HSMA 4

Module 3: Modelling Pathway and Queueing Problems

Session 3D: A Generic Emergency Department Model Using R

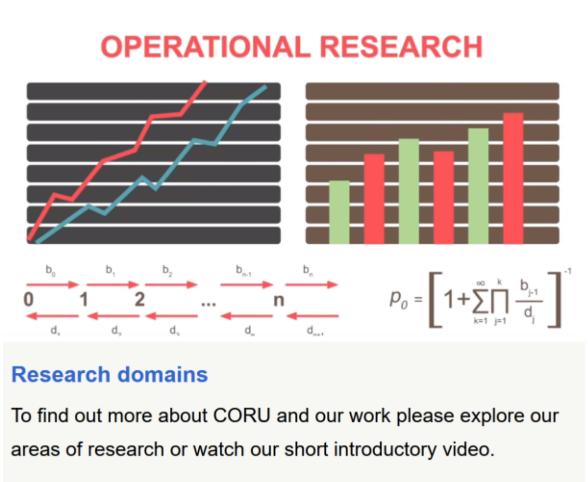
Dr Luca Grieco
UCL Clinical Operational Research Unit

08/12/2021

UCL Clinical Operational Research Unit (CORU)

- Operational Research applied to clinical and healthcare services
- Research unit founded in 1983
- At the forefront of clinical and healthcare O.R. in the UK and the world

http://www.ucl.ac.uk/clinical-operational-research-unit



- Congenital heart disease in children and adults
- Health protection
- Embedded operational researchers within health care teams
- Operational Research to support health services
- Data Science in healthcare and industrial applications

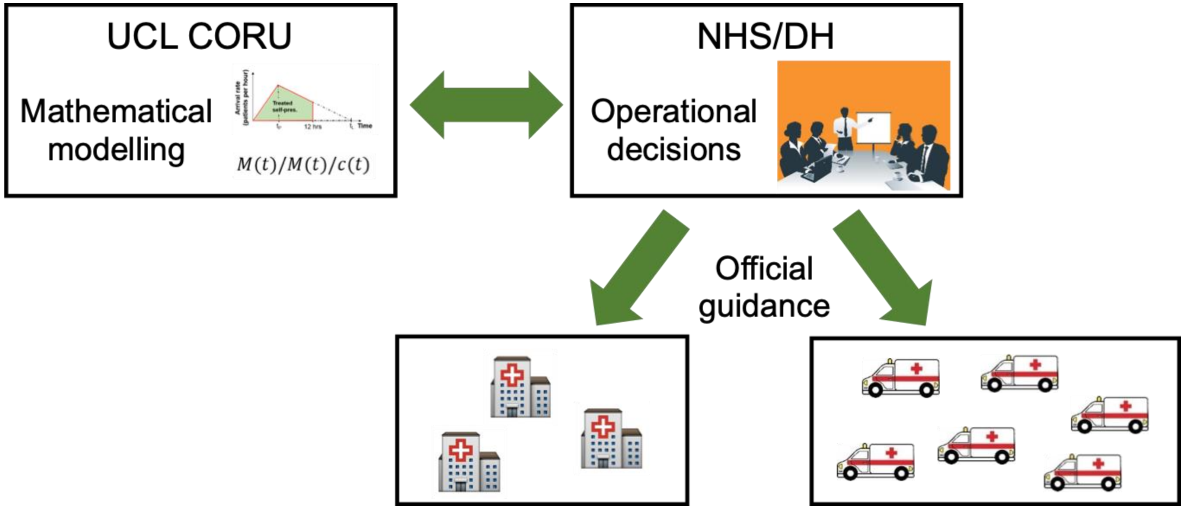
About me

l.grieco@ucl.ac.uk

- MSc Industrial Engineering
- PhD in Bioinformatics
- Joined CORU in 2014 as postdoc – now Lecturer
- Main research activity: apply quantitative methods to improve healthcare systems (single hospitals, NHS)

More info about me: *https://iris.ucl.ac.uk/iris/browse/profile?upi=LGRIE55*

- Optimisation
- Queueing theory
- Stochastic Simulations



Learning objectives for today

- Criteria for choosing a simulation vs analytical approach
- Quick ways to analyse queueing systems
- Conducting scenario analyses
- Interpreting model results

What we will do today

- Simulation vs analytical modelling
- Queueing Theory
- Networks of queueing systems
- Case study: an analytical model for Emergency Department overcrowding
- Practical exercises using R, or just pen & paper

Let’s start simple

A doctor visits, on average, 1.5 patients per hour

Some expert told us that visit time follows an Exponential distribution

Using stochastic simulation, let’s determine:

- The expected visit duration (though this should be obvious)
- The chances (probability) that a visit will take no more than 1 hour

Let’s use RStudio

Solution file: example_exponential.R

Let’s start simple

A doctor visits, on average, 1.5 patients per hour

Some expert told us that visit time follows an Exponential distribution

Using simulation, we have obtained:

- The expected visit duration (though this should be obvious)

  ➡ 0.66… hours

- The chances (probability) that a visit will take no more than 1 hour

  ➡ 77. … %

Though often acceptable, results might (slightly) vary with simulation runs

Solution file: example_exponential.R

But, did we need simulation to do that?

Probabilities distributions are studied a lot by mathematicians!

Let's check what we know about the Exponential distribution:

https://en.wikipedia.org/wiki/Exponential_distribution

In our example, exponential distribution with rate $\lambda = 1.5$

- Expected visit duration (mean value): $\frac{1}{\lambda} = \frac{1}{1.5} = 0.666666 \approx 0.67$
- Probability of visit ≤ 1 hr: $1 - e^{-\lambda \cdot 1} = 1 - e^{-1.5} = 0.7768698 \approx 77.69\%$

We could have actually done that without using R (or Python)!

Exercise 1

1. Check the Wikipedia page of the Normal distribution (https://en.wikipedia.org/wiki/Normal_distribution): look at the probability density function (PDF) and observe how its shape changes at varying values of the theoretical mean value μ and variance σ^2

2. Pick one of those examples and, using R, generate N numbers according to the Normal distribution with the corresponding parameters (use the R function *rnorm()* - note, this function asks you for the standard deviation, which is the square root of the variance)

3. Plot histograms of the data generated above, for different values of N (say 1000, 10000, 1000000)

4. Compare your histograms with the plots of the theoretical distribution shown on Wikipedia

5. Compute the average of your data and compare it with the theoretical mean μ of the Normal distribution

Solution file: exercise1.R

It's very easy to make it too complex

A doctor can visit, on average, 1.5 patients per hour

Some expert told us that visit time follows an Exponential distribution

After the visit, the patient needs to book a follow-up appointment. A nurse does the bookings with a rate of 3 per hour. We assume that the appointment booking time follows a Normal distribution with standard deviation 1.

Assuming no queues, what's the expected time a patient takes to go through the whole process?

- Very easy to do it using a simulation approach
- Ready-to-use formulas very difficult to determine in this case (probability distribution of $X + Y$, with X following an Exponential distribution and Y following a Normal distribution)

Simulation vs Analytical modelling

There aren't strict rules to determine which modelling approach to use

Things to consider:

