

Using queuing theory to analyse the Government's 4-h completion time target in Accident and Emergency departments

L. Mayhew · D. Smith

Received: 5 January 2007 / Accepted: 11 July 2007 / Published online: 3 October 2007
© Springer Science + Business Media, LLC 2007

Abstract This paper uses a queuing model to evaluate completion times in Accident and Emergency (A&E) departments in the light of the Government target of completing and discharging 98% of patients inside 4 h. It illustrates how flows through an A&E can be accurately represented as a queuing process, how outputs can be used to visualise and interpret the 4-h Government target in a simple way and how the model can be used to assess the practical achievability of A&E targets in the future. The paper finds that A&E targets have resulted in significant improvements in completion times and thus deal with a major source of complaint by users of the National Health Service in the UK. It suggests that whilst some of this improvement is attributable to better management, some is also due to the way some patients in A&E are designated and therefore counted through the system. It finds for example that the current target would not have been possible without some form of patient re-designation or re-labelling taking place. Further it finds that the current target is so demanding that the integrity of reported performance is open to question. Related incentives and demand management issues resulting from the target are also briefly discussed.

Keywords Queuing theory · Accident and emergency departments · Government targets

L. Mayhew (✉) · D. Smith
Faculty of Actuarial Science and Insurance,
Cass Business School,
106 Bunhill Row,
London EC1Y 8TZ, UK
e-mail: Lesmayhew@blueyonder.co.uk

D. Smith
e-mail: d.a.smith@city.ac.uk

1 Introduction and background

In the NHS plan published in 2000 the Government committed to a range of improvements in the delivery of health care services. In terms of Accident and Emergency (A&E) services, the NHS Plan said:

“By 2004 no-one should be waiting more than four hours in Accident and Emergency from arrival to admission, transfer or discharge. Average waiting times in accident and emergency will fall as a result to 75 minutes. if they (patients) need a hospital bed they should be admitted to one without undue delay”¹.

This policy was spelt out in a subsequent publication, “Reforming Emergency Care”². As a step towards this, it was decided that all A&E departments should achieve a somewhat reduced standard of 90% of completions within 4 h and that hospitals should be measured on this basis during the last week of March 2003.

Following discussions with the medical profession, it was accepted that the eventual target aim should be less than the original 100% on the grounds that there will always be a minority of patients that fall outside the range due to their condition or special circumstances and it would not be practical or desirable to force the system to deliver something outside of its control.

This was accepted by the Government such that 98% of completions within 4 h is now the standard. As an indication of the impact of this policy, the Health Care Commission records that in the last quarter of 2005, 97

¹ Para 12.10 National Health Service (NHS) plan 2000, CM 4818-I, Department of Health.

² Reforming Emergency care, 2001, Department of Health.

A&E providers achieved the required target of 98% in 4 h; 54 under achieved the target (i.e. their performance was greater than or equal to 95% but less than 98%) and 8 ‘significantly’ underachieved the target (i.e. they achieved less than 95%).

The National Audit Office (NAO) for example records that in 2003, 23% of patients spent over 4 h in A&E³ as compared with 2% enshrined in the target. The achievement of the 98% target by so many A&E departments thus appears to be a massive step forward. Based on the queuing model described in this paper, it implies that most patients are now being discharged inside an average of 1 h instead of 3 or more hours that was the norm just a few years ago.

The Government gives reasons for the improvements as better management, more resources, changes to work flow, faster admissions to beds on wards and a stronger commitment to removing bottlenecks (e.g. in reception, triage, and undertaking diagnostic tests). However, as we show in this paper, not all the gains can be attributed either to increased efficiency or more resources.

Some of the gains have also resulted from a re-designation of patients with the effect that they are discharged from A&E earlier than they would have been under old arrangements. The use of ‘medical assessment units’ into which some patients are transferred is a good example of this; but whether patients will have noticed the difference is open to question if it represents simply a re-labelling.

The introduction of ‘payment by results’⁴ has further encouraged some hospitals to push patients through A&E even more quickly so benefiting from higher inpatient ‘tariffs’. The possibility of such perverse incentives was not part of the original aim behind A&E targets, which were primarily a response to patients’ concerns, and may have encouraged the manipulation of data.

The way targets have been achieved is more than just about good management. Targets are the cornerstone of the star rating system for assessing the performance of hospitals. It is not an exaggeration to suggest that the stakes for not meeting them are very high and if missed can significantly affect hospital reputations and future funding.

To measure A&E performance involves a significant investment in terms of data collection and analysis. Patients have to be clocked in and out and a running tally of patients approaching the 4-h wait needs to be maintained. For many purposes it is easier to report on an average rather than a distributional measure of the type ‘ $x\%$ in y h’.

³ Improving Emergency Care in England, 2004, NAO HC1075 Session 2003–2004.

⁴ This is a national system of standard tariffs for hospital procedures such as A&E attendance or overnight stay on a medical ward, or for an operation in which hospitals charge commissioning health authorities for the work carried out.

In this paper we set out to achieve three objectives: firstly show that A&E departments can be represented as a queuing process, secondly show that completion time distributions can be predicted from completion time averages (and vice versa), and thirdly that Government targets are to a considerable extent unrealistic and results have to some extent been manipulated.

2 A&E as a queuing process

In mathematical terms A&E workflow is a classic example of a queuing process—patients arrive, wait, are treated and then leave. Literature on queuing theory goes back many years and is included in chapters of many text books (see Taha, 2007 [17], Bronson and Naadimuthu, 1997 [4]). Examples of applied work in A&E using queuing theory are less evident in the literature than theoretical papers; however, good examples involving simple queuing models include [6, 16].

