10

System Design

Jack Stout, B.A.

The most powerful force influencing an EMS system's ability to convert available dollars into clinical performance and response time reliability is system design. The prime directive of EMS is to provide every critical patient the best possible chance of survival without disability, given state-of-the-art prehospital care and available financial resources.

Price of Entry

EMS advocates unwilling or unable to grapple with the realities of economic efficiency, worker productivity, and the need to generate more high quality service with limited financial resources cannot be taken seriously by public officials.

Before the United States was forced to tighten its economic belt, EMS advocates could ask, "What's a life worth?" Not anymore. Public needs of equal or perhaps even greater importance (for example, prenatal care, infant nutrition, drug abuse prevention, the battle against AIDS, law enforcement, and deficit control) now compete vigorously for the limited dollars available. The country is facing a far more challenging question, "Is this the best we can do with the money we've got?"

To answer this question, EMS leaders must compare systems. For the comparison to be more than a whitewash, they must risk comparing their system's current clinical and economic performance—not with systems known to be inferior—but with the handful of renegade systems operating at the outer limits of economic efficiency.

Secret Ingredient

High-performance EMS systems enjoy the advantages of superior system design discussed in this chapter; but design alone cannot explain the startling 300% differences throughout the industry in output per dollar. Something more is involved. The secret ingredient of high-performance EMS systems was discovered by a member of the Michigan Ambulance Association. During an overview of advanced system status management (SSM) practices he exclaimed, "Why hell, there's no magic in this; it's just a bunch of hard work." Indeed.

We Are Doing It For . . . Whose Benefit?

EMS system designers and managers cannot avoid the conflict between decisions made for the patient's benefit and those made for the system's convenience. Following are four examples:

- Faced with highly predictable 400% fluctuations in time-of-day and day-of-week demand for prehospital care (usually coincident with the worst possible traffic), ambulance services staffed at or near a constant level around-the-clock cannot claim, "We are doing it for the patient."
- Faced with massive and surprisingly cyclical shifts in geographic demand concentration by time-of-day and day-of-week (that is, buildings are fixed but people move around) and the predictable AM/PM rush hour reversal of traffic flow in most urban areas, ambulance services employing fixed-post deployment methods cannot claim, "We are doing it for the patient."
- Fire departments that, in spite of declining demand for fire suppression services, refuse to provide trained first response using existing personnel or overload a limited number of costly rescue crews while millions of dollars in engine company resources stand idle cannot claim, "We are doing it for the patient."
- Volunteer services failing to guarantee fully staffed paramedic response, while pointing

with pride to 100% volunteer staffing and token user fees or no fees at all (that is, angels of mercy don't charge a fee) cannot claim, "We are doing it for the patient."

The hard truth is that the conventional wisdom of EMS system design (for example, tiered response, the specialized production strategy, 24 hour shifts in urban settings, use of fixed-post locations, the every-man-for-himself dispatch method still used by many volunteer organizations) made convenience to the system and respect for its traditions a higher priority than meeting the needs of patients. For the convenience of the system, patients have been inconvenienced—in some cases to death.

Some EMS systems designed and managed for their own convenience have attempted to offset their inefficiencies through ever-increasing local tax support. Developing a tolerance for large even massive dollar dosage, low-performance systems have become addicted to local tax support with a bizarre result. Today an imperfect but unmistakable inverse correlation exists between the level of local tax support (that is, per capita per year subsidy) and the quality of patient care. The most startling example, as of 1992 America's most heavily subsidized urban ambulance service is also its least effective—that is, the District of Columbia (DC) ambulance service.

Whether or not the latest newly appointed medical director of the DC system will overcome the system's long-standing affinity for flawed design and inefficient production methods, the fact will remain that throughout their tenure, the DC system's traditional design and production methods consumed cash at the fastest rate in the industry, while producing an internationally embarrassing level of service.

Facing the political liability of a poorly performing EMS system design, elected officials too often fail to diagnose the *system itself* as the cause and treat the symptoms. Although it is doubtful that too much money causes quality of prehospital care to deteriorate, chronic symptoms of faulty system design are often misdiagnosed as financial malnutrition and treated with monetary nutrients. Over time, the EMS system designs most resistant to dollar therapy have become the most dependent. Figure 10-1 (based on 1991 data) illustrates the point.

The question is not whether an EMS system design is *capable* of consuming large sums of money, any design can. The question is whether, at any given level of financial support, one design generates more service of higher quality than other designs. This chapter presents the major elements of EMS system design that contribute to or detract from the system's ability to convert available dollars into quality prehospital care.

Fundamentals of EMS System Design

System design refers to the EMS system's underlying framework of legal, organizational, business and medical oversight structures, and financing strategy.

Limitations

System design is critically important but not all-powerful. The following are its limitations:

1. Talented and motivated people can produce good results in a bad system design, but not for extended periods of time.
2. Incompetence can produce poor results in even the best system design.
3. Talented people tend to be attracted to system designs more likely to nurture and showcase their individual talents.
4. Talented people have options because they are talented. In general, our industry's most talented managers choose to avoid employment in EMS systems that tend to nullify rather than demonstrate their abilities.
5. Good system design makes excellence possible and superior performance probable, but guarantees neither.
6. Bad system design makes excellence impossible and inferior service probable.
7. Sound system design cannot guarantee clinically appropriate and economically efficient performance; but poor system design can make consistent lifesaving performance extremely unlikely even impossible.

What System Design Determines

The term system design is uncomfortably vague. The first step in defining the concept is identifying the *direct effects* of an EMS system's design. Any EMS system's design establishes 8 major structural attributes and 28 secondary features. The 8 most important are as follows:

1. **Geographic scope.** The geographic scope of the system's *primary service area* (monojurisdictional or multijurisdictional) affects economies of scale and can determine whether less economically desirable areas can be served at all. (Note: The allocation of market rights in other utility industries deliberately grafted difficult-to-serve areas onto more economically desirable markets so a nationwide network of otherwise impossible telephone and electric power systems could develop. Although the