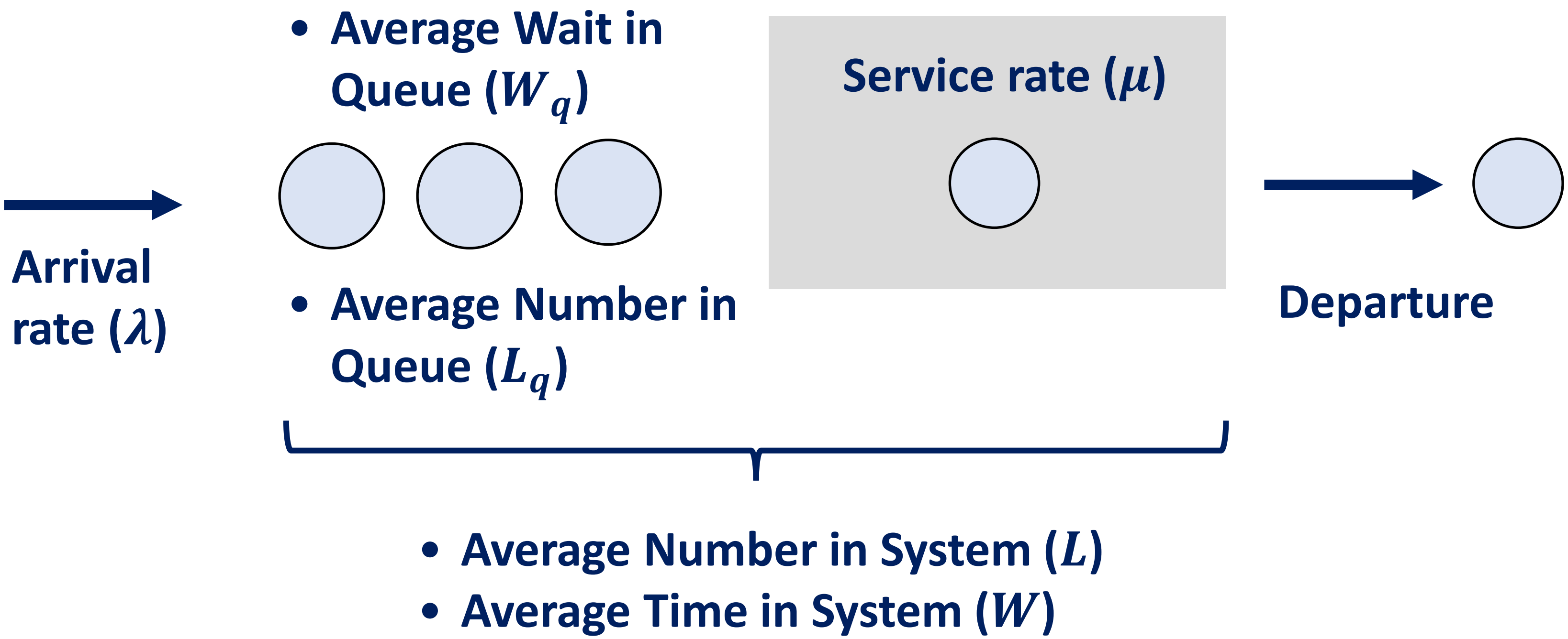
Analytical modelling

- Ready-to-use formulas (exact solution) – just change parameter values to get results in different conditions
- Limits on assumptions that can be made – e.g. X+Y easy if both Exponential or both Normal; very difficult if one Exponential and one Normal

Stochastic simulation

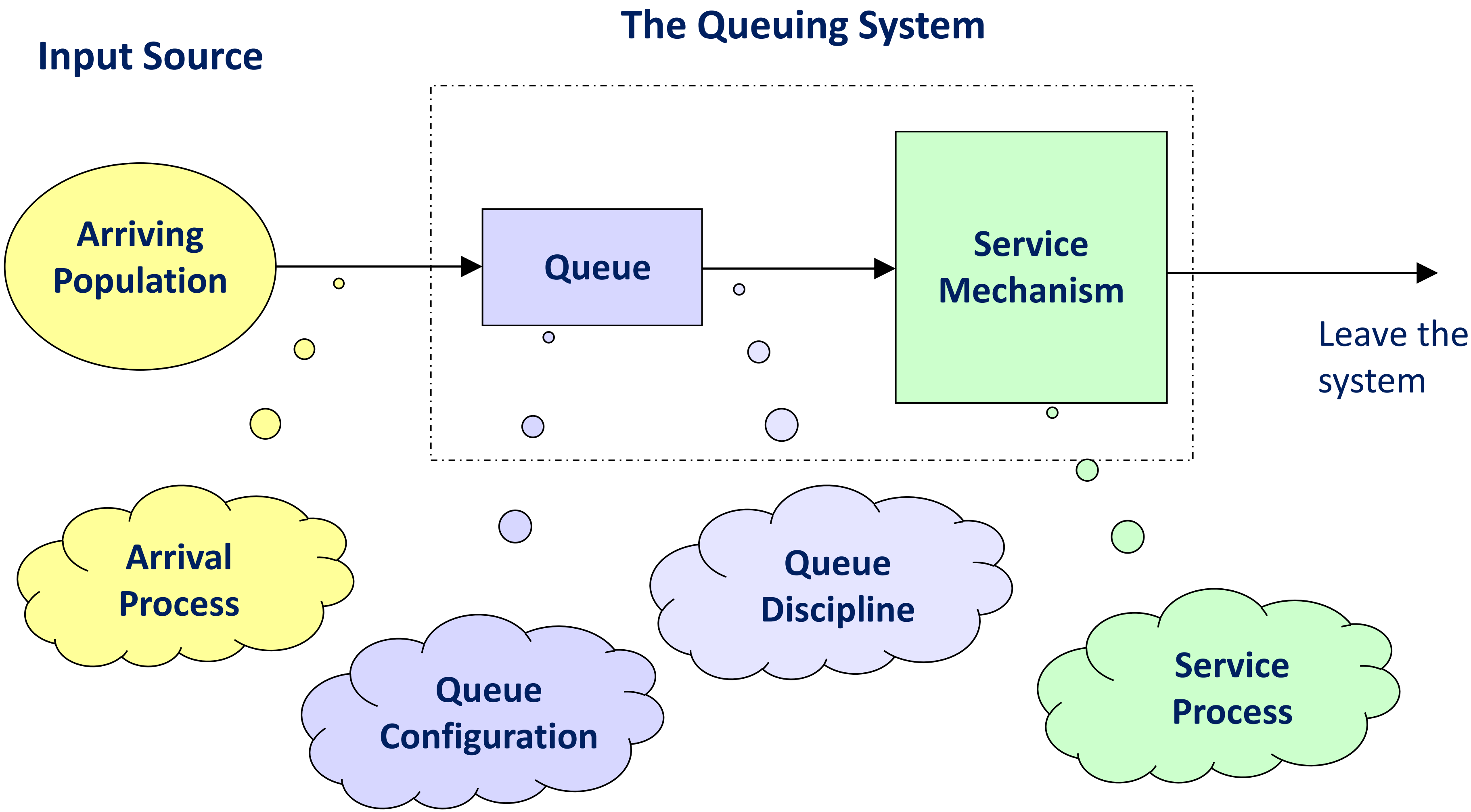
- We can model very complex systems without being worried of how difficult the underlying maths would become (though beware of computational complexity!)
- Need to run the model many times to account for variability and compensate for the fact that randomisation leads to (slightly) different results when repeating the analysis

Queueing systems



- Customers (people or items) arrive at an average rate λ, are served by a server for an average length of time ${}^{1}/_{\mu}$, and then leave
- We can compute performance measures such as the average waiting time in queue, the average time in the system, the average number (of customers) in the queue and the average number in the system

Aspects of a queueing system



Queueing Theory studies the aspects of queueing systems using mathematics

Arrivals

- Arriving population
  - Finite or infinite
  - Homogeneous or heterogeneous (e.g. different types of customers)
- Arrival process
  - Probability distribution of inter-arrival times (time elapsed between two consecutive arrivals)
  - Possible arrivals in groups (“batch arrivals”)

Queue

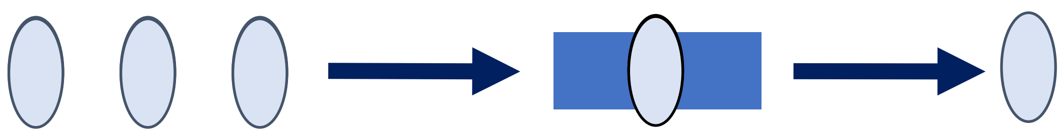
- Queue configuration
  - Finite or infinite queue length
  - Single or multiple queues (based on customer types)

- Queue discipline
  - How customers are selected from the queue:
    - First come first served
    - Shortest processing time
    - Earliest due date
  - Customer priority classes

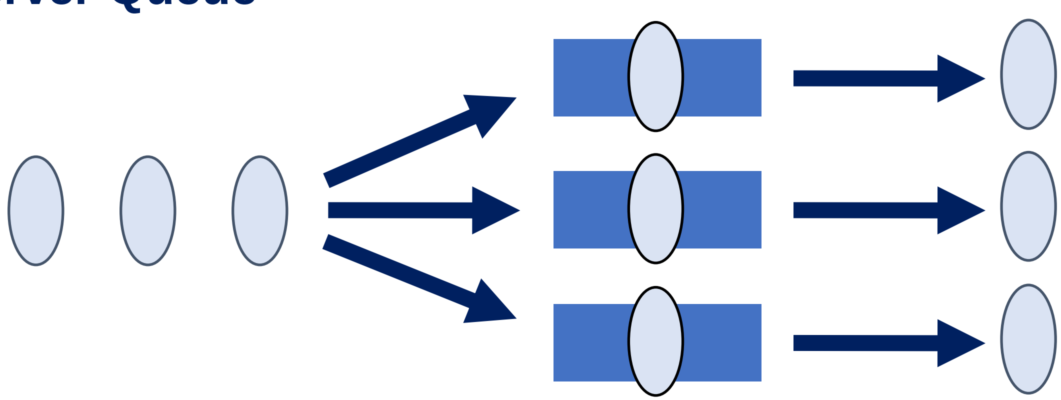
Service

- Service process
  - Probability distribution of service times
  - Number of available servers

