Control of Patient Flow in Emergency Departments, or Multiclass Queues with Deadlines and Feedback

Junfei Huang (NUS & Northwestern U)

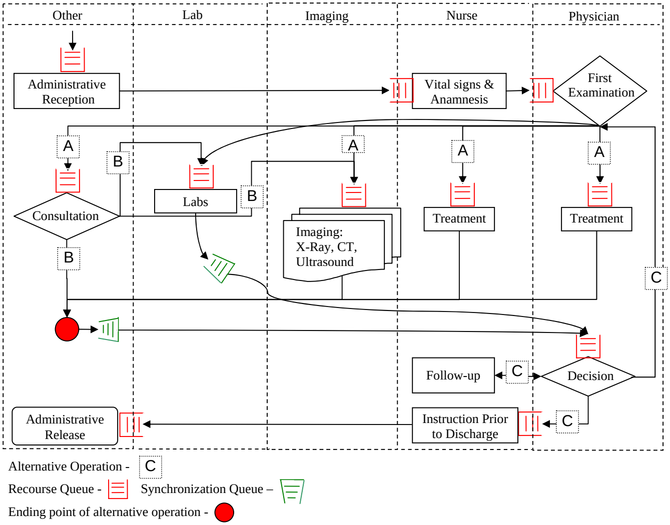
Jointly with Boaz Carmeli (IBM) and Prof. Avishai Mandelbaum (Technion)

http://ie.technion.ac.il/serveng/References/references.html

https://sites.google.com/site/junfeih/

INFORMS 2012 Annual Meeting

Patient Flow in Emergency Department (ED)



(Armony M., et al. (2011): Patient Flows in Hospitals: A Data-Based Queueing-Science Perspective.)

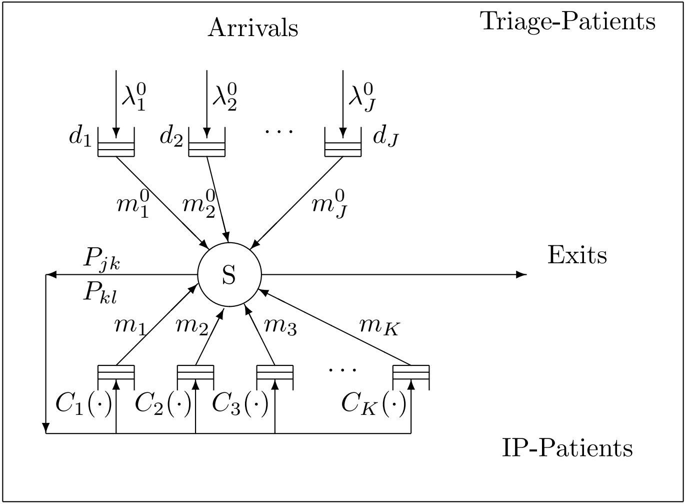
Emergency Department (ED)

- **Feedback** (Yom-Tov and Mandelbaum (2012)):

Physician Type	Patient Type	Average number of visits
1	1, 7	3.9698
2	2, 5	2.9904
3	3, 6	2.9700
4	4	2.9904

- Clinical (**quality**) vs. Operational (**efficiency**):
  - **Quality**: **Deadlines** on *time-till-first-treatment*;
  - **Efficiency**: Congestion costs.

Model Structure



Model Description

- S physicians (small and fixed);
- J classes of triage patients, $j(\in \mathcal{J})$-triage patients:
  - arrival rate λ_j^0;
  - mean service requirement m_j^0, $M_{\mathcal{J}} = (m_j^0)$;
  - after service, transfer to k-IP patient with probability P_{jk}; $P_{\mathcal{J}\mathcal{K}} = [P_{jk}]$; (*triage-to-IP* transition matrix)
  - have deadline d_j:
    - denote by $\tau_j(t)$ the age of the head-of-the-line patient, then $\tau_j(t) \leq d_j$;
- K classes of in-process (IP) patients, $k(\in \mathcal{K})$-IP patients:
  - no exogenous arrivals;
  - mean service requirement m_k, $M = (m_k)$;
  - after service, transfer to l-IP patient with probability P_{kl}; $P = [P_{kl}]$; (*IP-to-IP* transition matrix)
  - incur queueing cost at rate $C_k(Q_k(t))$ (or $C_k(W_k)$);

Problem Formulation:

- Cumulative cost till t: $\int_0^t \sum_{k\in\mathcal{K}} C_k(Q_k(s))\mathsf{d}s$;
- Deadline constraints: $\tau_j(t) \le d_j$, $j \in \mathcal{J}$;
- Constraint optimization problem: for any $T > 0$,

$$\min_{\Pi} \quad \int_0^T \sum_{k\in\mathcal{K}} C_k(Q_k(s))\mathsf{d}s$$
$$\text{s.t.} \quad \tau_j(t) \le d_j, \quad \forall j \in \mathcal{J} \quad \text{and} \quad 0 \le t \le T.$$

- Does this problem have feasible solution?
  - $\tau_j, j \in \mathcal{J}$ random, d_j deterministic;
  -
    - relax the 'feasibility' – 'asymptotically feasible';
    - relax the 'optimality' – 'asymptotically optimal'.

Effective Mean Service Time

- $M^e_{\mathcal{J}} = (m^e_j)$ – **effective mean service time** of Triage patients:

$$M^e_{\mathcal{J}} = M_{\mathcal{J}} + P_{\mathcal{J}\mathcal{K}}[I - P]^{-1}M;$$

  m^e_j is the expected *total* service time required by j-triage patients throughout their ED stay.

- *Traffic intensity*:

$$\rho = \frac{1}{S}\sum_{j\in\mathcal{J}} \lambda^0_j m^e_j,$$

  assume $\rho \approx 1$ (heavy-traffic);

- $M^e = (m^e_k)$ – **effective mean service time** of IP patients:

$$M^e = [I - P]^{-1}M;$$

  m^e_k is the expected *total* service time required by k-IP patients throughout their ED stay.

Proposed Policy

Choose *any* one of the triage classes (conceivably the least d_j, say d_1). Then a physician that becomes idle at time t adopts the following guidelines:

- Serve triage patients if $\tau_1(t) \geq d_1 - \epsilon$, where ϵ is small relative to d_1 (e.g. $d_1 = 30$ minutes while $\epsilon = 3$ minutes);
- Given that a triage patient is to be served, choose the head-of-the-line patient from the class with index

$$j \in \operatorname*{argmax}_{j \in \mathcal{J}} \frac{\tau_j(t)}{d_j};$$

- Given that an IP patient is to be served, choose the head-of-the-line patient from the class with index

$$k \in \operatorname*{argmax}_{k \in \mathcal{K}} \frac{C_k'(Q_k(t))}{m_k^e}.$$

Literature Review

Only those closely related:

- Generalized $c\mu$ ($Gc\mu$) policy:
  - van Mieghem (1995);
  - Mandelbaum and Stolyar (2004);
- Due-date:
  - van Mieghem (2003);
  - Plambeck, Kumar and Harrison (2001);
- Feedback:
  - Reiman (1988) and Dai and Kurtz (1995);
  - Klimov's model.

Asymptotic Framework

- A sequence of systems, indexed by $r \uparrow \infty$:
- Triage patients:
  - Arrival rate λ_j^r;
  - Service requirement m_j;
  - Markovian routing, $P_{\mathcal{J}\mathcal{K}} = (P_{jk})_{J\times K}$;
- IP patients:
  - Internal arrival;
  - Service requirement m_k;
  - Markovian routing, $P = (P_{kl})_{K\times K}$;
- Independence assumption;
- *Traffic intensity*: $\rho^r = \sum_{j\in\mathcal{J}} \lambda_j^r m_j^e$;

Asymptotic Framework (Cont.)

- (Conventional) *heavy traffic condition:* as $r \to \infty$,

$$r(\rho^r - 1) \to \beta,$$
$$\lambda_j^r \to \lambda_j, \quad j \in \mathcal{J},$$

  for some $\beta \in \mathbb{R}$ and $\lambda_j > 0$.

- Deadlines for triage patients, d_j^r, $j \in \mathcal{J}$, satisfy

$$\frac{d_j^r}{r} \to \widehat{d}_j, \quad \text{as } r \to \infty, \quad \text{for all} \quad j \in \mathcal{J},$$

  where $\widehat{d}_j > 0,\ j \in \mathcal{J}$.

Asymptotic Framework (Cont.)