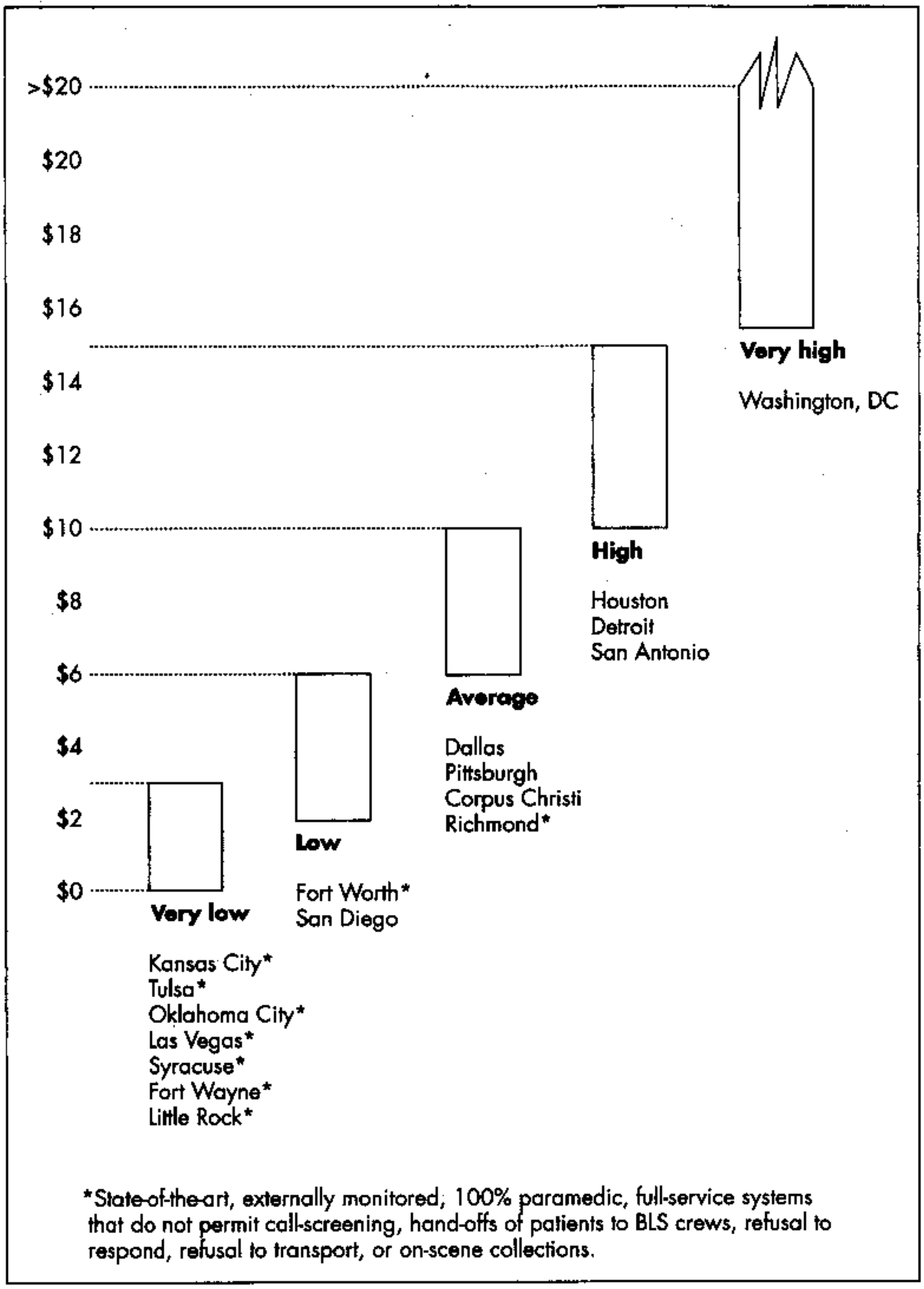


Figure 10-1. Local tax EMS sibsidy per capita per year.

ambulance service industry exhibits every characteristic of a utility industry, EMS regulators have thus far failed to learn from the success of other utility industries.)

2. **Standards setting and enforcement.** The process whereby performance standards are established (if established), monitored, and enforced is in the long run the most important aspect of an EMS system's design. In general, if this element is well-designed, it will eventually expose and correct defective aspects of the system's design.
3. **Division of functions.** Division of functional responsibilities among participating organizations (for example, EMT-D level firefighter; first response with paramedic-level private ambulance service) impacts funding requirements and revenue options, affects distribution of lia-

bility, and can enhance or dilute organizational accountability for performance results.

4. **Production strategies.** Production strategies employed (for example, the specialized production strategy vs. the flexible production strategy, engine company first response vs. rescue squad first response) establish the system's limit of worker productivity and thus more directly impact production costs and funding requirements than any other single factor including wages.
5. **Market allocation.** Selection of organizations to participate within the system (for example, bid competition, accident of history, political influence, default) sets the stage for all that follows. Oddly, only a few EMS systems deliberately award EMS market rights and responsibilities to the best-qualified organizations willing to do the job.
6. **Consequences of chronic failure to perform.** As 75 years of Soviet experience has revealed, organizations whose role in society is perceived to perpetual rather than contingent on performance, just do not try hard enough. If and how a poorly performing organization can be replaced by one more qualified is a key element of EMS system design. Not surprisingly, this feature is least likely to be applied where it is most clearly defined and readily available.
7. **Business structure.** The sources, amounts, routes, and contingencies of dollar flow into and within the system determine which organizational behaviors will be financially supported or rewarded and which will not. In many lower performance systems, these contingencies of financial reinforcement are either irrelevant to the system's reason for being or downright contrary (for example, budget increases justified by deficient performance). Market test comparisons with other systems serving similar markets and known for their efficiency can be a powerful stimulant where complacency prevails. In short the system's business structure provides or fails to provide the critical link between the public interest and organizational motivation.
8. **Caliber of management required.** Not all EMS system designs require the same caliber of management to extract the system's maximum potential. In general, policymakers must choose between low-performance designs capable of approaching their full but modest potential even if lead by managers of mediocre ability versus high-performance designs capable of achieving superior results, but requiring leadership talents well above the norm.

The last of the preceding items is in some ways the most interesting. Just as a high-performance automobile or aircraft operated at or near its limits demands a highly skilled driver or pilot, an EMS system with high-performance potential demands a higher caliber of management. Furthermore, in the same way a high-performance machine may be dangerous in the hands of a modestly skilled operator, a high-performance EMS system design can be deadly and expensive if managed by persons of modest ability and motivation.

Thus the chief advantage of low-performance EMS system designs is that their maximum performance, modest though it may be, can often be achieved by managers of limited ability. For example, what could be easier to manage than a system using level coverage staffed with 24-hour shifts and fixed-post locations? However, even where substantial subsidies are available the maximum performance of inferior designs falls far below that achieved by systems of superior design.

Origin of Current Design

Most prehospital system designs are *not* the product of policy decisions deliberately selecting certain features and rejecting others with full understanding of each option's demonstrated advantages and pitfalls. Rather, most system designs have simply evolved through a process of short-term, issue-driven policy accretion or were loosely patterned after a television series.

