Background

Emergency Medical Service providers have a need to efficiently ration limited medical resources to patients. Patients are assigned treatment priorities by a process called triage. Traditional triage protocols are static in terms of medical resource restrictions. However, in daily emergencies, the system experience continuous changes in availability of medical resources and call volume. When there are multiple types of patients as well as medical units, a proper match between patient-server is critical to improve service quality.

Objectives

- Explore how to reassign priorities dynamically considering the change in the medical resource availability to enhance service quality of EMS
- Derive practical rules-of-thumbs to assign medical resources and a heuristic method to solve the model in time

Setting

Resources: Ambulances are differentiated by the types of treatment they can provide.
- Advanced Life Support (ALS) vehicles staffed by paramedics to serve urgent calls (rate \( \mu_A \))
- Basic Life Support (BLS) vehicles with staffed by EMTs to serve less serious calls (rate \( \mu_B \))

Patients: Calls arriving at rate \( \lambda = \sum \lambda_i \) have
- accident types \( i \in \{1, \ldots, m\} \)
- pre-assigned priorities \( p \in \{1,2,3\} \)

Information: The true urgency of an arriving call with priority 2 is known to the decision maker but only probabilistically on its call type, with parameter \( \alpha^t = \Pr(\text{urgent} \mid \text{call type}=i,priority=2) \)

Reachability: Since ambulances are spatially located, only a subset can respond in a timely fashion. The probability that an arriving call is served successfully in-time when \( s^A \) ALSs (\( s^B \) BLSs) are busy is modeled as a concave non-increasing function \( f^A(s^A)/f^B(s^B) \).

Prioritization:

- Ambulances are differentiated by the type of patient they serve:
  - Basic Life Support (BLS) vehicles with staffed by EMTs to serve less serious calls (rate \( \mu_B \))
- Advanced Life Support (ALS) vehicles staffed by paramedics to serve urgent calls (rate \( \mu_A \))

- Explore how to reassign priorities dynamically considering the change in the medical resource availability to enhance service quality of EMS
- Derive practical rules-of-thumbs to assign medical resources and a heuristic method to solve the model in time

Model

The problem involves sequential decision making and is modeled as a undiscounted finite-horizon discrete Markov Decision Process.

Markov Decision Process model

State \( s = (s^A, s^B) \):
- \( s^A(s^B) \): number of busy ALS(BLS) servers

Transition at most 1 event happens among:
- An urgent call arrives \( (s^A, s^B) \to (s^A + 1, s^B) \)
- A less urgent call arrives \( (s^A, s^B) \to (s^A, s^B + 1) \)
- An ALS server finishes service \( (s^A, s^B) \to (s^A - 1, s^B) \)
- A BLS server finishes service \( (s^A, s^B) \to (s^A, s^B - 1) \)

Action Based on the of type(\( i \)) and priority of an arriving call, assign either an ALS \( (\alpha^t > \bar{\alpha}_t(s)) \) or BLS \( (\alpha^t = \bar{\alpha}_t(s)) \), depending on the system congestion \( s^A, s^B \).

Reward expected utility from serving a call:

<table>
<thead>
<tr>
<th>Urgent</th>
<th>Not urgent</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha^t = 0 )</td>
<td>( U_{HA} )</td>
</tr>
<tr>
<td>( \alpha^t = 1 )</td>
<td>( U_{HB} )</td>
</tr>
</tbody>
</table>

Given \( U_{HA} > U_{LA}, U_{HA} > U_{LB}, U_{HA} > U_{HB} \) to penalize over-service and under-service.

Finite-time discrete MDP is solved by backward induction to maximize total expected reward.

\[ U_i(s_i) = \sup_{\alpha} \{ R_i(s_i, \alpha) + \sum_j P_j(s_i, \alpha) U_{i+1}(j) \} \]

Structural Results

Theorem 1. (Type Independence) The decision-making for a specific call type \( i \) can be made regardless of decision for other call types.

Theorem 2. (Optimality of threshold-type policy) A threshold value \( \alpha^t \) can be specified such that it is optimal to send ALS server to type \( i \) call if and only if \( \alpha^t > \bar{\alpha}_t(s) \), if

\[ f^A(s^A)/f^B(s^B) > \frac{U_{HA} - U_{LB}}{U_{HA} - U_{LA}} \]

Theorem 3. (Optimality of monotone policy) Optimal action is non-decreasing in \( s^A \) and non-increasing in \( s^B \). If \( f^A(s^A)/f^B(s^B) \) are concave and service rate is common between two server types. In addition, cost-to-go function \( U_i(s^A, s^B) \) is concave and submodular.

Future Work

Extensions:
- Develop a practical model that makes priority reassignment less frequently (e.g. every hour) allows multiple events to happen between decision making points.
- Provide heuristic algorithms to solve the extended model applying structural results from the base model

Computational study:
- sensitivity analysis on change of benefiting from patient-resource match
- sensitivity analysis on vehicle mix

Evaluation: Statistical evaluation on number of over-served and under-served patients

Acknowledgement

This work was funded by the National Science Foundation [Award 1444219]. The views and conclusions contained in this document are those of the author and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the National Science Foundation.

References


Soovin Yoon and Laura A. McLay, yoon57@wisc.edu, lauramclay@engr.wisc.edu
Department of Industrial and Systems Engineering, University of Wisconsin-Madison