There are examples of A&E papers that use descriptive statistics rather than queuing theory as such, but are effective nonetheless in identifying key workflow issues (e.g. Aharonson-Daniel et al. [1], Barlow 2002 [2]). However, in general queuing theory has advanced beyond the capability of management information systems to support highly theoretical models using different structures, assumptions and probability distributions for arrival and server behaviour. Resultant models have their origins in subjects such as engineering or mathematics and are designed for specific purposes (e.g. Bučar et al., 2004 [5] or Van Vuuren et al. [18]).

A&E modelling has turned towards simulation techniques in the belief that these are better and more flexible at capturing systems’ complexity and dynamics (e.g. Fletcher et al. [7], Brailsford [3], Lane et al. [10]). Such models are necessarily highly specified, but also make detailed demands on data and can be sensitive to parametric changes (see [9] for a detailed review). The problem we wish to address requires that our model is both mathematically stable and based on routinely collected management information and so a different approach is needed.

A queuing model which accurately predicts completion times and is credible still needs to be specified and underpinned in terms of processes so that elements of the model such as assumptions and structure are transparent. This is normally taken to mean that individual elements must reflect system processes explicitly, for example in terms of arrival rates and server times (i.e. average treatment durations) in every possible substage.

In order to get a good predictive model it is not necessary to replicate exactly the system on the ground providing the model is fit for purpose [15]. A key feature of

our approach is that we endogenise the arrival and server rates in order to concentrate on the overall time spent in A&E. This method was used by Mayhew (1987) [11] applying similar techniques to social security workflows⁵, and by [8] in a separate health care example. In both cases state probabilities are assumed to be independent of any initial conditions.

Rather than producing a detailed state-of-system snapshot, we consider the aggregate performance of an A&E department at monthly intervals, as if it behaved as a simple queue in equilibrium. As in Mayhew [11] we assume Poisson arrival rates and exponential server rates for deriving the model but use observed distributions of completion times for calibration purposes (see below). We find that the resultant model is flexible and fits the data very well, but we make no claim that this is best possible model of its type there is, as this was not our objective.

Even so to specify and calibrate such a model requires a considerable amount of data covering an extended period in an A&E department, preferably exhibiting wide variation in completion times in order to adequately validate the model. With recent improvements in completion times and a convergence in performance among UK A&E departments this ideal is arguably becoming harder to achieve, which is why we focussed our attention on an extended historical period from 2002 to 2006 spanning the introduction of A&E targets.

A project funded by the Department of Health called Nu-Care involved a detailed examination of patient flows, the use of waiting areas, staff resources and completion times [12]. This study met the requirement that it spanned the period when completion times made rapid improvements towards the eventual tighter targets and so the data were particularly suitable for building a model of this kind⁶. However, a detailed description of Nu-Care and its findings are beyond the scope of this paper.

We use Nu-Care data for calibrating a model that relates actual average completion times to the national target. We base our overall findings on workflow data involving around 150,000 observations over an extended 4-year period, using monthly data generated by the incumbent computerised management information system. This clocks

people in and out by time and date of arrival (registration) and departure (home discharge, admission, or death).

While the management information system provided basic data, detailed data were unavailable on individual processes, such as time spent waiting for blood test results or on resources deployed at a specific point in time. To determine this information the Nu-Care study carried out two surveys six months apart monitoring the flow and pathways of around 2,500 patients over two 7-day periods⁷. These data enabled us to check on whether our model accurately reflected, for example, the split in the flows of patients between what we describe later in the paper as ‘short’ (minor) or ‘long’ (major) treatments.

Our first simplification was to imagine the workflow as a series of stages. These stages could include initial clinical assessment in triage, diagnostic tests including treatment and then eventual discharge. Some patients experience only one stage and others more than one but what constitutes a ‘stage’ is not always clear and can vary, since each can often be broken down into several substages and where one begins and ends may be blurred.

We found there was a key difference between patients who are discharged home and those that are admitted as an inpatient or are referred to another health care provider based on typical completion times. This led to a mathematical model with two queues or streams arranged in parallel. One stream, those discharged home, has one ‘stage’ and those admitted or referred, two ‘stages’. However, the model in this paper is more refined in that we introduce several stages as well as an additional pathway based on an analysis of the information contained in the surveys (see above).

It was not our intention to represent exactly every aspect of workflow at the subsystem level as such a model would have been highly complex and would have added little value to our objectives. Our strategy was to use a model calibrated on the basis of three months data in 2002 to predict results in subsequent months and years, and thereby improve our confidence that we were capturing and mapping the underlying processes sufficiently accurately for the purposes of evaluating the government target.

By focusing on results in a macro way rather than micro-modelling every detail of the processes involved we are able to produce practical and verifiable results. In the paper we begin with an explanation of the model, how it was constructed, calibrated and then validated. Subsequent sections discuss how the model can be used to monitor

⁵ The approach has several parallels with this paper in that the results were used to inform national turn round targets for social security benefit administration conducted at that time in over 400 local benefit offices.

⁶ Data extracts were analysed and information fed back to the A&E manager and consultant clinician on delays and bottlenecks on a regular basis. Examples of management actions included rescheduling staff to fit patient arrival profiles, re-organising ‘triage’ to make it faster, and elimination of process duplication. As result average completion times were brought down from 5:42 h:min to 2:50 h:min between March 2002 and March 2003.

⁷ Patient pathways were ‘bar-coded’ and then analysed to identify of bottlenecks and capacity constraints. A good example was the time spent on blood or urine tests. At the time this was based on a batch system with samples transferred to the pathology laboratory. Subsequent adoption of near patient blood testing helped to reduce delays.