Fortunately a growing number of today's working EMS systems are products of informed selection of structural features that deliver the best possible chance of survival without disability using limited financial resources—the "Prime Directive." The track records of these high performance EMS systems are defining the future of EMS.

Economic Efficiency

Just as nothing impacts a system's ability to turn dollars into service more powerfully than design, nothing impacts patient care more powerfully than the system's ability to convert dollars into service. Economic efficiency and clinical effectiveness are more than related; they are absolutely interdependent.

Measuring Economic Efficiency

By shortening one or two legs of the triangle in Figure 10-2, even a poorly structured, badly managed system

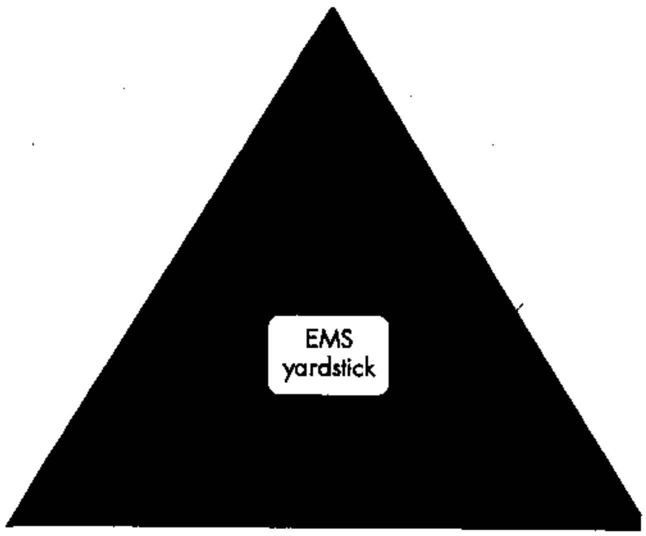


Figure 10-2: Measuring economic efficiency.

can perform reasonably well on one or two measures, creating an appearance of competence when viewed from a favorable angle. The consequences of such distortion harm the patient, the taxpayer, or frequently both. For example, an unskilled manager can limit spending by allowing clinical quality and response time reliability to deteriorate. Similarly, given enough money, even the most pathetic management team can generate something of value.

The challenge is to simultaneously generate clinical excellence, response time reliability, and economic efficiency. Meeting that challenge qualifies any system and its personnel as high-performance.

Economic efficiency among EMS systems today varies by more than 300%. That is, when compared in terms of total system cost per patient transport, total system cost per capita for service area population, or combined subsidy-cost per capita and user-fee level, some system designs consistently require more than 3 times the financial resources than others. With few exceptions, differences in system design account for differences in economic efficiency; the combination of poor system design and severe budget limitations is deadly.

Learning to Distinguish the Best from the Rest

As in most facets of human endeavor, we stand to learn the most from those in the lead. Some EMS systems ingesting huge sums of money produce service of limited quality, quantity, and reliability. Other EMS systems feed on scarce financial resources yet consistently produce high-quality service and impressive response time reliability. In nearly every case, system design accounts for the difference.

After nearly two decades testing our industry's 27 basic prehospital system designs (and hundreds of variations), design features can be isolated that are more common and sometimes universal among better performing systems and less common or nonexistent among systems performing at low levels of efficiency.

Levels of Efficiency

For purposes of examination, EMS systems can be divided into the following four categories:

- High quality with above average cost (San Antonio),
- Low quality with below average cost (most common),
- Low quality with above average cost (Washington, DC), and
- High-performance EMS systems—above average service with below average cost (Las Vegas; Syracuse; New York; Tulsa/Oklahoma City; Reno, Nevada; Fort Wayne, Indiana; Little Rock, Arkansas; Fresno, California; Fort Worth, Texas; Kansas City, Missouri; the ambulance service component of Pinellas County EMS, Florida; Clark County, Washington; the Acadia System serving rural Louisiana; the East EMS System serving rural eastern Texas; and a handful of others).

The systems in the high-performance category share key features of system design rarely associated with less cost-effective systems. Before valuable lessons can be learned from high-performance systems, it is necessary to distinguish them from the rest. Just as the acid test of a quality product lies not in its initial performance but in its long-term reliability, most EMS designs perform reasonably well while new and shiny. The question is, how will the system perform when the new wears off?

Economic efficiency and clinical performance are directly interdependent and certain system design characteristics influence and define efficiency potential. It is also necessary to understand that three broad objectives, when combined, furnish a basis for predicting and explaining the operational consequences of policy decisions affecting system design. Those three objectives are as follows:

- To establish a framework for comparing the structure and performance of any EMS system with those of any other EMS system regardless of system design.

- To develop a basic working knowledge of three measures essential to judging and comparing economic efficiency.
- To identify EMS system design elements that contribute to or detract from the system's ability to convert dollars into service.

First a method of defining, examining, and classifying EMS systems must be established that can be fairly and productively applied to any EMS system regardless of its design.

The Great Conceptual False Start of 1973

Propelled by the force of more than $300 million in federal grants, the federal government's 15 components of EMS system design sent the nation on a conceptual wild-goose chase from which we are still recovering (see p. 10).

People began to believe that systematic thinking was logical thinking. In fact, it is quite possible to think illogically in a very orderly and systematic way.

Lacking fundamental elements such as medical direction, financial structure, legal structure, a method of allocating market rights and responsibilities, or even a framework for describing performance expectations the 15 components fell short of furnishing a conceptual basis for rational EMS system development. After brief national recognition primarily for their ability to secure and spend federal grant funds the shining EMS stars of the 1970s faded. Almost without exception the high-performance systems of today were not recipients of federal EMS grants.

The 15 components failed in four ways to furnish a conceptual foundation on which a nationwide network of quality prehospital care systems could be built. First the 15 components were not developed from the patient's (that is, the customer's) point of view. Second, as a systems approach (the rage of the 1970s), the 15 component framework violated key principles of systems engineering. Third the 15 components failed to identify and organize essential elements of an effective EMS system into a useful framework. Fourth the 15 components assumed a hopelessly oversimplified view of a complex industry.

With the advantage of hindsight a far more effective conceptual framework has developed. A framework evolved from the patient's point of view encompassing essential elements of input, structure, output (process), and outcome.

EMS System Defined from the Patient's Point of View

An EMS system consists of those organizations, individuals, facilities, and equipment whose participation is required to ensure timely and medically appropriate response to each request for prehospital care and medical transportation.

