by Jack Stout

Measuring Response Time Performance

by Jack L. Stout

You've probably seen the commercials where comedian Rich Hall promotes Pizza Hut's promise of a lunch-time pizza within 5 minutes or your next pizza is on the house. I guess they figured we'd rather watch Rich Hall than a Pizza Hut attorney reading off the terms of a limited warrantee, e.g.:

> *"WHEREAS, certain purchasers of prepared foods, hereinafter referred to as 'Noon-time Customer,' may be subject to restriction by certain externally imposed time limitations; and WHEREAS, such time limitations may cause an effective reduction in the benefit or value to said Noon-time Customer of foods not available for consumption by said Noon-time Customer within the window of opportunity defined by said time limitations; NOW, THEREFORE, it is hereby agreed . . . etc."*

I think you'd agree, Rich Hall is better. Even so, the commercials could be even funnier. Say this guy (Rich Hall) walks in, orders his "5-minute" pizza, and 6 minutes later the waiter delivers the goods. Hall looks up at the waiter and says, "Hah! You're 1 minute late, so my next pizza's free." The waiter replies, "Not so fast, hotshot, we never said we'd deliver 'em *all* in 5 minutes. The last guy got his pizza in 4 minutes, yours came in 6, so the *average* is 5 minutes. The joke's on you, buddy!" As the waiter walks away, Hall jumps on his back, they fall to the ground, Hall's teeth chomping on the waiter's left ear . . . camera fades out. Pretty darn funny.

Or try this: Same guy walks in, orders his "5-minute Pizza," and 20 minutes later (just as he gets up to go back to work), the waiter brings his pizza. The guy says to the waiter, "My lunch hour's over so wrap it up to go, and give me my chit for a free one next time." The waiter replies, "No deal, smart mouth, you got your pizza in 5 minutes *from the time we put it in the oven* (big emphasis on that italicized part). The joke's on you, buddy!" Same exit scene, fadeout. Equally funny.

We don't need a legal definition to know that Pizza Hut's promise does not refer to an "average response time." Why don't we? Because from the only point of view that really matters – that of the individual customer – the "average response" time is meaningless. For the same reason, we also know (without needing to be told) that the Pizza Hut clock starts as soon as we've placed our order. What happens between the time we place our order and the time we get our pizza may be of interest to Pizza Hut, but not to us. As customers, we have little interest in learning *why* we didn't get our pizza on time. We just want our pizza.

In the EMS industry, there is no such widespread agreement about what "response time" is or how it should be measured. I think I know why. Public and private EMS providers prefer to be evaluated, not from the perspective of the patients they exist to serve, but from the far more comfortable point of view of their own internal operations. The draft report of a recent survey of the 30 most populous U.S. cities, conducted by the District of Columbia's Productivity Management staff, concluded in part that "There is no standard basis for the calculation of response time . . . (in the EMS industry)."

---

*Jack Stout has been at the forefront of innovations in the design and implementation of EMS systems for the past dozen years. If you have a question, a problem, or a solution related to the public/private interface in prehospital care, address your letter to "Interface," **JEMS**, P.O. Box 1026, Solana Beach, CA 92075.*

The purpose of this "Interface" is to initiate development of just such an industry standard, and to provide an interim standard for use by elected officials, members of the press and those in our industry whose commitment is to serve – not the *average* patient – but *every* patient.

Documenting Times

Every response to an EMS request proceeds in stages from the receipt of the request through delivery of the patient and return of the responding unit(s) to "available" status. Times must be captured and recorded at each of these stages, for two separate purposes: first, to evaluate system performance from the *patient's* point of view; and second, to look within the process to analyze and diagnose the causes of poor response time performance, so that corrections can be made.

To fully serve both purposes, times should be routinely documented for all EMS calls at each of the stages listed on the following chart (in systems without 9-1-1 access, the second listed item, "Time Call Transferred," is ommitted):

