Cass means business

Cass Business School
City of London

Faculty of Actuarial Science and Insurance

Using Queuing Theory to Analyse Completion Times in Accident and Emergency Departments in the Light of the Government 4-hour Target

Les Mayhew and David Smith

Actuarial Research Paper No. 177

December 2006

ISBN 978-1-905752-06-5

Cass Business School
106 Bunhill Row
London EC1Y 8TZ
T +44 (0)20 7040 8470
www.cass.city.ac.uk

“Any opinions expressed in this paper are my/our own and not necessarily those of my/our employer or anyone else I/we have discussed them with. You must not copy this paper or quote it without my/our permission”.

Using queuing theory to analyse completion times in accident and emergency departments in the light of the Government 4-hour target

L Mayhew and D Smith

Les Mayhew Cass Business School, Faculty of Actuarial Science and Insurance
David Smith Cass Business School, Faculty of Actuarial Science and Insurance
Lesmayhew@blueyonder.co.uk
d.a.smith@city.ac.uk
Cass Business School
106 Bunhill Row
London EC1Y 8TZ
UK
December 2006

Using queuing theory to analyse completion times in accident and emergency departments in the light of the Government 4-hour target

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 hours. It illustrates how flows though an A&E can be very accurately represented as a queuing process, how the outputs of a queuing model can be used to visualise and interpret the 4-hour hours Government target in a simple way and how queuing models 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. It finds 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. 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 and that a different approach is needed. Related incentives and demand management issues resulting from this Government target are also briefly discussed.

Les Mayhew Cass Business School, Faculty of Actuarial Science and Insurance
David Smith Cass Business School, Faculty of Actuarial Science and Insurance

Introduction and background

Accident and Emergency (A&E) services are the main way patients in the UK access urgently needed medical care. However, long waiting times have resulted in widespread criticism over a period of years and were repeatedly the most important complaint in patient satisfaction surveys.

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 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], which set the target into a context of wider reforms aimed at improving services. As a step towards this, it was decided that all A&E departments should achieve a somewhat reduced standard of 90% of completions within four hours and that hospital trusts should be measured on this basis during the last week of March in 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 its control.

[1] Para 12.10 NHS plan 2000, CM 4818-I, Department of Health
[2] Reforming Emergency care, 2001, Department of Health

This was accepted by the Government with the result that 98% of completions within four hours 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 trusts achieved the required target of 98% in 4 hours; 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 achievement of the 98% target by so many trusts appears, at first sight, to be a massive step forward. The NAO for example records that in 2003, 23% of patients spent over 4 hours in A&E[3] as compared with 2% enshrined in the target. Based on the work reported in this paper, it implies that most patients are now being discharged in an average of one hour instead of three or more hours that was the norm just a few years ago.

The reasons for the improvement include 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, not all the gains can be attributed either to increased efficiency or more resources.

Improvements in completion times have also resulted from, in many cases, a re-designation of patients with the effect that they are discharged from A&E earlier than they would have been discharged 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 notice the difference is open to question if it represents simply a re-labelling.

A second concern is that introduction into the NHS of 'payment by results' may have encouraged some trusts to push patients through A&E even more quickly so benefiting from the higher inpatient 'tariffs' as compared with A&E tariffs. The possibility of perverse incentives such as these was not in the original aim behind the introduction of A&E targets which were primarily a response to patients' concerns and may have encouraged the manipulation of data.

More fundamentally, as A&E completion times have improved so demand has increased. Some of this additional demand will be 'genuine' but some will be demand diverted from other medical centres (e.g. GPs) and some will have been supply-induced as a result of service improvements. Government policy encourages the creation of 'urgent care' and 'walk-in' centres that deal with more routine cases at lower unit cost. The re-routing of demand in this way may therefore have further ramifications for the achievement of A&E targets in the future.

The way in which targets have been achieved, the prospects for further improvements and the sustainability of current performance is therefore clearly more than just about good management. For example, the achievement of the A&E target is a cornerstone of the star rating system for assessing the performance of acute trusts by the Health Care Commission. It is not an exaggeration to suggest that the stakes for not meeting

[3] Improving emergency care in England, 2004, NAO HC1075 Session 2003-2004.

the A&E target are very high and if missed can significantly affect the reputation and resources available to an acute trust in the future.

As services develop local NHS providers need to strike a balance across a range of needs and so the question of appropriate demand management via all the various patient channels that are available in the field of urgent care becomes an issue as well. Consideration of these issues is outside the scope of this paper, although the analysis it contains will be relevant in any discussion to the future development of A&E services and targets.

Queuing theory

To measure A&E performance involves a significant investment and overheads in terms of data collection and analysis. Patients have to be clocked in and out and a running tally of patients approaching the four-hour wait needs to be maintained. A&E departments with good management information are better able to manage work flow as a result but only if the data can be turned round quickly and analysed in real time.

A queuing model can save valuable time by providing analytical short cuts thereby improving the timeliness of management interventions and be helpful in the process of external scrutiny (e.g. via the Health Care Commission). In practice it may only be possible to do some of these things because of data and other limitations.

In this paper we consider how:

- flows though an A&E can be very accurately represented as a queuing process
- the outputs of a queuing model can be used to visualise and interpret the 4-hour hours Government target in a simple way
- queuing models can be used to assess the practical achievability of even tighter A&E targets in the future
- queuing models can be used to understand how completion times can be altered through re-designation of stages in the A&E process
- given the proper interpretation of data, underpinning information systems can improve performance over time
- queuing models can be used as a starting point for checking the credibility of reported performance

Representing A&E workflow as a queuing process

In mathematical terms, A&E workflow in an A&E department is a classic example of a queuing process - patients arrive, are treated and then leave. In theory, a queuing model can help illuminate the relationship between resources and waiting times, provide a method for understanding and monitoring performance, identify bottlenecks, and be used a general planning tool for estimating floor space and other requirements.

To calibrate such a model requires a considerable amount of data covering many patients over an extended period in a department exhibiting a wide variation in monthly performance. With recent improvements in completion times and a

convergence in performance among A&E departments this ideal is arguably becoming harder to achieve.

A project funded by the Department of Health called the NU-Care project involved a detailed examination of patient flows, the use of waiting areas, staff resources and completion times (Mayhew and Carney-Jones, 2003). This study has the practical advantage that it spanned the period when completion times made rapid improvements towards the eventual tighter targets and so the data were particularly suitable for calibrating a model of this kind.

We build on the data from that study to calibrate a queuing model that gives an almost perfect fit to the data and enables us to accurately relate achieved average completion to the national target. This may be compared with alternative approaches for modelling such processes but in far more detail using simulation models (e.g. Brailsford et al, 2004).

Queuing models range in complexity according to the arrival pattern, the order in which patients are treated, the existence of parallel or sub queues e.g. for X-rays or blood tests and so forth. We develop a tool that can be used by non-mathematicians in a range of A&E departments for monitoring and managing performance.

One of the features of queues that often surprises is the speed with which they can get out of control because there are too few resources to deal with them or they are being managed badly. Queuing theory shows there is a narrow margin between queues that are under control and those that are not. The lesson from the NU-Care project is that queues can be brought under control and waiting times reduced with appropriate organisational and management strategies.

The first simplification is to imagine the workflow as a series of stages. These stages could include initial clinical assessment, diagnostic tests including treatment and then eventual discharge. In practice, we know that some patients experience only one stage and others more than one. What constitutes a ‘stage’ however is not always clear, since each can often be broken down into several sub-stages so that where each one begins and ends is blurred.

We found that there is a key difference between patients who are discharged home and those that are admitted as an inpatient or referred. This leads to an initial 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’.

We found that splitting the queues in to further streams with different numbers of stages improved the goodness of fit only slightly when using a basic model. A feature of this approach therefore is that we *infer* the number of stages and the workflow characteristics through consideration of the aggregate distribution properties of the data. For later models we have introduced stages through inference about the workings of the department but as explained it is impossible to have a direct link with the actual stages that a patient will go through.

This method was, for example, successfully employed in an application to social security queues (Mayhew, 1987) – the main difference is that social security deals

with benefits and an A&E department with patients. A queuing model was also used to represent bed occupancy in hospitals, which bears some resemblance to the present approach (Gorunescu et al, 2002). Note that it is possible that two different queuing models making slightly different assumptions could provide equally good 'fits' to the data. We make no claims that this is the most accurate and most general model that exists, or that it correctly represents every aspect of the queuing process.