Single Server Queue

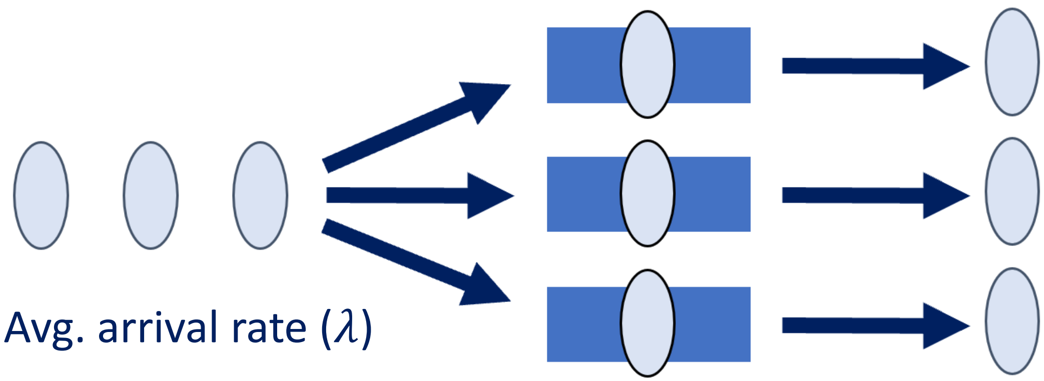


Multiple Server Queue



A first basic notion: traffic intensity

A measure of average resource utilisation in a queueing system



Avg. arrival rate (λ)

Avg. service rate (μ)

Number of servers (s)

Traffic intensity

$$\rho = \frac{\lambda}{s \cdot \mu}$$

$$= \frac{\textit{Rate of arrivals}}{\textit{Rate of clearance}}$$

For instance, for a ward we could measure the average utilisation of beds using ρ:

$\boldsymbol{\lambda}$	**LoS = 1 / $\boldsymbol{\mu}$**	$\boldsymbol{s}$	$\boldsymbol{\rho}$	
10 patients/day	4 days	50 beds	0.8	< 1 → Not too utilised
15 patients/day	4 days	50 beds	1.2	> 1 → Over-utilised
10 patients/day	4 days	32 beds	1.25	> 1 → Over-utilised

<u>On average though</u>!

Typical assumptions in Queueing Theory

- Inter-arrival times following an Exponential distribution with rate λ
- Service times following an Exponential distribution with rate μ
- A given number of servers s

These are also known as **M/M/s** queueing systems

We have some ready-to-use formulas giving us information on the long-term behaviour of these queueing systems

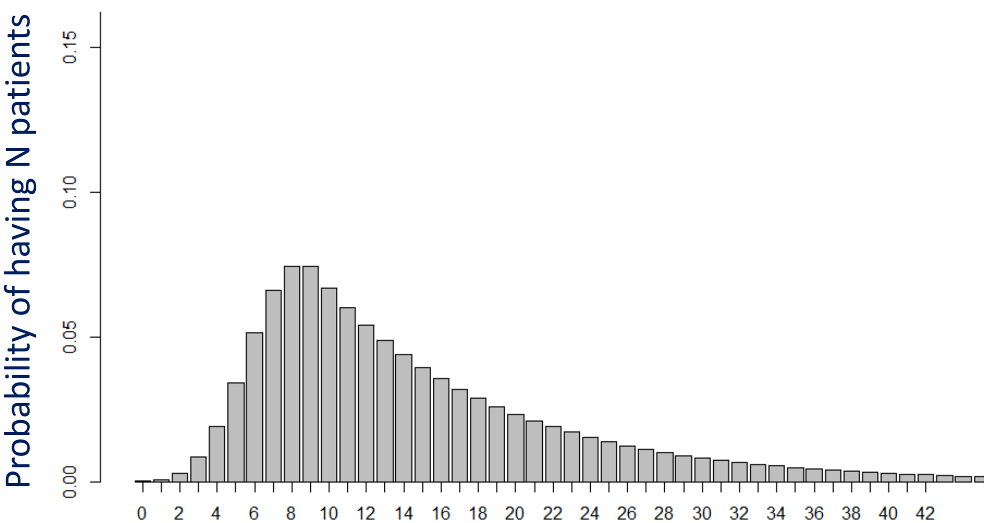
These formulas are only valid under the condition that the traffic intensity is less than 1 ($\rho < 1$)

If this condition is not met, then the system cannot be stable on the long term, i.e. the queue will “explode” over time (customers arriving too quickly for the system to deal with them)

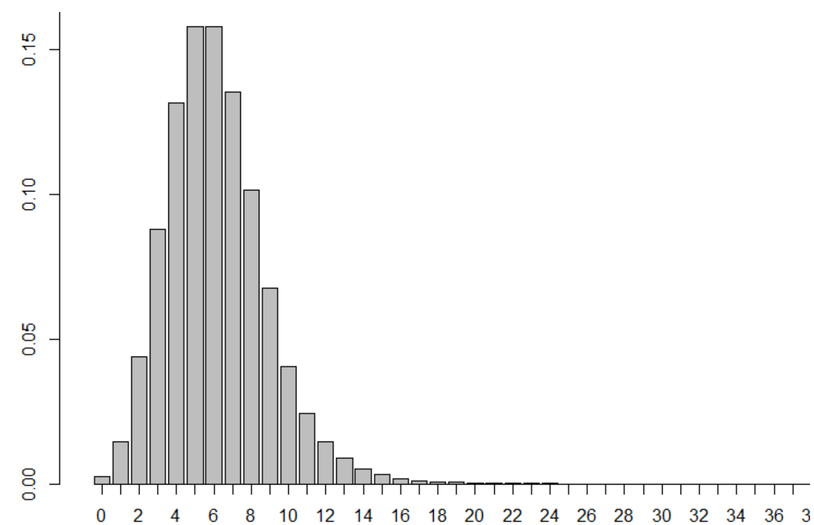
Useful formulas for M/M/s systems

Traffic intensity less than 1 ($\rho < 1$)

We have formulas for the probabilities of having 0, 1, 2, 3, …. customers in the system at any time in the long term



λ	LoS = 1 / μ	s	ρ
4.5 pat/day	2 days	10 beds	0.9



λ	LoS = 1 / μ	s	ρ
3 pat/day	2 days	10 beds	0.6

Useful formulas for M/M/s systems

Traffic intensity less than 1 ($\rho < 1$)

Let's just see the case of M/M/1 (single server)

	In the queue	In the system (queue + service)
Expected number of customers	$L_q = \dfrac{\rho^2}{1-\rho}$	$L = \dfrac{\lambda}{\mu - \lambda}$
Expected time spent by a customer	$W_q = \dfrac{\rho}{\mu - \lambda}$	$W = \dfrac{1}{\mu - \lambda}$

Formulas a bit more complex for $s > 1$, but you'll get the code!

Exercise 2

In your groups, discuss the following two cases and try and answer the questions asked. Compare the two models and the results obtained. In both cases, assume both the inter-arrival times and the service times follow an Exponential distribution.

Case 1 – Drop in blood test

Your clinic provides blood tests. You expect, on average, 100 patients to drop in every hour. On average it takes about 7.5 minutes to take a blood sample. You can deploy up to 20 phlebotomists due to space constraints. How many phlebotomists would you deploy at any given time?

Case 2 – Intensive care beds

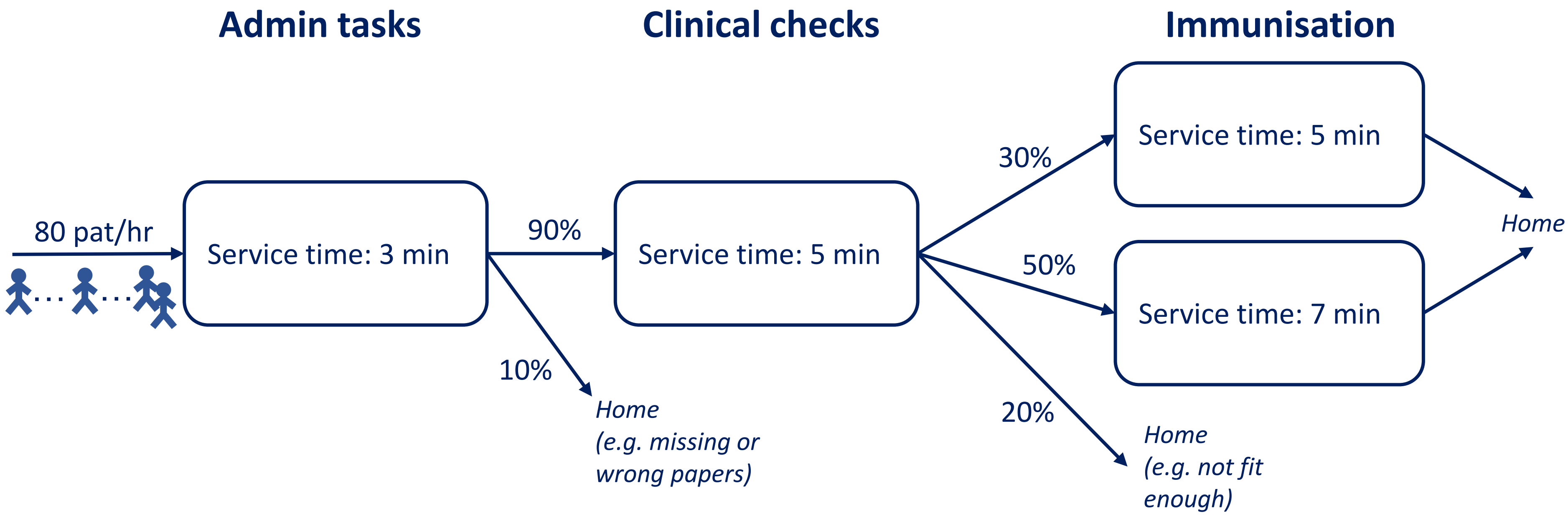
You manage an intensive care unit characterised by an average demand of 5 patients per day. Average length of stay for a patient in the unit is 2.5 days. The unit can host at most 20 staffed beds. How many beds would you want to have for the unit?