Stage of Response	Comments
Time Call Received (by 9-1-1 or other center).	Time-stamp logged the moment callback number, location, and nature of emergency are known, even if additional conversation continues.
Time Call Transferred to Amb./First Responder Dispatchers	If transfer to first responder and ambulance dispatchers is not simultaneous, times for each must be recorded separately. (In well-designed systems, EMS dispatchers continue caller inquiry using priority dispatching protocols and giving pre-arrival instructions *after* this time-stamp has been entered.)
Time First Responder Unit Alert	This is the moment the selected first response crew, if sent, is alerted to respond (e.g., by pager, etc.).
Time Ambulance Alert	This is the moment the selected ALS or BLS ambulance crew is alerted to respond. (In multi-tiered systems, reporting methods should distinguish ALS response times from BLS response times.)
Time First Response Unit En Route	This is the moment the selected first responder vehicle actually starts en route to scene.
Time Ambulance Unit En Route	This is the moment the ambulance vehicle actually starts en route to the scene.
Time First Responder Arrives On Scene	This is the moment the first responder unit arrives upon the emergency scene.
Time Ambulance Arrives On Scene	This is the moment the ambulance arrives upon the emergency scene. (Again, distinguish ALS from BLS in multi-tiered systems.)
Time Ambulance Departs Scene	This is the moment the ambulance with patient on board, departs from scene for hospital.
Time Ambulance Arrives Hospital	Self-explanatory.
Time Ambulance Available at Hospital	This is the moment the ambulance unit, still at the hospital, is ready for posting or assignment to another call.
Time Ambulance Available at Post	This is the moment the ambulance unit arrives at its assigned post location. (In high-productivity EMS systems, the new post may often be a post other than that from which the unit was originally dispatched.)

Defining Response Time Performance

From the patient's point of view, ambulance response time can only be defined as the interval between the moment callback number, location, and chief complaint are first made known to the 9-1-1 center and the moment the first ambulance unit arrives at the scene.

Even though additional information is often gathered by the 9-1-1 complaint-taker or ambulance dispatcher, and while pre-arrival instructions may also be given, the response time clock is properly started as soon as the "system" has acquired enough information to initiate a response.

Ambulance response time is not: the time from unit alert to arrival at the scene; nor the time from 9-1-1 center notification of the EMS provider to the time a unit is dispatched; nor the time from completion of the full telephone inquiry to the time of arrival at the scene; nor the time from call received to the time a lone paramedic arrives in a supervisor's car; neither is the ALS clock stopped by arrival of a BLS unit. Ambulance response time is nothing other than the interval between when the "system" first gained enough information to initiate a response and the time a properly equipped and staffed ambulance arrives at the scene.

The Most Common Response Time Scams

Many EMS providers, public and private, have for years operated without

external monitoring and verification of response time performance. As a result, an arsenal of "scams" has been developed to create the illusion of reliable response time performance even in EMS systems whose actual performance threatens the lives of substantial percentages of patients. Below is a partial listing and brief discussion of the most popular response time reporting scams.

Scam #1: The Average Response Time: The most common and perhaps the most dangerous and misleading response time scam is the reporting of "average" response time performance. By concentrating deployment upon areas experiencing the most frequent requests for service, a sufficient number of very quick times (e.g., from zero to 3 or 4 minutes) can be produced to offset (and cover up) a substantial frequency of life-threatening performance. "Average" response-time statistics tell us what happened to about one-half the patients. But what about the other half?

More scrupulous providers (and less gullible public officials) insist upon evaluations based upon "response time frequency distribution" statistics, rather than upon response time averages. *That is, they insist upon knowing the percentage of calls which result in an acceptable response time performance.* If your city's "average" response time is "acceptable," then by definition, only about half of your citizens are receiving an "acceptable" level of service – the other half are not. (The "median" and "mean" are usually pretty close together in EMS response time distributions.) The following is an example of a paramedic ambulance response-time standard used by several cities in contracting for services:

> The contractor shall produce a *maximum* (not average) response time of 8 minutes from time-call-received (not time-unit-dispatched) to on-scene arrival by a paramedic crew, on not less than 90 percent of all presumptively defined life-threatening emergency requests. Buyer considers this 8 minute maximum/90 percent standard to be the poorest acceptable response time performance for a major urban area – not the goal or ideal.

It is important to understand that no EMS provider can pursue the goal of maximizing the number of calls within a defined response time limitation, *and* simultaneously pursue the goal of the lowest possible average. That is because pursuing the goal of an acceptable level of service to *every* patient served – not just the *average* patient served – requires a totally different and far more complex method of staffing, deployment and constant redeployment than does the pursuit of a superficially impressive overall average performance.

Scam #2: Provider-Friendly Response Time Definition: Next to the "average response-time scam," the use of more flattering methods of calculating response-time results is probably the most common method of covering up bad management and life-threatening performance. (For example, the District of Columbia recently shaved a full three minutes off its response-time "average" with a single flick of the statistics.)

