SEPTEMBER 1974 TR-08-74

ANALYZING THE PROCESS OF SCREENING CALLS FOR EMERGENCY SERVICE

BY

KEITH A. STEVENSON

AND

THOMAS R. WILLEMAIN

TECHNICAL REPORT

"INNOVATIVE RESOURCE PLANNING IN URBAN PUBLIC SAFETY SYSTEMS"

NATIONAL SCIENCE FOUNDATION GRANT GI38004
RESEARCH APPLIED TO NATIONAL NEEDS
DIVISION OF ADVANCED PRODUCTIVITY, RESEARCH, AND TECHNOLOGY

OPERATIONS RESEARCH CENTER
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS 02139

17619

ANALYZING THE PROCESS OF SCREENING CALLS FOR EMERGENCY SERVICE

BY

KEITH A. STEVENSON

AND

THOMAS R. WILLEMAIN

Technical Report No. 08-74

"Innovative Resource Planning in Urban Public Safety Systems"
National Science Foundation Grant GI38004
Research Applied to National Needs
Division of Advanced Productivity, Research, and Technology

Operations Research Center
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

FOREWORD

The research project, "Innovative Resource Planning in Urban Public Safety Systems," is a multidisciplinary activity, supported by the National Science Foundation, and involving faculty and students from the M.I.T. Schools of Engineering, Architecture and Urban Planning, and Management. The administrative home for the project is the M.I.T. Operations Research Center. The research focuses on three areas: 1) evaluation criteria, 2) analytical tools, and 3) impacts upon traditional methods, standards, roles, and operating procedures. The work reported in this working paper is associated primarily with category 2, in which a set of analytical and simulation models are developed that should be useful as planning, research, and management tools for urban public safety systems in many cities.

The work reported herein was supported by the National Science Foundation under Grant GI38004.

Richard C. Larson
Principal Investigator

ACKNOWLEDGEMENTS

The authors thank Dr. Lawrence Rose and Dr. Geoffrey Gibson for conversations which stimulated our interest in this topic, Dr. Richard Larson for his helpful review of the manuscript, and Ms. Debbie Brooks for her expert technical typing.

ABSTRACT

The increasing demand for urban emergency services raises the possibility that the quality of service provided might be improved by a better matching of resources to needs through a process of screening. Because of the risk of errors on the part of the screener, there is a natural reluctance on the part of those responsible for providing these services to undertake such a program. In this paper we provide a methodology for characterizing the quality of a screening program and establishing the conditions under which the introduction of screening can improve service. Screening is also compared to adding response units as an alternative method for improving service.

While it is probably impossible to determine the actual performance of screeners theoretically, it is possible to analyze mathematically a rather simple process called "categorical screening." We have determined the optimal categorical screening policy under two conditions: "loss screening," in which screened calls receive secondary rather than primary service; and "priority screening," in which screened calls are assigned low priority in any queues that form. The fact that a screening method as crude as categorical screening can improve service suggests that trained personnel should be able to do much better.

Table of Contents

List of Figures

1. INTRODUCTION

Providers of any emergency service must be conscious of the quality of their service, the costs of the services and the morale of their personnel. It is a frequent observation that these three aspects of emergency service are suffering in the face of a general increase in demands of a non-emergent nature being placed on emergency services [1]. Increased demand creates pressure for expansion of services, with corresponding increases in costs; while the allocation of resources to non-emergent problems increases the likelihood that proper service will not be available for true emergencies and leads to complaints by emergency service personnel that their special skills are being wasted. In Boston only 35% of patients transported by the city's emergency ambulance service were classified as emergent in a review of emergency room records [2].

Emergency ambulance attendants, for example, tend to feel resentful toward someone who requests their services for a routine trip to a hospital for a scheduled clinic visit. The result of this resentment may be a barrier of suspicion erected between the emergency service agency and the public it serves, with potentially serious consequences to an individual whose situation is less than obviously emergent. This was demonstrated by a recent incident in Boston in which a patient who had used the emergency service for apparently routine ambulance trips suffered inordinate delays in receiving emergency ambulance service when it was really needed [3].

At the same time, the threshold for invoking emergency aid is set too low among some people, it is also set too high in others, especially for people with medical emergencies. Mogielnicki, et al. [4] found that residents in Cambridge, Mass. with truly emergent problems bypassed the formal emergency ambulance system entirely in getting to the emergency room. In order to

reach out to these people, many communities have implemented the 911 emergency telephone number or at least have established a well-publicized telephone number for centralized emergency dispatch. This is likely to have two results: the improved accessibility stimulates the use of emergency services by those with true emergencies who might otherwise have responded less appropriately, but also increases the number of calls which are not emergent.

Given the two problems of over-use of emergency services by those with non-emergent problems and under-use by those with true emergencies, one potentially useful strategy is both to improve citizen access to the system and to establish some form of screening (or "triage") mechanism. This has been done, for instance, in New York City, where calls for emergency ambulances received through 911 are transferred to nurse-screeners who verify the need for an ambulance (and also provide medical advice and information). Other cities appear to have established various informal screening mechanisms, apparently under pressure from emergency personnel, which focus more on control of perceived non-emergent calls than on proper handling of true emergencies.

The notion that priority should be given to patients most seriously in need of care is appealling so that triage has enjoyed support from emergency medical planners, and although formal triage is rare most emergency rooms employ some sort of informal triage process [5]. The pressure for formalized emergency medical screening prior to the dispatch of an ambulance was increased last year by one of the requirements of the Emergency Medical Services Systems Act which is almost the sole source of federal funding for planning, operating and improving emergency medical services systems. Systems eligible for funding are expected to include the 911 system and to

"utilize emergency medical telephonic screening" (defined to be a "communications system [that] has the capability of redirecting requests for assistance that appear to be non-emergent in nature") [6].

Unfortunately, there are absolutely no guidelines available to those who will be responsible for this screening process which could indicate what the consequences of screening will be in differently configured systems (urban vs. rural; low demand vs. high demand); what cost-effective alternatives to screening might exist; what kinds of screening decision rules should be employed; and how the performance of the screening process can be monitored and evaluated. The need for a clear analysis of this problem is intense because of the legal, political, ethical and emotional problems involved in the denial of service; and because there is a potential conflict in the benefits afforded by screening to the call and to the emergency service agency: the caller would like to recieve service, but the agency would like to keep its non-emergent workload down.

Although this report does not address every aspect of the screening problem, it is a first attempt to outline some of the issues unambiguously, and to analyze the screening process mathematically so as to determine the conditions under which screening of calls can improve emergency service.

Most of the report is formulated in terms of the screening of calls for emergency ambulance service, but the results are general and apply to other services, such as hospital emergency departments or police. We shall analyze two types of screening arrangements. In the first arrangement, which we call "loss screening," we assume the existence of a two-level response system which seeks to reserve its "primary" service for true emergencies and uses its "secondary" or backup service to handle both non-emergency calls and those true emergencies which arise when the primary service is saturated.

In the second arrangement, which we call "priority screening," all calls are responded to by the same vehicles, in the same way, but the order in which calls are answered is such that true emergencies move to the head of any queue for service that develops.

In both arrangements the analysis is similar. With no screening, a true emergency might receive poor service, either because it must be given to the secondary service when all the primary ambulances are busy (in the first case) or because it must wait its turn for service (in the second case). If the problem were caused by the allocation of resources to non-emergency cases, then screening might help avoid these problems. However, screening introduces the possibility of a different problem: misclassification of calls. If a true emergency is wrongly classed as non-emergent, it may receive worse service in a system with screening than in one without. If screening decisions were never wrong, screening would always improve service. Mistakes, however, are inevitable, and our goal is to find the conditions under which--on the average--the errors made in screening are compensated by the reduction it brings about in the demand for service from non-emergencies.

2. LOSS SCREENING

2.1 Performance Measures

Under loss screening, a call for service will be answered by either the primary or the secondary service. By assumption, the primary service is a better trained, better equipped, specialized service appropriate for true emergencies. Thus, the system can be said to have failed the true emergency if the call is directed to the secondary, i.e., if the service is provided by a secondary rather than a primary server. Such an occurence will be called an "under-response" and the probability of under-response should be the most important measure of the efficiency of the service (we assume that differences in response time between primary and secondary are negligible). An under-response occurs either when an emergency call is misclassified as a non-emergency, or when the emergency call is given to a secondary ambulance because all the primaries are already busy. For instance, if a heart attack is handled by an unequipped volunteer ambulance rather than a mobile coronary unit, or if a family dispute is handled by a patrolman rather than a family crisis unit, then a system under-response is said to have occured. We shall assume that there are always sufficient secondary units available to answer all calls directed to the secondary service [7].

While under-response is the more important measure of service quality, another measure of interest to the service provider is over-response, which occurs when the primary service handles a non-emergent case, potentially jeopardizing an impending emergent case and causing some frustration to the operating personnel.

To analyze the probabilities of over- and under-response, define

P_u = probability of under-response to a true emergency.

P_o = probability of over-response to a non-emergency.

P_{av} = probability that primary service is available to handle a case.

P_{fn} = probability that the screener will make a false negative error, i.e., will class a true emergency as a non-emergency.

P_{fp} = probability that the screener will make a false positive error, i.e., will class a non-emergency as a true emergency.

P_e = probability that a given call is a true emergency.

Note that with no screening, we can say $P_{fn} = 0$ and $P_{fp} = 1$. Ideally, with perfect screening $P_{fn} = 0$ and $P_{fp} = 0$, but this will not be the case in practice. See Figure 1 for a representation of the path of a call through the system.