- Control policies $\pi^r = \{T_j^r, T_k^r\}$;
- Diffusion-scaled age processes:

$$\widehat{\tau}_j^r(t) = r^{-1}\tau_j^r(r^2 t), \quad j \in \mathcal{J}.$$

- Diffusion scaled queue length processes:

$$\widehat{Q}_k^r(t) = r^{-1}Q_k^r(r^2 t), \quad k \in \mathcal{K}.$$

- Cumulative queueing cost:

$$\mathcal{U}^r(t) := \int_0^t \sum_{k\in\mathcal{K}} C_k\left(\widehat{Q}_k^r(s)\right)\mathsf{d}s.$$

Asymptotic Compliance

Definition

A family of control policies is said to be asymptotically compliant if, for any fixed $T > 0$, as $r \to \infty$,

$$\sup_{0 \le t \le T} \left[\widehat{\tau}_j^r(t) - \widehat{d}_j \right]^+ \Rightarrow 0, \quad j \in \mathcal{J}.$$

Asymptotic Optimality

Definition

A family of control policies $\{\pi_*^r\}$ *is said to be asymptotically optimal if*

- *it is asymptotically compliant and*
- *for every* $t > 0$ *and every* $x > 0$,

$$\limsup_{r\to\infty} \mathbb{P}\left\{\mathcal{U}_*^r(t) > x\right\} \leq \liminf_{r\to\infty} \mathbb{P}\left\{\mathcal{U}^r(t) > x\right\},$$

$\{\mathcal{U}_*^r\}$*: corresponding to* $\{\pi_*^r\}$*;*
$\{\mathcal{U}^r\}$*: corresponding to any asymptotically compliant policies.*

The Proposed Policies

Choose *any one* of the triage classes, say $1 \in \mathcal{J}$. In the rth system, a physician that becomes idle at time t adopts the following guidelines:

- Serve triage patients if $Q_1^r(t) \geq \lambda_1^r d_1^r$;
- Given that a triage patient is to be served, choose the head-of-the-line patient from the class with index

$$j \in \underset{j \in \mathcal{J}}{\operatorname{argmax}} \frac{\tau_j^r(t)}{d_j^r};$$

- Given that an IP patient is to be served, the physician uses a policy ensuring (for any $T > 0$)

$$\max_{l,k \in \mathcal{K}} \sup_{0 \leq t \leq T} \left| \frac{C_l'(\widehat{Q}_l^r(t))}{m_l^e} - \frac{C_k'(\widehat{Q}_k^r(t))}{m_k^e} \right| \Rightarrow 0.$$

Alternative Policies for Triage

- *Shortest-Deadline-First* policy: when the triage classes are chosen to be served at time t, the physician chooses the head-of-the-line patient from the class with index

$$j \in \underset{j \in \mathcal{J}}{\operatorname{argmin}} \left(d_j^r - \tau_j^r(t) \right);$$

Examples of Policies for IP

- G: a $K \times K$ irreducible matrix:
  - all components of GM^e being non-zero;
- H: the K-dimensional vector, $H_k = 1/(GM^e)_k$;
- When the IP classes are chosen to be served at time t, the physician chooses the head-of-the-line patient from class

$$k \in \underset{k \in \mathcal{K}}{\operatorname{argmax}}\, H_k \left(GC' \left(\widehat{Q}^r(t) \right) \right)_k .$$

- Two special cases:
  - If $G = I$, $k \in \operatorname{argmax}_{k \in \mathcal{K}} \frac{C'_k\left(\widehat{Q}^r_k(t)\right)}{m^e_k}$ – (modified) $Gc\mu$;
  - If $G = I - P$, the policy conjectured in Mandelbaum and Stolyar (2004);

Main Results

Theorem

Our proposed family of control policies is asymptotically optimal.

Intuition and Proofs

- $A(t)$: total potential workload brought into the ED;
- $T(t)$: amount of workload served;
- $W(t) = A(t) - T(t)$: total potential workload left:
  - minimized by work-conserving policy;
  - **invariant** to any work-conserving policy;
  - conditional on the queue length processes,

$$W(t) \approx \sum_{j \in \mathcal{J}} m_j^e \times Q_j(t) + \sum_{k \in \mathcal{K}} m_k^e \times Q_k(t).$$

- Making $\sum_{j \in \mathcal{J}} m_j^e \times Q_j(t) \approx \sum_{j \in \mathcal{J}} \lambda_j m_j^e \tau_j(t)$ close to the upper bound $\sum_{j \in \mathcal{J}} \lambda_j m_j^e d_j$;
- Making Q_k minimize the cost rate.