Pizza Hut's "5-minute promise" clearly and inarguably implies that the clock starts when you have ordered your pizza and stops when your pizza arrives. EMS response time is no more of a mystery. The EMS clock starts when you have "ordered your EMS unit" and stops when your unit arrives.

All other time-stamped events are recorded only for purposes of diagnosing and correcting causes of poor response-time performance. For example, you'll never know if long "out-of-chute" times are causing response time delays unless you calculate the times between "time unit alert" and "time unit en route."

Scam #3: Stalling on "Time Call Received." Just a few days ago, I was quietly observing the EMS dispatch center of a major U.S. city when I watched a dispatcher carry on an extended conversation with a caller, deliberately stalling to avoid logging "time call received: until a unit became available. To avoid this temptation, dispatch computer software should be designed to automatically "stamp" time-call-received the moment callback number, location, and nature of problem have been entered, even though additional information may continue to be gathered and entered.

In both manual and computer-assisted dispatch systems, periodic comparisons of time-stamped tape recordings of telephone inquiries with response-time data related to the same calls will eventually catch the cheaters. (Note: Routine daily synchronization of tape-recorder clocks with time-clocks or dispatcher-computer clocks is essential to effective verification.)

Scam #4: Abuse of "Exemption Provisions." Nearly every EMS system incorporates into its standards certain defined circumstances under which the run shall be excluded from reported response-time statistics – i.e., conditions under which poor response-time performance is clearly beyond the provider's reasonable control. In some systems, the list of these exemptions is so long that the net effect is to exclude

nearly every late run (e.g., the night was too dark, the sun too bright, etc.). In our own work, we usually limit such exclusions by contract as follows:

"The contractor shall be exempt from the response-time performance requirements, and from late run penalty assessments only as follows:

A. Runs shall be exempt which occur during periods of severe weather conditions which could reasonably be expected to substantially impair contractor's response-time performance, provided it shall be the contractor's responsibility to document said conditions, the time period affected, and the affected runs, and to apply for approval to an external regulating body not under the provider's control whose decision shall be final.

B. Excess runs shall be exempt which occur during periods of unusual system overload defined for these purposes as a period of time during which more than (insert number) *emergency* calls are simultaneously in progress. Response times to calls *in excess* of that number, and which originate during the same time period, shall be excluded from response time calculations and exempt from late-run penalties. (Note: the number inserted into that blank is established by analysis of peak load demand fluctuations. In general, the provider is responsible for covering the average of the highest peak load demand levels found to occur during each of a series of consecutive 10-week periods.)

C. The response time and late-run penalty requirements of this contract shall be suspended during a declared disaster, locally, or in a neighboring jurisdiction which has requested assistance from the provider.

D. In cases of multiple-response incidents (i.e., where more than one ambulance is sent to the same incident), only the response time of the first arriving ambulance shall be counted for purposes of measuring contractor's response time performance.

E. No other causes of poor response-time performance, such as traffic congestion, vehicle failure, faulty address-match data from the 9-1-1 computer, road construction, blocked railroad crossings, or other such causes, shall be allowed as exemptions to these response time requirements and late run penalty provisions."

Scam #5: Using BLS Units to Stop the ALS Clock. Some multi-tiered systems have deliberately or accidentally designed their data and reporting systems so as to fail to distinguish between the arrival of BLS vs. ALS units dispatched to the same call. While the BLS level is fine for a 1st responder unit, it is no substitute for the timely arrival of a full-fledged ALS capability.

Calculations From Manual Dispatch Cards

Where manual dispatch card systems are still in use, it is usually not possible to capture time-stamps in seconds as well as minutes. For example, a call with an actual start time of 17:00:59 and an actual arrival time of 17:07:00 would be reported at a "7-minute response time," even though the actual response time was 6 minutes, 1 second. In the same system, a call with an actual start time of 17:00:00 and an actual arrival time of 17:07:59 would also be reported as a "7-minute response time," even though the actual response time was only 1 second short of a full eight minutes.

Therefore, where manual time clocks are still in use, response times reported at "7-minute response times" actually include a range of performance between greater than 6 minutes and less than 8 minutes.

But not all of the runs with response times within that range will be reported as "7-minute runs." Some calls with actual response times over 7 minutes but less than eight will be reported as "8-minute runs" because the "7-minute" and "8-minute" categories overlap each other.