To find the probability of under-response, P_u, note that an under-response occurs either when the screener makes a false negative error or when no mistake is made but the primary service is unavailable to an emergency call. Thus,

$$P_u = P_{fn} + (1 - P_{fn}) \cdot (1 - P_{av}) \quad . \tag{1}$$

Over-response can occur only when the screening officer makes a false positive error while primary service is available. Thus,

$$P_o = P_{fp} \cdot P_{av} \quad . \tag{2}$$

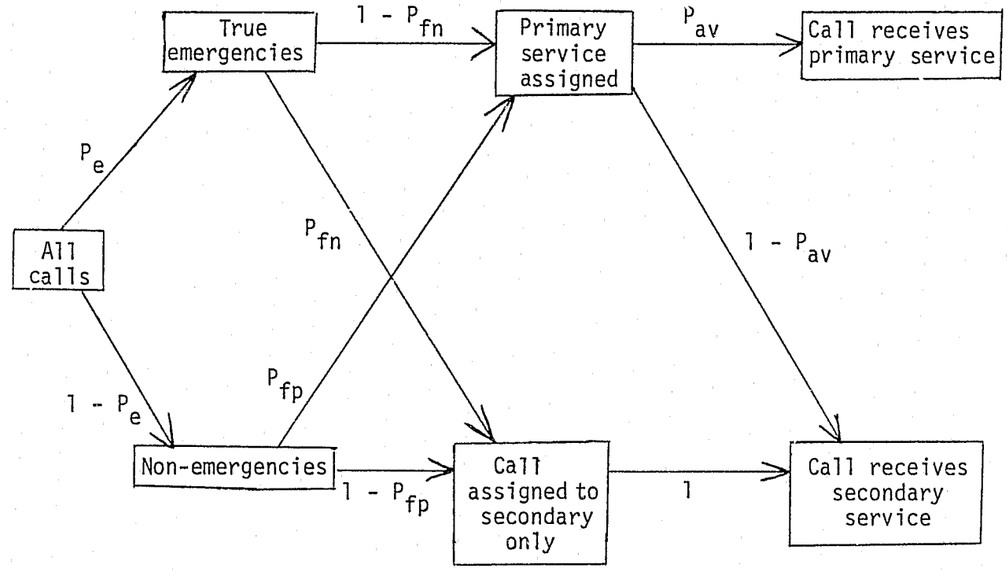


Figure 1: The Screening Process

Through equations (1) and (2) we have characterized the performance of a system in terms of the quality of the screening process (P_{fp},P_{fn}) and the operational readiness of the system (P_{av}).

The probability that primary service is available, P_{av}, can be computed from a knowledge of demand, the screening process, and the size of the service.

Let

N = number of service units (e.g., ambulances) in the primary service.

T = average time required to service one call, in hours.

C = average number of calls received by the screener per hour. (C is assumed to be a constant during a working shift for the units.)

F = fraction of calls which the screener attempts to assign to the primary service.

$$F = P_e(1 - P_{fn}) + (1 - P_e)P_{fp} \quad . \tag{3}$$

If we made the reasonable assumptions that calls arrive at random (in a Poisson manner) and that successive service times are independent with finite mean then reference to any good text on queuing theory will show that

$$P_{av} = 1 - \frac{(CTF)^N/N!}{\sum_{k=0}^{N}(CTF)^k/k!} \quad . \tag{4}$$

Thus, using equations 1 to 4, we can compute the probabilities of under- and over-response for any screening program, as characterized by P_{fn} and P_{fp}.

2.2 Analysis of the Single Server Case (N = 1)

Analysis of the conditions on the false negative and false positive error rates is algebraically tedious for a primary service with more than one service unit (or "server"), but relatively easy for the use of a single-server primary. For this case, equation (4) specializes to

$$P_{av} = \frac{1}{1 + CTF} \quad . \tag{5}$$

Thus, recalling equations (1), (2) and (3)

$$P_u = P_{fn} + (1 - P_{fn}) \cdot \frac{CT[P_e(1 - P_{fn}) + (1 - P_e)P_{fp}]}{1 + CT[P_e(1 - P_{fn}) + (1 - P_e)P_{fp}]}$$

$$P_o = P_{fp} \cdot \frac{1}{1 + CT[P_e(1 - P_{fn}) + (1 - P_e)P_{fp}]} \quad .$$

Manipulating these equations, we derive the following linear conditions on the screening parameters. For the introduction of screening to reduce under-response compared to the situation with no screening, we require

$$P_u < \frac{1}{1 + CT} \quad ,$$

or

$$\left[\frac{1 + CT(1 - P_e)}{CT(1 - P_e)}\right] P_{fn} + P_{fp} < 1 \quad . \tag{6}$$

For screening to reduce over-response we require

$$P_o < 1 - \frac{1}{1 + CT} \quad ,$$

or

$$\left[\frac{CTP_e}{1 + CTP_e}\right] P_{fn} + P_{fp} < 1 \quad . \tag{7}$$

These two relationships are shown in Figure 2. Note that there are three regions, corresponding to three classes of screening programs. Those in the upper right of Figure 2 are characterized by such high error rates that no screening is preferable on both counts. Those in the lower left have error rates sufficiently low that both under-response and over-response are

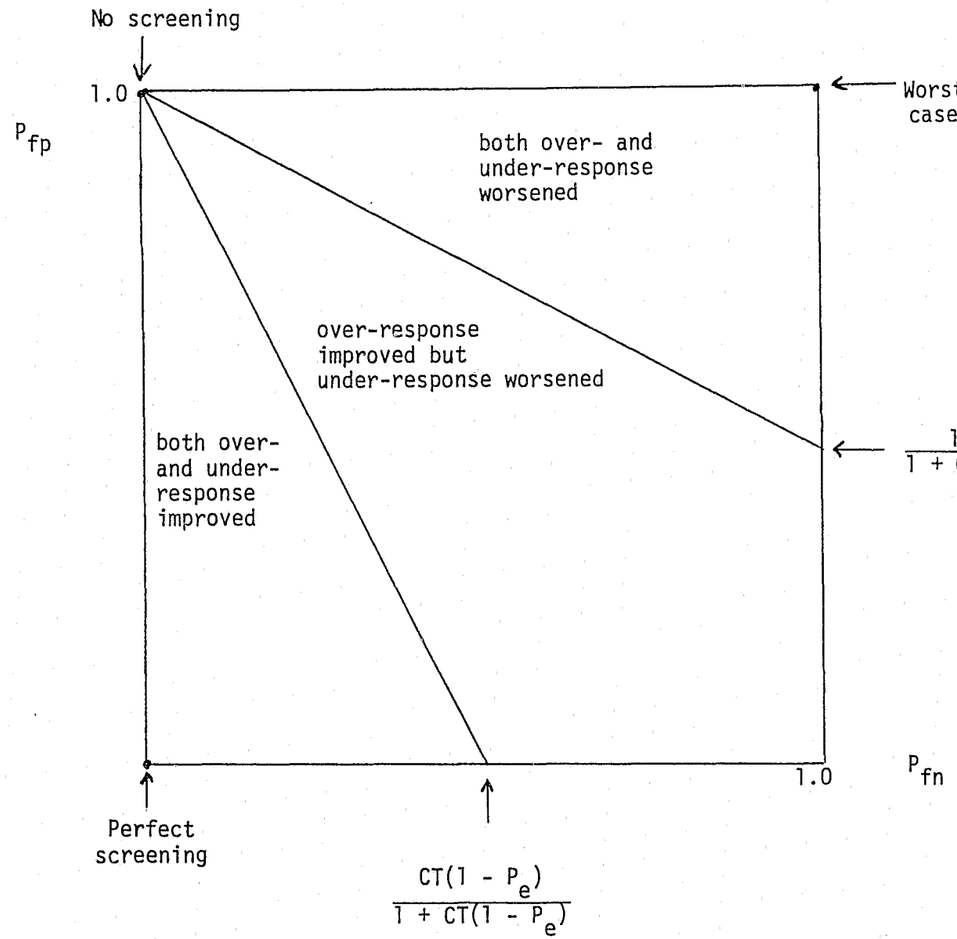


Figure 2: Single-server Case:
Impact of P_{fp} and P_{fn} on Over- and Under-response

improved by screening. Those in the middle region "crack down" sufficiently against "abusers" so that over-response is improved, but deny primary service to so many true emergencies that under-response is worsened by screening. This middle region is intriguing, for in it an agency can satisfy its internal needs by protecting itself from spurious calls, while actually providing poorer service to the public, as measured by the more important yardstick of under-response probability. Thus, the commonly held view that a screening program to weed out non-emergency cases will benefit the truly emergency cases is not necessarily true. In contrast, however, Figure 2 illustrates that all screening programs designed to lessen the probability of under-response will also lessen the probability of over-response. When the public benefits, the service provider benefits, but not necessarily vice-versa.

The degree of error tolerable in the screening process is reflected in the range of values of P_{fn} and P_{fp} contained within the triangular area in the lower left of Figure 2. This area grows with the quantity $CT(1 - P_e)$, which represents the non-emergent workload of the system. Thus, screening becomes a more plausible strategy when the number of calls per hour, C, increases, when the time to service a call, T, increases, and when the proportion of all calls which are truly emergent, P_e, decreases.