Every EMS system regardless of design has two types of input; patients and money and several types of output: services such as patient assessment, extrication, defibrillation, and medical transportation. Linking input and output the system converts dollars into service. Depending on efficiency, some EMS systems generate small volumes of low quality service from a large dollar flow and others generate large volumes of high quality service from a small dollar flow.

Impact on patient outcome, the most important result of all this activity, is a function of two related but very separate factors. First, efficacy of system protocols (that is, priority dispatch protocols, medical protocols, and system status management protocols) and system compliance with those protocols. These two factors must be dealt with separately, because a defect in one factor cannot be corrected by acting on the other.

Relationship of the System to Its Environment

Past definitions of EMS systems developed from the viewpoint of the system itself. For example, systems based on the specialized production strategy define routine transport service and the financial and other resources involved as outside the system, thus avoiding comparison with systems based on the flexible production strategy. The specialized production strategy employs two, three, or even four types of ambulances each specializing in specific types of calls. The flexible production strategy employs a single ambulance staffed and equipped to respond to any type of call. Either strategy may incorporate the use of one or more types of first responder units (Figure 10-3).

Some EMS system definitions assume that callers reliably triage themselves into emergency and routine categories, dialing 9-1-1 in emergencies and a 7-digit telephone number for routine transport. The underlying assumption is that callers who do not dial 9-1-1 are not a responsibility of the EMS system because they failed to follow the access rules. Again the system escapes accountability for a huge segment of the patient population including related costs and outcomes.

The term "EMS" reflects the original assumption that emergency patients can be distinguished from non-emergency patients and that specialized production dedicated to each type of work will provide better service at a lower cost. This once conventional wisdom failed to account for patterns of demand for prehospital care and has since proved false.

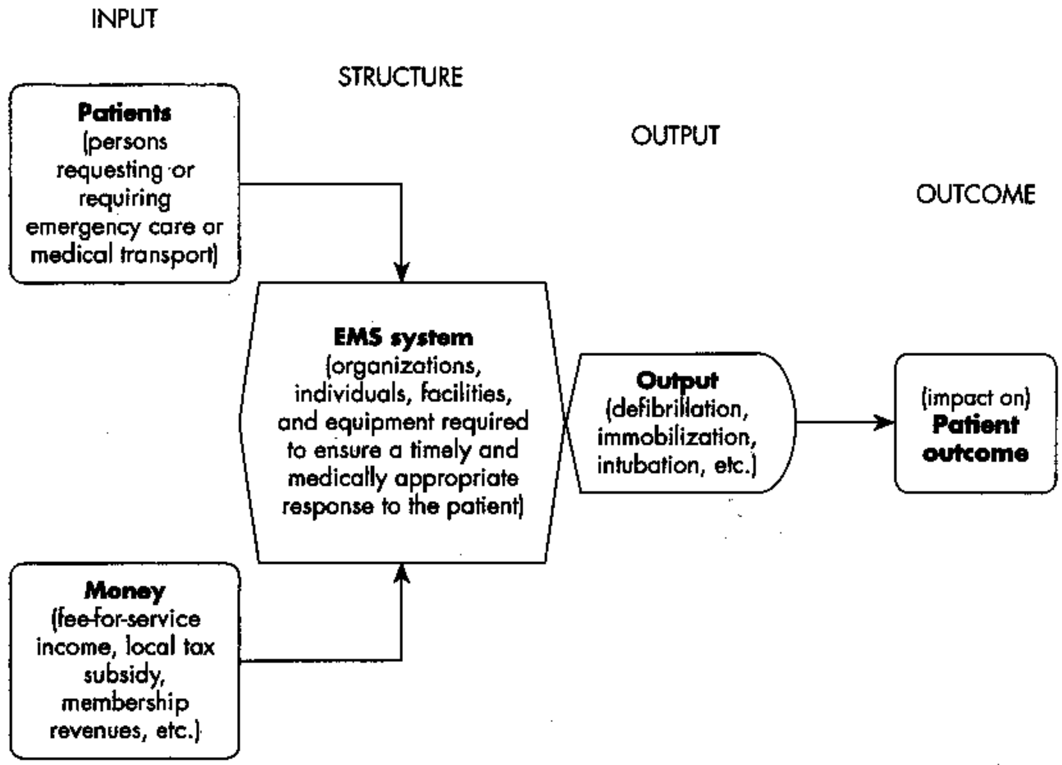


Figure 10-3. Relationship of the "system" to its environment.

Any conceptual framework subordinating the patient to the system fails to furnish a basis for fair comparison of different system types. Because the needs of the patient are not securely anchored to the system's purpose, such conceptual frameworks are not stable platforms for comparing advantages and disadvantages of system designs.

Alternatively, defining the EMS system from the patient's point of view avoids bias and apples-to-apples comparisons become possible. Before a patient-centered conceptual framework can replace the 15 components of 1973 the patients must be identified.

If an EMS system serves patients, then there must be a clear understanding of who these patients are. The following are the facts regarding consumer requests for medical assistance and transportation where a universal 9-1-1 exists for an extended period of time:

1. Slightly more than 50% of all medical requests enter the system through 9-1-1 access.
2. Slightly less than 50% of all medical requests enter the system through 7-digit telephone numbers.
3. When paramedic crews respond, 20% to 30% of 9-1-1 medical requests are found to require paramedic skills.
4. When paramedic crews respond, about 12% of 7-digit number ambulance requests require paramedic skills.
5. About 70% to 80% of 9-1-1 medical requests do not require paramedic service.
6. About 88% of non-9-1-1 medical requests do not require paramedic service.
7. Many bona fide emergency calls do not require ALS, whereas some bona fide non-emergency calls require ALS en route (for example, scheduled interfacility transport of a patient requiring advanced en route support—an increasingly common type of call) (Figure 10-4).

Thus in creating a conceptual framework for comparing EMS systems of diverse design (that is, a framework biased in favor of the patient but blind to system design), the system's response to every patient must be tracked including those failing to follow the system's telephone access rules and those the system excludes from its field of vision. If the decision to organizationally and operationally separate emergency service production from routine transport production is clinically and financially sound, only a comparison of total system performance—clinical and financial, emergency and routine—can test the proposition conclusively.

