Toward a Simulation-Based Real-Time Decision-Support System for Emergency Departments

Yariv N. Marmor[1], Segev Wasserkrug[2], Sergey Zeltyn[2], Yossi Mesika[2], Ohad Greenshpan[2], Boaz Carmeli[2], Avraham Shtub[1], Avishai Mandelbaum[1]

[1]Technion – Israel Institute of Technology
[2]IBM Haifa Research Labs

Winter Simulation Conference 2009

Motivation - ED overcrowding

- **Staff (re)scheduling** (off-line) using simulation:
  - Sinreich and Jabali (2007) – maintaining **steady utilization**.
  - Badri and Hollingsworth (1993), Beaulieu et al. (2000) – **reducing Average Length of Stay** (ALOS).
- Alternative **operational** ED **designs**:

  King et al. (2006), Liyanage and Gale (1995) – aiming mostly at reducing **ALOS**.
- Raising also the patients' view: **Quality of care**

  Green (2008) – **reducing waiting times** (also the time to first encounter with a physician).

Simulation support: short- to long-term

Part 1 (short-term): Decision-Support system (ED Staffing) in real-time (hours, shift).

Part 2 (medium) (Staffing) Over mid-term (weeks).

Part 3 (long): Fitting an efficient operational model to a given ED Environment.

Part 4 (long): Benefits of using real-time patients tracking (e.g. RFID) in the ED.

Here we focus on Part 1 – ED staffing in real-time, over a single shift (intraday staffing).

Part 1: Decision-Support system for Intraday staffing in real-time

Real-Time staffing: Objectives

- [Gather **real data** in **real-time** regarding current state]
- **Complete the data** when necessary via simulation.
- **Predict short-term** evolution (workload) via simulation.
- Corrective **staffing,** if needed, via simulation and mathematical models.

- All the above in **real-time** or close to real-time

Research framework and basic ED simulation model

- Our **ED** admits 80,000 patients annually:
  - ~ 60% classified as **Internal**.
  - ~ 40% classified as **Surgical or Orthopedic**.

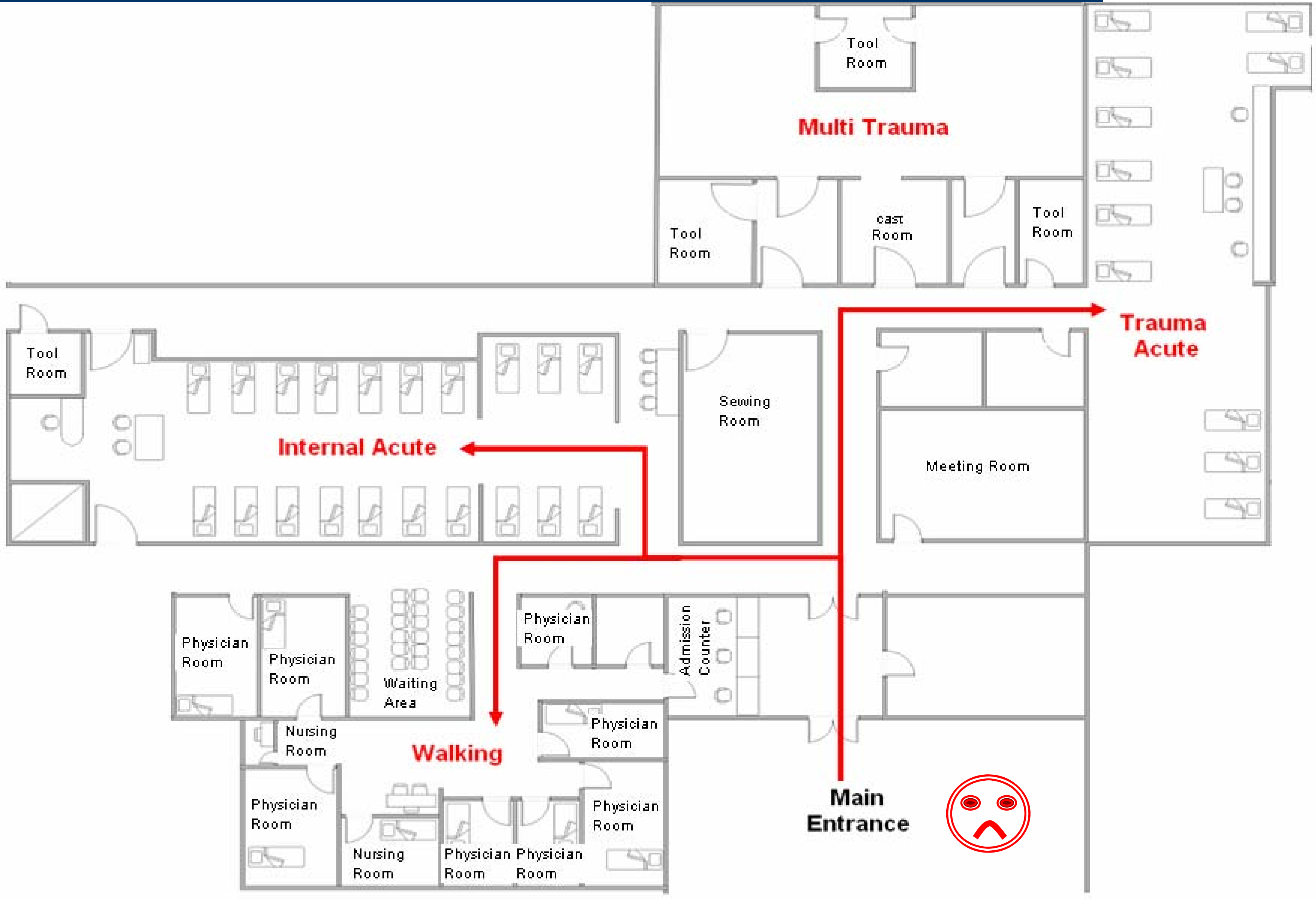
- The ED has three major physical areas:

(1) Internal acute

(2) Trauma acute

(3) Walking.

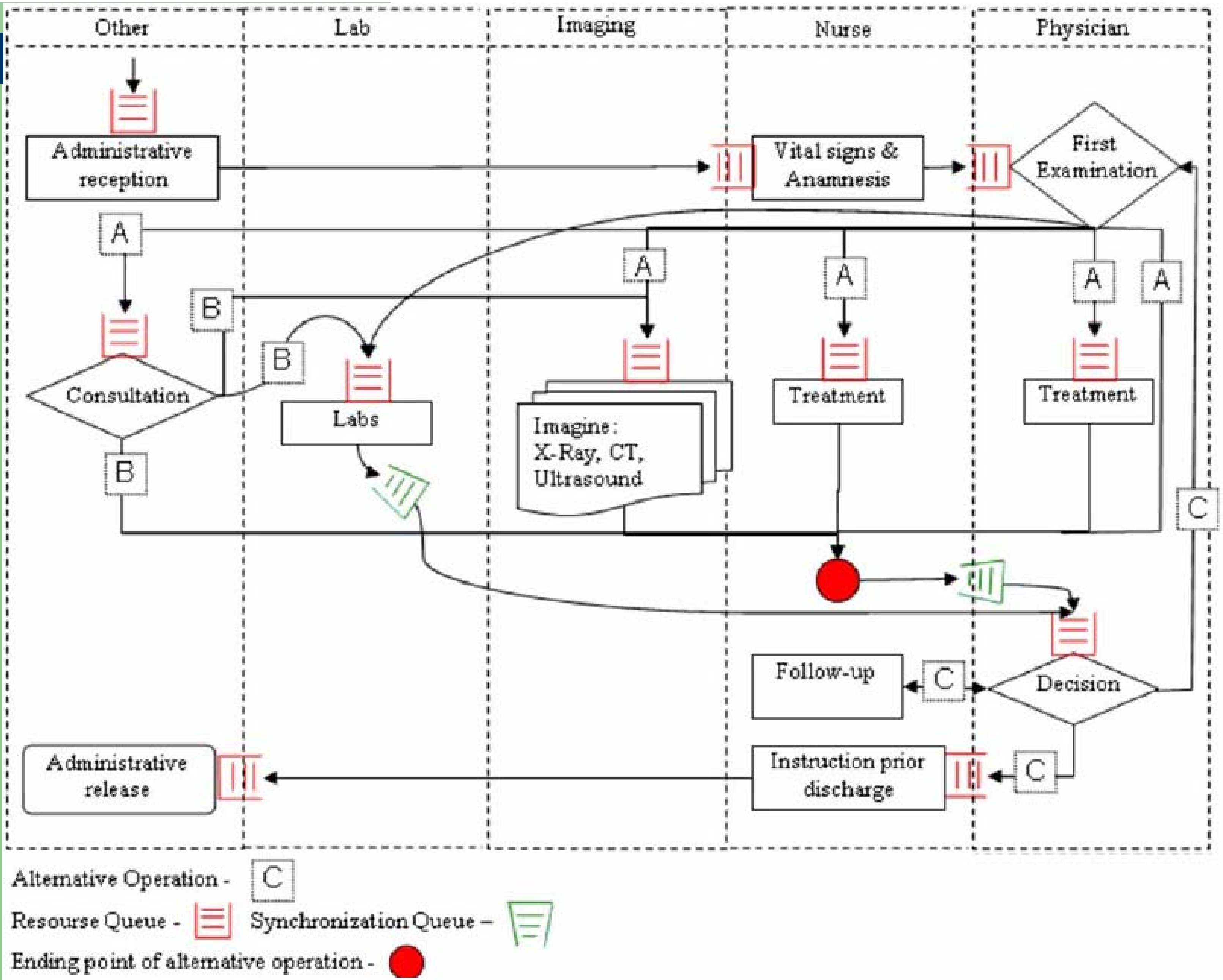
Research framework and basic ED simulation model