2.3 Analysis of the Multi-Server Case ($N > 1$)

As noted earlier, the multi-server case is algebraically complex and must be solved numerically. One general statement can be made: the criteria for successful screening (i.e., for reduction in P_u) are more stringent for a larger system than for a smaller system with the same potential demand per server, CT/N. This is illustrated in Figure 3, which compares a busy system with one primary server to a ten server system facing a proportionately equal

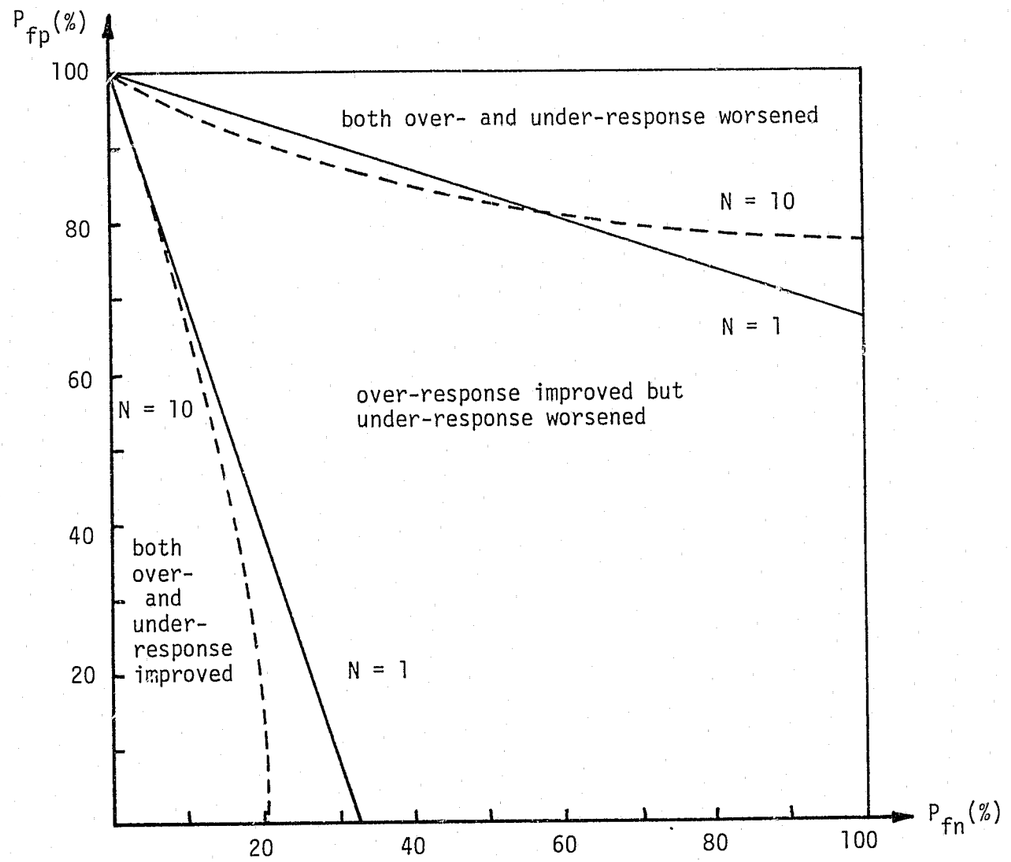
demand. In the smaller system, for instance, a screening process characterized by $P_{fn} = 0.2$, $P_{fp} = 0.2$ would reduce the likelihood of under-response to a true emergency, while the same screening process in the larger system would increase under-response.

2.4 Screening as an Alternative to Larger Systems

When demand for emergency service increases, providers of the service can respond either by attempting to control supply or to control demand, or both. A supply-oriented response is to increase the size of the primary service (i.e., increase N). A demand-oriented response is to screen out non-emergent demand. The same methods used above to compare screening against no screening can be used to find the number of primary servers needed with and without screening to provide the same level of under-response.

An illustrative example is shown in Figure 4, which lists for various pairs of screening parameters the number of primary servers required to match the performance of a very busy ten-server system without screening. Those screening programs characterized by low error rates permit the same level of service to be provided by fewer servers. For instance, the performance of the ten server system without screening can be matched by an eight server system coupled with a screening program characterized by $P_{fn} = 0.05$, $P_{fp} = 0.40$.

Note, however, that as the quality of screening programs decreases, so does the potential saving in the number of servers, until the screening process introduces so many errors that more than ten servers are required to compensate for the errors made in screening. The program characterized by $P_{fn} = 0.15$, $P_{fp} = 0.60$ requires eleven servers, for instance. When the rate of false negative errors is sufficiently high, it becomes impossible to match



Note: Example constructed for: $P_e = 0.50$

$\frac{CT}{N} = 1.00$

N = Number of Primary Units.

Figure 3: Multi-Server Case: Impact of P_{fp} and P_{fn} on Over- and Under-Response

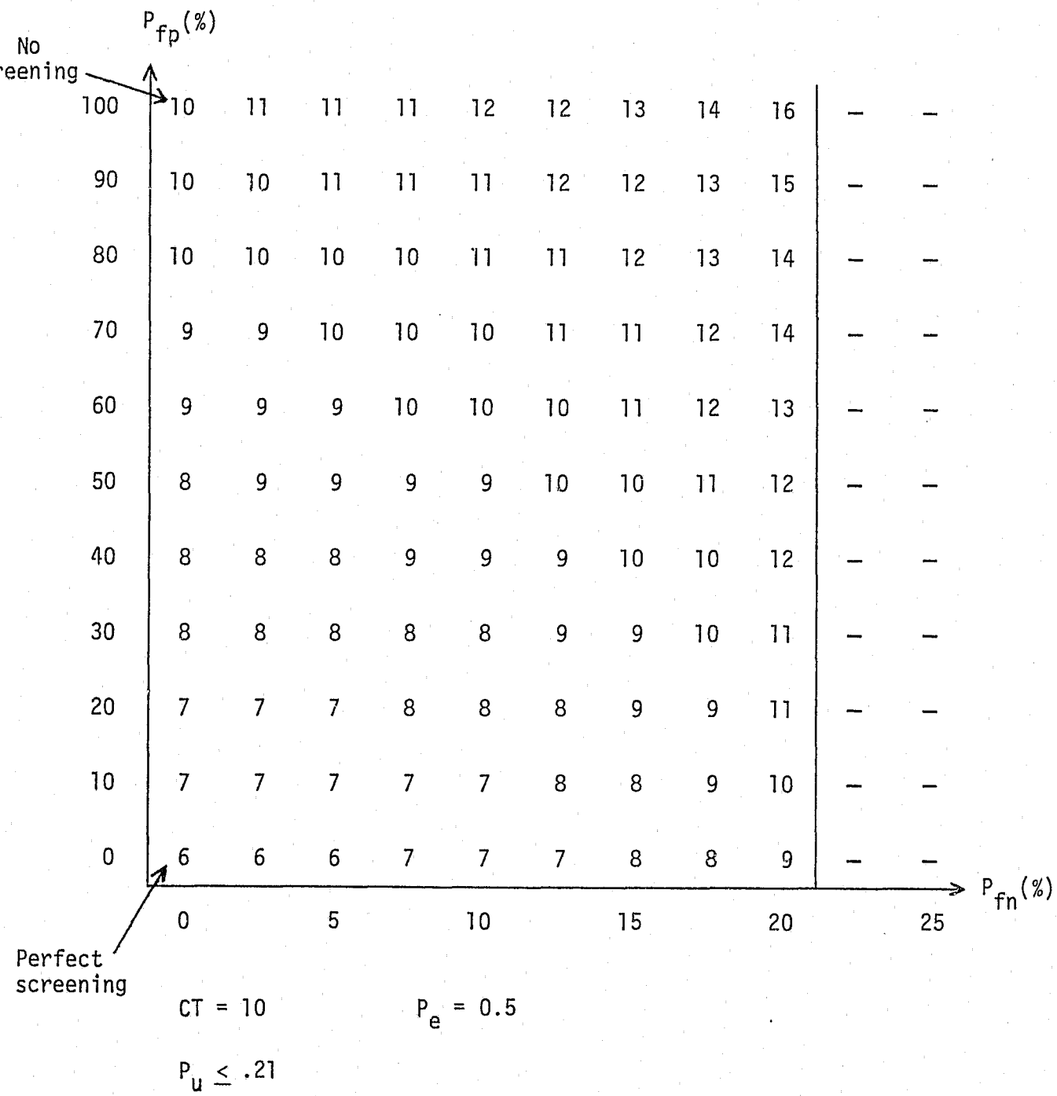


Figure 4: Effect of Screening Errors on the Number (N) of Units Needed to Match the Performance of a Ten-Server System without Screening

the ten server system without screening, no matter how many servers are used with screening. In the example shown in Figure 4, this occurs for $P_{fn} > 0.20$. It is important to observe that the ability of any given screening program to substitute for expansion of the primary server depends on how busy the service is and on how many calls are truly emergent. The results in Figure 4 pertain to an extremely heavily loaded system (CT/N = 1.0). Should the loading on that system be reduced by half (CT/N = 0.5), it happens that only screening programs with virtually no false negative errors lead to savings in the number of servers. The results for any particular case can be obtained using equations (1), (3) and (4).

2.5 A Useful Reformulation

Conceptually, the preceding analysis is complete. However, the formulation in terms of false positive and false negative error rates is not convenient for implementation, monitoring and evaluation since information on false negative cases can only be had by reviewing cases served by the secondary service. It may often happen that different agencies provide the primary and secondary service, and monitoring the quality of the screening process may be inordinately difficult owing to the different procedures and record-keeping policies of the providers. We now present a reformulation of the analysis from a more practical point of view for the primary service.

Define two new parameters to characterize the screening process:

P_{ok} = probability that a given call will initially be assigned ("ok'd") by the screener to receive primary service. This probability is distinct from the probability that primary service is actually given, since receipt of service depends on the availability of a server.

P_{eg} = conditional probability that a particular case is truly emergent, given that it actually gets primary service. We would expect that under a careful screening process $P_{eg} > P_e$; and we assume that primary server unavailability is independent of the true need of any patient whom the screener attempts initially to assign to the primary service.

Use of these parameters presents two advantages over use of false negative and false positive error rates. First, P_{eg} can be determined by retrospective review of only those cases handled directly by the primary service (e.g., review by physicians of hospital emergency room records); this should facilitate identification of cases and location of records. Second, P_{ok} can be easily determined from a simple log of calls handled by the screener and has a direct operational interpretation--P_{ok} represents the "degree" of screening, a value near 1.0 indicating that few calls are deliberately deflected to the secondary service.