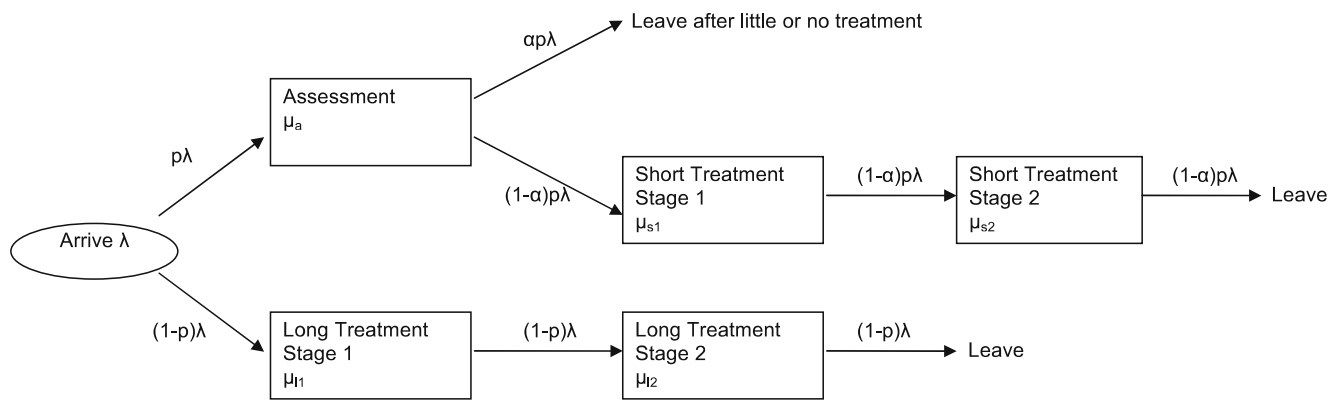


Fig. 1 Diagram of the model

progress against government targets, and how results can be manipulated to give an impression of improved performance (i.e. shorter completion times). A final section discusses the policy implications of our findings.

3 The initial model

Consider initially the total time spent in the department by a patient and make two further simplifying assumptions: (1) that the average time spent in each stage is the same; (2) arrivals are random with inter-arrival times specified by a Poisson process. The probability of the total time spent in A&E equalling z may be considered to be the sum of s random variables as follows:

$$z = \tau_1 + \tau_2 + \tau_3 \dots + \tau_s$$

Where τ_i is the time spent in stage i and s is the number of stages.

Let the system be characterized by an exponentially distributed arrival rate with parameter λ and service times exponentially distributed at each stage with parameter μ then the probability density function of z can be shown to be:

$$p(z) = \frac{z^{s-1}(\mu - \lambda)^s \exp(-z(\mu - \lambda))}{(s - 1)!}$$

i.e. the distribution is a gamma distribution.

This is when the queue has reached a stable state, but if $\lambda > \mu$, the queue is unstable and grows indefinitely⁸.

⁸ Whilst this remains a theoretical possibility and queues do intermittently build up rapidly at certain times of day, our primary interest lies in average system performance measured over longer periods because this is how performance is actually measured and judged.

Since our main interest is average completion times and the distribution around the average for stable queues, we may write this equation more conveniently in terms of t , the average completion time.

$$p(z) = \frac{\left(\frac{zs}{t}\right)^s \exp\left(\frac{-zs}{t}\right)}{z(s - 1)!}$$

where $t = \frac{s}{\mu - \lambda}$.

This p.d.f. has the cumulative distribution function:

$$P(z) = 1 - \exp\left(\frac{-zs}{t}\right) \sum_{i=0}^{s-1} \left(\frac{zs}{t}\right)^i / i!$$

The model is hence similar to that applied by Mayhew [11] to social security workflows. However, we found that in applying it to an A&E department, it did not capture the processes, and hence completion times, with sufficient accuracy. There were two key problems: firstly completion times differed significantly depending on whether a patient was discharged home or admitted to a ward and so could not be captured accurately without splitting the flows into separate categories or workflows. The second problem was that individual stages could not be adequately represented by the assumption that the service time distribution parameter μ was the same at each stage.

A more detailed examination of the processes showed that it was possible that a patient with acute problems will be processed very quickly as a priority, whereas another patient with non-severe symptoms will need to wait far longer for treatment. As it is time taken to complete treatment that we are interested in we relabelled these two paths as ‘short treatment’ and ‘long treatment’. The raw data also indicated that some patients are processed in under 10 min and so to model these patients it was decided that for ‘short treatment’ there should be an initial stage where people are assessed (i.e. triage). If the assessment is that they can be discharged as not needing the care provided only by a hospital, then they exit the system whereas other patients move into the ‘short treatment’ path. This route can also be seen as patients who

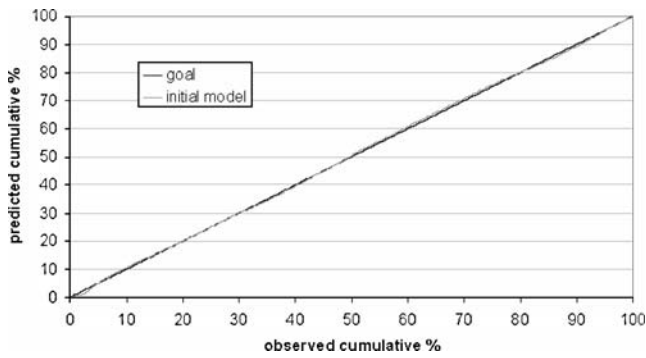


Fig. 2 Comparison of the quality of fit generated by the initial model compared to observed data. Perfect agreement would be on the diagonal line

arrive with very minor needs at particularly quiet times and hence are processed quickly.

The patients deemed as needing no or very little treatment in effect are processed through a very short one stage queue as shown in Fig. 1, which now comprises three paths—‘no/little’, ‘short’ and ‘long’. The actual paths of this model appeared to be sensible and on the basis of the survey undertaken reflected the processes that patients find themselves going through in hospital. We can imagine for example that a person arriving with obvious problems will not need to be assessed and can immediately go to the long treatment path. A patient who is not severely ill will be assessed first and then may either go home or will enter the short treatment path.