Solution file: exercise2.R

Queueing networks

Can Queueing Theory help when we have a network of queueing systems?

Let's consider an example about a COVID-19 mass vaccination centre:



- Suppose the Admin task step can be modelled as an M/M/s system
- Also suppose all service times above follow an Exponential distribution

Queueing networks

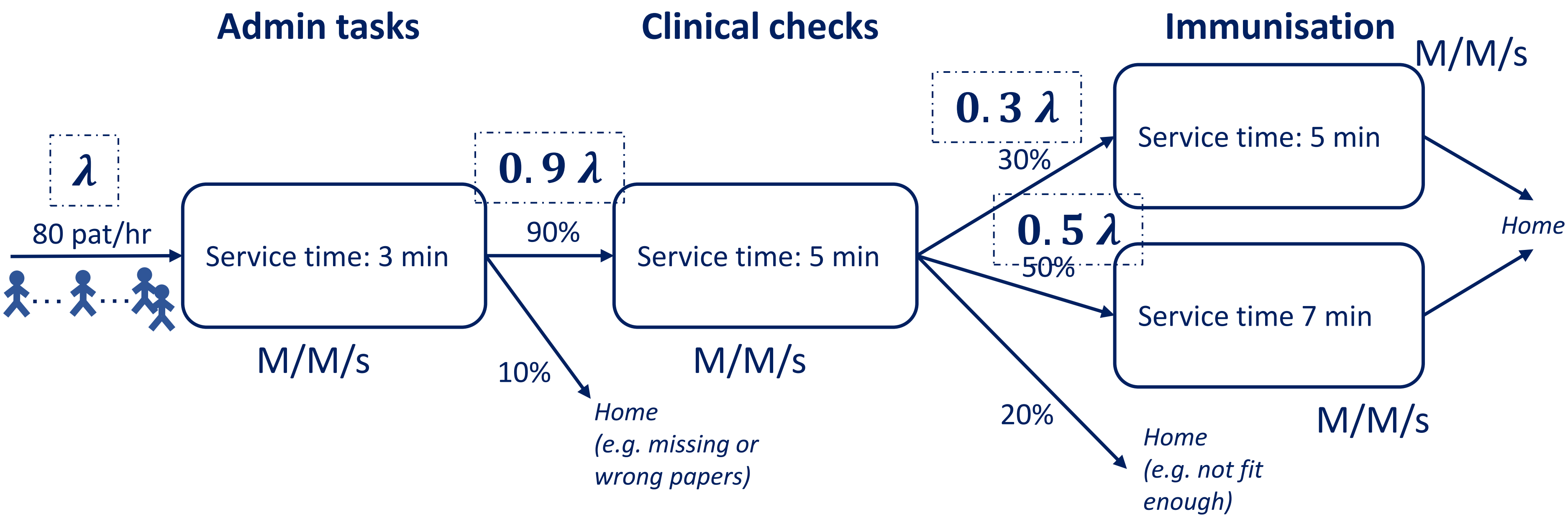
Admin tasks

Inter-arrival times: Exponential distribution
Service times: Exponential distribution
--> M/M/s

We also need $\rho < 1$

- Theory tells us that patients will leave this step with inter-departure times that have the same distribution as the inter-arrival times, with the rate proportional to the chance of following each pathway
- **So, we can use the Exponentially distributed departure rates from this step as arrival rates to the following step!**

Queueing networks



We can analyse the 4 steps above separately, as independent queueing systems!
(e.g. decide about the capacity for each queue independently as we did in Exercise 2)

What if we change our assumptions?

- Exponential distribution assumptions might be unrealistic sometimes (e.g. lognormal distribution often more appropriate for service times)
- In general, difficult to derive formulas for other probability distributions
- Also, in some systems, traffic intensity varies a lot over time so the formulas we saw earlier cannot be used directly

→ Stochastic simulation (seen with Dan)

→ Approximation methods (today)

University College London Hospitals NHS
NHS Foundation Trust

Collaboration for Leadership in
Applied Health Research and Care
North Thames

Case Study:
Reducing Emergency Department overcrowding

UCLH Embedded Research Team

Project overview

What is Emergency Department (ED) overcrowding

ED function disrupted mainly due to number of patients waiting to be seen at different steps of their visit. This negatively impacts ED performance targets, such as the 4-hour target.

Aim

To determine which evidence-based interventions are likely to be most effective and feasible for UCLH, given its context-specific problems and capacity for change.

Features of the project

- Timeline: April 2016 – November 2017
- Multidisciplinary study: ethnographic research, analytical modelling and soft systems methodology
- Strong involvement of ED staff

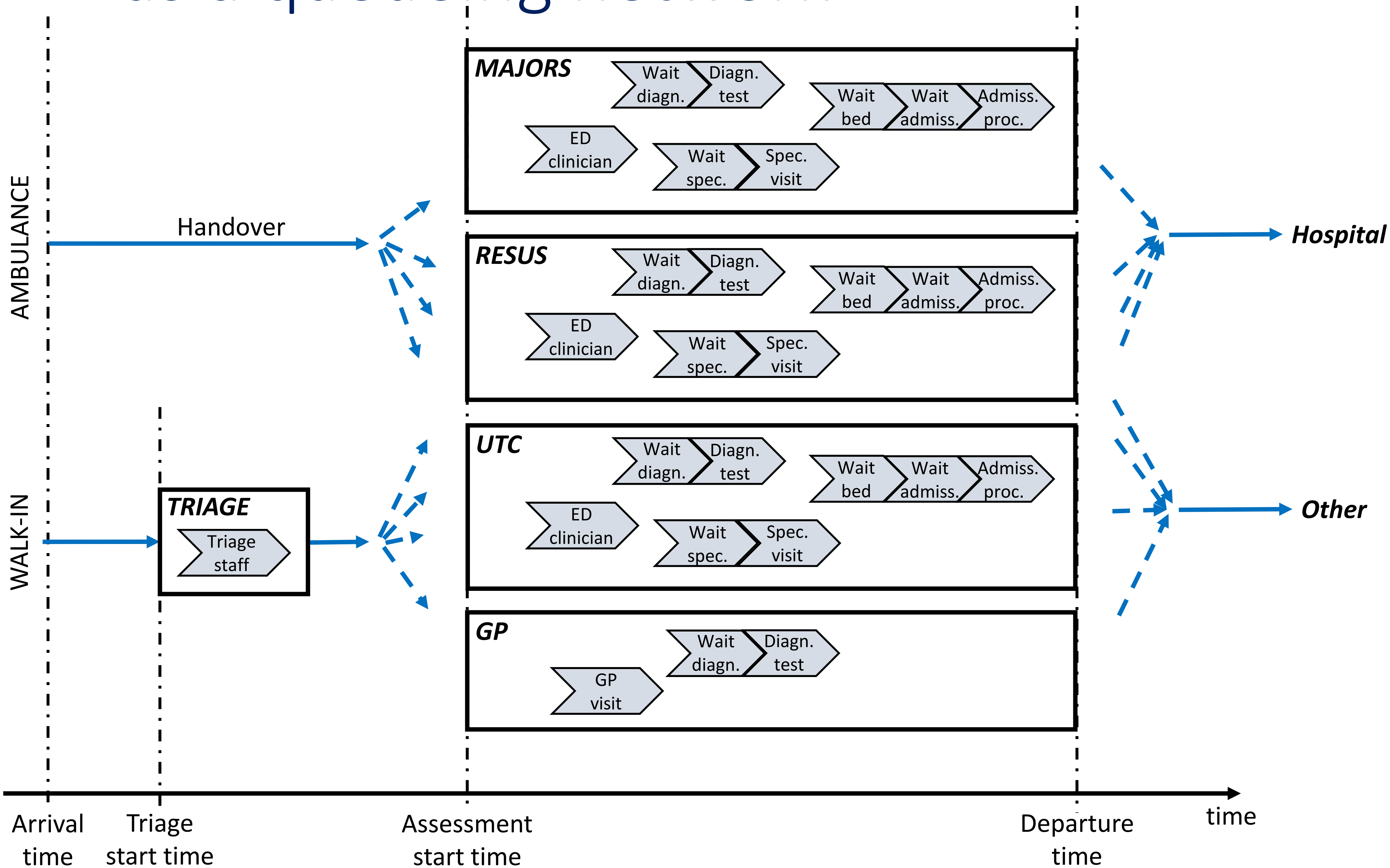
Approach

Qualitative research
Organisational practices and perceptions of staff through interviews with ED and non-ED staff, observations and staff shadowing.

Mathematical modelling
Using queuing theory to explore the intrinsic limits and drivers of ED performance. Informed by the qualitative research, discussions with ED staff and data analysis.