In summary, we have argued that trying to model each potential stage accurately will lead to a model that is very complex and unstable. By focusing on results in a macro way rather than micromanaging the queues we found that it produces practical and generalisable results. In the following sections, we explain how the model was designed and calibrated.

The original model

We consider a queuing model of the type in which there is one or more stages through which patients pass before they are discharged from A&E (see Figure 1).

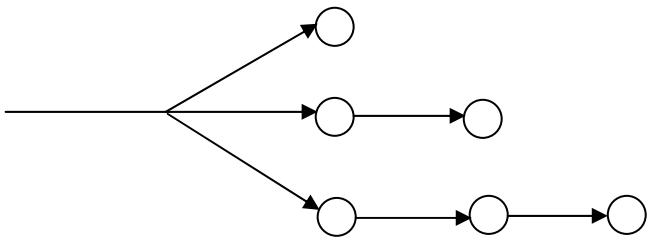

Figure 1: Depiction of a queuing system with different sub-queues and stages

Patients arrive and are initially sorted into queues depending on the severity of their condition. The number of stages that patients pass through will depend not only on severity but also standard clinical protocols depending on the presenting symptoms. Some patients, called 'absconders', leave before being seen or treated.

Over a period of time, workflows tend to follow a pattern and are quite stable features of the system. For example the proportion of patients discharged home is in the order of 60% and those admitted or referred around 40%. A small percentage (<1%) are dead on arrival or die in the department.

Consider the total time spent in the department by a patient and make two further simplifying assumptions: (i) that the average time spent in each stage is the same; (ii) 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:

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

where τ_i is the time spent in stage i.

Assume that the system is characterized by an exponentially distributed arrival rate with parameter λ and exponentially distributed service times 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 indefinitely. 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 *t*.

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

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

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

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

Figure 2 shows the probability of different completion times based on models with sequential numbers of stages (1, 2, 3…7) and completion time averages (1, 2, 3, …7 hours). For example, the curve furthest to the left is a one-stage model with a completion average of 1 hour, and the curve furthest to the right is a 7-stage model with a completion average of 7 hours. As is seen, the model can deal with a widespread range of possible queuing behaviour. The empirical question is to determine the appropriate number of stages by fitting the theoretical distribution to actual distributions of completion times and known averages. Before we do that, however, we need to consider how the information produced by the model will be used.

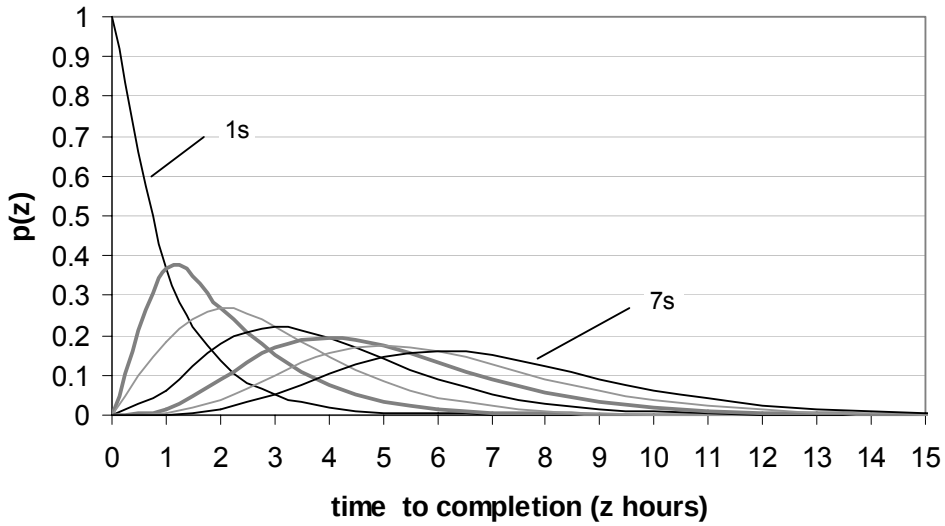

Figure 2: Distributions of completion times based on different completion averages and numbers of stages.

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 hours; today it is 98% in 4 hours. This type of specification has the obvious attraction over averages because averages are sensitive to extremely long waits or completion times.

We therefore need a convenient method of moving between averages and distributions. An example would be one that links the target of *x%* clearance in *y hours* to an average *t,* or which relates the average t to the work still outstanding after a given time *z* in the system.

Consider a simple case in which there is only 1 stage (*s* = 1), it can be shown that the average completion time is related to the cumulative distribution around the average by:

$$t = \frac{-z}{\ln(1 - P(z))}$$

We may plot this equation for different values of *t* and *z* to obtain the result in Figure 3, which we call a ready-reckoner. By reading off the average (follow direction of arrows), we can determine the time taken to clear a given percentage of cases. In this example, a 4 hour average completion time would equate to 70% of completions in just less than 5 hours.

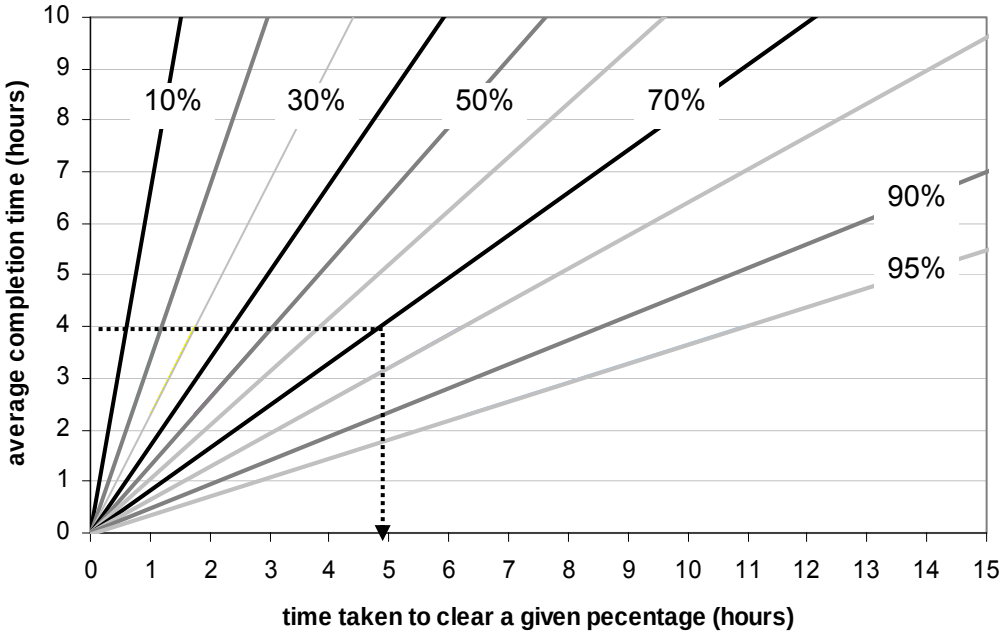

Figure 3: A ready-reckoner for a queuing system with one stage (s=1)

Whilst this ready-reckoner achieves its purpose, it is inaccurate to the extent that it represents only one of several possible sub-queues with different numbers of stages. In a typical day, only the number of patients is counted and not the numbers of stages they pass through. Therefore when we observe the completion time distribution for all patients we are really observing the aggregate effects of several queues conflated within one completion time distribution.

Thus, we need a model of a form that is a weighted probability distribution in which the weights represent the proportion of patients in each queue. If there are two parallel queues, one with one stage and the other with two, the composite or hybrid probability distribution will be as follows:

$$P_c = pP_1(z) + (1-p)P_2(z)$$

This model is represented diagrammatically in Figure 4.

In the diagram λ represents the arrival rate of people into A&E and p represents the split between those patients that are subsequently discharged and those that are admitted as an inpatient or are referred. The process rate for those who are discharged is given the value μ_1 and the process rate for both stages of those who become inpatients or 'referreds' is μ_2. It should be noted that we are interested in the queue once it has reached a steady state hence the arrival rates for both stage 1 and stage 2 of the patients who are referred or become inpatients must be the same. It should also be noted that for the model we are more interested in the time taken to go through the individual paths, $t_\alpha = \dfrac{s_\alpha}{\mu_\alpha - p_\alpha \lambda}$ rather than the values of λ, μ_1 and μ_2 i.e. after

selecting a value for λ we can adjust the values of μ_1 and μ_2 to get suitable values for t_1 and t_2.