The Functions of The Proposed Policies

- Triage patients – making triage classes well-behaved:
  - One class embodies enough information for all classes;
  - when one class is close to the deadline, all other classes are also close to the deadlines;
- Threshold policy: Making $\tau_1^r(t)$ closest to d_1^r for all t;
  - Then all classes are close to their deadlines;
- IP patients: $\widehat{Q}_k^r(t)$, $k \in \mathcal{K}$, asymptotically solve:

$$\begin{aligned} \min \quad & \sum_{k\in\mathcal{K}} C_k(\widehat{Q}_k^r(t)) \\ s.t. \quad & \sum_{k\in\mathcal{K}} m_k^e \widehat{Q}_k^r(t) = (\widehat{W}^r(t) - \sum_{j\in\mathcal{J}} \lambda_j m_j^e \widehat{d}_j)^+. \end{aligned}$$

Verify via KKT condition[1].

[1] KKT (Karush - Kuhn - Tucker) condition is used to help solving non-linear programming problems.

State Space Collapse

Theorem

Under the proposed policy, $\widehat{Q}^r \Rightarrow \widehat{Q}$*, where* $\widehat{Q}$ *satisfy*

- $\frac{\widehat{Q}_j(t)}{\lambda_j \hat{d}_j} = \frac{\widehat{Q}_{j'}(t)}{\lambda_{j'} \hat{d}_{j'}}$, $j, j' \in \mathcal{J}$;
- $\sum_{j \in \mathcal{J}} m_j^e \widehat{Q}_j(t) = \min(\widehat{Q}_w(t), \widehat{\omega})$;
  - $\widehat{\omega} = \sum_{j \in \mathcal{J}} \lambda_j m_j^e \widehat{d}_j$;
  - $\widehat{Q}_w(t)$ *is a reflected Brownian motion;*
- $\widehat{Q}_k(t)$, $k \in \mathcal{K}$, *satisfy*

$$\frac{C_k'(\widehat{Q}_k(t))}{m_k^e} = \frac{C_{k'}'(\widehat{Q}_{k'}(t))}{m_{k'}^e}, \quad k, k' \in \mathcal{K};$$

$$\sum_{k \in \mathcal{K}} m_k^e \widehat{Q}_k(t) = (\widehat{Q}_w(t) - \widehat{\omega})^+.$$

Sample-Path Little's Law

- FCFS among each class;
- ω_k^r: virtual waiting time, $\widehat{\omega}_k^r(t) = r^{-1}\omega_k^r(t)$;
- τ_k^r: age, $\widehat{\tau}_k^r(t) = r^{-1}\tau_k^r(t)$;

Proposition

$$\lambda_k^r \widehat{\omega}_k^r - \widehat{Q}_k^r \Rightarrow 0, \quad k \in \mathcal{K},$$
$$\lambda_k^r \widehat{\tau}_k^r - \widehat{Q}_k^r \Rightarrow 0, \quad k \in \mathcal{K}.$$

- $Q_k^r(t + \omega_k^r(t)) = E_k^r(t + \omega_k^r(t)) - E_k^r(t) \approx \lambda_k^r \omega_k^r(t)$,
  $Q_k^r(t) = E_k^r(t) - E_k^r(t - \tau_k^r(t)) \approx \lambda_k^r \tau_k^r(t)$;
- **Snapshot principle**: $Q_k^r(t + \omega_k^r(t)) \approx Q_k^r(t)$;

Sojourn Time

- FCFS among each IP-class;
- $h \in \mathbb{Z}_+^K$: visit vector:
  - h_k: time of visits to k-IP class before leaving the system;
  - j-feasible;
- $W_{jh}^r(t)$: *sojourn times* of the next j-triage patient arriving after t with visit vector h,

$$\widehat{W}_{jh}^r(t) = r^{-1} W_{jh}^r(r^2 t).$$

- **Snapshot principle**: $\widehat{W}_{jh}^r(t) \approx \widehat{\omega}_j^r(t) + \sum_{k \in \mathcal{K}} h_k \widehat{\omega}_k^r(t)$;

Proposition

$$\widehat{W}_{jh}^r - \frac{\widehat{Q}_j^r}{\lambda_j^r} - \sum_{k \in \mathcal{K}} \frac{h_k}{\lambda_k^r} \widehat{Q}_k^r \Rightarrow 0,$$

$$\widehat{W}_{jh}^r - \widehat{\tau}_j^r - \sum_{k \in \mathcal{K}} h_k \widehat{\tau}_k^r \Rightarrow 0.$$