Research framework and basic ED simulation model

- Generic simulation tool (Sinreich and Marmor ,2005).
- ED activity-resource chart:

Research framework and basic ED simulation model



Part 1: **Intraday staffing in real-time**

Estimation of current ED state

- **Goal** – Estimate **current** ED state (using simulation):
  - For each patient: **type** (e.g. internal, ….) and **status** in the ED process (e.g. X-ray, Lab,…)

  [status un-extractable from most currently installed ED IT systems]

- Data description:
  - **Accurate** data - arrival and home-discharge processes.
  - **Inaccurate (censored)** data - departure times for delayed ED-to-Ward transfers (recorded as departures but are still in an ED bed).
  - **Unavailable** data – all the rest (e.g. patients status).
- Method to estimate present state:

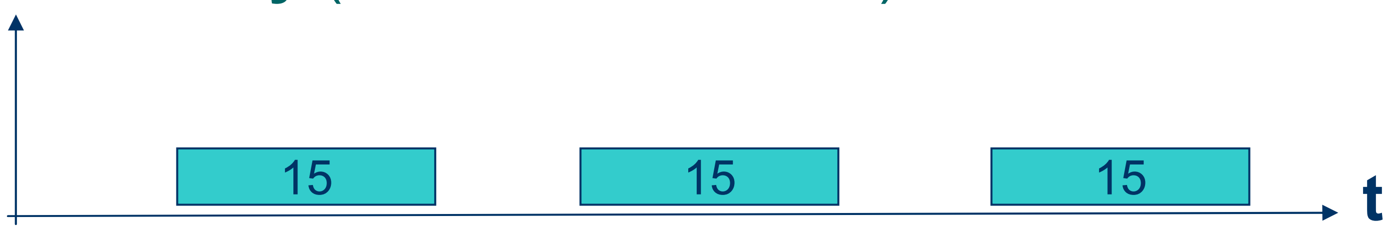
  Run ED simulation from “t=-∞”; keep replications that are consistent with the observed data (# of discharged)

Required staffing level – short-term prediction

Staffing models:

- **RCCP** (Rough Cut Capacity Planning) – Heuristic model aiming at **operational-efficiency** (resource utilization level).

- **OL** (**Offered Load**) - Heuristic model aiming at balancing high levels of **service-quality** (time till first encounter with a physician) and **operational-efficiency** (resource utilization).



OL: Offered-Load (theory)

In the simplest **time-homogeneous** steady-state case*:

R - the **offered load** is:

$$\boxed{R = \lambda * E(S)}$$

λ – arrival rate,

$E(S)$ – mean service time,

*Little's Law

"**Square-Root Safety Staffing**" rule: (Erlang 1914, Halfin & Whitt ,1981):

$$\boxed{n \approx R + \beta\sqrt{R}}$$

$\beta > 0$ is a "tuning" parameter.

Gives rise to **Quality** and **Efficiency-Driven** (QED) operational performance: carefully balances <u>high</u> service-quality (time to first-encounter) with <u>high</u> resource-efficiency (utilization levels).