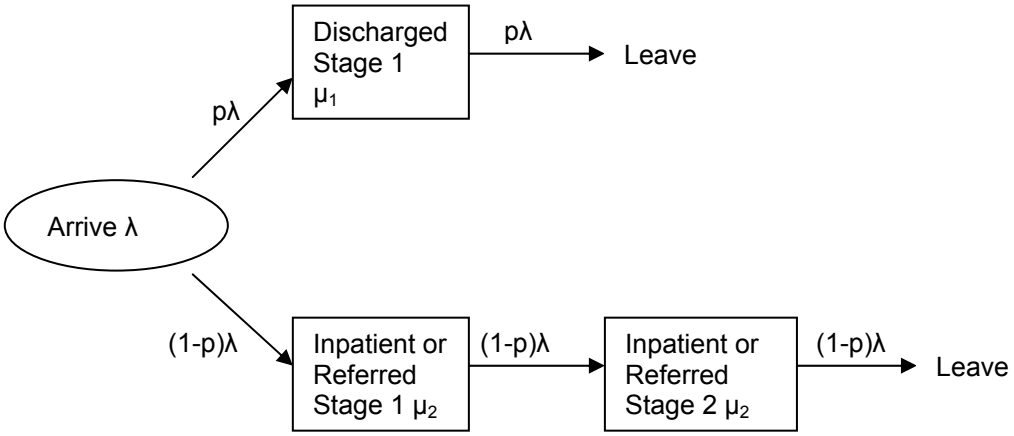

Figure 4: Diagram of the original model

In order to characterise and identify the correct distribution, we need to determine firstly how many processing stages are implicit in an observed distribution of completion times, and secondly the value of the weights (in this case *p* and *1-p*). We adopted the following simple procedure. Using the observed cumulative distribution of completion times and actual average completion time, we compared the predicted distribution based by systematically varying the set of weights for a 1,2 and 3 stage model. We then plotted the observed and predicted values to see how closely they matched over the z-range. A sample of the results is shown in Figure 5.

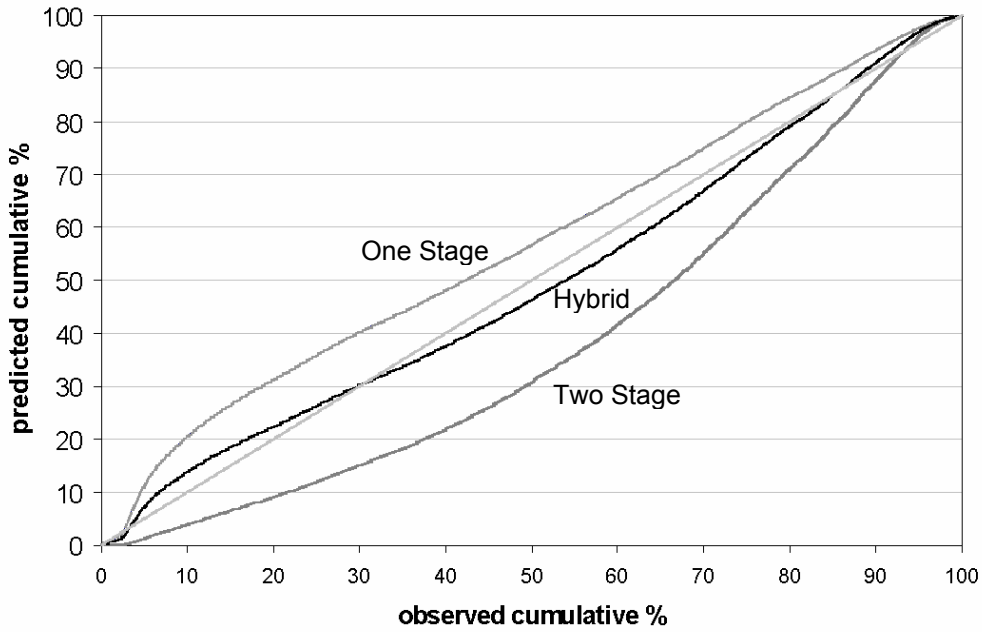

Figure 5: Comparison of the quality of fit generated by a 1 stage, 2-stage and hybrid model. The best 'fit' is the hybrid model with 60% of flows through a 1-stage queue and 40% with a 2-stage queuing model. Perfect agreement would be on the diagonal line.

By experimentation we found that the best results from this model are obtained using two queues in parallel with 60% of flows through a 1-stage queue and 40% through a 2-stage queue. It turns out that these weights are almost identical to actual percentage flows of patients categorised into those discharged home and those admitted or referred. This model is labelled 'hybrid' in Figure 5 and the closeness of the fit to the diagonal line is an indication of how well the model fits the data.

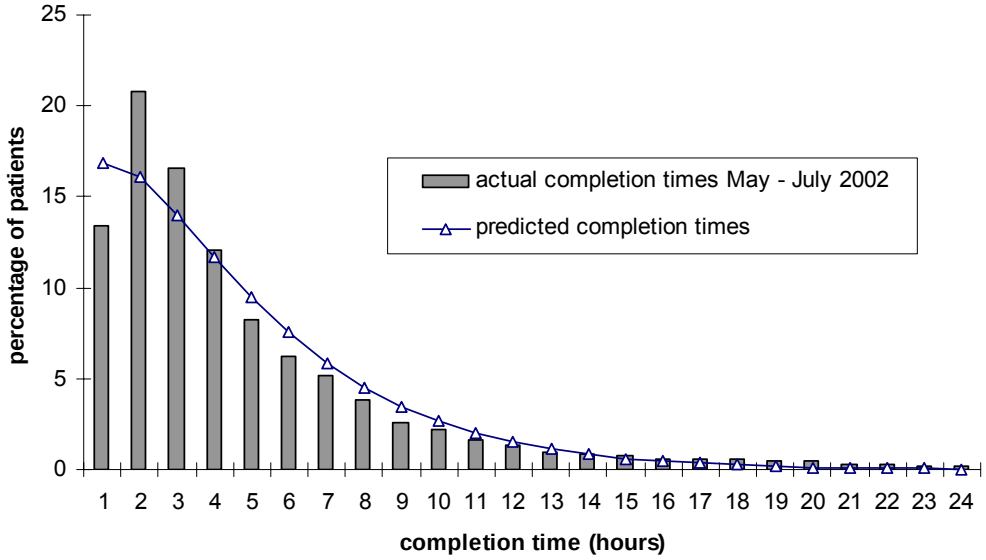

Figure 6: Comparison of observed and predicted completions times based on hybrid model and May- July 2002 data.

If we plot the actual data and the predicted completion times according to their relative frequency we obtain the results shown in Figure 6 which is taken from completion times of over 19,000 A&E patients in May - July 2002. The results indicate a reasonably good fit over the range, although the quality of fit is poorer in the 1-2 hour range. This difference, an over-estimate of up to 1-hour and an under-estimate between 1 and 2 hours can be traced to the 'triage' bottleneck, which patients must pass through following registration.

Accepting that this was likely to be the best possible representation using this first model, we recalculated the ready-reckoner accordingly using the hybrid model deriving two variants, which represent two sets of solutions to the equation.

$$P_c = pP_1(z) + (1-p)P_2(z)$$

The first variant shown in Figure 7 establishes, for a given average completion time, the time taken to complete a given percentage of patients. The second variant shown in Figure 8 establishes, for a given average completion time, the percentage of patients outstanding after a given time in the A&E Department.

To give an example, suppose the average completion time is 4 hours. The dotted line in Figure 7 indicates , in this case, 70% of patients would be cleared in 5 hours. The comparable result in Figure 8, also based on a 4-hour average, shows that 30% of

patient would still be outstanding after 5 hours and so the two ready-reckoners are complementary.

Compared with Figure 3, the 1-stage model, the time to clear the same percentage based on the hybrid model is therefore similar. However, larger differences can occur depending on the choice of average completion time and percentile.