Each pair (P_{ok},P_{eg}) corresponds uniquely to a pair (P_{fn},P_{fp}). Using the laws of conditional probability we can write

$$P_{fn} = \frac{P_e - P_{eg}P_{ok}}{P_e} , \qquad (8)$$

and

$$P_{fp} = \frac{P_{ok}(1 - P_{eg})}{1 - P_e} , \qquad (9)$$

so that

$$F = P_e(1 - P_{fn}) + (1 - P_e)P_{fp} = P_{ok} . \qquad (10)$$

Thus, knowing the characteristics of a screening system in terms of P_{ok} and P_{eg}, we can convert to the earlier characterization and compute the probabilities of under- and over-response using equations (1) and (2).

As before, analysis is impractical for multi-server systems but rather easy for a single primary server. The conditions under which screening will improve under- and over-response, analogous to equations (6) and (7) are now, respectively,

$$(1 + CT)P_{ok}P_{eg} - (CTP_e)P_{ok} - P_e > 0 \qquad (11)$$

and

$$(1 + CT)P_{ok}P_{eg} - (1 + CTP_e)P_{ok} + (1 - P_e) > 0 . \qquad (12)$$

Similarly, we can characterize the performance of the various pairs (P_{ok},P_{eg}) as was done previously in Figure 2; this is shown in Figure 5. The relatively narrow crescent in the upper portion of the figure represents those screening programs which improve service. (Note that because of the nature of the variable transformations in equations (8) and (9), Figure 5 contains two regions which are physically impossible to reach.) Numerical calculations indicate that the crescent characterizing successful screening programs shrinks considerably in size as N, the number of primary servers, increases at the same value of CT.

A word is in order about estimation of the parameters. Unlike P_{ok} and P_{eg}, which can be estimated from data readily available to the primary server, P_e requires a search for data throughout the secondary service as well. In practice, calls deflected to the secondary service will probably be relatively expensive to investigate; thus, whereas P_{ok} can be logged continuously and P_{eg} might be obtained through frequent sample surveys, estimation of P_e may have to be done on the basis of an infrequent sample. Of course, experienced personnel might subjectively estimate P_e, but any fixation on their part with "abusers" could lead to underestimates of P_e, which in turn would result

in inappropriately lax criteria for judgement of a screening program, since the crescent in Figure 5 (and the lower left region in Figure 2) grows as P_e decreases. The parameter P_{av} could be simply logged by the screener.

Thus for each incoming call the screener would record his decision about the nature of the call (emergency/non-emergency) and would record the state of the primary system (available/unavailable). These data would be used to estimate P_{ok} and P_{av} respectively. Retrospective reviews of those cases actually handled by the primary service would provide the estimate of P_{eg}. Finally, a retrospective sample survey of all calls would estimate P_e.

Note: Example constructed for: $P_e = 0.20$
$CT = 1.00$

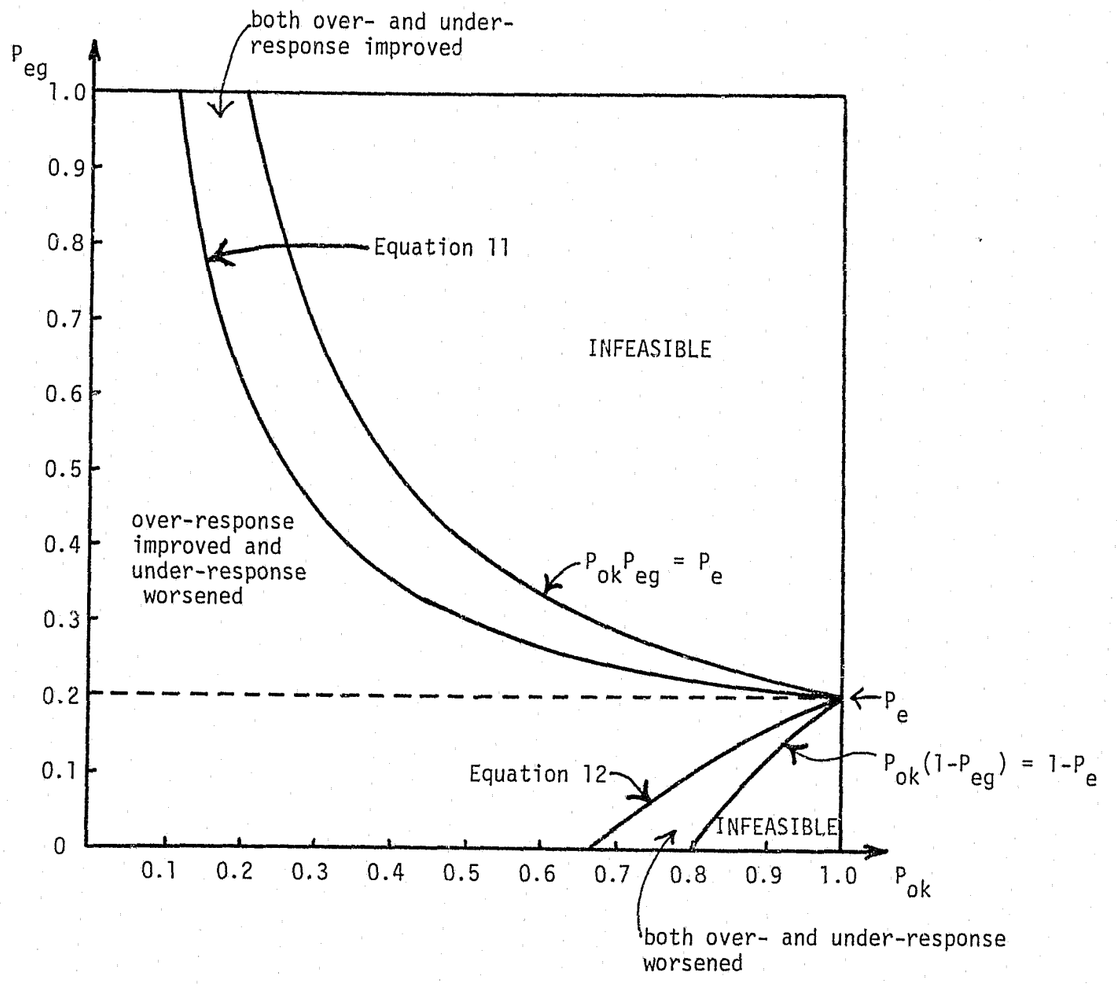


Figure 5: Single-Server Case: Impact of P_{ok} and P_{eg} on Over- and Under-Response

3. PRIORITY SCREENING

3.1 Performance Measures

Priority screening is descriptive of a rather different type of emergency service. Now we envision a service lacking a secondary level--all calls are eventually taken care of by the primary servers. Whereas in the loss system, any call arriving when the primary service was saturated was "lost" to the secondary service and never permitted to queue, now we permit a queue to form should calls arrive when the primary is unavailable. In this case there is no question about an improper match between the needs of the caller and the type of server responding to the call, so over- and under-response are not appropriate measures. Instead, it is natural to be concerned about the delays encountered by truly emergent cases.

Poor service would be provided if truly emergent cases were forced to wait for non-emergent cases to be serviced ahead of them. If calls are responded to strictly in the order in which they arrive, all cases encounter the same average delay regardless of priority. One could decrease the average delay for truly emergent cases by identifying them and placing them at the head of the queue, trading more prompt service for true emergencies against longer delays for non-emergencies. However, if one misclassifies an emergency as a non-emergency, then his low priority status will produce a longer average delay then he would have encountered without screening. One appropriate measure of the value of screening calls into high and low priority is the ratio of the average delay encountered by a true emergency under the screening program to that encountered without screening. Letting D be this delay without screening and D_s be that with screening, define the ratio

$$r = \frac{D_s}{D} . \tag{13}$$

Successful screening will have $r < 1$.

3.2 Analysis of the Single Server Case (N = 1)

To analyze the single server case, we make use of well known results for priority queues [8]. We must assume only that $CT < 1$, that calls arrive in a random (Poisson) manner, that successive service times are independent with finite mean, and that high priority calls are served ahead of low priority calls but never interrupt a low priority service once it has begun. For simplicity, we will further assume that all calls take the same time to be serviced on average. As before we let P_{ok} represent the probability that a given call will be cleared for high priority service, and P_{eg} represent the probability that a call assigned high priority by the screener is truly emergent.

Without screening, the average delay is

$$D = \frac{k}{1 - CT}$$

where k is a constant. All high priority calls endure an average delay of

$$D_H = \frac{k}{1 - CTP_{ok}} \tag{14}$$

and all low priority calls a delay,

$$D_L = \frac{k}{(1 - CTP_{ok})(1 - CT)} . \tag{15}$$

Let Q be the probability a given emergent call is classed as high priority. Then the average delay for a true emergency under screening is

$$D_s = D_H \cdot Q + D_L(1 - Q) . \tag{16}$$

Using the laws of conditional probability we can write

$$Q = \frac{P_{eg} \cdot P_{ok}}{P_e} . \tag{17}$$

Thus, we can write, using equations (14) through (17),

$$D_s = \frac{k}{1 - CTP_{ok}} \cdot \frac{P_{eg}P_{ok}}{P_e} + \frac{k}{(1 - CTP_{ok})(1 - CT)}\left(1 - \frac{P_{eg} \cdot P_{ok}}{P_e}\right) . \qquad (18)$$

Finally, we divide by D to find the ratio of delays

$$r = \frac{1 - (\frac{P_{eg}}{P_e})CTP_{ok}}{1 - CTP_{ok}} . \qquad (19)$$

The ratio r is less than unity (i.e., screening improves service) for any value of P_e such that

$$P_{eg} > P_e .$$

In addition, realizable systems require the further condition that

$$P_{ok}P_{eg} \leq P_e .$$

The ratio, r, is minimized when $P_{ok} = P_e$, $P_{eg} = 1$, i.e., when all true emergencies are properly classified. At best, the ratio drops to

$$r_{min} = \frac{1 - CT}{1 - CTP_e}$$

indicating that, as in loss screening, priority screening is most worthwhile when C and T are large and P_e is small.

