 4). The average difference for all RTCs when the system was regionalized and serving both emergency and transfer calls is $38,500, for the base cases: $18,000. These differences are produced by the 1,311 call increase generated by serving transfers (i.e., total annual calls are now 6,152). Zebrowski estimated, in his survey of private EMS firms, that the cost per call for transfers was $13.13 in 1980. The results from Table 4 indicate that each transfer call on average generates an additional cost of $29.37 in the regional case and $13.73 in the base case. This provides evidence for splitting EMS into separate emergency and transfer services. This segregation would reduce peak demand pressures on the emergency system and cut costs. A transfer service could be designed to operate at a much lower cost than the additional costs generated from serving transfers with EMS units.

Conclusions

The objective of this study was to estimate the benefits of regionalizing EMS. For a variety of response time constraints, a peak demand criterion and an average speed assumption, these savings were estimated. Substantial benefits were shown to exist. Demand, cost, and quality models were incorporated into the location model in order to optimize a regionalized system. The location model developed in this study represents an initial effort to endogenize peak demand considerations through an explicit assessment of excess capacity. While the selected approach is conservative, it is expected that these results will provide an impetus for a discussion of the relevance of a regionalization strategy for improving the efficiency of EMS. Implementation studies would improve upon the results presented by allowing for precise descriptions of peak demand distribution, site cost differences and more accurate service time estimates. Other delivery systems such as helicopters could also be incorporated. Consideration should also be given to an analysis of privately supplied EMS, an increasingly attractive alternative to many fiscally stressed communities. However, any benefit-risk estimates must be coupled with consideration of the complex administrative and legal aspects of a regional service.

While the results of this research show that benefits are available through cooperative provision of EMS, this does not imply that the barriers to implementation of such cooperative arrangements are not significant. Consequently, the analytical focus is placed directly on small regions. Financial incentives alone may not be sufficient to overcome the inhibitions to partnership. In particular, EMS services are traditionally locally supplied, poor information on quality and performance can

engender controversy, and cost savings can be difficult to assess due to the ambiguous nature of the quality measures used for EMS. Further, though this study does provide the information that incentives do exist, institutional resistance could be high due to possible reductions in EMTs if regionalization were implemented. Financial barriers appear not to be so important unless new locations, other than those presently existing in communities, would need to be constructed. The difficulty in creating a regional system will require discussions of risk, cost, revenues, and control. Attention must be focused on these issues to resolve some very thorny cooperative problems.

References

Bell, Colin E. and David Allen, "Optimal Planning of an Emergency Ambulance Service," *Socio-Economic Planning Service,* 3(1969):95-101.

Daberkow, S. G., "Demand and Location Aspects of Emergency Medical Facilities in Rural Northern California," Ph.D. thesis, University of California, Davis, 1976.

Daberkow, S. G. and G. A. King, "Demand and Location Aspects of Emergency Medical Facilities in Rural California," Giannini Foundation Research Report No. 329, University of California, March 1980.

Department of Public Health, " 105 CMR 170.000 Regulations for the Implementation of Massachusetts General Laws lllc, Governing Ambulance Services and Coordinating Emergency Medical Care," October 1982.

Dunlop and Associates, Inc., "Economics of Highway Emergency Ambulance Service," U.S. Department of Transportation, National Highway Traffic Safety contract FH-11-65451, 1968.

Hakimi, S., "Optimum Locations of Switching Centers and the Absolute Centers and Medians of a Graph," *Operations Research,* 12(1962):453-58.

Revelle, C. and R. Church, "The Maximal Covering Location Problem," *Papers of the Regional Science Association,* 32(1974): 101-14,

Stevenson, K,, "Operational Aspects of Emergency Ambulance Services," Technical Report No. 61, Operations Research Center, Massachusetts Institute of Technology, Cambridge, May 1971.

Toregas, C. and C. Revelle, "Optimal Location under Time or Distance Constraints," *Papers of the Regional Science Association,* 28(1972):133-42.

Toregas, C., R. Swain, C. Revelle and L. Bergman, "Location of Emergency Service Facilities," *Operations Research,* 19(1971): 1361-73.

Willemain, T. R., "The Status of Performance Measures for Emergency Medical Services," *Journal of the American College of Emergency Physicians,* (1975): 143-51.

Wilson, James M., "Risk-Benefit Analysis of the Regionalization of Emergency Medical Services," unpublished M.S. thesis, Department of Agricultural and Resource Economics, University of Massachusetts, Amherst, 1985.

Zebrowski, Chuck, Testimony, Public Hearing on 114.3 CMR 27.00 Ambulance Services, Boston, May 5, 1983.