Thus in defining the system from the patient's point of view, all patients are included regardless of how they access the system, whether system managers accept or reject responsibility for their care, and whether such patients ultimately require heroic

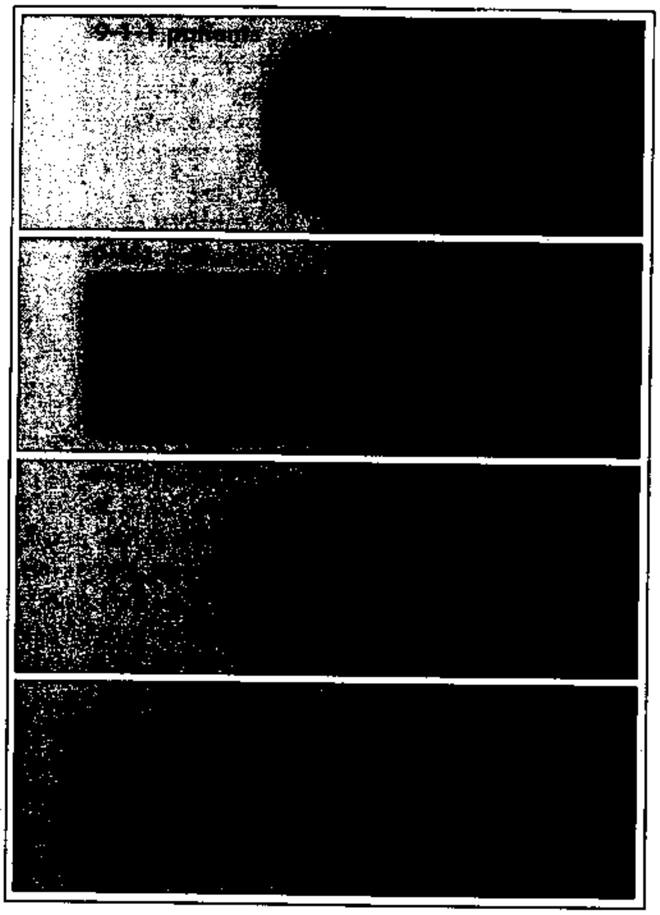


Figure 10-4. Predicting patient needs. All we know for certain is that some requests are via 9-1-1, and some are via 7-digit numbers. Figure based on data from full-service prehospital care systems serving a combined patient population of more than 300,000 patients annually. (Courtesy Wendy Jones, M.D., FACEP.)

intervention or just simple kindness and safe medical transport.

The system's most important input is the patient. The other input every system must have is money. Each of the organizations required to generate a timely and medically appropriate response to each patient's request requires funds to fuel its operation. Again the EMS system is a machine for converting dollars into service.

EMS systems regardless of design deal with two universal inputs, patients and money. Similarly, every EMS system regardless of design is a structure comprised of organizations, individuals, equipment, some type of communications networks and facilities to house functions such as fleet maintenance, training activity, control center operations, crew quarters, billing and collection activity, and other support services. Along the same line, every system has some type of information system, mechanisms of financing its operations, and a formal or informal legal structure defining organizational roles and responsibilities. Finally, almost every system has or purports to have some type of medical oversight structure, external or internal.

EMS System Matrix

The EMS System Matrix that follows provides a two-dimensional framework for documenting and comparing any system's structure and performance with that of any other regardless of design. The most important task of every EMS system is to generate a planned, coordinated, and medically appropriate response to every patient in need of its services. Thus the matrix is deliberately driven from its left vertical axis (that is, a comprehensive, sequential list of system response phases from prevention through indirect medical evaluation of the response and ongoing quality improvement).

Documenting an EMS system using the matrix is a formidable task regardless of the system's design; and therein lies its value. Using the matrix, any EMS system can be documented in detail and compared fairly with any other regardless of differences in design (Figure 10-5).

A few comments should be made regarding use of the matrix. Medical oversight items are the only elements appearing on the vertical and horizontal axes. As outputs, direct and indirect medical control require the full-range of structural components (such as a responsible organization, a legal basis, and funding) for which provision is made horizontally.

The three most important duties of medical oversight are formulating system performance specifications (that is, standards and protocols governing output performance), monitoring compliance with specifications, and action ensuring compliance as needed. Although most high-performance systems employ unified medical oversight (that is, a single medical authority in charge of all output categories), some systems still do not. Thus provision is made vertically for listing a separate medical oversight entity or specifying "none" for each output category.

Comparing Clinical Capability

To document and compare clinical capability, it is necessary to examine written protocols, personnel certification requirements, equipment standards, monitoring and quality improvement practices, patient outcome measures, and other aspects of the following system outputs on the left axis of the EMS matrix:

OUTPUTS (services)	STRUCTURE									SYSTEM STANDARDS AND PROTOCOLS
	A. Responsible organization	**B.** Personnel	**C.** Equipment	**D.** Communications	**E.** Facilities	**F.** Info-systems	**G.** Finance (sources, amounts)	**H.** Legal basis	**I.** Medical oversight	
1. Prevention and early recognition										
[redacted]										
3. Complaint-taking function										
[redacted]										
5. First Response dispatch										
[redacted]										
7. First Responder services (rescue)										
[redacted]										
9. Direct medical control										
[redacted]										
11. Indirect medical control										

Figure 10-5. EMS system matrix.

1. Prevention and early recognition (for example, seat belt awareness, feetfirst first time water safety, early recognition of cardiac symptoms)
2. Bystander action and system access
   - A. CPR instruction
   - B. Telephone
     - 1. Emergency (9-1-1)
     - 2. Routine (7-digit number)
3. Complaint-taking function (in 9-1-1 systems)
4. Telephone interrogation and prearrival instructions
5. First response dispatch
6. Ambulance dispatch
7. First responder services
   - A. Rescue and extrication
   - B. Initial medical support and assistance during transport
8. Ambulance services
   - A. Life-threatening emergency calls (presumptively classified)
   - B. Non life-threatening emergency calls (presumptively classified)
   - C. Routine transport calls (presumptively classified)
   - D. Interfacility transfers
   - E. Helicopter transport (scene flights)
9. Direct medical control
   - A. Electronic instructions
   - B. Hands-on care by scene physicians
10. Receiving facility interface
    - A. Patient exchange procedures
    - B. Participation in quality assurance
    - C. Equipment exchange arrangement
    - D. Information exchange arrangement
    - E. Selection of hospital destination

11. Indirect medical control
    A. Internal or external
    B. Advisory or authoritative
    C. Integrated or fragmented
    D. Qualifications of physicians
    E. Level of funding and staff support

Response Time Reliability

Just as a system generates reliable or unreliable clinical performance, it also generates reliable or unreliable response time performance. Figure 10-6 shows ambulance response time distributions for two systems claiming "8-minute response time reliability." Given what has been learned recently about the pathophysiology of cardiac arrest, it is clear that Oklahoma City's 8-minute *average response time* delivered life-threatening service to most patients when these data were collected. (That operator has since been replaced.)