3.3 Analysis of the Multi-Server Case (N > 1)

For priority screening, we can analyze the multiserver case simply [8], provided we make one further assumption, that the length of each service time is exponentially distributed and independent of other service times.

We now only require that CT < N. Since the equations for the multiserver priority queue are nearly identical to those for the single server, differing only in the constant k, it follows that the ratio of delays has nearly the same form as equation (19)

$$r_N = \frac{1 - \frac{P_{eg}}{P_e} \cdot \frac{CT}{N} \cdot P_{ok}}{1 - \frac{CT}{N} \cdot P_{ok}} . \qquad (20)$$

The same remarks hold true for equation (20) as were made for equation (19). In particular, any pair (P_{ok},P_{eg}) having $P_{eg} > P_e$ will lead to some improvement in delay for true emergency cases, and the greatest improvement will occur at $P_{ok} = P_e$, $P_{eg} = 1$, with

$$r_N(min) = \frac{1 - CT/N}{1 - CTP_e/N} .$$

4. ANALYSIS OF A SIMPLE SCREENING RULE

4.1 Categorical Screening

To this point we have structured the screening problem, noted the general conditions favorable to screening, characterized the screening process by the parameters P_{ok} and P_{eg}, and provided in equations (11), (19) and (20) a way to test whether any particular combination of P_{ok} and P_{eg} is an improvement over no screening at all. We have not addressed the problem of predicting the values of P_{ok} and P_{eg}. The parameter P_{ok} is fairly easy to adjust in practice, but we know little about the value of P_{eg} that will be paired with any particular value of P_{ok}. Depending on the skill of the screener it could in theory range from zero to one, although we would expect to see $P_{eg} > P_e$; i.e., the screener should improve over a "random" screening policy which arbitrarily selects a fraction P_e of the calls and assigns them high priority without regard to their nature.

It does not seem possible to obtain further information about the combinations of P_{ok} and P_{eg} characterizing an actual screening process without resort to experiment. It is possible, however, to analyze one particularly simple and inexpensive screening rule which we will call "categorical screening."

Categorical screening simply classifies each call for service as one of a number of categories of calls and makes the same yes/no decision about all calls in that category. For instance, calls for emergency ambulance service may be classified by chief complaint, with certain chief complaints always given high priority, irrespective of the individual details of any particular case. Thus all calls mentioning chest pain might be assigned to primary service and all calls mentioning abdominal pain might be deflected to secondary service.

Although a screening officer might use more information and more sophistication in making his decisions, it is worthwhile to study categorical screening for three reasons. First, it provides detailed insight into the functional relationships between P_{ok} and P_{eg}. Second, it probably provides a lower bound on the actual performance of a screener. Third, it is a simple and inexpensive algorithmic approach to screening which would warrant implementation in its own right, provided it can be shown to improve on no screening at all.

4.2 Analysis of Categorical Screening

Let all calls for service be divided into I mutually exclusive and collectively exhaustive categories. For each category i define

f_i = fraction of all calls which are type i.

e_i = fraction of all type i calls which are truly emergent.

$$x_i = \begin{cases} 1 \text{ if type i calls are given high priority, or are initially assigned to the primary service,} \\ 0 \text{ if type i calls are given low priority, or are assigned to the secondary service.} \end{cases}$$

The set of numbers $\{f_i\}$ and $\{e_i\}$ might be estimated from review of past calls for service. The set of numbers $\{x_i\}$ are the screening decision variables and are fixed on the basis of the $\{f_i\}$ and $\{e_i\}$ in categorical screening.

Using the definitions of f_i, e_i, P_{ok} and P_{eg} it follows that

$$P_{ok} = \sum_{i=1}^{I} f_i x_i \,, \tag{21}$$

and

$$P_{eg} = \sum_{i=1}^{I} f_i e_i x_i / \sum_{i=1}^{I} f_i x_i \quad . \tag{22}$$

Now we can express the probability of under-response in a single-server loss screening system as

$$P_u = 1 - \frac{P_{eg}P_{ok}}{P_e(1 + CTP_{ok})} \qquad \text{(using eq's (1),(5),(8) and (9))}$$

$$= 1 - \frac{\sum_{i=1}^{I} f_i e_i x_i}{P_e(1 + CT\sum_{i=1}^{I} f_i x_i)}$$

$$= \frac{1 + \sum_{i=1}^{I}(CTf_i - \frac{f_i e_i}{P_e})x_i}{1 + \sum_{i=1}^{I} CTf_i x_i} \qquad (23)$$

Likewise, in an N-server priority screening system, the ratio of delays is

$$= \frac{1 - \frac{P_{eg}}{P_e} \cdot \frac{CT}{N} \cdot P_{ok}}{1 - \frac{CT}{N} P_{ok}} \qquad (20)$$

$$r_N = \frac{1 - \frac{CT}{P_e N}\sum_{i=1}^{I} f_i e_i x_i}{1 - \frac{CT}{N}\sum_{i=1}^{I} f_i x_i} \qquad (24)$$

The optimal categorical screening policy will be that choice of $\{x_i\}$ which minimizes either equation (23) or equation (24). This is a nonlinear integer programming problem, but one with a structure that can be exploited to produce an optimum categorical screening policy with little effort, as seen in the next section.

4.3 An Algorithm for Categorical Screening

4.3.1 Single-Server Loss Screening

The following section will be easier to follow if we define the two sets R and S so that

$$R = \{i : x_i = 1\}, \text{ and}$$

$$S = R^c = \{i : x_i = 0\};$$

and we rewrite equation 23 in terms of R:

$$P_u = \frac{1 + \sum_{i\in R} f_i(CT - {}^{e_i}/_{P_e})}{1 + \sum_{i\in R} CTf_i} \cdot \qquad (23a)$$

Now P_u can be reduced, or at least not increased by adding the category $j\varepsilon S$ to the set R if

$$P_u' - P_u \leq 0 \,, \qquad (25)$$

where P_u' would be the new probability of under-response if category j were assigned to the primary service.

Let $P_u = \frac{M}{N}$, and rewrite (25) as

$$\frac{M + f_j(CT - {}^{e_j}/_{P_e})}{N + CTf_j} - \frac{M}{N} \leq 0$$

$$\text{i.e., } MN + Nf_j(CT - {}^{e_j}/_{P_e}) \leq MN + MCTf_j$$

which implies $e_j \geq CTP_e(\frac{N - M}{N})$

$$\text{i.e.,}\quad e_j \geq CTP_e(1 - P_u) \tag{26}$$

$$\text{i.e.,}\quad e_j \geq \frac{CT\sum_{i\in R} f_i e_i}{1 + CT\sum_{i\in R} f_i} \tag{27}$$

Note:

(i) It is easily shown that if (27) holds, then

$$e_j \geq \frac{CT[\sum_{i\in R} f_i e_i + f_j e_j]}{1 + CT[\sum_{i\in R} f_i + f_j]}, \tag{28}$$

so that it follows that if more than one category has the same fraction of truly emergent patients, and that value of e_j satisfies (27), then all of these categories should be added to R.

(ii) From equation (26), any category j such that $e_j \geq CTP_e$ is automatically a member of R.

Exploiting the above results we can contruct a simple algorithm for minimizing P_u using categorical screening which requires no more, and usually less (because of equation 26) than I iterations.

Algorithm

(1) Arrange the I categories in descending order of e_i so that $e_1 \geq e_2 \geq e_3 \geq \ldots \geq e_I$.

(2) Include in R, the set of categories which receive primary service, all i such that $e_i \geq CTP_e$.

(3) Assuming that (2) above results in the inclusion of categories 1

through s, add category s+j to R (starting from s+1) if

$$e_{s+j} \geq \frac{CT\sum_{i\in R} f_i e_i}{1 + CT\sum_{i\in R} f_i} \tag{29}$$

where $j = 1,2,\ldots,I-s$

and $R = \{i : i = 1,2,\ldots,s+j-1\}$.

(4) If $s+k \in S$ is the first category for which (29) does not hold, then the optimal categorical screening policy is realized with the set

$$R^* = \{i : i = 1,2,\ldots,s+k-1\}$$

and results in an under-response probability of P_u^*,

where $$P_u^* = \frac{1 + \sum_{i=1}^{s+k-1} f_i(CT - e_i/P_e)}{1 + CT\sum_{i=1}^{s+k-1} f_i} .$$

In order to show that P_u^* is optimal, we will demonstrate that P_u^* cannot be reduced by adding another category to R*, or deleting a category from R*, or exchanging categories between R* and its complement.

(1) Addition: All categories, m, excluded from R* are such that

$$e_m \leq e_{s+k} < \frac{CT\sum_{i\in R^*} f_i e_i}{1 + CT\sum_{i\in R^*} f_i} ;$$

and therefore, adding m to R* would increase P_u^* .

(2) Deletion: All categories, n, included in R* are such that

$$e_n \geq e_{s+k-1} \geq \frac{CT\sum_{i\in R^*} f_i e_i}{1 + CT\sum_{i\in R^*} f_i} \quad ;$$

and therefore deleting n from R* would not decrease P_u^* .