Waiting-Time Costs

- FCFS among each IP-class;
- Cumulative waiting costs:

$$\widetilde{\mathcal{U}}^r(t) := \sum_{k\in\mathcal{K}} \int_0^t C_k\left(\widehat{\omega}_k^r(s)\right) \mathsf{d}\bar{\bar{E}}_k^r(s);$$

- Threshold policy between triage and IP does not change;
- The policy determining priorities among triage does not change;
- If the IP classes are chosen to be served at time t, the physician uses a policy ensuring that, for any $T \geq 0$,

$$\max_{l,k\in\mathcal{K}} \sup_{0\leq t\leq T} \left| \frac{C_l'\left(\frac{\widehat{Q}_l^r(t)}{\lambda_l^r}\right)}{m_l^e} - \frac{C_k'\left(\frac{\widehat{Q}_k^r(t)}{\lambda_k^r}\right)}{m_k^e} \right| \Rightarrow 0.$$

Sojourn-Time Costs

- Routing matrix P (IP-to-IP transition matrix) is upper-triangular (WLOG, assume each patient has a deterministic routing vector);
- Denote by $\mathcal{C}_0$ the *starting classes* of any route;
- Denote by $\mathcal{C}_k$ all classes on a route starting with $k \in \mathcal{C}_0$;
- Classes in $\bigcup_{k \in \mathcal{C}_0} \mathcal{C}_k \backslash \{k\}$ are *subsequent classes*;
- Congestion cost:

$$\widetilde{\mathcal{S}}^r(t) := \sum_{k \in \mathcal{C}_0} \int_0^t C_k \left(\sum_{k' \in \mathcal{C}_k} \widehat{\omega}^r_{k'}(s) \right) \mathsf{d}\bar{\bar{E}}^r_k(s);$$

Sojourn-Time Costs: Asymptotically Optimal Policy

- Threshold policy between triage and IP does not change;
- The policy determining priorities among triage does not change;
- If the IP classes are chosen to be served at time t, the physician
  - Gives higher priority to subsequent classes;
  - For the starting classes, ensuring that, for any $T \geq 0$,

$$\max_{l,k\in\mathcal{C}_0} \sup_{0\leq t\leq T} \left| \frac{C_l'\left(\frac{\widehat{Q}_l^r(t)}{\lambda_l^r}\right)}{m_l^e} - \frac{C_k'\left(\frac{\widehat{Q}_k^r(t)}{\lambda_k^r}\right)}{m_k^e} \right| \Rightarrow 0.$$

An ED case study

- Data is from an Israeli ED; cost is on sojourn times;

Number of IP visits	1	2	3	4	5
Proportion	0.28	0.30	0.28	0.11	0.03

A& D Status	Admitted	Discharged
Proportion/Cost function	0.40, t^2	0.60, $2t^2$

- **No information**: The nurses can **NOT** estimate the number of IP visits or the A&D status;
- **Partial information**: The nurses can estimate the *number of IP visits* (costs can be reduced by 18.01%);
- **Complete information**: The nurses can estimate **both** the number of visits and the A&D status (costs can be reduced by 26.8%);

An ED case study

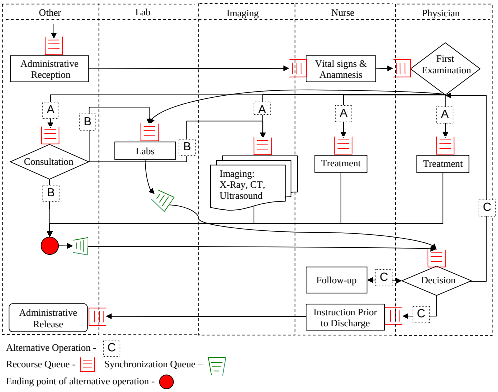
- Data is from an Israeli ED; cost is on sojourn times;

Number of IP visits	1	2	3	4	5
Proportion	0.28	0.30	0.28	0.11	0.03

A& D Status	Admitted	Discharged
Proportion/Cost function	0.40, t^2	0.60, $2t^2$

- **No information**: The nurses can **NOT** estimate the number of IP visits or the A&D status;
- **Partial information**: The nurses can estimate the *number of IP visits* (costs can be reduced by 18.01%);
- **Complete information**: The nurses can estimate **both** the number of visits and the A&D status (costs can be reduced by 26.8%);
- Good news: A good trained nurse can estimate both kinds of information very accurately!

Future Directions



- Adding delays between transfers;
- Time varying arrival rate;
- Adding global constraint on sojourn times;
- Adding abandonment (LWBS, LAMA).