Offered-Load (theory), time-inhomogeneous

Arrivals are better modeled by a **time-inhomogeneous** Poisson process, with arrival rate $\lambda(t)$; $t \geq 0$:

OL is calculated as the number of busy-servers (or served-customers), in a corresponding system with an **infinite** number of servers (Feldman *et al.* ,2008).

For simple model (not the ED):

$$\boxed{R(t) = E\left[\int_{t-S}^{t} \lambda(u)du\right] = \int_{-\infty}^{t} \lambda(u) P(S > t-u) du}$$

S - (generic) service time.

Offered Load (theory): time-inhomogeneous

QED-staffing approximation, achieving service goal α:

$$n_r(t) = R_t + \beta_t \sqrt{R_t}$$

$$1 - \alpha = P(W_q > T) \approx h(\beta_t) e^{-T\mu\beta_t \sqrt{n_r(t)}}$$

$n_r(t)$ - recommended number of resource r at time t, using OL.

α - fraction of patients that start service within T time units,

W_q – patients waiting-time for service by resource r,

$h(\beta_t)$ – the Halfin-Whitt function (Halfin and Whitt ,1981),

Offered Load methodology for ED staffing

- **∞ servers**: simulation run with "infinitely-many" resources (e.g. physicians, or nurses, or both).
- **Offered-Load**: for each resource *r*, and each hour t, calculate the number of busy resources (= total work).
- Use this value as an estimate for the offered load *R(t)* of resource *r* at time *t* (averaging over simulation runs).
- **Staffing**: for each hour *t* we deduce a recommended staffing level $n_r(t)$ via the formula:

$$n_r(t) = R_t + \beta_t \sqrt{R_t}$$
$$1 - \alpha = P(W_q > T) \approx h(\beta_t) e^{-T\mu\beta_t \sqrt{n_r(t)}}$$

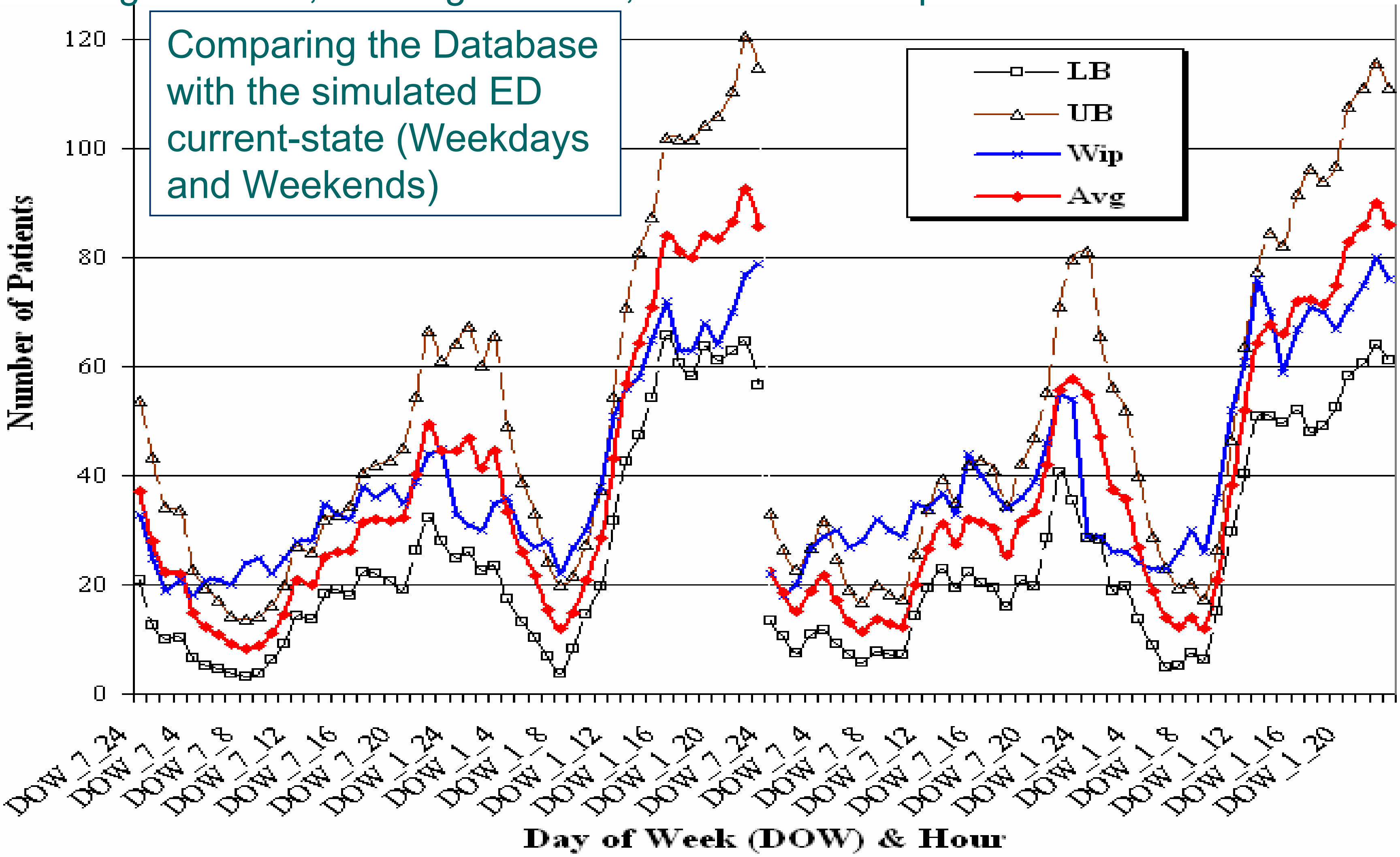
Methodology for short-term forecasting and staffing