(3) Exchange: In order to decrease P_u^* by creating a new set R which includes category $m \notin R^*$ and excludes category $n\in R^*$, we require

$$e_n < \frac{CT[\sum_{i\in R^*} f_i e_i - f_n e_n + f_m e_m]}{1 + CT[\sum_{i\in R^*} f_i - f_n + f_m]} \quad . \tag{30}$$

Equation (30) is simply the condition that category n cannot profitably be introduced into the new set R. We can easily show that for (30) to hold implies a contradiction.

Note that (i) $$e_n > e_m \quad ; \tag{31}$$

(ii) $$e_n \geq \frac{CT\sum_{i\in R^*} f_i e_i}{1 + CT\sum_{i\in R^*} f_i} \quad ,$$

or $$e_n + e_n CT\sum_{i\in R^*} f_i - CT\sum_{i\in R^*} f_i e_i \geq 0 \quad . \tag{32}$$

But (30) implies

$$CTf_m(e_m - e_n) > e_n + e_n CT\sum_{i\in R^*} f_i - CT\sum_{i\in R^*} f_i e_i \geq 0 \qquad \text{(by 32)}$$

i.e., $e_m \geq e_n$, a contradiction of (31).

It now follows from equations (27) and (28) that no screening will reduce the probability of under-response if $e_I \geq \frac{CTP_e}{1 + CT}$.

4.3.2 Multi-Server Priority Screening

Using equation (24) it is straightforward to derive an identical algorithm that will yield the optimal categorical screening policy for an N-server priority system. By analogy with equation (27), categories to receive priority service satisfy the condition

$$e_j \geq P_e\left[\frac{1 - {}^{CT}/_{P_e N}\sum_{i\in R} f_i e_i}{1 - {}^{CT}/_{N}\sum_{i\in R} f_i}\right] \tag{33}$$

where $R = \{i : i = 1,2,\ldots,j-1\}$.

Obviously any category j such that $e_j \geq P_e$ is automatically included in R.

4.3.3 Multi-Server Loss Screening: The Continuous Analog

Unfortunately, the companion condition to (33) for more than one primary server in a categorical loss screening system is not so easily derived. An instructive and useful approach is to consider the solution to the continuous analog of the discrete categorical screening problem. Imagine that the number of categories I increases without limit, so that the sums in equations (23) and (24) become integrals, and the discrete functions f_i and e_i become the continuous functions f(y) and e(y) over the interval (0,1]. Arrange the categories so that e(y) is a monotonic non-increasing function, as called for in the discrete optimization algorithm. Let s be the screening threshold, with

$$x(s) = 1 \qquad 0 < y \leq s \quad ,$$

$$x(s) = 0 \qquad s < y \leq 1,$$

i.e., all categories with $e(y) \geq e(s)$ receive priority service.

Now the optimization problem for the single-server loss system becomes

$$\min_{0<s\leq 1} P_u(s) = \frac{1 + \int_0^1 [CTf(y) - \frac{f(y)e(y)}{P_e}]x(y)dy}{1 + \int_0^1 CTf(y)x(y)dy}$$

$$= \frac{1 + \int_0^s [CTf(y) - \frac{f(y)e(y)}{P_e}]dy}{1 + \int_0^s CTf(y)dy} .$$

Differentiating with respect to s and setting the derivative equal to zero, one finds the condition on s for minimum under-response

$$e(s) = \frac{CT \int_0^s f(y)e(y)dy}{1 + CT \int_0^s f(y)dy} . \qquad (34)$$

which is, of course, the continuous analog of equation (27).

Now although the solution is tedious, it is also possible to solve the continuous analog to the multi-server categorical loss screening problem. We need to find the value of s such that the probability of under-response is minimized, i.e.,

$$\min_{0<s\leq 1} P_u(s) = 1 - \frac{1}{P_e} \int_0^s f(y)e(y) \left\{ 1 - \frac{[CT \int_0^s f(y)d_y]^N / N!}{\sum_{k=0}^{N} \frac{[CT \int_0^s f(y)d_y]}{k!}} \right\} dy$$

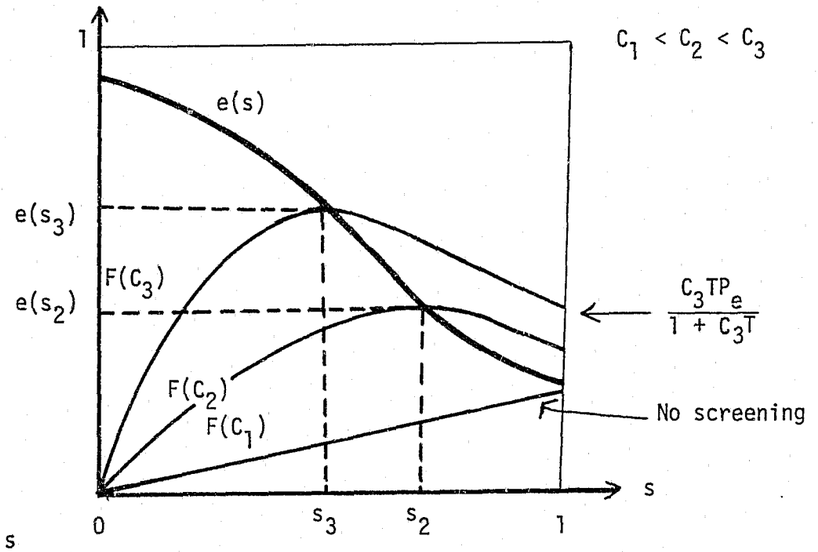
Differentiating with respect to s and setting the derivative equal to zero, we find the condition for optimal screening is that s satisfies

$$e(s) = [CT \int_0^s f(y)e(y)dy][P_{av}(s,N) - P_{av}(s,N-1)] \qquad (35)$$

where $$P_{av}(s,N) = 1 - \frac{\frac{(CT)^N}{N!}[\int_0^s f(y)dy]^N}{\sum_{k=0}^{N} \frac{(CT)^k}{k!}[\int_0^s f(y)dy]^k}$$

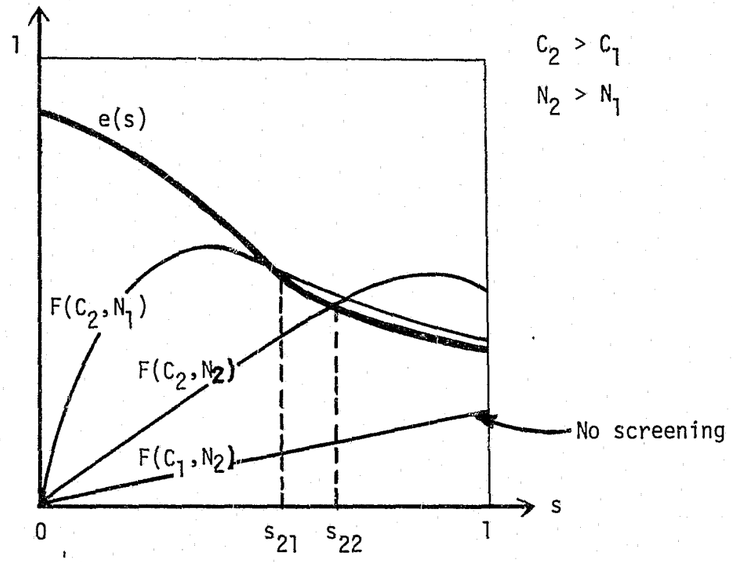
Equation (35) reduces to equation (34) when N = 1. Some basic properties of equation (34) are illustrated qualitatively in Figure 6(a). The left-hand side is a monotone non-increasing function of s by algorithmic construction, while the right-hand side is unimodel with the maximum occurring at the solution to equation (34), provided one exists. The maximum occurs at decreasing values of s for increasing values of CT, implying that as the demand for service increases, the screening process becomes more selective, assigning fewer categories to the primary service. Notice that if CT falls sufficiently, screening is abandoned.

Unfortunately, with N > 1, equation (35) does not have the same properties as equation (34). While the right-hand side of (35) is unimodel, the solution to (35) does not coincide with the maximum. In general, the solution will occur at larger values of s as the size of the primary service increases (at constant CT/N), indicating that fewer calls need to be screened out in larger systems to improve under-response. Similarly, the solution will occur at larger values of s as the call rate, C, decreases. These properties are illustrated qualitatively in Figure 6(b). Although we cannot prove the result, we conjecture by analogy that for the discrete multiserver categorical screening screening problem the condition for



$$F(C) = \frac{CT \int_0^s f(y)e(y)dy}{1 + CT \int_0^s f(y)dy}$$

Figure 6(a): Continuous Analog for Single-Server Categorical Loss Screening with Three Different Call Rates



$$F(C,N) = [CT \int_0^s f(y)e(y)dy][P_{av}(s,N) - P_{av}(s,N-1)]$$

Figure 6(b): Continuous Analog of Multi-Server Categorical Loss Screening with Different Call Rates and Numbers of Units

including category j in the set R is that

$$e_j \geq CT \sum_{i \in R} f_i e_i [P_{av}(s,N) - P_{av}(s,N-1)] \qquad (36)$$

where
$$P_{av}(s,N) = 1 - \frac{(CT)^N/N! [\sum_{i \in R} f_i]^N}{\sum_{k=0}^{N} \frac{(CT)^k}{k!} [\sum_{i \in R} f_i]^k}$$

and
$$R = \{i : i = 1,2,\ldots,j-1\}.$$

Using this condition the algorithm in section 4.3.1 can be used to find the optimal policy. As before, all $e_j \geq CTP_e$ are automatically included in R.

Notice that at s = 1, (35) reduces to

$$e(1) = CTP_e[P_{av}(1,N) - P_{av}(1,N-1)], \qquad (37)$$

and so we would infer by analogy that if e_I under categorical screening exceeds the right-hand side of (37), then no amount of categorical screening could reduce under-response.