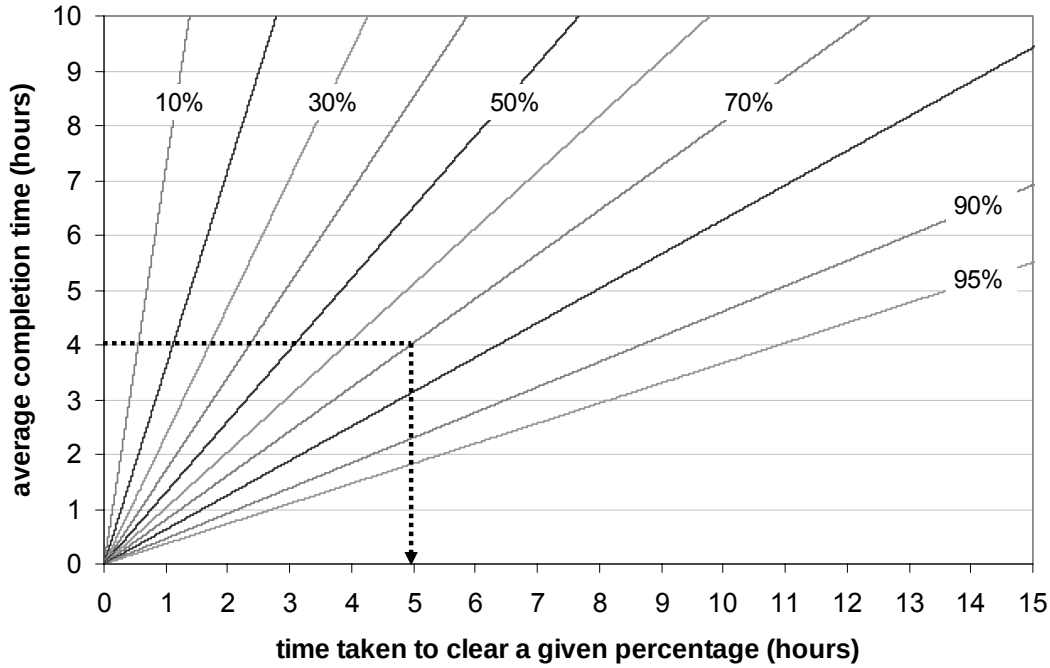

Figure 7: The time take to clear a given percentage of patients based on the hybrid model.

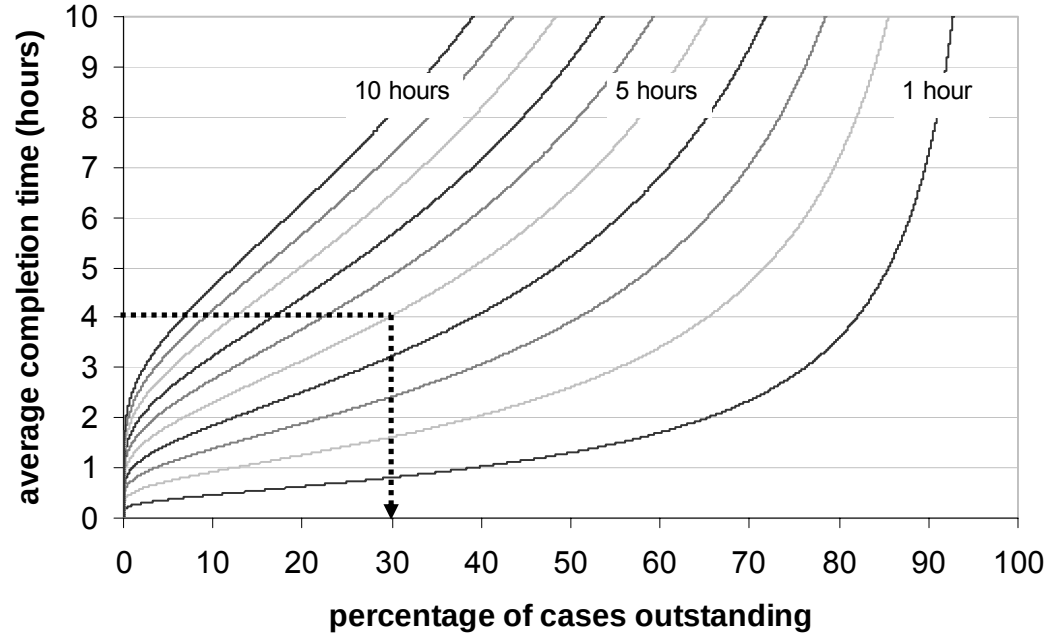

Figure 8: The percentage of cases still outstanding after the given number of hours in A&E, for a given average completion time.

Alternative model

As can be seen above, the original model offers a reasonable amount of accuracy over the data range. However, the simplistic nature of the structure of the model means that no extra accuracy can be achieved. It was decided that improvements could be made if the model structure could become more flexible with one obvious area of constraint being that both stages in the two-stage model have to be of the same length. This constraint means that the distribution is a simple gamma distribution but obviously not all stages in the process will take the same amount of time. For example, the length of time waiting for X-rays to be taken does not have to equal the length of time waiting for these X-rays to be analysed and treatment prescribed. The model was thus changed so that the process time of the two processes could be different. The model can be shown diagrammatically as in Figure 9.

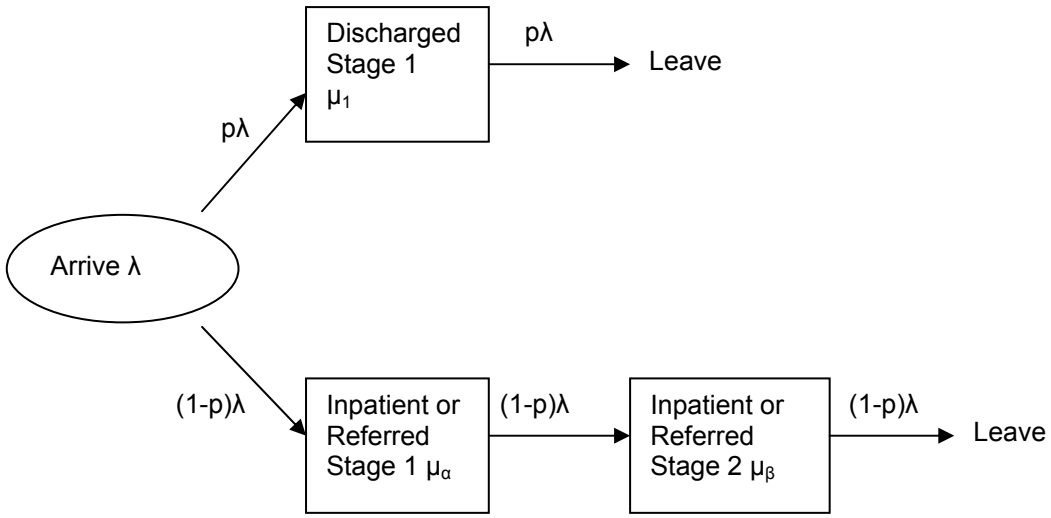

Figure 9: Diagram of alternative model 1

It can be seen that there is very little change to the model, the only difference being that μ_2 has been replaced by the process rates μ_α and μ_β. It should be noted that, once again, because the model is in steady state the arrival rate and exit rate must be the same for all processes.

One problem with allowing the rates to vary in the second path is that the distribution is no longer gamma but is now a hypo-exponential distribution. These distributions are not used as frequently in queuing theory as the distribution function is not as simple as that of the gamma distribution. However, for only two stages the distribution is still relatively straight forward.

As proved in Ross (1997), pp245-247:

Z = total time in for patients who become inpatients or 'referreds' = $T_1 + T_2$

$$T_1 \sim \exp(\mu_\alpha - (1-p)\lambda) \quad T_2 \sim \exp(\mu_\beta - (1-p)\lambda)$$
$$T_1 \sim \exp(\mu_1) \qquad T_2 \sim \exp(\mu_2)$$

$$
\begin{aligned}
f_{T_1+T_2}(z) &= \int_0^t f_{T_1}(s)\, f_{T_2}(t-s)ds \\
&= \int_0^t \mu_1 \exp[-\mu_1 s]\mu_2 \exp[-\mu_2 (t-s)]ds \\
&= \mu_1\mu_2 \exp[-\mu_2 t]\int_0^t \exp[-(\mu_1-\mu_2)]ds \\
&= \frac{\mu_1}{\mu_1-\mu_2}\mu_2 \exp[-\mu_2 t](1-\exp[-(\mu_1-\mu_2)t]) \\
&= \frac{\mu_1}{\mu_1-\mu_2}\mu_2 \exp[-\mu_2 t] + \frac{\mu_2}{\mu_2-\mu_1}\mu_1 \exp[-\mu_1 t]
\end{aligned}
$$

$$
\begin{aligned}
F(Z) &= 1-\int_z^\infty f_{T_1+T_2}(t)dt \\
&= 1-\int_z^\infty \left(\frac{\mu_1}{\mu_1-\mu_2}\mu_2 \exp[-\mu_2 t] + \frac{\mu_2}{\mu_2-\mu_1}\mu_1 \exp[-\mu_1 t]\right)dt \\
&= 1-\left[\frac{-\mu_1}{\mu_1-\mu_2}\exp[-\mu_2 t] - \frac{\mu_2}{\mu_2-\mu_1}\exp[-\mu_1 t]\right]_z^\infty \\
&= 1-\left(\frac{\mu_1}{\mu_1-\mu_2}\exp[-\mu_2 z] + \frac{\mu_2}{\mu_2-\mu_1}\exp[-\mu_1 z]\right) \\
&= 1-\left(\frac{\mu_1 \exp[-\mu_2 z] - \mu_2 \exp[-\mu_1 z]}{\mu_1-\mu_2}\right)
\end{aligned}
$$

On first appearances this function appears to remove all references to expected total time taken in the system and the arrival rate of patients into the system. These in fact are included as they provide constraints on the values of μ_1 and μ_2.