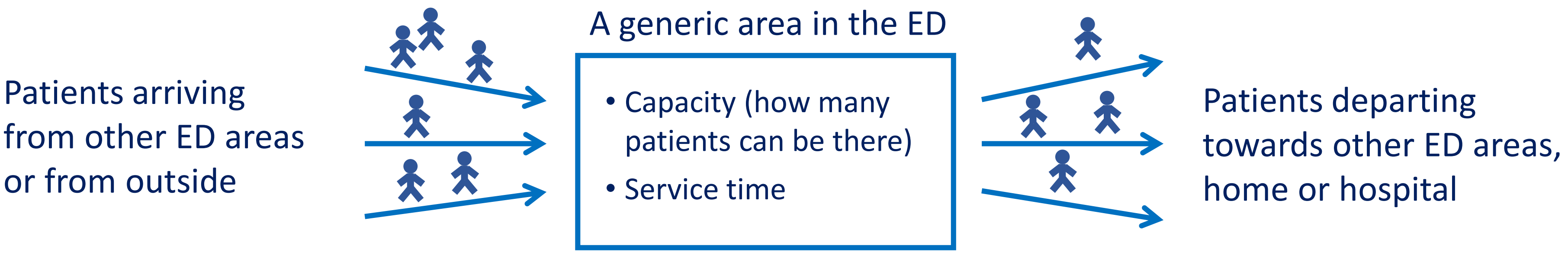
Evidence-based interventions
Identified from literature. Their impact and feasibility for UCLH was assessed based on our findings from the qualitative research and mathematical modelling.

ED as a queueing network



ED as a queueing network

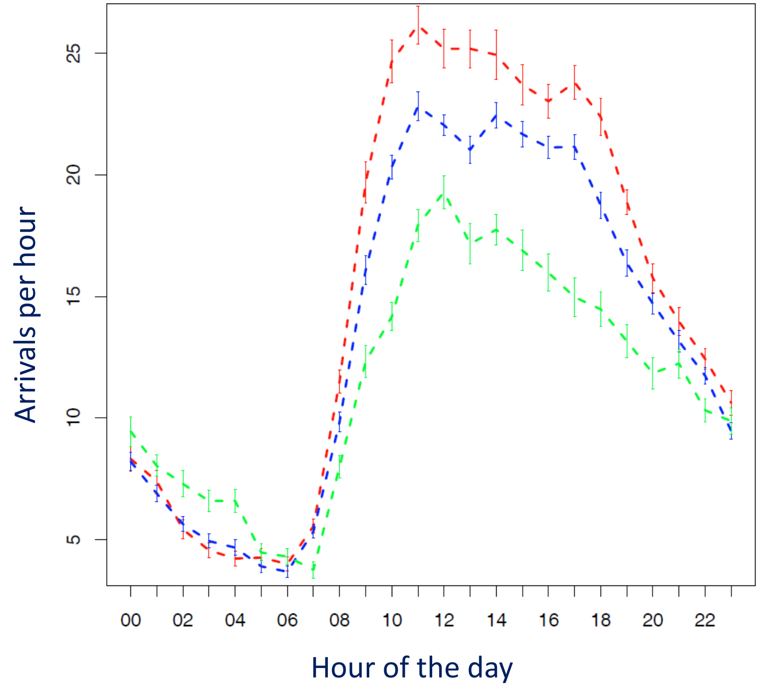
- Modelling framework based on Queueing Theory



- Output from the model:
  - % patients leaving within 4 hours
  - time to initial assessment (triage)
  - time to treatment (first seen by ED clinician or specialist)
  - total sojourn time
  - number of patients in the ED at a given time

Dealing with time-varying parameters

- $\rho > 1$ very often
- Interested in what happens in the system at different times in the day



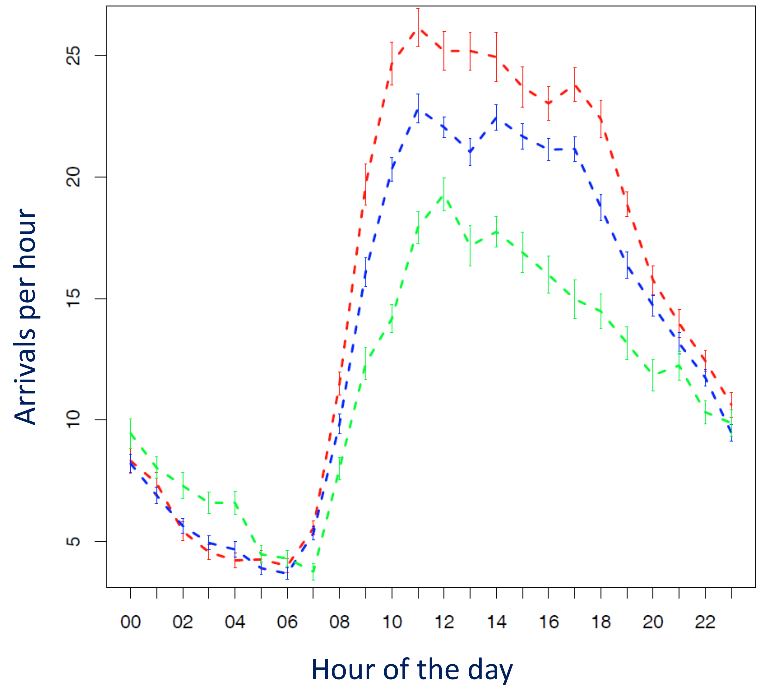
Approximation model of ED accounting for time-varying parameters: $M(t)/M(t)/s(t)$

λ_1	λ_2	λ_3	λ_{24}
M/M/s	M/M/s	M/M/s	M/M/s

time

Dealing with time-varying parameters

- $\rho > 1$ very often
- Interested in what happens in the system at different times in the day

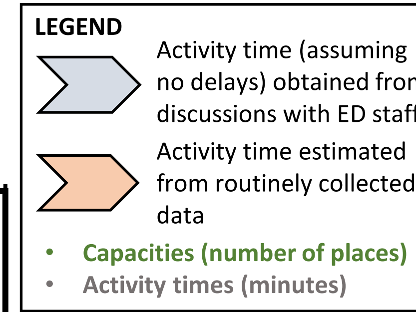


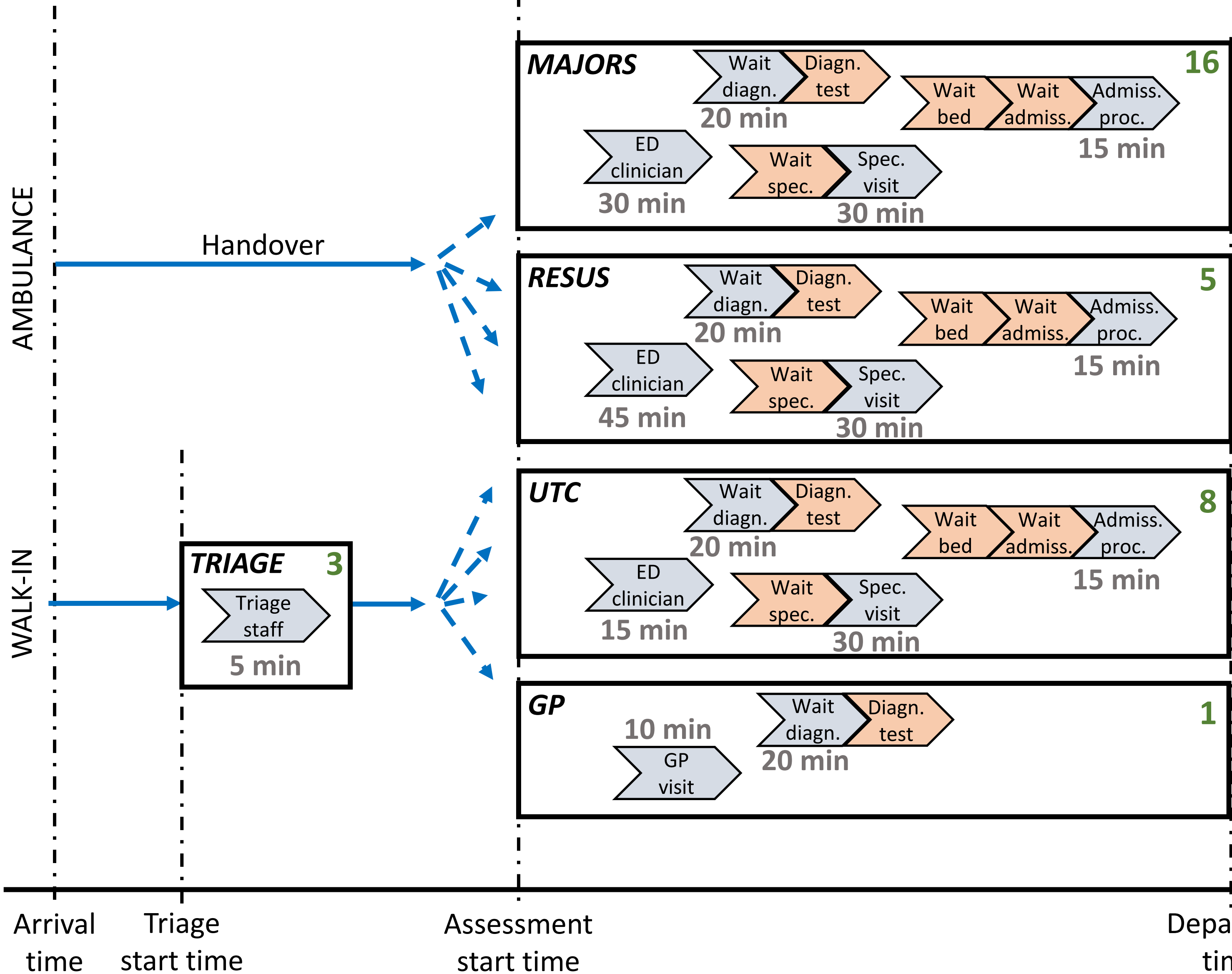
Approximation model of ED accounting for time-varying parameters: $M(t)/M(t)/s(t)$