4.3.4 Multi-Server Priority Screening: The Continuous Analog

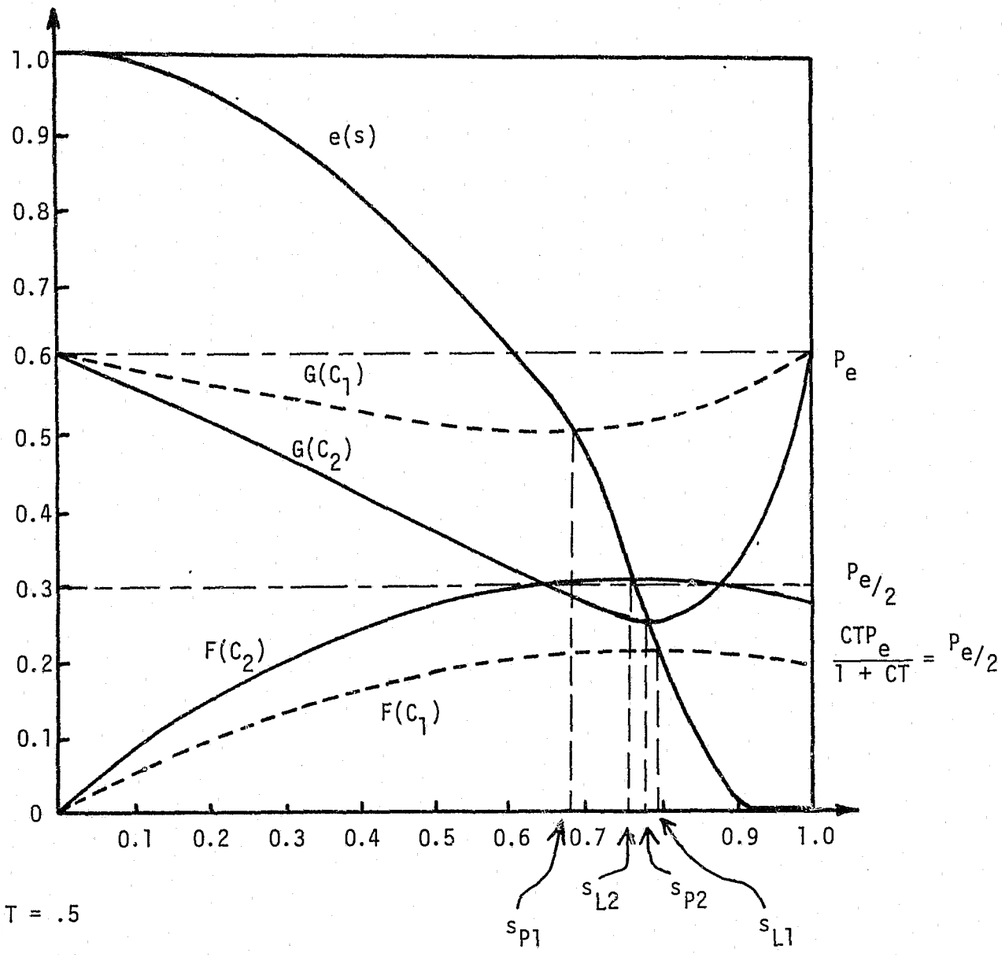
The continuous analog for the optimization of priority screening is easily written as

$$\min_{0<s\leq 1} r_N(s) = \frac{1 - \frac{CT}{P_e N} \int_0^s f(y)e(y)dy}{1 - \frac{CT}{N} \int_0^s f(y)dy} \qquad (38)$$

The solution occurs when

$$e(s) = P_e \frac{1 - \frac{CT}{P_e N} \int_0^s f(y)e(y)dy}{1 - \frac{CT}{N} \int_0^s f(y)dy} \qquad (39)$$

Note: When $e(s) = {}^{P_e}/_2 = .305$
$s = 0.76$



$C_1T = .5$

$C_2T = .9$

$$F(C) = \frac{CT\int_0^s f(y)e(y)dy}{1 + CT\int_0^s f(y)dy}$$

$$G(C) = \frac{P_e - CT\int_0^s f(y)e(y)dy}{1 - CT\int_0^s f(y)dy}$$

Figure 7: Single-Server Case: Loss and Priority Screening (Hypothetical Continuous Example)

It can be shown that the solution to equation (39) occurs where the derivative of the RHS is zero

A graphical representation of the solutions to equation (34) and (39) for N = 1 with hypothetical functions e(s) and f(s) is shown in Figure 7. Notice that the solutions, s_L for loss screening and s_P for priority screening, occur at the points at which the derivatives of the right-hand sides of the respective equations equal zero. Is is interesting that as the call rate increases more callers are given secondary service under loss screening (i.e., s_L moves to the left), while more callers are given priority service under priority screening (i.e., s_P moves to the right). Although the behavior of the loss screening process confirms one's intuition--screening becomes more selective as the system load increases--the apparently anomolous movement of s_P deserves clarification. Under priority screening, as the call rate increases, very long delays are imposed on all callers assigned low priority. Since priority screening involves minimizing delays suffered by true emergencies, the point s_P moves so as to reduce the number of true emergencies in the low priority class as the total call rate increases, i.e., s_P moves to the right.

If the functions represented by the right hand sides of equations (34) and (39) intersect, the intersections always occur at the value $\frac{P_e}{2}$. If there is a unique point of intersection, then $s_L = s_P = s^*$ and the critical fraction of emergencies, $e(s^*) = {}^{P_e}/_2$.

4.4 A Sample Calculation for Categorical Loss Screening

To illustrate the potential impact of categorical screening, we will use data derived from the emergency medical service in Boston [2]. During one week in 1972, Boston Health and Hospitals Department ambulances transported patients with 15 different types of problems. The problems and

Table 1

Rank(i)	Problem	% Emergent(e_i)	% of all calls(f_i)
1	Coma, unconsciousness	100%	2%
2	Seizures	100	12
3	Stroke	78	2
4	Vomiting blood, rectal bleed	75	4
5	Chest pain	62	2
6	Overdose, intoxication	56	6
7	Breathing difficulty	48	8
8	Nausea, vomiting	36	4
9	Trauma	19	34
10	Psychiatric	6	2
11	Abdominal pain	1	4
12	Neck, back, shoulder pain	0	4
13	Fainting, dizziness	0	8
14	Arm or leg pain (no trauma)	0	4
15	Vague or undefined	0	4

Overall Emergent P_e = 35%

DATA FOR EXAMPLE

based on Boston Health and Hospitals data, 1972 [2]

estimates of e_i and f_i for each problem are listed in Table 1.

Assume that such a mix of problems had been imposed on a single-server loss screening system operating with CT = 1, that is, with very heavy demand. Successively dropping categories from the bottom of the list in Table 1 produces the operating curve shown in Figure 8. The corresponding values of P_u and P_o are shown in Figure 9. Note in Figure 9 that screening contributes relatively little in this example to reduction of the probability of under-response, but has a major impact on the probability of over-response. The "optimal" categorical screening strategy occurs at point A in Figure 8, corresponding to screening out about 60% of the calls.

It is interesting to note, though, that categorical screening at point A may be unacceptable, in practice, compared to screening at point B. Screening at point B certainly leads to greater probability of over-response, but the more important degredation in probability of under-response is less then one percent, and screening at point B differs from that at point A only in permitting the trauma category to receive primary service. The trauma category was dropped in the "optimal" categorical policy because only 19% of calls in that category were truly emergent. While this does reserve the ambulance for categories more likely to contain true emergencies, it is difficult to conceive of an agency never sending the best ambulances to crime or accident victims. In effect, the number of categories in this example would have to be increased before a realistic categorical screening program could be implemented.

Naturally an important element of any screening program is the accuracy with which the caller reports the incident to the screener. Anecdotal evidence from both police and emergency ambulance services indicates that some members of the public will exaggerate the seriousness of an incident in order to