Total expected time in system = t = $\frac{1}{\mu_1} + \frac{1}{\mu_2} = \frac{1}{\mu_\alpha - (1-p)\lambda} + \frac{1}{\mu_\beta - (1-p)\lambda}$

The easiest parameters to measure and hold constant are the arrival rate and the total time in the system. When this is the case we can calculate μ_β in terms of μ_α. If we let $\lambda_2 = (1-p)\lambda$ then

$$
\begin{aligned}
\frac{1}{\mu_\beta - \lambda_2} &= t - \frac{1}{\mu_\alpha - \lambda_2} \\
\frac{1}{\mu_\beta - \lambda_2} &= \frac{t(\mu_\alpha - \lambda_2) - 1}{\mu_\alpha - \lambda_2} \\
\mu_\beta - \lambda_2 &= \frac{\mu_\alpha - \lambda_2}{t(\mu_\alpha - \lambda_2) - 1} \\
\mu_\beta &= \frac{\mu_\alpha - \lambda_2}{t(\mu_\alpha - \lambda_2) - 1} + \lambda_2
\end{aligned}
$$

or $\mu_2 = \dfrac{\mu_1}{t\mu_1 - 1}$

With further constraints that

$\mu_\alpha > \lambda \Rightarrow \mu_1 > 0$
$\mu_\beta > \lambda \Rightarrow \mu_2 > 0$

Fitting the model to the data

The new model was fitted to the original data. This was done by starting with the parameters from the original model and then changing the parameters relating to the processes for the referred patient path to get a better fit. This was achieved partly by sight and partly by iterative means to get a fit.

Below is the diagram of the fit with the original model as well.

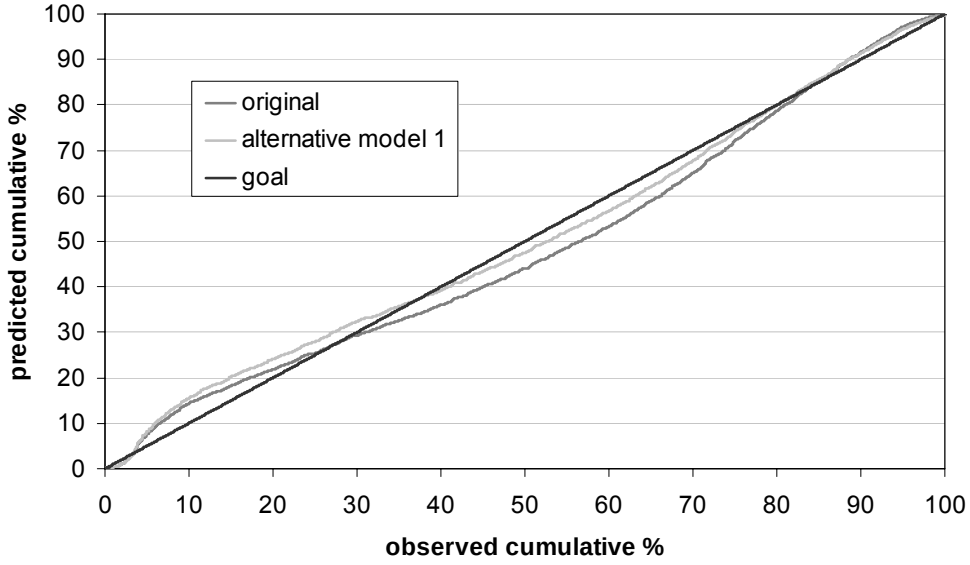

Figure 10: Comparison of the quality of fit generated by the original and alternative model. Perfect agreement would be on the diagonal line.

As can be seen the added flexibility of the new model has allowed us to 'stretch' the shape of the curve so that it now fits better for the range of patients between 35% and 80%. However, this has been at the expense of accuracy early on in the distribution (between 10% and 30%) as the fit moves away from the observed data at this point.

While the fit is better there are two main problems:

1. The fit has now deteriorated at earlier durations.
2. Although the fit overall is better it can be argued that the extra complexity is not worth it, especially given point 1.

As a result we have not produced the same graphs as for the original model but instead tried to improve the fit at early durations with a new model.

Alternative model 2

The problem with the early durations (i.e. those patients processed within 2 hours) is that the shape of the curve is wrong. Too many patients were expected to be processed in under 1 hour and not enough patients were expected to be processed between one and two hours. By looking at the shapes of curves in Figure 2 it can be seen that the problem lies with using a one stage model for the patients who are discharged. This shape would imply that more patients are processed in the first hour than the second, which is not what is observed. Our new model therefore changes the path for discharged patients from a one stage to a two stage process.

Once again this makes sense when applied to the way that patients will be processed. Any tests will mean that patients will first wait for the tests and then wait for the test results and any actions that need to be taken. A two stage process should therefore be a more suitable model.

The time taken to process a patient will be short, with a total mean time of less than 3.5 hours. It was therefore decided that the process rate for the two stages for discharged patients should be equal allowing the mathematics to be kept simple. The new model is shown in Figure 11.

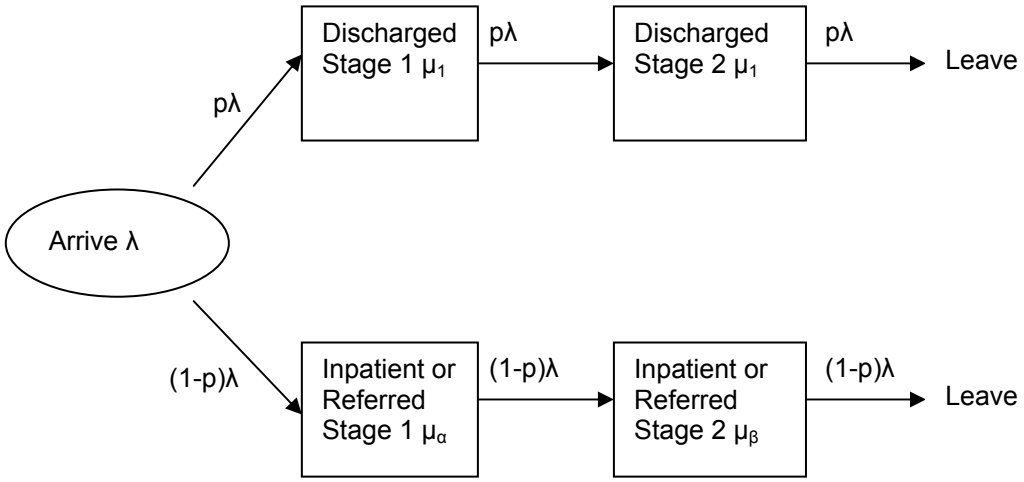

Figure 11: Diagram of alternative model 2

The mathematics of this model are very similar to the one above but with a gamma distribution replacing the exponential distribution.

Hence the time spent in the system for discharged patients is

$$P(z) = 1 - \exp(\frac{-z2}{t}) \sum_{i=0}^{i=1} \left(\frac{z2}{t} \right)^i / i!$$

where $t = \dfrac{2}{\mu_1 - p\lambda}$.

The new model was plotted against the observed rates with the results shown in Figure 12.