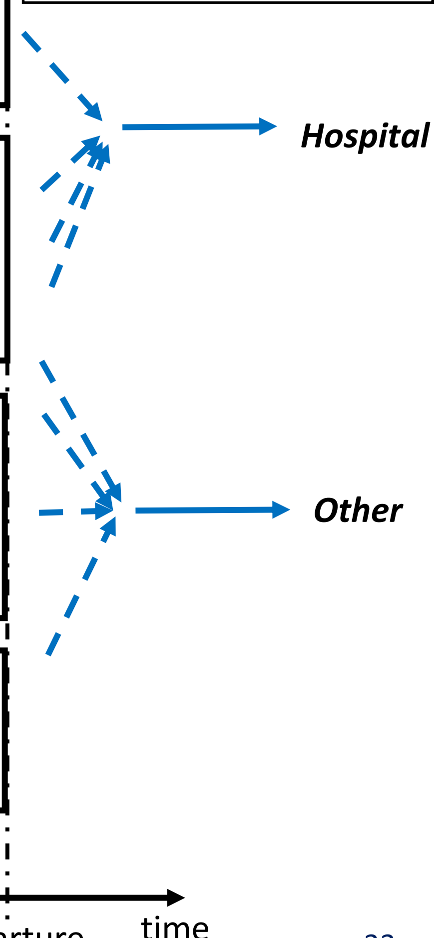
λ_1	$\lambda_2 + b_1$	$\lambda_3 + b_2$	$\lambda_{24} + b_{23}$
M/M/s	M/M/s	M/M/s	M/M/s