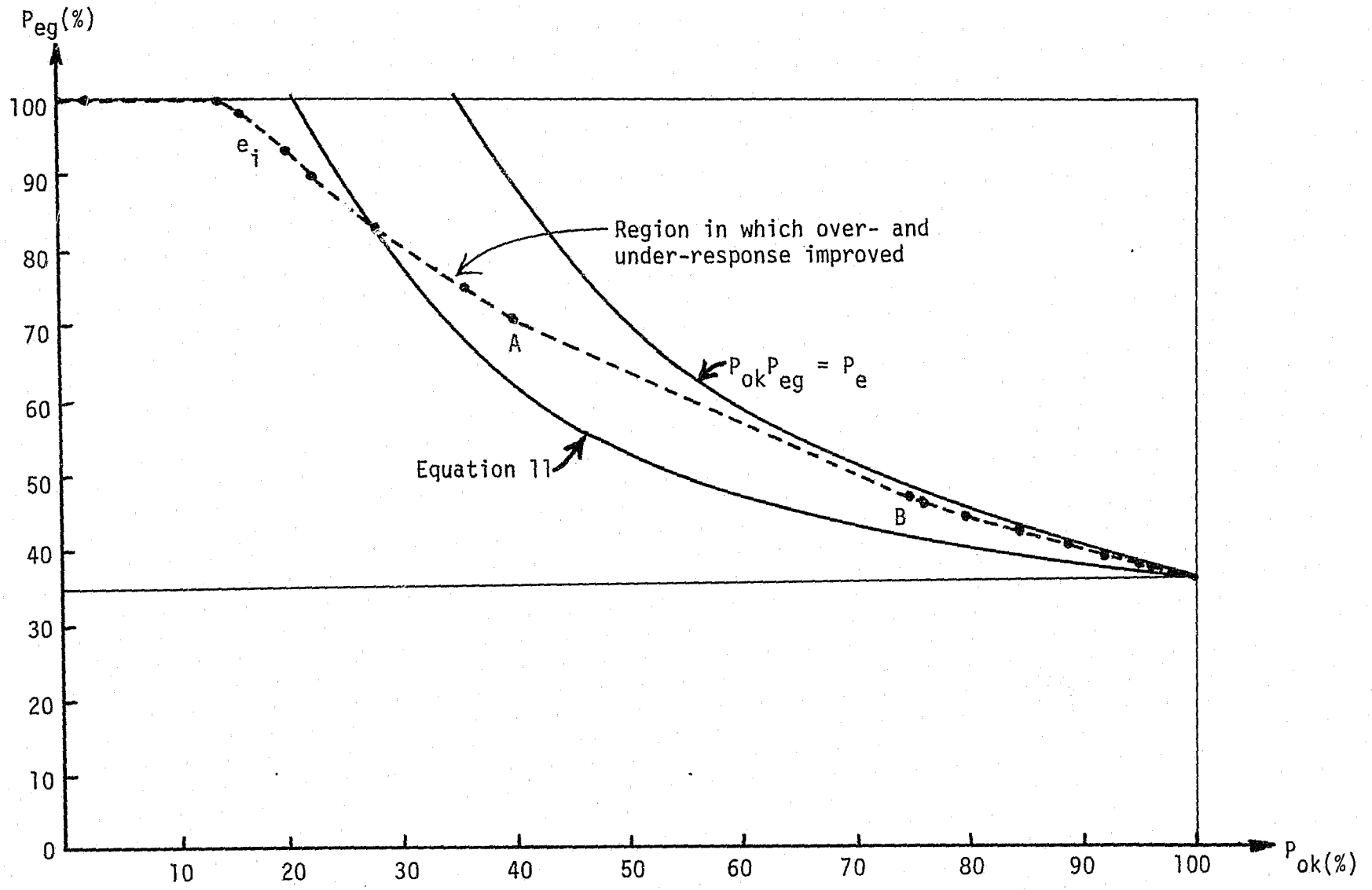


Figure 8: Operating Points Under Various Screening Policies in "Boston" Example

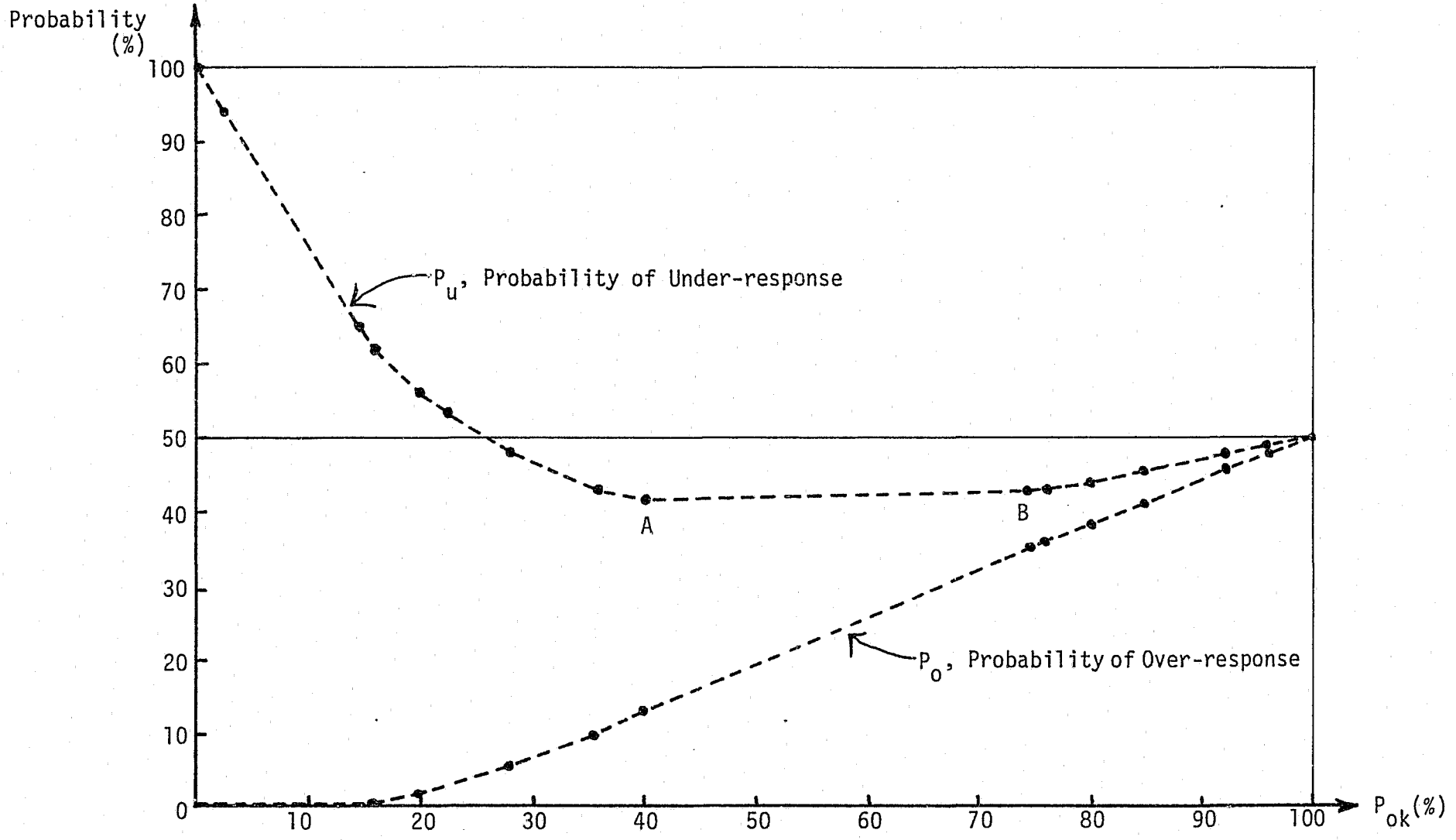


Figure 9: Performance of Screening System Under Various Screening Policies in "Boston" Example

ensure the maximal response for themselves. Although examples of this kind of deceit are very disturbing to screeners and service providers, their impact on measures like under-response seem unlikely to be severe enough to negate the value of screening. For one thing this behavior will generally be limited to unscrupulous frequent users of the service, who constitute a small minority of all users. Furthermore their impact on the screening process will be to increase P_{fp}, the probability of a false positive error, which, as we saw in Figure 3, is the less critical screening variable. Only if the process is operating close to its error limits will public manipulation of the dispatcher threaten the utility of screening. Its impact can be incorporated into the preceding analytical framework by treating the parameter e_i, the fraction of truly emergent members of category i, as a random variable.

5. SUMMARY

Increases in demand for emergency services, especially by those not faced with true emergencies, degrade the level of service provided to those truly in need, frustrate emergency service personnel and create pressure for greater investment in emergency services. At the same time, the fact that many truly emergent cases never use emergency services motivates attempts to improve citizen access to these services. However, improved citizen access is likely to exacerbate the problem of non-emergent use. A combination of better citizen access together with the screening of calls for emergency service is a response aimed at better matching emergency resources to emergency needs.

The process of screening carries the risk of error on the part of the screener. Screening will improve the level of service only when its deflection of spurious demand is not outweighed by erroneous deflection of true emergencies. The criteria for successful screening identified here include reduction of the likelihood that a true emergency will not be given specialized service (for "loss screening"), and reduction in the average delay in initiating service (for "priority screening"). We noted that some screening programs have the potential to provide poorer service to the public while simultaneously meeting internal service goals regarding "abuse," but that any program improving service to the public will also improve the operating position of the service provider.

We have provided a methodology for characterizing the quality of a screening operation and verifying that the screening can in fact improve service. Conditions favorable to screening include a heavy demand relative to system size, a small percentage of truly emergent cases, and rather small probabilities of classifying patients' conditions wrongly. We have also

provided a comparison between call screening and the expansion of the system without screening (Figure 4). At heavy system loading only if the probability of misclassifying an emergent patient is (well) below 20% is screening preferable to system expansion as a way of meeting increases demand.

While it does not appear possible to determine without experiment the actual performance of screening personnel, it is possible to analyze mathematically the particularly simple process called "categorical screening," in which calls for service are accepted solely on the basis of the general nature of their problem, regardless of individual details. We have presented an algorithm for determining the optimal categorical screening policy and have illustrated its use in an example. The fact that such a crude screening methodology can improve service suggests that trained personnel should be able to do much better.

REFERENCES

1. Rosen, P., M. Segal, L. Coppleson and B. Fauman, "A Method of Triage Within an Emergency Department," Journal of the American College of Emergency Physicians, March-April 1974, p. 85.

2. Kleinman, J., R. J. Weiss and M. M. Tanner, Emergency Medical Services in the City of Boston, Harvard Center for Community Health and Medical Care, Boston, 1972.

3. The Boston Sunday Globe, 8/4/1974, p. 1.

4. Mogielnicki, R. P., K. A. Stevenson and T. R. Willemain, Patient and Bystander Response to Medical Emergencies, Technical Report 05-74, Innovative Resource Planning Project, Laboratory of Architecture and Planning, M.I.T., Cambridge, MA, July 1974.

5. Graves, H. B., "ACEP Surveys Hospital Triage Systems," Journal of the American College of Emergency Physicians, November-December, 1972, p. 31.

6. Federal Register, Vol. 39, No. 62, Part 3, March 29, 1974, pp. 11758-11766.

7. Stevenson, K. A., Operational Aspects of Emergency Ambulance Services, M.I.T. Operations Research Center Technical Report 61, May 1971.

8. Cobham, A., "Priority Assignment in Waiting Line Problems," Operations Research, 2 (1954), p. 70.

END