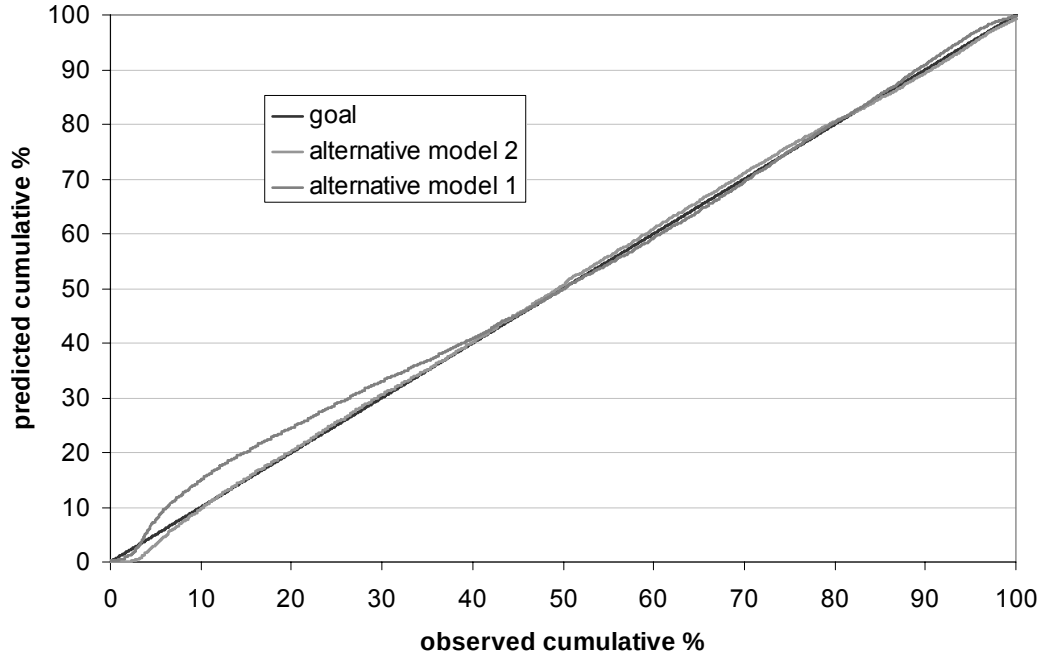

Figure 12: Comparison of the quality of fit generated by alternative model 1 and alternative model 2. Perfect agreement would be on the diagonal line.

The effect of this change in the model is far greater than the first alternative model. Patients who are processed quickly are far more accurately modelled as the two stage model is more suitably shaped than the original one stage. However, this is not true for patients who are processed very quickly (the first 5% of patients). This problem addressed below.

It should also be noted that the model is more accurate for the end of the distribution as well i.e. after 85%. However, this accuracy was reached by decreasing the mean time for the discharged patient path compared to the previous models and increasing the mean time for the inpatient and referred path. There is a problem though in the fact that these new expected times are quite different to the average times for discharged and inpatients. We therefore need to revaluate the purpose of the model.

Definition of treatment

It was stated at the beginning of the paper that our objective was to model results rather than the micro stages that the patients passed through as this created too much complexity and also created a very unstable model. It is possible therefore to change the paths from the original titles to ones of ‘small treatment’ and ‘large treatment’. Note that these words do not actually indicate the severity of the patient’s condition. This is because it is possible for two patients to have the same level of treatment but for one to be discharged afterwards whereas the other is referred (based on age, general health, etc).

It is also 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. It is further possible that a patient will be quickly 'patched up' and then referred for more extensive treatment. For clarity, the two paths will therefore be referred to for the rest of the paper as 'small treatment' and 'large treatment'. This definition conforms quite closely to the split between 'minors' and 'majors' often used in A&E departments to distinguish severity and treatment areas, and so it is helpful to keep this terminology in mind in what follows.

Alternative model 3

As noted above alternative model 2 creates a new problem because we are using the gamma model for patients having small treatment. Those patients who are processed very quickly through this path are modelled incorrectly as it is impossible for the required number of patients to pass through the two stages quickly enough. If we try to fit the model to these patients then the length of time in small treatment is too small on average and hence the model overstates the number of patients who will be processed in the first few hours. To try and correct this problem alternative model 3 was formulated.

The raw data we have on times taken to process patients indicates that some patients are processed in under 10 minutes. To model these patients it was decided that for 'short treatment' there should be an initial stage where people are assessed. If the assessment is that they can be dismissed as not needing the care provided only by a hospital then they exit the system whereas other patients move into the original 'short treatment' path. This route can also be seen as patients who 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 which fits the early data well as shown in Figure 13. The actual paths of this model appear to be sensible and reflect the processes that patients find themselves going through in hospital. We can imagine that a person arriving with obvious problems will not need to be assessed and can immediately go to the large treatment path.

A patient who is not acute will be assessed first and then may either go home or will enter the short treatment path. However, as discussed in the previous model, the data fit far better if we assume that large treatment does not necessarily mean acute and small treatment only minor ailments. In fact this fits reality rather better and allows for acute emergencies that are treated on arrival. This has the further advantage of keeping the model simpler than it would otherwise have been.

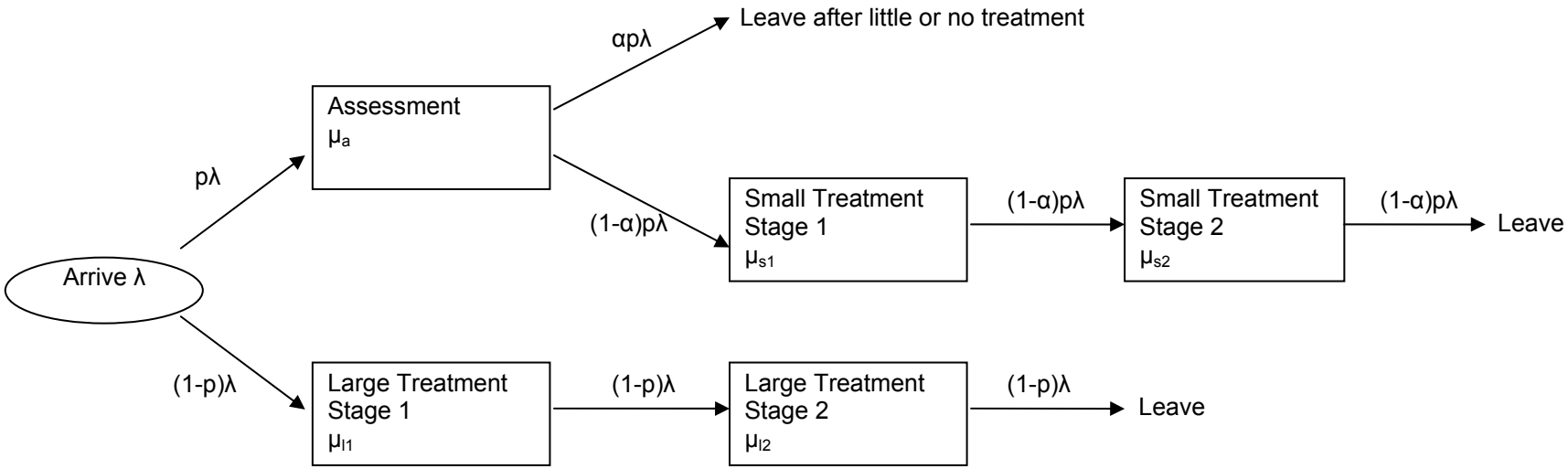

Figure 13: Diagram of alternative model 3

Expected time spent in the three paths

No/Little Treatment Path

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

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

where $t = \frac{1}{\mu_a - p\lambda}$.

Short treatment path

The short treatment path is now a three stage hypo-exponential distribution. 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
$$\mu_1 = \mu_a - p\lambda$$
$$\mu_2 = \mu_{s1} - (1-\alpha)p\lambda$$
$$\mu_3 = \mu_{s2} - (1-\alpha)p\lambda$$

(This distribution is derived in the appendix.)

Long treatment path

The long treatment path remains 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
$$\mu_1 = \mu_{l1} - (1-p)\lambda$$
$$\mu_2 = \mu_{l2} - (1-p)\lambda$$

Fitting the model to the data

The best fit for the data was found (again using sight and iterative means) and the plot is shown in Figure 14. As can be seen, the new model improves on the previous one for the first 10% of patients processed. However, the first few percent of patients to be processed is still poorly replicated but as these times tend to be very volatile on a monthly basis this is not a major concern. The improvement in the model can be seen

by recalculating Figure 6 using the new model. Figure 15 highlights the improvement achieved in the modelling of patients with completion times of less than 4 hours.