4 Modelling the expected time spent in the three paths

To derive the expected times in the three paths we assume that the system is characterized by an exponentially

distributed arrival rate with parameter λ and service times for the various stages exponentially distributed with the relevant parameter μ_i where i refers to a particular stage (see Fig. 1). The derivation of the completion time distributions applying to different path and hence stage combinations is explained further in [13].

4.1 No/Little treatment path

The no/little treatment path is a simple exponential distribution. The probability distribution function is therefore:

$$P(z) = 1 - \exp\left(\frac{-z}{t}\right)$$

where $t = \frac{1}{\mu_a - p\lambda}$ and p is the percentage of patients who go to assessment.

4.2 Short treatment path

The short treatment path is a three stage hypo-exponential distribution (for derivation of the hypo-exponential distribution see [13], pp 245–247). The probability distribution function is:

$$F(Z) = 1 - \left(\frac{\mu_2 \mu_3 e^{-\mu_1 t}}{(\mu_2 - \mu_1)(\mu_3 - \mu_1)} + \frac{\mu_1 \mu_3 e^{-\mu_2 t}}{(\mu_1 - \mu_2)(\mu_3 - \mu_2)} + \frac{\mu_1 \mu_2 e^{-\mu_3 t}}{(\mu_1 - \mu_3)(\mu_2 - \mu_3)} \right)$$

with

$$\begin{aligned} \mu_1 &= \mu_a - p\lambda \\ \mu_2 &= \mu_{s1} - (1 - \alpha)p\lambda \\ \mu_3 &= \mu_{s2} - (1 - \alpha)p\lambda \end{aligned}$$

where α is the percentage of patients who go home with little or no treatment.

Fig. 3 Comparison of observed and predicted completions times based on the initial model and May–July 2002 data

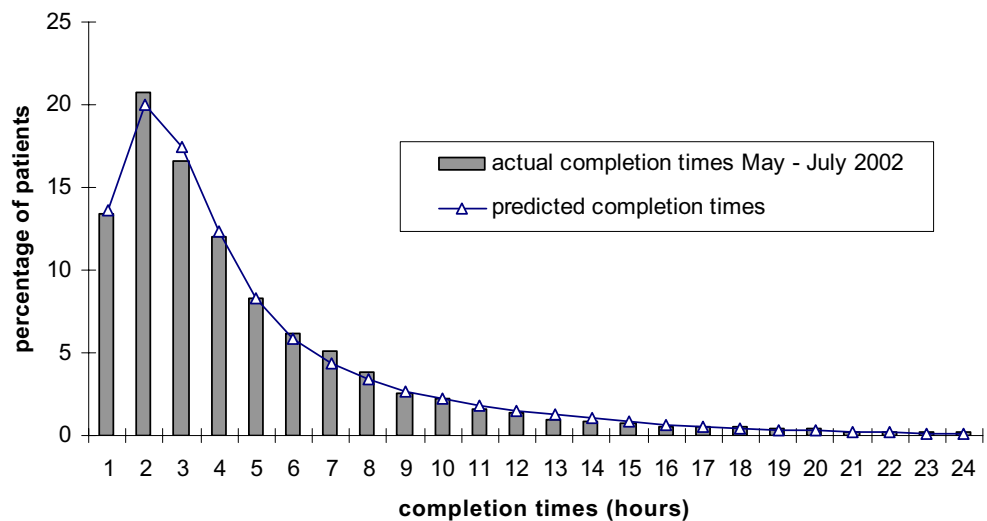


Table 1 The values of the model determined by observations from May–July 2002

Pathway	Parameter	Values
Little or no treatment	Percentage of patients	6
	Average time (hours)	0.4
Short treatment	Percentage of patients	54
	Average time (hours)	2.53
Long treatment	Percentage of patients	40
	Average time (hours)	7.3
Total average time (hours)		4.2

4.3 Long treatment path

The long treatment path is a two stage hypo-exponential distribution with a probability distribution function:

$$F(Z) = 1 - \left(\frac{\mu_1 \exp[-\mu_2 z] - \mu_2 \exp[-\mu_1 z]}{\mu_1 - \mu_2} \right)$$

with $\mu_1 = \mu_{l1} - (1 - p)\lambda$
 $\mu_2 = \mu_{l2} - (1 - p)\lambda$

5 Fitting the model to the data

We fitted the model using standard iterative techniques. We minimised the squared differences between the observed and expected cumulative patients, concentrating on the range between 5 and 99%. The parameters of the model were constrained to reflect the percentage of patients entering the different paths and the total time spent in the paths (using the dataset from the survey).

While we are most interested in the patients who are taking the longest time to have their treatment completed,

Fig. 4 A comparison of the time take to clear a given percentage of patients based on the initial model