time

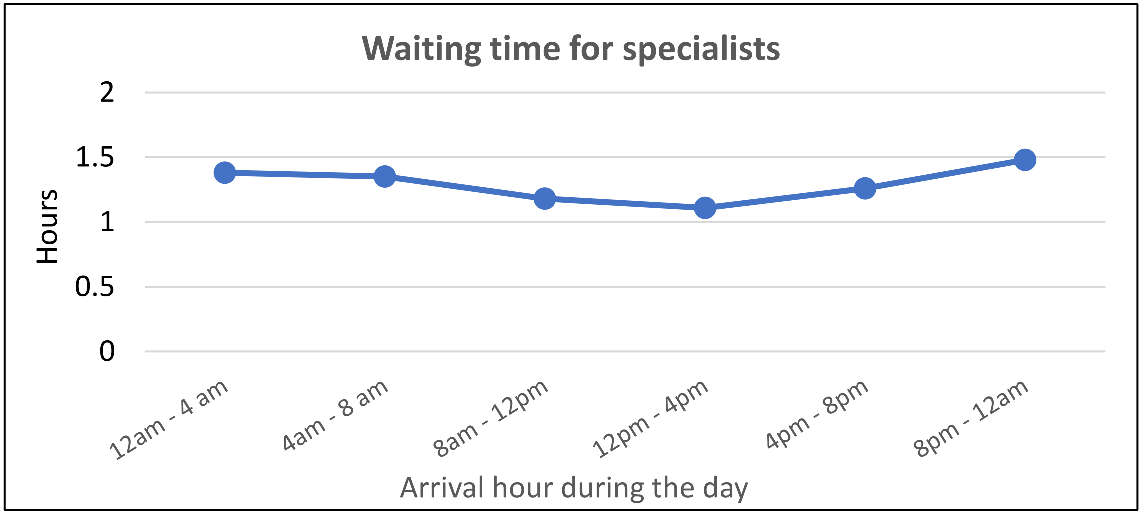
Capacities and activity times

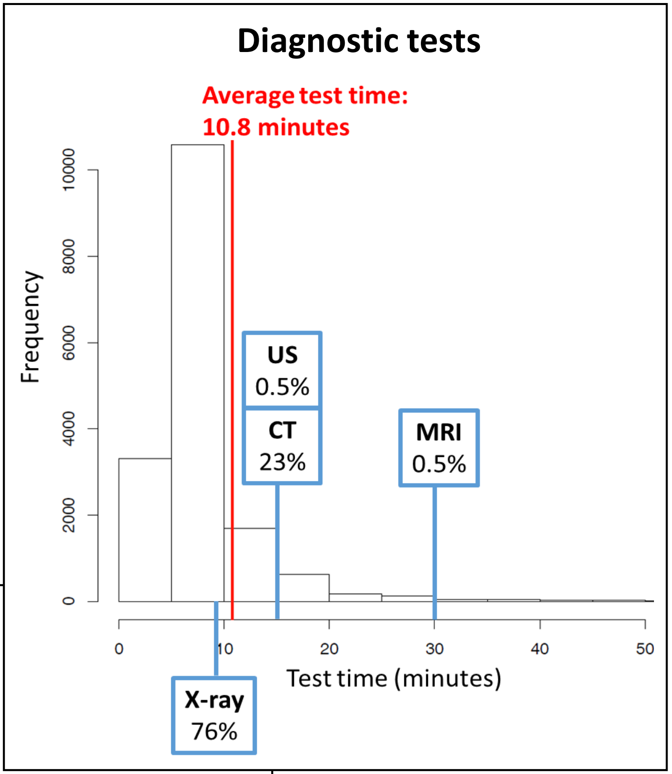


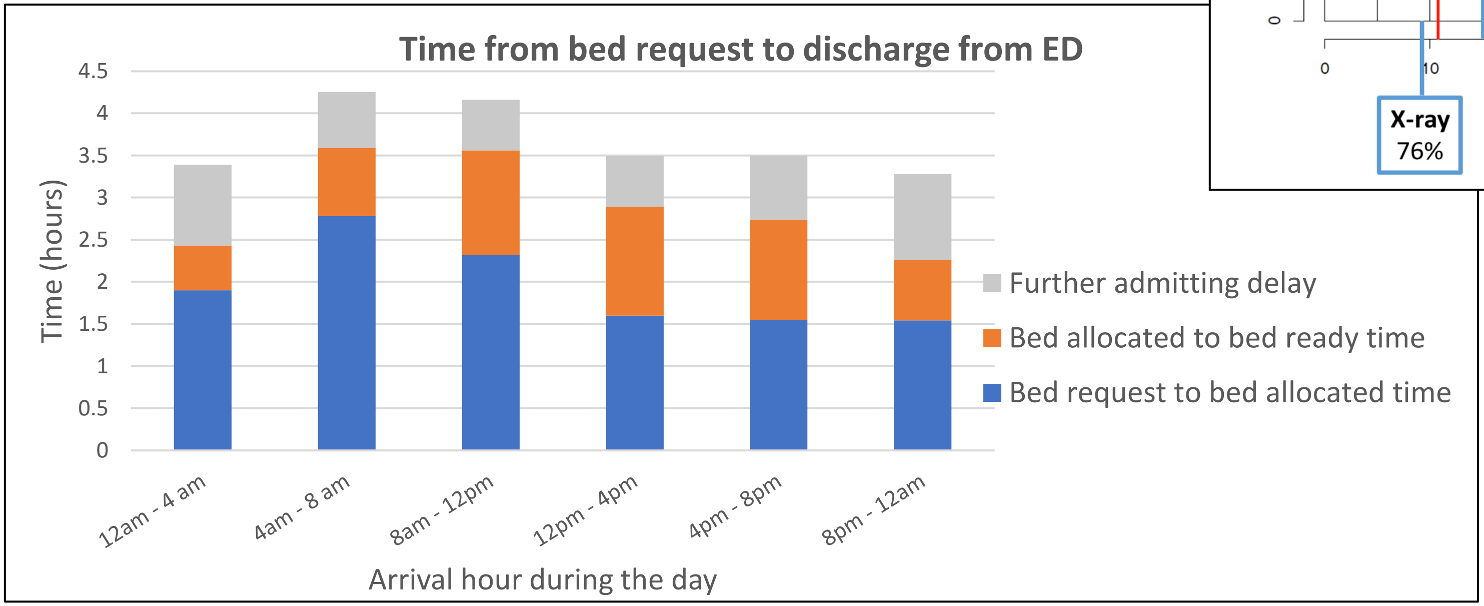




Exogenous waiting times

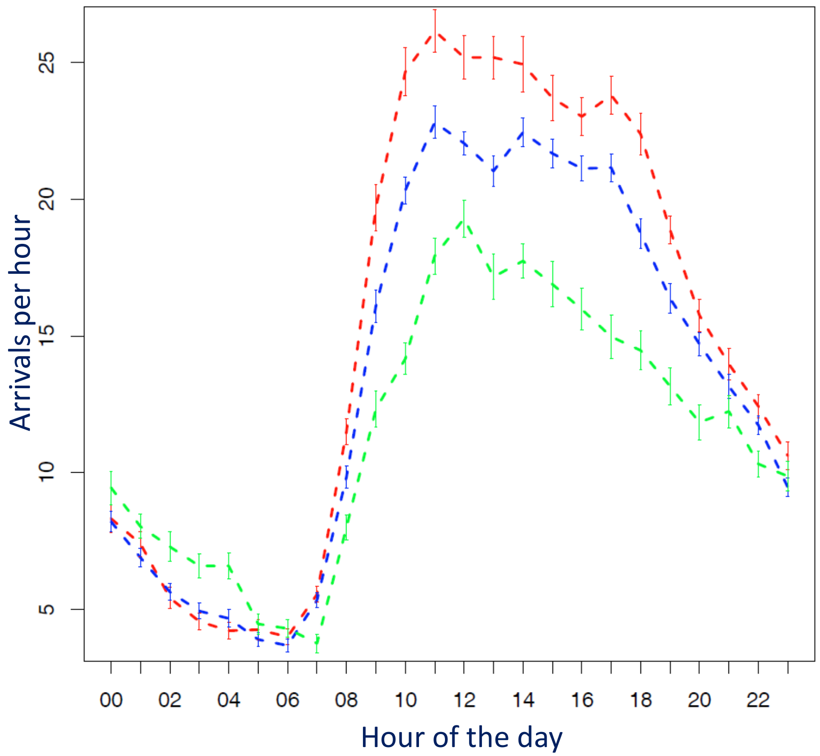






Arrival rates during the day

Estimated from routinely collected data

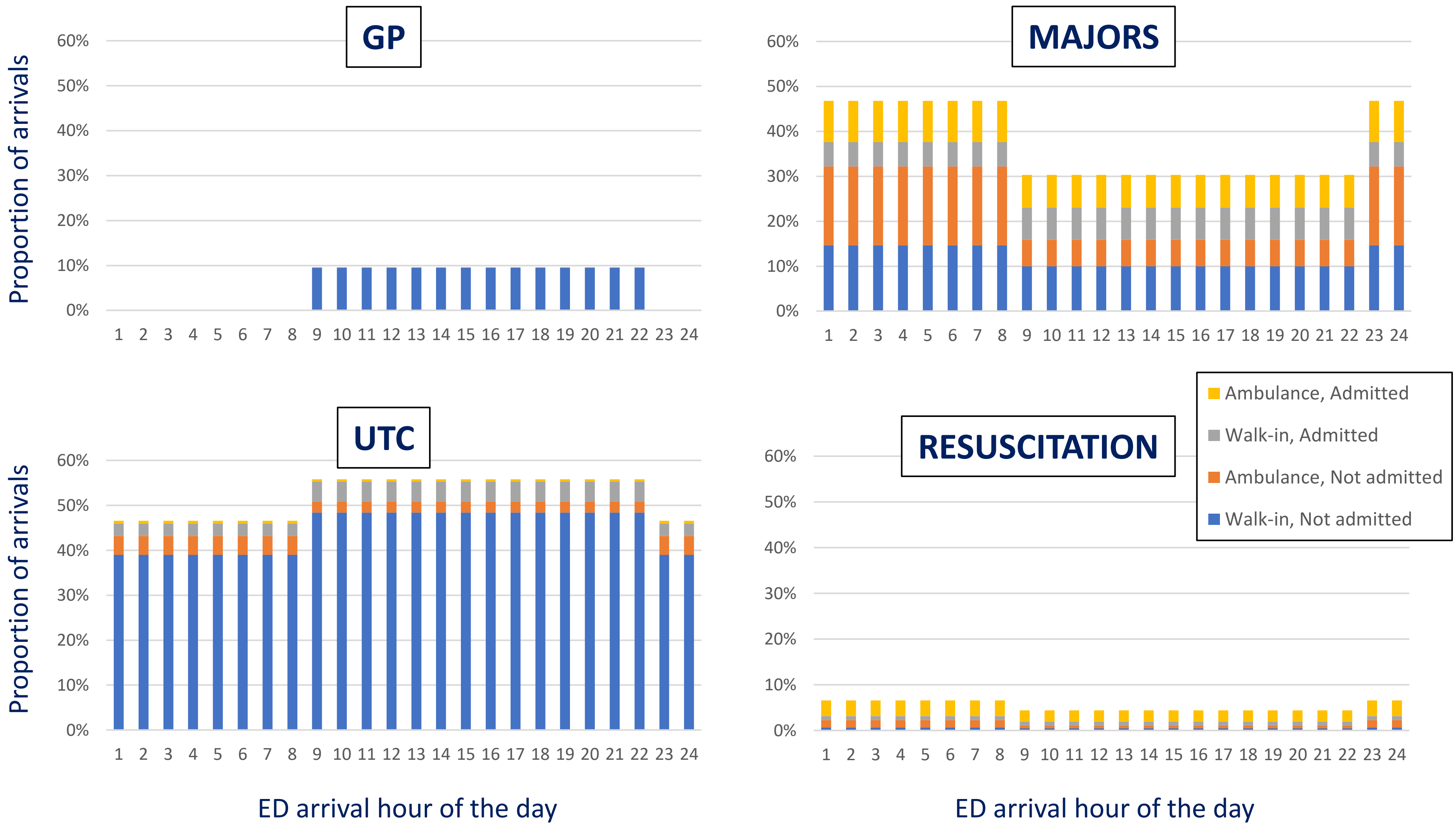


Low arrival rates → 25% of days with smallest number of arrivals

Reference arrival rates → 50% of days with intermediate number of arrivals

High arrival rates → 25% of days with biggest number of arrivals

Patient mix


GP
MAJORS
UTC
RESUSCITATION
Proportion of arrivals
Proportion of arrivals
60%
50%
40%
30%
20%
10%
0%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
ED arrival hour of the day
ED arrival hour of the day
Ambulance, Admitted
Walk-in, Admitted
Ambulance, Not admitted
Walk-in, Not admitted

Exercise 3

- Consider the case and model described above
- Imagine the ED manager approached you saying that they are struggling with a constantly overcrowded ED with only 82% of patients leaving the department within 4 hours (the target is 95%)
- They would like to know what intervention(s) would improve the situation. Possible interventions they mentioned are: 1) reducing arrival rates for patients with minor injuries (they would like to reduce UTC arrival rates by 50%); 2) doubling beds for MAJORS and RESUS patients; 3) trying to halve delays caused by unavailability of hospital beds and specialists.
- Working in your groups:
  - ➢ Discuss the feasibility in practice of the above interventions
  - ➢ Test the interventions, either alone or in combination, using the R Shiny app provided
  - ➢ What advice would you give to the ED manager? What should they prioritise?

Model analysis and results

Scenario		% Patients leaving ED within 4 hrs				
		GP	UTC	Majors	Resus	Overall
Baseline scenario		100%	86%	75%	66%	82%
ED arrival rate *(Current: 12.5 patients per hr)*	Low	100%	95%	82%	68%	90%
	High	100%	77%	62%	60%	73%
Average waiting time for a hospital bed *(Current: 160 mins)*	60 mins	100%	98%	91%	82%	95%
	Halved	100%	96%	89%	80%	93%
	Doubled	100%	72%	45%	45%	64%
Average waiting time for a specialist assessment *(Current: 74 mins)*	30 mins	100%	92%	80%	68%	88%
	Halved	100%	91%	80%	68%	87%
	Doubled	100%	80%	69%	63%	77%
Average waiting time for diagnostic tests *(Current: 20 mins)*	Halved	100%	91%	77%	67%	86%
	Doubled	100%	81%	73%	64%	79%
Average waiting time for admission (e.g. waiting for a porter) *(Current: 5 mins)*	Halved	100%	87%	76%	66%	83%
	Doubled	100%	86%	74%	65%	82%

Model analysis and results

Scenario		% Patients leaving ED within 4 hrs				
		GP	UTC	Majors	Resus	Overall
Baseline scenario		100%	86%	75%	66%	82%
ED arrival rate *(Current: 12.5 patients per hr)*	Low	100%	95%	82%	68%	90%
	High	100%	77%	62%	60%	73%
Average waiting time for a hospital bed *(Current: 160 mins)*	60 mins	100%	98%	91%	82%	95%
	Halved	100%	96%	89%	80%	93%
	Doubled	100%	72%	45%	45%	64%
Average waiting time for a specialist assessment *(Current: 74 mins)*	30 mins	100%	92%	80%	68%	88%
	Halved	100%	91%	80%	68%	87%
	Doubled	100%	80%	69%	63%	77%
Average waiting time for diagnostic tests *(Current: 20 mins)*	Halved	100%	91%	77%	67%	86%
	Doubled	100%	81%	73%	64%	79%
Average waiting time for admission (e.g. waiting for a porter) *(Current: 5 mins)*	Halved	100%	87%	76%	66%	83%
	Doubled	100%	86%	74%	65%	82%