In contrast, Tulsa's 8-minute *fractile response time* (at the 90% level of reliability) delivered a bona fide chance of survival more than 90% of the time. Because cardiac arrest is the most time critical of all calls, response time standards for first response and ambulance service are both geared to meeting the needs of patients experiencing cardiac arrest.

Average response time is not only a dangerously misleading indicator of response time reliability, it is also a clinically inappropriate goal. Deployment practices producing the most impressive-sounding average statistics are quite different from those generating 8-minute—90% fractile reliability or a less stringent rural equivalent. It is impossible to pursue the lowest possible average response time *and* the highest percentage fractile reliability; a choice must be made.

The type of report shown in Table 10-1 typifies performance by the ambulance component of the Fort Wayne, Indiana, EMS system, and is widely used by high-performance systems to track response time reliability.

Response Time From the Patient's Point of View

Response time is discussed as though what it is and how it is measured are understood. The following questions and answers summarize what paramedic response time is and is not.

Q: When does the paramedic clock start?

A: When the first request for help is received. It does *not* start when an extended telephone interrogation has been completed, when the responding

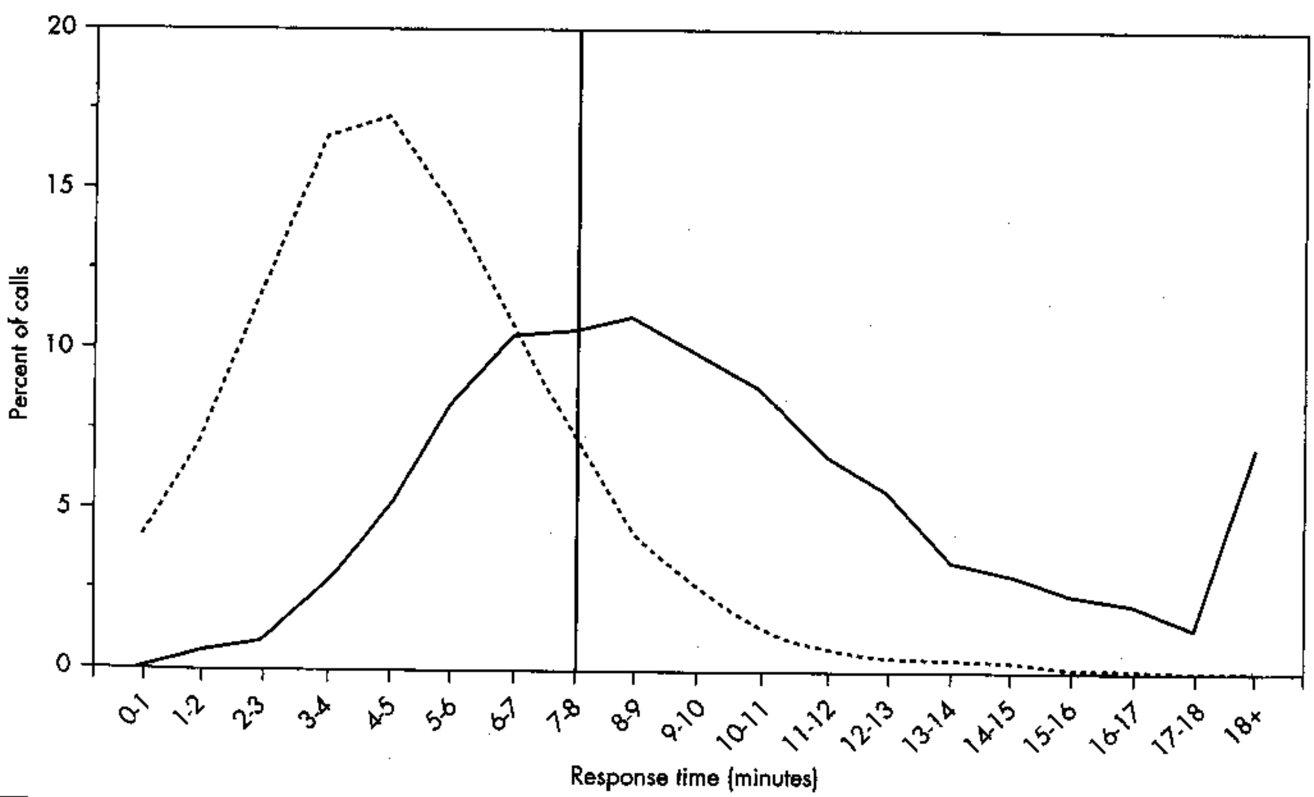


Figure 10-6. Ambulance response time distributions for two systems (Tulsa and Oklahoma City) claiming "8-minute response time reliability."

unit is alerted, when the crew acknowledges the dispatch, or when the unit starts en route.

Q: When does the paramedic clock stop?

A: When a fully staffed and equipped paramedic unit arrives on the scene. Some systems stop the clock not on arrival at the scene, but on arrival at the patient. To time-stamp this moment in large-volume systems a "patient contact" button may be programmed on a portable radio with automatic number identification capability and interfaced to the (CAD). Where automated vehicle tracking (AVT) is available the CAD also time-stamps electronic convergence of AVT and scene. It does *not* stop when a volunteer crew responds to the station, when the responding crew acknowledges the dispatch, when the unit starts en route, when the unit is within sight of the scene, when a First Responder crew arrives on the scene, when a supervisor arrives in a "flycar", or when an EMT-A mutual aid provider arrives on the scene.

Equal Opportunity

Although a fractile response time standard equalizes response time reliability among various zones, districts, or neighborhoods of the service area, it still leaves room for medically and politically dangerous geographic pockets of chronic poor performance. If the areawide standard is already stringent, requiring that standard within each zone or district will raise the systemwide standard to an unattainable level. The solution is institution and enforcement of the following standard: there shall be no chronic pattern of response time discrimination against any neighborhood, district, or zone as defined by local ordinance.

Table 10-1. Fractile Response Time Distribution

Response Time (Min.)	Runs (No.)	Total (%)	Cumulative (%)
<1	4	2.0	2.0
1-<2	8	4.0	6.0
2-<3	18	9.0	15.0
3-<4	28	14.0	29.0
4-<5	36	18.0	47.0
5-<6	40	20.0	67.0
6-<7	34	17.0	84.0
7-<8	18	9.0	93.0
8-<9	6	3.0	96.0
9-<10	2	1.0	97.0
10-<11	1	0.5	97.5
11-<12	0	0.0	97.5
12-<13	0	0.0	97.5
13-<14	2	1.0	98.5
14-<15	1	0.5	99.0
15-<15	2	1.0	100.0

How costly is response time equality among neighborhoods? The city with the longest-standing and highest expectations of response time equality among neighborhoods is Kansas City, Missouri. There 87% citywide paramedic reliability at the 8-minute level is cause for alarm, practically a crisis, and grounds for declaration of major default. The elected official from each of Kansas City's six districts expects monthly reports demonstrating that the constituents are getting their fair share of response time reliability. Adding the neighborhood standard to Kansas City's already stringent 8-minute—90% fractile standard increased annual production costs of an unusually efficient provider between $500,000 and $700,000.