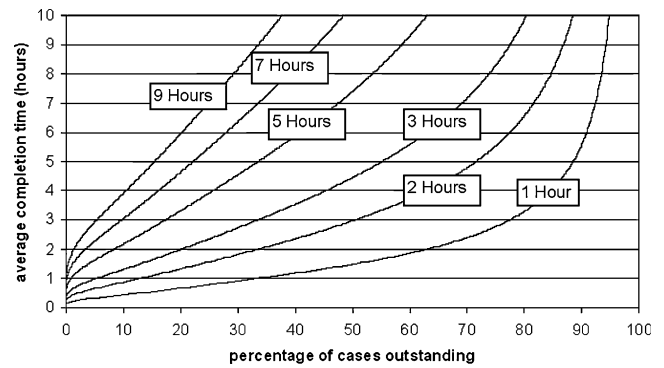
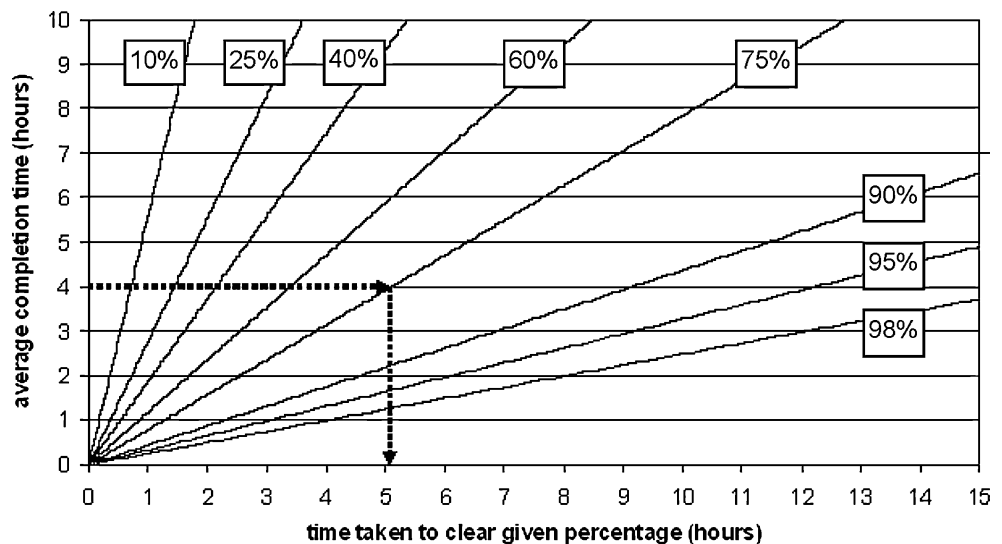


Fig. 5 The percentage of cases still outstanding after the given number of hours in A&E, for a given average completion time

the data for the last 1% of patients as measured by completion times is very volatile with some patients taking many hours to complete their treatment. To model the last 1% of patients accurately was therefore impossible and in any case added little value to the analysis.

On a monthly basis the process times for the 5% of patients ranked by completion time was also highly volatile and as we have less interest in these patients we were content not to attempt to fit the model as accurately for this group as compared with the rest.

Figure 2 shows the fit of the initial model to the observed data by looking at completed treatments and Fig. 3 shows how the model captures the shape of processing times that were observed.

6 Average time for the different paths

At this stage it is useful to present the values that the model is generating for the average completion time for the

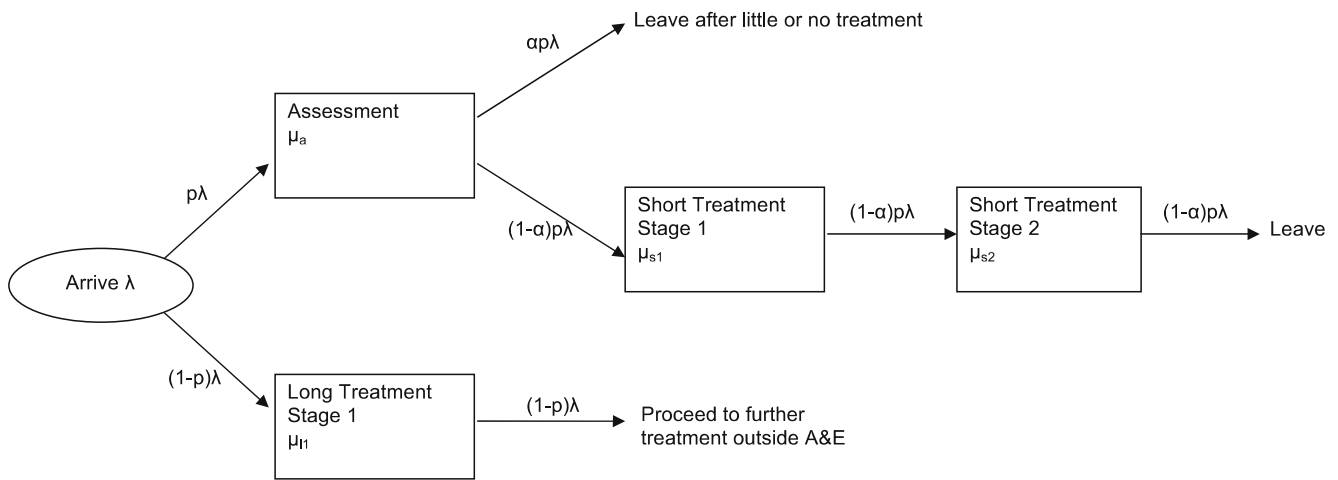


Fig. 6 Diagram of the re-designated model

different paths and the percentage of patients that enter each pathway and this is shown in Table 1.

7 Ready-reckoners

It has become custom and practice to express completion time targets, not as averages but as the percentage of patients to be dealt with in a given time. For example the national standard in emergency care in March 2003 was 90% in 4 h; currently (as at May 2007) it is 98% in 4 h. This type of specification has an obvious attraction over averages because averages are sensitive to extremely long waits or completion times.

We therefore needed a convenient method of moving between average completion times and the inferred probability distributions. This can be done by using ready-reckoners which plot various percentiles calculated using different average completion times. Ready-reckoners were

calculated for the initial model for a selection of percentiles which are shown in Fig. 4. The dotted line indicates that if the average completion time for patients is four hours then it will take 5.1 h (5 h 6 min) to process 75% of patients.

The second variant shown in Fig. 5 establishes, for a given average completion time, the percentage of patients outstanding after a given time in the A&E Department and is complementary to the first ready-reckoner. It too conveys useful management information about completion times but expressed in different but compatible units.