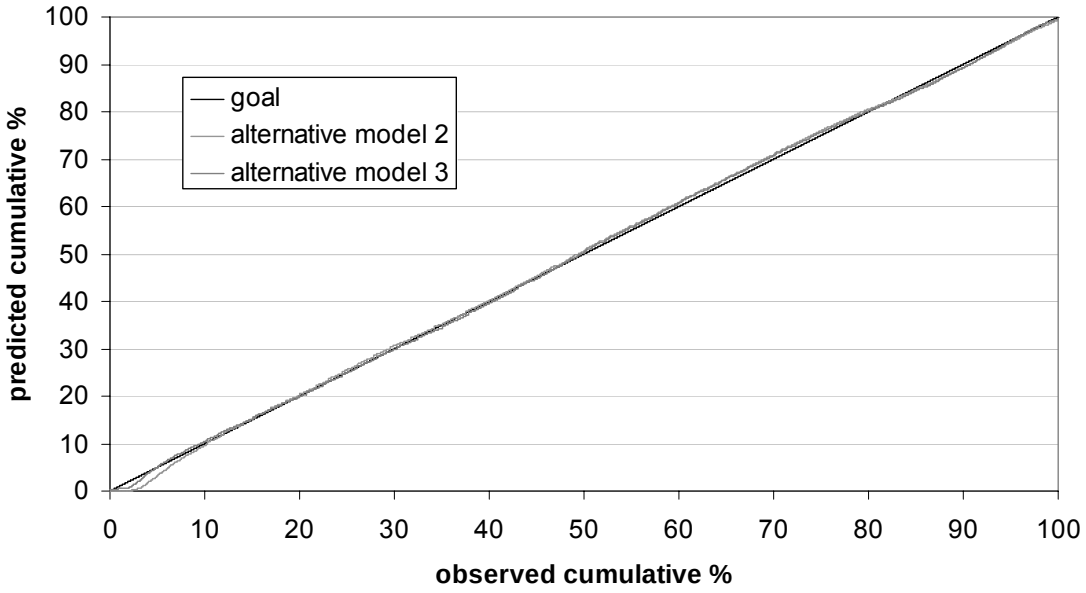

Figure 14: Comparison of the quality of fit generated by alternative model 2 and alternative model 3. Perfect agreement would be on the diagonal line.

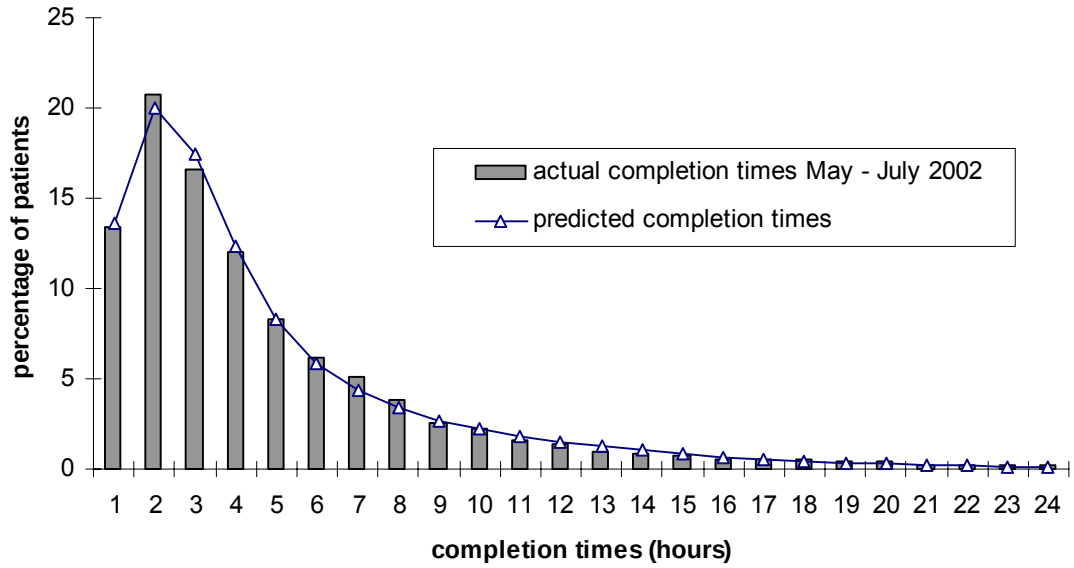

Figure 15: Comparison of observed and predicted completions times based alternative model 3 and May - July 2002 data.

Ready reckoner: alternative model 3

For this final model we recalculated the ready reckoners for the first model (originally shown in Figures 7 and 8). These graphs are given below in Figures 16 and 17.

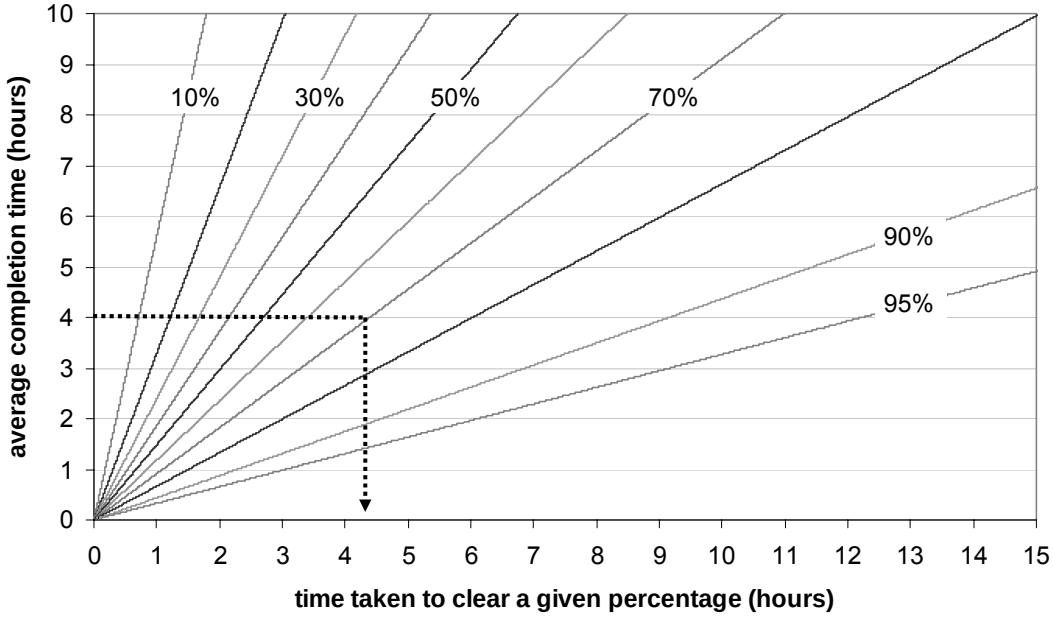

Figure 16: The time take to clear a given percentage of patients based on the alternative model 3.

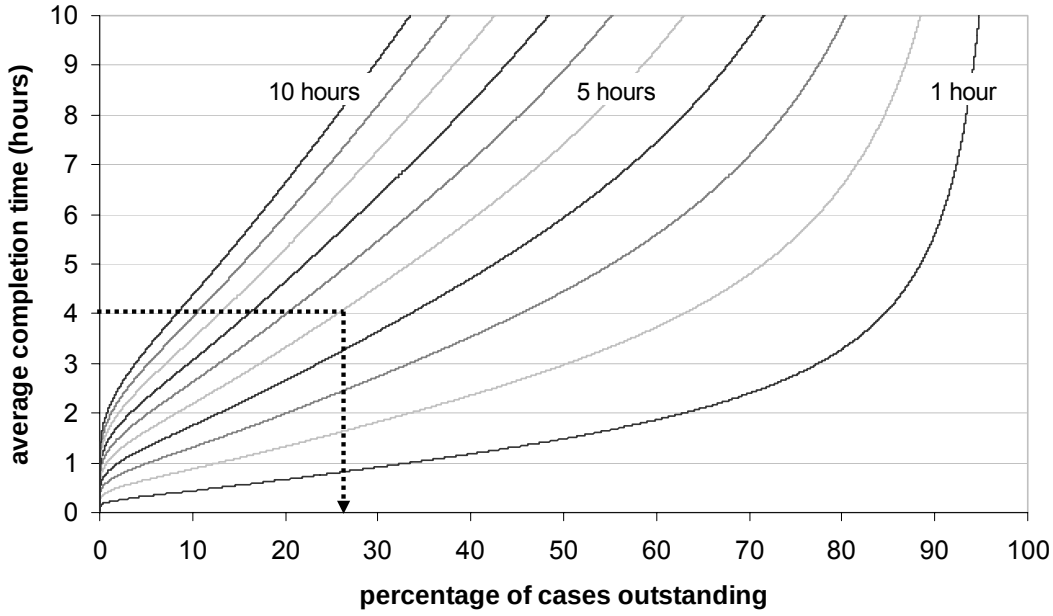

Figure 17: The percentage of cases still outstanding after the given number of hours in A&E, for a given average completion time.