- Hospital bed availability appears to be the external factor driving ED overcrowding, and it strongly influences performance in MAJORS/RESUS

Model analysis and results

Scenario		% Patients leaving ED within 4 hrs				
		GP	UTC	Majors	Resus	Overall
Baseline scenario		100%	86%	75%	66%	82%
ED arrival rate *(Current: 12.5 patients per hr)*	Low	100%	95%	82%	68%	90%
	High	100%	77%	62%	60%	73%
Average waiting time for a hospital bed *(Current: 160 mins)*	60 mins	100%	98%	91%	82%	95%
	Halved	100%	96%	89%	80%	93%
	Doubled	100%	72%	45%	45%	64%
Average waiting time for a specialist assessment *(Current: 74 mins)*	30 mins	100%	92%	80%	68%	88%
	Halved	100%	91%	80%	68%	87%
	Doubled	100%	80%	69%	63%	77%
Average waiting time for diagnostic tests *(Current: 20 mins)*	Halved	100%	91%	77%	67%	86%
	Doubled	100%	81%	73%	64%	79%
Average waiting time for admission (e.g. waiting for a porter) *(Current: 5 mins)*	Halved	100%	87%	76%	66%	83%
	Doubled	100%	86%	74%	65%	82%

- Hospital bed availability appears to be the external factor driving ED overcrowding, and it strongly influences performance in MAJORS/RESUS
- ED performance is also very sensitive to arrival rates, though variations in arrival rates seem to mainly affect UTC

Model analysis and results

Scenario		% Patients leaving ED within 4 hrs				
		GP	UTC	Majors	Resus	Overall
Baseline scenario		100%	86%	75%	66%	82%
ED arrival rate *(Current: 12.5 patients per hr)*	Low	100%	95%	82%	68%	90%
	High	100%	77%	62%	60%	73%
Average waiting time for a hospital bed *(Current: 160 mins)*	60 mins	100%	98%	91%	82%	95%
	Halved	100%	96%	89%	80%	93%
	Doubled	100%	72%	45%	45%	64%
Average waiting time for a specialist assessment *(Current: 74 mins)*	30 mins	100%	92%	80%	68%	88%
	Halved	100%	91%	80%	68%	87%
	Doubled	100%	80%	69%	63%	77%
Average waiting time for diagnostic tests *(Current: 20 mins)*	Halved	100%	91%	77%	67%	86%
	Doubled	100%	81%	73%	64%	79%
Average waiting time for admission (e.g. waiting for a porter) *(Current: 5 mins)*	Halved	100%	87%	76%	66%	83%
	Doubled	100%	86%	74%	65%	82%

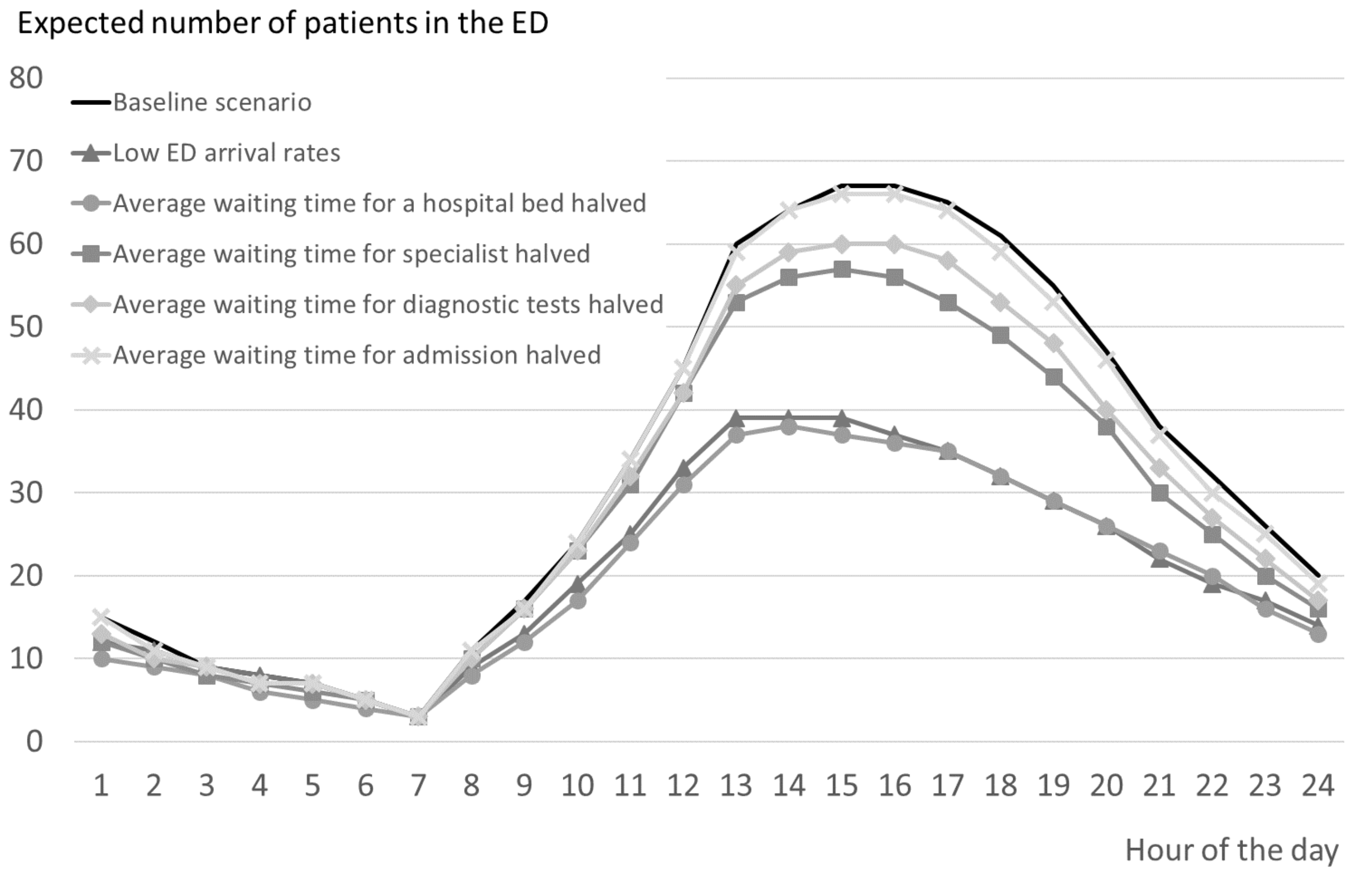
- Hospital bed availability appears to be the external factor driving ED overcrowding, and it strongly influences performance in MAJORS/RESUS
- ED performance is also very sensitive to arrival rates, though variations in arrival rates seem to mainly affect UTC
- Therefore, delays in UTC seem to be more related to ED causes (capacity issues), whereas delays in MAJORS/RESUS seem to be heavily influenced by hospital bed availability

Model analysis and results

Scenario		% Patients leaving ED within 4 hrs				
		GP	UTC	Majors	Resus	Overall
Baseline scenario		100%	86%	75%	66%	82%
ED arrival rate *(Current: 12.5 patients per hr)*	Low	100%	95%	82%	68%	90%
	High	100%	77%	62%	60%	73%
Average waiting time for a hospital bed *(Current: 160 mins)*	60 mins	100%	98%	91%	82%	95%
	Halved	100%	96%	89%	80%	93%
	Doubled	100%	72%	45%	45%	64%
Average waiting time for a specialist assessment *(Current: 74 mins)*	30 mins	100%	92%	80%	68%	88%
	Halved	100%	91%	80%	68%	87%
	Doubled	100%	80%	69%	63%	77%
Average waiting time for diagnostic tests *(Current: 20 mins)*	Halved	100%	91%	77%	67%	86%
	Doubled	100%	81%	73%	64%	79%
Average waiting time for admission (e.g. waiting for a porter) *(Current: 5 mins)*	Halved	100%	87%	76%	66%	83%
	Doubled	100%	86%	74%	65%	82%

- Hospital bed availability appears to be the external factor driving ED overcrowding, and it strongly influences performance in MAJORS/RESUS
- ED performance is also very sensitive to arrival rates, though variations in arrival rates seem to mainly affect UTC
- Therefore, delays in UTC seem to be more related to ED causes (capacity issues), whereas delays in MAJORS/RESUS seem to be heavily influenced by hospital bed availability
- Other external factors seem to be less critical to ED performance compared to bed waiting time and arrival rates

Model analysis and results



- Bed waiting time and arrival rates seem to be critical in terms of physical congestion of the ED as well
- Changes in exogenous waiting times can lead both to reduction in the number of patients within the ED and to shifts of peak ED occupancy during the day

Learning objectives for today

- Criteria for choosing a simulation vs analytical approach
- Quick ways to analyse queueing systems
- Conducting scenario analyses
- Interpreting model results

Thank you!