But My Town is Harder

In rural Arizona, *low* population density is the excuse for extended response times, and in New York City *high* population density serves the very same purpose. Wherever response time reliability is deficient the managers have developed persuasive arguments accounting for the deficient performance. As one manager put it, "If God ran this system on the budget I've got, His response times wouldn't be any better."

It is true that call-density-per-square-mile, road systems, shape of the service area, caliber of mutual aid, traffic congestion, placement of hospitals, weather conditions, population fluctuations, and other factors define the limits of response time reliability achievable from any given operating budget. In other words, these factors determine the maximum productivity level or unit-hour utilization ratio (U/UH) at any given level of response time reliability. One of several indicators monitored by managers of high-performance systems, the U/UH ratio is calculated by dividing the number of patients transported during a given period by the number of ambulance unit-hours produced during that same period. In metropolitan areas, U/UH ratios from 0.33 to more than 0.5 are routinely generated by high-performance systems meeting the 8-minute—90% standard of reliability. However, experience has shown that the most important factor in determining maximum, realistic response time reliability at any given funding level is system status management (SSM).

Comparing System Costs

Having established a basis for comparing internal structures, clinical performance, and response time

Next estimate the current total average bill (base rate mileage and all add-on charges) for ambulance service in the service area. Note that number, but do not plot it on the chart yet. Estimate the current per capita per year local tax subsidy of the system's ambulance service component. Using these two numbers (the current total average bill and the current per capita per year subsidy), locate where the current total average bill intersects the current per capita subsidy and mark the spot. This point defines the system's current actual position on its subsidy-price line, the second of the two points required to draw the system's current subsidy-price trade-off line.

Now for the moment of truth. Connect the two marks, extending the left end beyond the plotted point to intersect the vertical axis. The result is the subsidy-price trade-off line. At this point many feel they have somehow misunderstood the instructions. Probably not. If the system is an urban system generating paramedic ambulance service with 8-minute—90% reliability and you plot both ends of the subsidy-price trade-off line on Figure 10-7 without extending either axis, your system is more efficient than 90% of all urban ambulance services. For example, to plot the current subsidy-price trade-off line for Washington, DC's ambulance service, it is necessary to more than triple the sample chart's vertical axis and double its horizontal axis; for New York City the vertical axis would be doubled.

Design Factors Affecting Efficiency Potential

Although a high-performance EMS system is the product of "a thousand little things done right," design characteristics of four key elements impact the system's ability to convert dollars into service. Those key elements and the most significant design characteristics of each follow.

Service Area Definition

In the same way and for many of the same reasons that it has become impossible to operate a clinically superb and financially stable stand-alone 20-bed hospital, it is increasingly difficult to operate a small ambulance service of high clinical quality. The minimum for bona fide high-performance operation is a service area population of approximately 150,000 people exclusively served (emergency and routine) by a single ambulance service provider. The average cost curve continues to decline as the population increases to about 1.2 million, still served by a single firm (Figure 10-8).

Where local populations are insufficient to generate optimum economies of scale, multijurisdictional systems (ideally, but not necessarily, contiguous) are often the answer. For example, Tulsa's Emergency Medical Services Authority, the first public utility model, handles about 68% of all patient transports in the state of Oklahoma including much rural area without local tax support.

On the other hand, a large urban service area can be made to function as poorly economically as if it were a sparsely populated rural area. By allowing multiple firms to share the same geographic market, each firm's service area remains equally large, but call-density per square mile resembles that of a rural area.

Medical Oversight

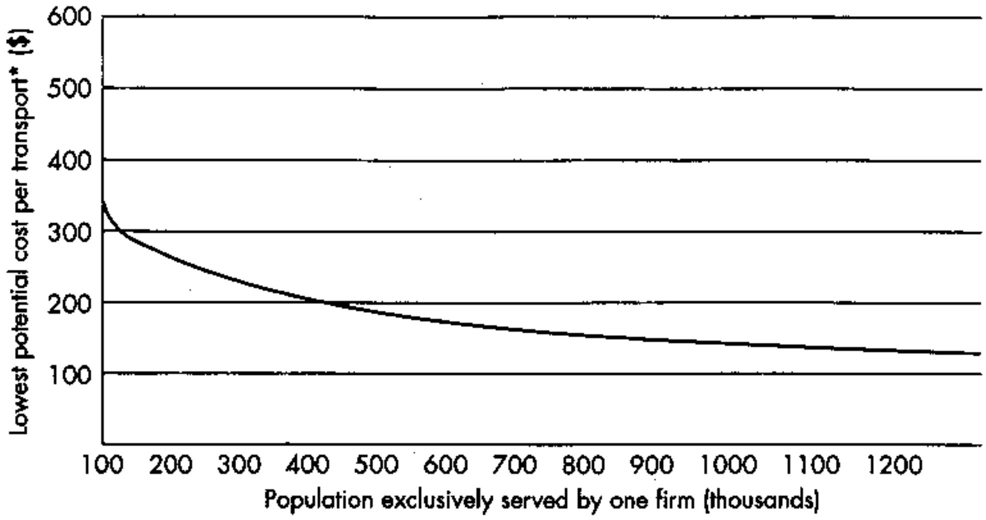
High-performance EMS systems are driven by externally imposed and enforced clinical and response time standards. In this regard, the characteristics of effective medical oversight are as follows:

External versus internal. The medical director is neither hired nor compensated by any organization whose work is the subject of medical oversight. The medical director's authority is independent of and superior to the organizations participating in the EMS system. Otherwise, it would be difficult or impossible for the medical director to fill the role of "referee" in disputes among participating organizations.

Advisory versus authoritative. On matters affecting the quality of patient care, the medical director does direct, not advise. (That is, the system design assumes that the delivery of prehospital care is the medical director's practice of prehospital emergency medicine.)

Scope of authority. A single medical director oversees all organizations and individuals participating in the EMS system. The scope of authority is systemwide and extends to all output components of the EMS matrix, all organizations in the system, and in multijurisdictional systems this authority is legally recognized by all jurisdictions served.

Funding. Effective medical direction requires a sustained commitment and continuing level of effort. The minimum funding required appears to be from $3 to $5 per patient transport (emergency and routine) systemwide. It is higher in smaller systems than larger systems with better economies of scale.