8 Re-designating patients

As was explained in the introduction, hospitals are having to meet ever more stringent A&E targets. A possible solution to the problem of bringing down completion times that do not meet the target is to reconsider the point at which a patient can be considered as having been dis-

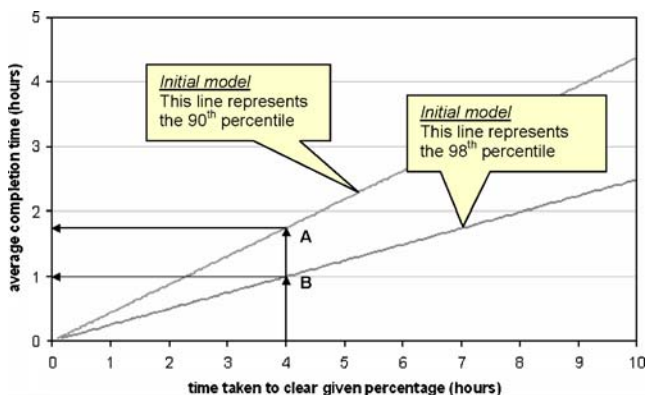


Fig. 7 Initial model showing how the average completion time changes when moving from the 90th to the 98th percentile represented by points A and B

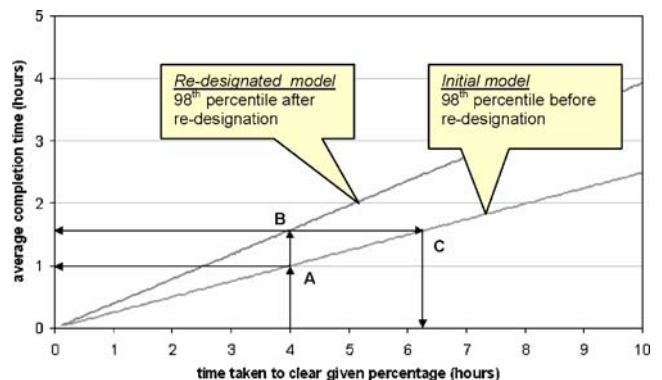


Fig. 8 A comparison of the time taken to complete the A&E treatment of 98% of patients based on the initial model and the re-designated model

Table 2 The time taken to clear a specified percentage of patients given the average time for the alternative model where model A is the initial model and B is the re-designated model

Average Time	90%		95%		98%		99%	
	A	B	A	B	A	B	A	B
1	2.29	1.37	3.06	1.83	4.03	2.55	4.74	3.13
2	4.58	2.73	6.12	3.67	8.05	5.10	9.47	6.25
3	6.87	4.10	9.18	5.50	12.08	7.65	14.20	9.38
4	9.16	5.47	12.23	7.33	16.10	10.20	18.93	12.51
5	11.46	6.84	15.30	9.18	20.15	12.76	23.69	15.65
6	13.74	8.21	18.36	11.01	24.18	15.31	28.42	18.78
7	16.03	9.57	21.42	12.84	28.20	17.86	33.15	21.90
8	18.32	10.94	24.48	14.67	32.23	20.41	37.89	25.03
9	20.61	12.30	27.54	16.51	36.25	22.96	42.62	28.16
10	22.90	13.67	30.59	18.34	40.28	25.51	47.35	31.28
11	25.19	15.04	33.65	20.17	44.30	28.07	48+	34.41
12	27.48	16.40	36.71	22.01	48+	30.61	48+	37.54

charged and therefore no longer classed as being part of ‘A&E’. For example, it can be claimed that while discharged patients spend their whole time in A&E, patients who are referred or become inpatients spend part of their treatment in a different classification i.e. the latter stages of their treatment in A&E are really the first stages of treatment in a different category.

Examples of this are defined in this paper as ‘re-designation’. As noted in the introduction to this paper a good example of this are ‘Medical Assessment Units’ in which patients arriving in A&E are kept under observation and assessment that may result in stays of longer than 4 h. By considering this change we can alter our model to try and show the effect on times that patients spend in A&E with and without re-designation. The appropriate way to model this arrangement is to modify the ‘long treatment’ path.

From Fig. 1 this is currently a two-stage process but we can easily argue that these patients can spend stage one in A&E with the second stage being reclassified as part of their new treatment. The model is therefore changed so that

the ‘long treatment’ path becomes a one-stage model so that a patient would be expected to take approximately half the time that was spent in the long treatment path of the initial model.

Again this is a slight simplification as the initial model had two separate stages for the long treatment path with independent processing rates. However, for the fitted model the difference between the two processing rates was very slight so we decided to have one stage with a process rate that gave an average time of half that spent in the long treatment path. This gave us our new model, called the re-designated model, which is represented in Fig. 6.

9 Issues arising in relation to the 4-h national target

By the end of 2003 A&E departments were expected to achieve a standard of 90% completion within 4 h. Since then the standard has been further tightened to 98%, a difference of 8 percentage points. Figure 7 is a partial ready-reckoner, based on the initial model that focuses on the two percentiles in question and shows the implications of this tightening of the target. Point A shows the average completion time required to meet the target of 90% in 4 h. This equates to an average completion time on the vertical axis of 1.75 h (1:45 h:min).

Tightening the target to 98% completions within 4 h implies that the average completion time must fall to point B which equates to an average of 0.99 h or 59.4 min. In other words, a change of 8% points in the target has caused the required average to reduce by 43% or just over 45 min. To achieve such a reduction represents a massive challenge in A&E terms especially when it is borne in mind that average completion times of 4 or 5 h were not uncommon just a few years ago. For example, it implies that patients