And because the "6-minute" and "7-minute" categories also overlap each other, some calls with actual response times over six minutes but less than 7 minutes will be reported as "6-minute runs.

Thus, response time statistics based upon data from manual time clocks are actually a series of overlapping categories. That is the "7-minute response time" category merely refers to a group of calls ranging in actual response time from >6 minutes to <8 minutes, while the "6-minute" category includes response times in the overlapping range between >5 minutes and <7 minutes, and the "8-minute" category includes response times in range between >7 minutes and <9 minutes – a range that also overlaps that of the "7-minute" category, and so on.

Comparisons Of Manual vs. Automated Data: When response-time statistics from EMS systems using manual record-keeping methods are compared with those of automated systems, it is necessary to allow for certain adjustments. The problem is this: a "10-minute maximum/90 percent reliability standard" based upon automated time-stamping is a substantially higher (more stringent) standard than a "10-minute maximum/90 percent reliability standard" based upon a manual-time clock. That is, with times recorded in minutes *and* seconds, all runs with response times reported as "$\leq$10 minutes" will actually have response times of 10 minutes or less; whereas, using times recorded in minutes only, a group of runs reported as "$\leq$10 minutes" will include some runs with actual response times greater than 10 minutes, but less than 11 minutes (because of the overlapping categories discussed above).

A reasonable solution when making such comparisons is to assume that a "10-minute maximum/90 percent reliability" standard measured in minutes and seconds is about the same as a "10½-minute maximum/90 percent reliability" standard measured in minutes only.

Regarding Use of Mutual Aid Providers

Question: Under what conditions should a response by a mutual aid provider from another jurisdiction be allowed to "stop the response time clock?" Another fairly common "scam" involves excluding from the data all runs "handed off" to a mutual-aid provider. Thus, when demand loads peak or an extended response time seems likely, the unscrupulous provider simply dumps the problem on a neighboring provider. Being outside the neighboring provider's own jurisdiction, the run is excluded from his reports as well. Like magic, the late run simply disappears from everyone's statistics.

Contracts developed for our own clients hold the local provider solely responsible for *all* calls originating from within his jurisdiction, regardless of who handled the call. The provider may negotiate and utilize mutual-aid agreements with neighboring providers, and may utilize services furnished by such neighboring paramedic providers toward fulfillment of response-time requirements, provided the following conditions are met:

1. The mutual-aid agreements with the neighboring paramedic providers must be reciprocal and fair to both jurisdictions;

2. Services rendered by the neighboring paramedic provider must be substantially medically equivalent to the level of care required of the local provider, (i.e., trading BLS for ALS is about as "mutual" as bringing a bag of corn chips as your contribution to a Baptist church potluck dinner);

3. The mutual-aid provider and its personnel must agree to cooperate with and participate in medical audits of the

runs in question, if asked;

4. To effect reliable coordination and accurate documentation of response times, provision must be made for direct radio contact between the neighboring paramedic provider's ambulances and the local control center. (Where digital data transmission is normally used to record unit arrival times, voice notification by an approved mutual-aid provider is sufficient for such documentation purposes.)

Equality of Service Among Neighborhoods

In addition to areawide response-time requirements, our clients also incorporate standards requiring reasonable equality of service among the various defined neighborhoods or "districts" of the community. Where local officials represent election districts (e.g., "councilmanic district"), it makes sense to require equality of service among those districts. Elsewhere, dividing the community into 8 or 10 zones based upon demographic differences will achieve the same result. (Note: These "zones" have nothing to do with unit deployment or dispatching method, but are used only to determine whether any neighborhood or area is getting a raw deal.)

In our latest project (Pinellas County, Florida) we have incorporated an additional financial incentive for achieving response time equality among the various areas. Late run deductions from payment (i.e., $10 per minute per call) are retrospectively doubled for runs in districts where the standard of equality hasn't been met.

The point is that, even where two communities experience identical overall response-time performance, the community receiving the most consistently equal response-time performance throughout its various neighborhoods is getting substantially better service, and its EMS provider is working both harder and smarter than the other provider to achieve the result.

Conclusion

It's been more than 15 years since the federal government first started injecting more than 300 million dollars in grant funds to "fix" our EMS systems. Over that period, our entire industry has talked about "response-time performance" as though we had some idea of what the term means. Response time performance is by far the most easily documented and evaluated of all parameters of EMS system performance. The public might well ask, if our industry hasn't yet resolved this most basic of issues, what else have we not done? Let's hope they don't ask. □