*Assumptions:
- 8-minute/90% paramedic response time reliability
- High clinical quality
- Medium productivity (0.25 to 0.40 U/UH ratio)
- Efficient unit-hour production ($50 to $80 per unit hour)
- Midwest United States cost of living
- Average difficulty of coverage

Figure 10-8. Impact of economies of scale: average cost curve.

First Responder Program

Delivered by existing fire department personnel with fire apparatus, firefighter first response is without question local government's best public service bargain (less than $27 marginal cost per response at the EMT-D level including training, fuel, medical supplies, accelerated vehicle maintenance and depreciation, and amortized costs of medical equipment). The characteristics of first response in high-performance systems include the following:

Production method. Engine company responds at the EMT-D level, no separate rescue units requiring additional personnel and vehicles. Many experts suggest that a rapid First Responder-D or EMT-D response with 10-minute—90% paramedic ambulance service will save more lives than an EMT-A response with 8-minute—90% paramedic ambulance service.[1] This is an easy trade economically and medically, that is, the amortized costs of automated external defibrillators and marginal training costs for first responders is less than the savings realized by replacing the 8-minute—90% standard with a 10-minute—90% standard. Note that the 8-minute—90% ambulance response time standard was developed before the value of early defibrillation was fully understood.

Utilization. Priority dispatch (for example, Clawson protocols) on all presumptively classified life-threatening calls, usually from 40% to 45% of 9-1-1 medical requests or about 25% of all ambulance calls (emergency and routine).

Service area. Unlike ambulance services that must respond and transport across geopolitical boundaries in conformance with medical-trade patterns, the most cost-effective form of first response is monojurisdictional. Thus in multijurisdictional EMS systems (for example, Pinellas County, Florida) each participating fire department is responsible for serving its sponsoring jurisdiction, and the ambulance service component operates regionally for best economies of scale.

Unit selection. In advanced multijurisdictional EMS systems, first response near geopolitical borders is dispatched on a "nearest-unit" basis, not on the basis of political affiliation. The Pinellas County, Florida, EMS system, with 17 participating fire departments and a single regional ambulance service provider, is highly advanced in this respect.

First—before what? No discussion of first response would be complete without raising the two following questions:

- How often are your First Responders first?
- Early defibrillation: earlier than what?

Even though fire department resources outnumber ambulance service resources, it has often proved difficult for first responders to routinely arrive before ambulance crews achieving an externally monitored 8-minute–90% standard of reliability. The Fort Worth Fire Department may deliver America's most reliable first response. As part of a system whose ambulances achieve the 8-minute standard with 90% reliability, the Fort Worth Fire Department first responders are first on the scene in approximately 80% of presumptively classified life-threatening calls.

Ambulance Service

Of all EMS system elements, ambulance service is the most costly (except where separate rescue units are used for first response), the most complex, and the most politically volatile. To make matters worse the proven principles of high-performance ambulance operation fly squarely in the face of the conventional wisdom of the 1970s.

Every ambulance service documenting externally monitored paramedic response time performance at the 8-minute—90% level of reliability while consuming a subsidy of less than $3 per capita per year exhibits the following characteristics:

Single-provider. Exclusive market rights to furnish emergency and routine ambulance service are granted to a single (often competitively selected) organization. "Cream skimming" competition is banned by local ordinance or state law. To put it more directly, elected officials designing an EMS system face a difficult but inevitable choice. They may enjoy the benefits of two of the following three features, but they cannot have all three:

- High quality ambulance service with clinically sound response time reliability.
- Retail competition and consumer choice within the market.
- Little or no local tax subsidy.

Flexible production strategy. Rather than operating specialized ambulance fleets (for example, paramedic units for really sick patients, EMT-B units for not-so-sick patients, and still other fleets providing "invalid coach" transport), high-performance systems employ a single fleet of paramedic units capable of handling any type of service request. Single-provider systems using advanced system status management practices (flexible production strategy) allow safe use of productivity levels or U/UH ratios as much as 300% higher than those found in systems relying on the specialized production strategy.

Peak-load staffing. Patient demand and traffic congestion tend to cycle predictably and coincidentally on a weekly basis. Thus rather than relying heavily on 24-hour shifts and constant staffing practices, high-performance systems carefully match supply with predictable demand and traffic pattern fluctuations for each of the 168 hours of the week, demand fluctuations are tracked at one or two standard deviations from the mean to achieve the 90% standard of reliability. Patterns of demand for EMS more closely resemble those for law enforcement than fire suppression. For the same reason urban police departments do not use 24-hour shifts, high-performance ambulance services in urban areas use few if any 24-hour shifts.

System status management. Because buildings are fixed and fires are rare, fixed-post locations make sense in deploying for fire suppression. In contrast, geographic patterns of demand for EMS cycle widely on an hourly basis with the movements of people and their changing patterns of activity. Just as high-performance EMS systems carefully study and learn the cyclical patterns of *volume* demand, they study and learn the cyclical patterns of *geographic* demand distribution and changes in traffic congestion patterns, designing and refining deployment strategies precision matched to patients' need for service. A medium-sized urban high-performance system may use more than 2000 system status plans (SSPs), one for each potential remaining level of production capacity for each of the 168 hours of the week.

Summary

Although most high-performance EMS systems employ a private contractor for delivery of paramedic ambulance services, every high-performance EMS system employs one or more fire departments in the critical role of first response. America's best and most cost-effective EMS systems blend the best capabilities of the public and private sectors. Even so, is it necessarily true that private is better when it comes to provision of paramedic ambulance service? The irrelevance of this question cannot be better illustrated than by the simultaneous truth of *all* four of the following statements:

1. Some of the best ambulance service is provided by government agencies.
2. Some of the worst ambulance service is provided by government agencies.

3. Some of the best ambulance service is provided by private firms.
4. Some of the worst ambulance service is provided by private firms.

In choosing an ambulance service provider, each jurisdiction selects from one or two local government agencies and several hundred private firms throughout the United States, several with national reputations and excellent track records serving multiple communities in several states. The rational choice is the best qualified organization willing to take on the task. If that organization's credentials and track record clearly demonstrate an ability to generate the required clinical and response time performance *from the financial resources available,* then whether that organization is a government agency or a private firm is only as important as the color of its stationery. Experience has revealed that much of what was assumed and believed during the 1970s simply was not true. Unfortunately the lessons were learned slowly.

REFERENCES

1. Lombardi G et al: Outcome of out-of-hospital arrest in New York City, *JAMA* 271:678-683, 1994.