Our simulation-based methodology for short-term staffing levels, over 8 future hours (shift):

1) Initialize the simulation with the **current ED state**.
2) Use the average arrival rate, to generate **stochastic arrivals** in the simulation.
3) Simulate and collect data every hour, over 8 future hours, using **infinite resources** (nurses, physicians).
4) From Step 3, calculate **staffing** recommendations, both $n_r(RCCP,t)$ and $n_r(OL,t)$.
5) Run the **simulation** from the current ED state with the **recommended staffing** (and existing staffing).
6) Calculate **performance** measures.

Simulation experiment – current state (# patients)

n=100 replications, Avg-simulation average, SD-simulation standard deviation, UB=Avg+1.96*SD, LB=Avg-1.96*SD, WIP-number of patients from the database



Part 1: **Intraday staffing in real-time**

Experiment – performance of future shift

Utilization:

I_p - Internal physician

S_p - Surgical physician

O_p - Orthopedic physician

N_u - Nurses.

Used Resources (avg.):

#Beds – Patient's beds,

#Chairs – Patient's chairs.

Service Quality:

%W - % of patients getting physician service within 0.5 hour from arrival (effective of α).

	Utilization						
Hour	Ip	Sp	Op	Nu	#Beds	#Chairs	%W
09-10	73%	1%	23%	55%	15.7	8.6	7%
10-11	93%	25%	59%	68%	23.5	17	33%
11-12	94%	59%	67%	72%	29.3	22.8	51%
12-13	90%	45%	81%	58%	33.2	30.3	53%
13-14	95%	68%	94%	71%	36.2	34.7	77%
14-15	90%	62%	76%	63%	34.2	33.3	70%
15-16	91%	51%	46%	51%	34.4	30.5	77%
16-17	100%	43%	41%	53%	34.6	27.6	69%
17-18	95%	58%	46%	57%	33.4	23.6	52%
18-19	90%	46%	52%	50%	32.4	23.9	31%
19-20	89%	64%	70%	58%	29.3	25.3	40%
20-21	79%	64%	75%	56%	26.5	20.6	39%
21-22	84%	46%	60%	45%	23.4	17	23%
22-23	66%	38%	51%	46%	20.2	13.9	20%

Simulation experiment – staffing recommendations

Staffing levels (current and recommended)