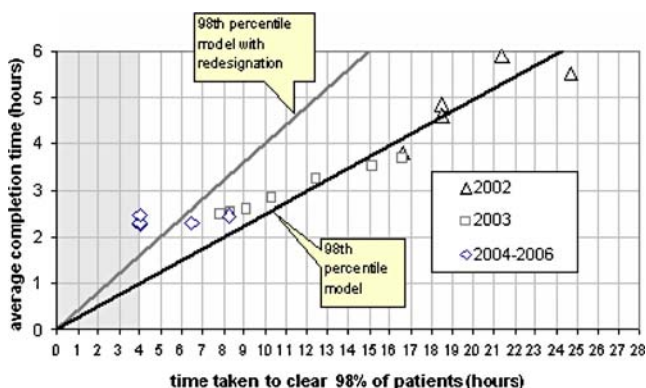
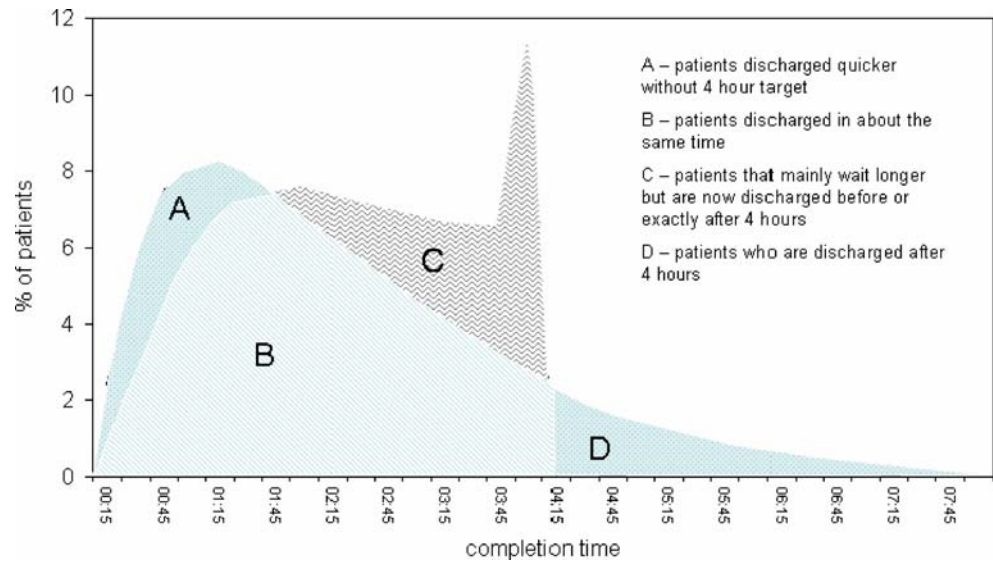


Fig. 9 A comparison of actual performance between 2002 and 2006 against the 98th percentile based on both model variants

Fig. 10 A comparison of completion times in 2 months in 2003 and 2006 before and after the introduction of the 98% in 4 h target



must be processed and discharged at almost twice the rate as compared with when the target was 90% in 4 h.

Let us now consider the re-designated model in which the long treatment path has been truncated after stage one following re-designation (Figs. 1 and 6 refer). Figure 8, by comparing the 98th percentiles in the initial model and the re-designated model, shows the difference this makes to the achievement of the 4-h target. Consider again point A. This corresponds to the average completion time required to complete the treatments of patients in A&E according to the original classification of processes in Fig. 8 i.e. 0.99 h.

As a result of re-designation, we find that the average can be allowed to increase to 1.57 h (1:42 h:min), whilst still achieving the target of 98% in 4 h (point B, Fig. 8). The time taken to treat 98% of patients under the original classification with this average completion time would be 6.31 h (6:19 h:min) which is, of course well, outside the target (point C, Fig. 8). The difference between the models is thus noticeable and becomes even more noticeable the higher the percentage that needs to be cleared. This is because the re-designated model reduces the chance of particularly long treatments as only one process has to be passed through now rather than two.

Table 2 provides a comparison of both the initial model and the re-designated model for average completion times of 1 to 12 h based on the 90th, 95th, 98th and 99th percentiles. As can be seen the results for each model confirm the previous finding, namely that they diverge and gains become larger as the average or the percentiles increase. So for example, given an average completion time of 5 h, 99% of patients would be completed in 15.65 h (15:39 h:min) with re-designation and 23.69 h (23:41 h:min) without re-designation.

10 Discussion

The stated aim of the NHS Plan (2000) is to complete the treatment of 75% of patients in 1 h with an eventual aim of clearing all patients (100%) inside 4 h from the time of arrival to discharge. The model in this paper shows that the original 75% standard would translate into an average completion time of 0.78 h (or 47.0 min). By using the variant that re-designates some patients, this underlying average completion time is increased to 1.16 h (or 69.8 min).

Today the target is to clear 98% of patients in 4 h. Meeting this target using the re-designated model is the equivalent of only 47.5% patient completion using the initial model. Thus the possibility of re-designating some patients is clearly a significant aid towards achieving the 98% target where re-designation has occurred. It seems highly improbable therefore that the original target could ever have been achieved without changing the basis for counting patients through the system. Headline improvements in completion times however has led to announcements from the Government minister responsible at different times proclaiming the progress made⁹.

This is not a criticism of re-designation if it results in patients being cared for in more appropriate surroundings but with a 'different label'. However, in comparing the 'old' with

⁹ In 2005 the then Health Minister said: "Across the NHS we are eradicating long waits for treatment. At the beginning of 2003, almost a quarter of patients spent more than four hours in A&E. That figure is now down to its lowest point ever. Our own data shows that this performance is being sustained week-in-week-out, even against a background of increasing attendances".

the ‘new’ is not comparing ‘like’ with ‘like’ and so true improvements are somewhat less than headline figures might suggest, certainly as perceived by patients. In view of the large difference in average completion time that results from a small change in the definition of the target (e.g. from a 90 to a 98% completion rate), it seems doubtful to us that the impact of this change was ever properly evaluated before its introduction. The original claim in the NHS 2000 plan (see introduction) that patients will spend on average 75 min in A&E also seems wildly optimistic in retrospect (which may be why it was abandoned), whereas a 1-h average borders on the impossible. These findings therefore raise concerns about the credibility of reported performance.

If we review the change through time the effects of re-designation are evident to see. Figure 9 shows a range of results for different months over a 4-year period from before targets were introduced up until they were tightened to 98% in 4 h. The solid lines represent the 98th percentiles with and without patient re-designation as predicted by the models. As is evident actual performance reflects the initial model quite well (calibrated using three months data in 2002—see thick solid line) in the period 2002 to 2003 but after that it starts to diverge. Why is this so?