	n (Current)				Offered Load				n (OL)				RCCP Load				n (RCCP)			
Hour	I_p	S_p	O_p	N_u	I_p	S_p	O_p	N_u	I_p	S_p	O_p	N_u	I_p	S_p	O_p	N_u	I_p	S_p	O_p	N_u
16-17	4	1	2	5	7.8	0.8	0.8	4.1	9	2	2	5	3	0.5	0.6	2.4	4	1	1	3
17-18	4	1	2	5	3.7	0.4	0.9	2.5	5	1	2	3	3.3	0.4	0.7	1.3	4	1	1	2
18-19	4	1	2	5	3.2	0.4	1.1	2.7	4	1	2	4	2.3	0.4	0.4	1.3	3	1	1	2
19-20	4	1	2	5	2.3	0.5	1.2	2.5	3	1	2	3	2.4	0.5	0.6	1	3	1	1	2
20-21	4	1	2	5	2.7	0.6	1.5	2.7	4	1	2	4	2.3	0.5	0.4	1	3	1	1	2
21-22	4	1	2	5	2.4	0.4	1.3	2.4	3	1	2	3	2.8	0.5	0.4	1.1	4	1	1	2
22-23	4	1	2	5	2.3	0.2	0.9	2	3	1	2	3	2.4	0.3	0.2	1	3	1	1	2

Simulation experiments – comparison

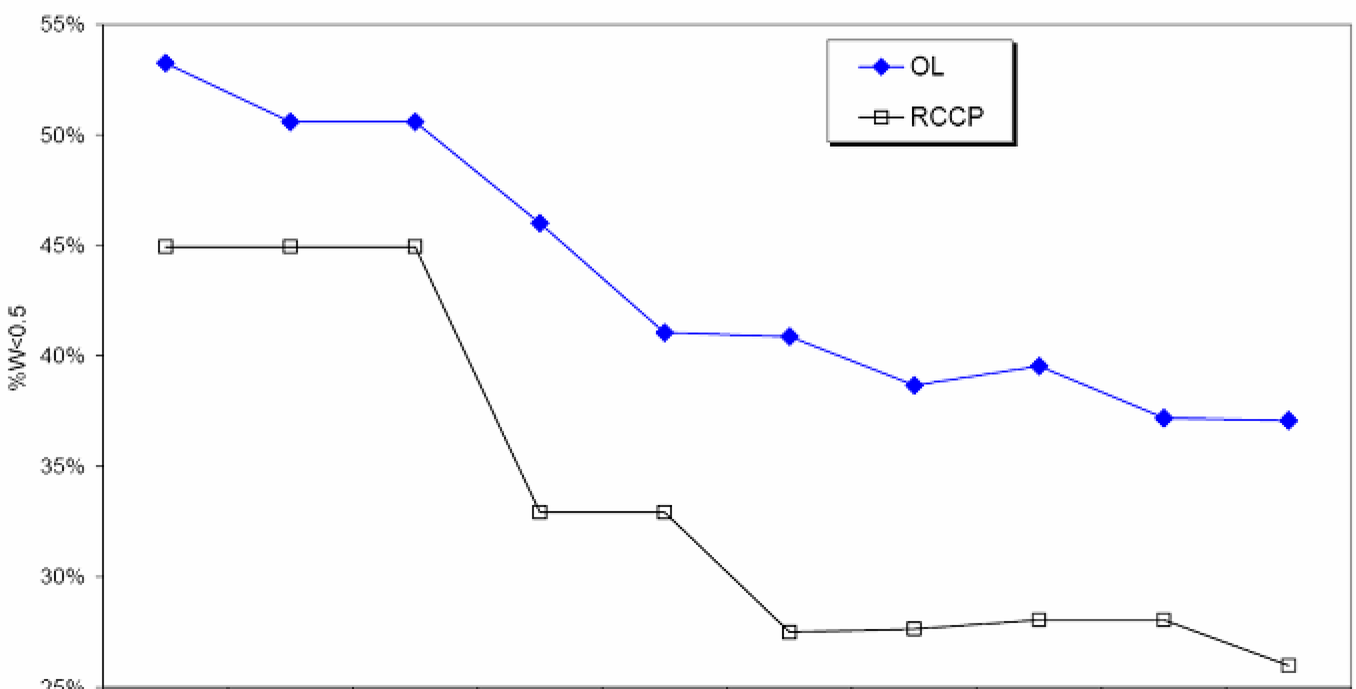
	Performance measures using **OL** recommendation							Performance measures using **RCCP** recommendation						
	Resource Utilization				#Beds	#Chair	%W	Resource Utilization				#Beds	#Chair	%W
Hour	Ip	Sp	Op	Nu				Ip	Sp	Op	Nu			
16-17	62%	38%	40%	58%	36	29	**56%**	90%	54%	60%	59%	38.3	35.3	**78%**
17-18	59%	33%	35%	67%	34.8	31.6	**36%**	82%	47%	65%	81%	39.3	40.2	**82%**
18-19	75%	49%	53%	76%	32.2	29.9	**46%**	80%	45%	69%	92%	40.6	46.2	**86%**
19-20	84%	48%	57%	80%	31.5	31.1	**38%**	72%	43%	79%	97%	42.3	52.2	**90%**
20-21	76%	52%	65%	71%	28.7	28.4	**38%**	68%	46%	85%	99%	43.4	57.7	**91%**
21-22	83%	49%	59%	75%	27.8	27.9	**42%**	55%	45%	89%	99%	44.7	62.4	**91%**
22-23	85%	45%	50%	73%	25.7	25.4	**50%**	63%	39%	87%	99%	45.9	64.9	**91%**

OL method achieved good service quality: *%W* is stable over time.

RCCP method yields good performance of resource utilization - near 90%.

Simulation experiments – comparisons

Comparing RCCP and OL given the same average number of resources



*	3.00	3.00	3.00	2.71	2.71	2.64	2.61	2.54	2.54	2.32
**	3.00	2.93	2.86	2.79	2.68	2.64	2.57	2.54	2.46	2.43
***	5%	9%	11%	14%	32%	34%	39%	51%	66%	90%

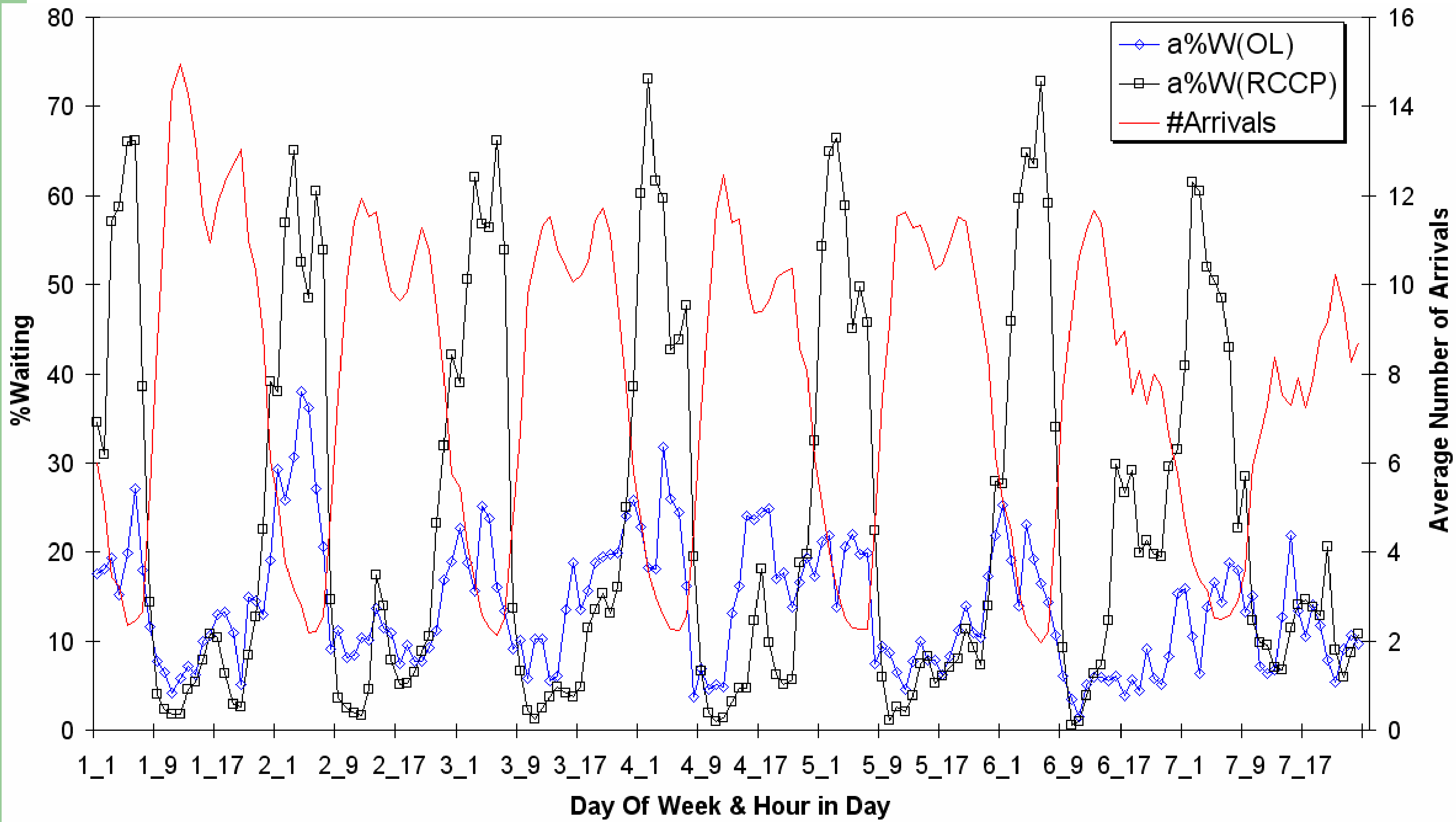
* avg number of resources per hour (RCCP)
** avg number of resources per hour (OL)
*** α