In 2003 the target was 90% in 4 h and in the following year it moved to 98%. A&E departments sought ways to meet performance such that clinical processes were not compromised. In practice they have found it very difficult to get below a 2-h average, as the chart suggests, but have managed to shift patients that stayed over 4 h partly through re-designation and data points from 2004 in Fig. 9 appear to reflect the shift in management regime that is the result.

Plainly even with re-designation a theoretical average completion time of 1.5 h that would be needed to meet the 98% target in 4 h is still stretching. Thus in any independent audit of A&E completion times it is important to look at the detail behind reported performance including the source data to check that it is genuine and not based on administrative convenience (for example, simply discharging people regardless after 4 h in the system). That this possibility occurs is easily demonstrated because completion time distributions would be truncated after 4 h i.e. there would be a cliff-edge effect in a typical graph showing the distribution of completion times and examples of this can be found in data from 2005 onwards.

Figure 10 based on two months in 2003 and 2006 with and without the 4 h target shows clearly this effect. For clarity the completion time distributions are divided into four categories: A—patients discharged quicker without the 4 h target; B—patients discharged in roughly the same time with or without the target; C—patients with the 4 h target that wait longer but still complete within 4 h; D—patients

without the target that spent more than 4 h in A&E. In this example, it is of interest to note that the average completion time *before* the introduction of the target was 1.78 h for patients discharged inside 4 h, whereas *after* the introduction it had increased to 2.23 h.

11 Conclusion

To conclude we suggest that a target should not only be demanding but that it should also fit with the grain of the work on the ground and not lead to disruptive practices otherwise the target and how to achieve it becomes an end in itself. One way to do this would be to vary the percentile in the target to suit different types of departments. So for example a department with a higher percentage of seriously ill cases, the target could be to discharge a lower percentage of patients in 4 h (say 90%), but in a walk in or urgent care centre¹⁰ 95% or higher (but never more than 98%).

A&E completion time targets appear to have had a beneficial effect in terms of improving services to patients compared to a few years ago, but the application of the targets leaves a considerable margin for doubt as to their integrity and the perverse incentives they create within the system. On a cautionary note, the practicality of a single target fitting all A&E and related services will come under increasing strain, as services are re-focused and become more specialised in terms of complex and less complex caseloads and it may be necessary to revise targets in any case.

The opportunity should be taken to make these targets more credible in order to avoid the distorting effects of targets that border on the impossible. Queuing models of the kind described in this paper can illuminate the appropriate choice of target but extended applications also seem promising. The Nu-Care study showed that as completion times fall patient arrival rates systematically increase, so rendering the target even more impossible¹¹. Given that demand and targets are related, there is a strong case for using targets in a more intelligent way to regulate and direct demand to more appropriate care settings, not necessarily within an A&E environment.

¹⁰ A centre for minor injuries and treatments with lesser clinical competencies than a fully fledged A&E department located in a major hospital.

¹¹ For example, NuCare found that as average completion times fell or rose by 10% throughput increased or fell by 3.7%.

Acknowledgement The authors are most grateful to three anonymous referees for their views and comments which we found extremely valuable in drafting this paper.

References

- Aharonson-Daniel L, Fung H, Hedley AJ (1996) Time studies in A&E departments—a useful tool for management. *J Manage Med* 10(3):15–22
- Barlow GL (2002) Auditing hospital queuing. *Manag Audit J* 17(7):397–403
- Brailsford SC, Lattimer VA, Tamaras P, Turnbull JC (2004) Emergency and on demand health care: modelling a large and complex system. *J Oper Res Soc* 55:34–42
- Bronson R, Naadimuthu G (1997) *Operations Research*, 2nd edition, MacGraw-Hill, New York London
- Bučar T, Nagode M, Fajdiga M (2004) Reliability approximation using finite Weibull mixture distributions. *Reliab Eng Syst Saf* 84:241–251
- Coates TJ, Michalis S (2001) Mathematical modelling of patient flow through an accident and emergency department. *Emerg Med J* 18:190–192
- Fletcher A, Halsall D, Huxham S, Worthington D (2006) The DH accident and emergency department model—a national generic model used locally. Lancaster University Management School Working Paper 2006/042. Lancaster, UK
- Gorunescu F, McClean SI, Millard PH (2002) A queuing model for bed occupancy management and planning of hospital. *J Oper Res Soc* 53:19–24
- Jun J, Jacobson S, Swisher J (1999) Applications of discrete-event simulation in health care clinics: a survey. *J Oper Res Soc* 50:109–123
- Lane DC, Monefeldt C, Rosenhead JV (2000) Looking in the wrong place for health care improvements: a system dynamics study of an accident and emergency department. *J Oper Res Soc* 51(5):518–531(14)
- Mayhew L (1987) Resource inputs and performance outputs in social security. *J Oper Res Soc* 38(10):913–928
- Mayhew L, Carney Jones E (2003) *Evaluating a New Approach for Improving Care in an Accident and Emergency Department*. Cass Business School (60 pp)
- Mayhew L, Smith D (2006) *Using Queuing Theory to Analyse Completion Times in Accident and Emergency Departments in the Light of the Government 4-hour Target*. Actuarial Research Paper 177, City University
- Ross S (1997) *Introduction to Probability Models*, 6th Edition. Academic Press, San Diego US
- Saaty TL, Alexander JM (1981) *Thinking with Models*. Pergamon Press, Oxford
- Siddharthan K, Jones WJ, Johnson JA (1996) A priority queuing model to reduce waiting times in emergency care. *Int J Health Care Qual* 9(5):10–16
- Taha H (ed) (2007) *Operations research—an introduction*. MacMillan, London, New York
- Van Vuuren M, Adan I, Resing-Sassen S (2005) Performance analysis of multi-server tandem queues with finite buffers and blocking. *OR Spectrum* 27:315–338