The simulation results are conclusive – OL is superior, implying higher quality of service, with the same number of resources, for all values of α.

Part 2: Intraday staffing over the mid-term

Mid-term staffing: Results

%W (and #Arrivals) per Hour by Method in an Average Week (α = 0.3)



Conclusions and future research

- Developed a staffing methodology for achieving both high utilization and high service levels, over both short- and mid-term horizons, in a highly complex environment (e.g. ED)
- More work needed:
  - Refining the analytical methodology (now the α is close to target around $\alpha = 50\%$).
  - Accommodate constrains (e.g. rigid shifts).
  - Incorporate more refined data (e.g. from RFID).

Part 3: Fitting an efficient operational model to a given ED environment, using Simulation and DEA

With Prof. B. Golany & Prof. A. Mandelbaum

Research problem: matching design to environment (long run)

Current practice: Priority queues at the ED are based on patients' urgency and illness type (e.g. Garcia *et al.*, 1995).

Problem: No account of operational considerations, e.g. relieving overcrowding by accelerating discharges (SPT).

Managerial solution: To use ED design to enforce operational preferences:

- Illness-based
- Triage
- Fast Track
- Walking-Acute

Conclusions (sample)

- There is no dominant operating model for all ED environments.
- EDs exposed to high volume of **elderly** patients need a **Fast-Track** lane for high-priority patients.
- Other EDs (Low volume of elderly patients) can use **Triage-based** priorities.
- If FT and Triage are not feasible options (e.g. no extra nurse is available for Triage or no room for FT), use **Walking-Acute** for differentiation.

- **Future Research:**
  - Adding operational models (e.g. **Output-based** approach and **Specialized-based** approach).

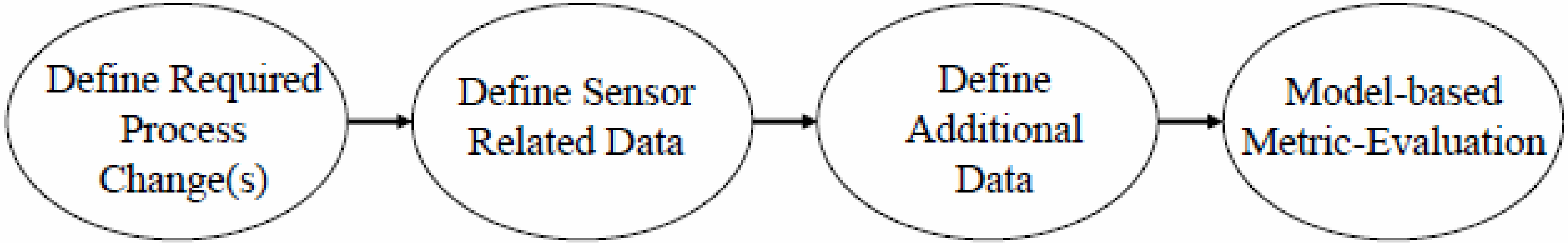
Part 4: long-term benefits of using real-time tracking (RFID) in the ED

With:

Prof. Mandelbaum Prof. Shtub, Dr. Wasserkrug, Dr. Zeltyn
(M.D. Schwartz – ED Manager, Tzafrir – IT Head)

Goal

Present a multi-stage methodology to evaluate the potential benefits of introducing RFID technology, supported by examples of its application (operational, clinical, financial).



Methodology Steps

Questions?