We can see that the time taken to clear the 70% of cases for a four hour average completion time has fallen from five hours to four and a half hours. A similar change is also seen in the second graph, which shows that less than 30% of cases are outstanding after five hours.

It can be noted in passing that although the model has become more complex the ready-reckoner still has straight lines for all percentiles. This is as a result of the

model being built on modified exponential and gamma models where the standard deviations are a constant ratio of the mean and hence increase proportionally as the mean time is increased.

Comparing models

It is useful at this stage to see how the fitted data for the different models compare with each other. As has been noted, as we have advanced through the various models there has been a change in fits such that the short treatment path has had a shorter average process time for alternative models 2 and 3, and the long treatment path has had a longer average process time for the same models. Table 1 gives the average waiting times for the various paths for the four models based on the data.

		Original Model	Alternative Model 1	Alternative Model 2	Alternative Model 3
Path 1	Percentage of Patients	60	60	60	6
	Average Time (hours)	3.5	3.5	2.2	0.4
Path 2	Percentage of Patients	40	40	40	54
	Average Time (hours)	5.25	5.25	7.3	2.53
Path 3	Percentage of Patients	0	0	0	40
	Average Time (hours)	0	0	0	7.3
Total Average Time (hours)		4.2	4.2	4.24	4.31

Table 1: A comparison of the implied average waiting times for patients in May – July 2002 for the four fitted models

As can be seen, the total average waiting time has increased for the models but not by a large amount. However, there has been a noticeable change in the predicted process times for the various paths. As has been discussed earlier, these changes can be attributed to the assumed changes in the processes we are modelling i.e. the change from defining small treatment to be the amount of time spent treating the patient rather than the severity of the ailment.

Ready-reckoners: comparing original to alternative model 3

The ready-reckoners from the original model and alternative model 3 can be brought into the same diagram to show the different results given for the selected percentiles as shown in Figure 18. For each of the selected percentiles the original model is shown with a black line whereas the new model is shown as a grey line. For the 10^{th}, 25^{th} and 90^{th}, 95^{th}, 98^{th} percentile the original model has a smaller time taken to clear the given percentage than the new model. This is reversed for the 40^{th}, 60^{th} and 75^{th} percentiles where the new model gives the smaller time. Where the old model gives the quicker time the difference between the models is shaded a darker grey than when the new model gives the quicker time. This switch between which model gives the quicker time is of course perfectly logical as the new model shows patients being processed at a slower rate for the first and last patients but at quicker rates over the middle range.

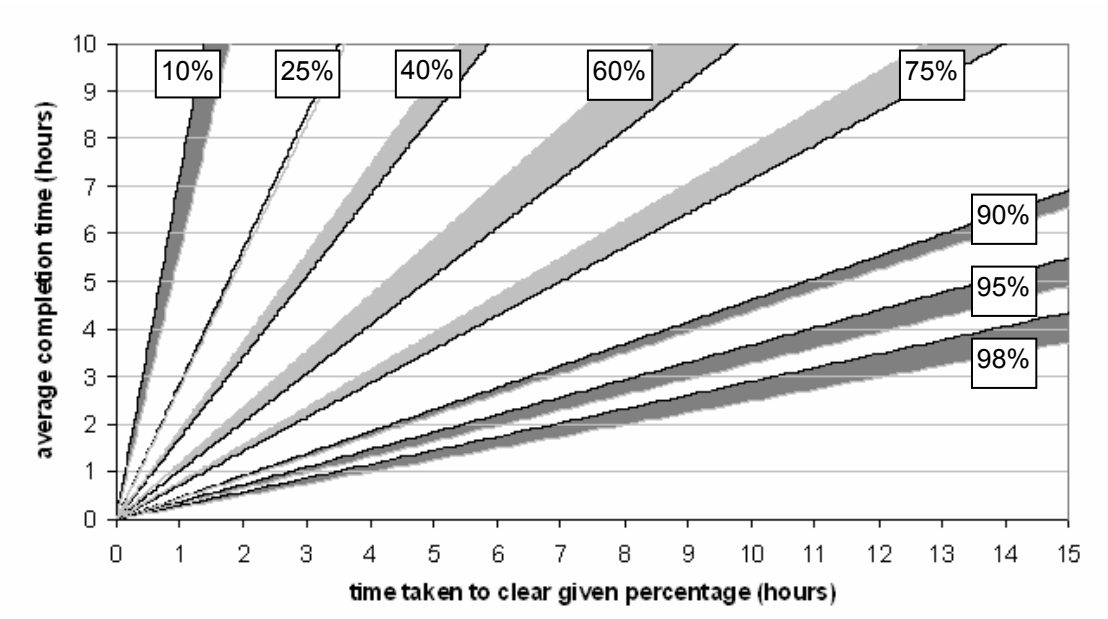

Figure 18: A comparison of the time take to clear a given percentage of patients based on the original model and alternative model 3.

However, some care must be taken when looking at these results as we have seen that the original model uses a slightly lower average completion time for the same set of data. It is therefore not strictly correct to look horizontally along the data points but the difference is not material and it is more important to get a feel for how the new model changes the expected results at the different cumulative levels.

We can also compare the other ready-reckoner where we are interested in the percentage of patients still outstanding after a particular period of time.

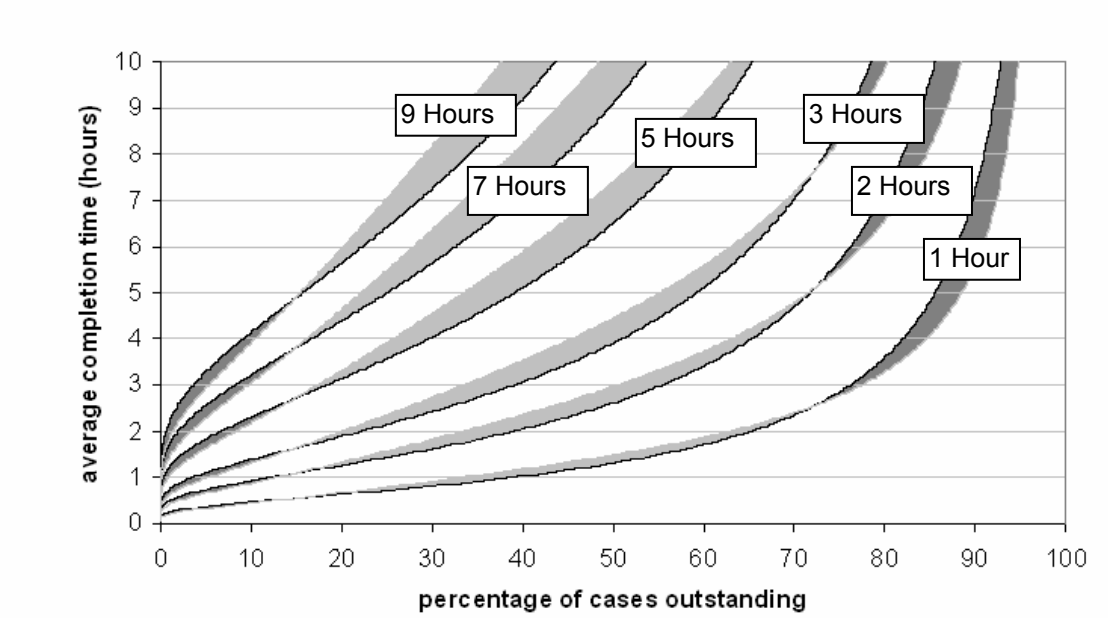

Figure 19: The percentage of cases still outstanding after the given number of hours in A&E, for a given average completion time.

The graph plotting average completion times and percentage of cases outstanding was also calculated for the new model and is shown above with the original model. This is shown in Figure 19. Once again the plots shown are sensible as they diverge between approximately the 15th and 75th percentile (which corresponds to 85% of patients processed and 25% of patients processed), which is where the major difference between the models occurs.

Suitability of the model on a monthly basis

So far we have concentrated on using all the data we have between the months of May and July (2002). However, our objective is to fit a model that will allow us to predict how quickly patients will be processed in the future. It is therefore necessary to check how well our model predicts the monthly processing of patients taking each month in turn. Below are three graphs (Figures 20 to 22) showing the expected results according to the model compared with the observed results. Note that for each month the same set of parameters has been used.

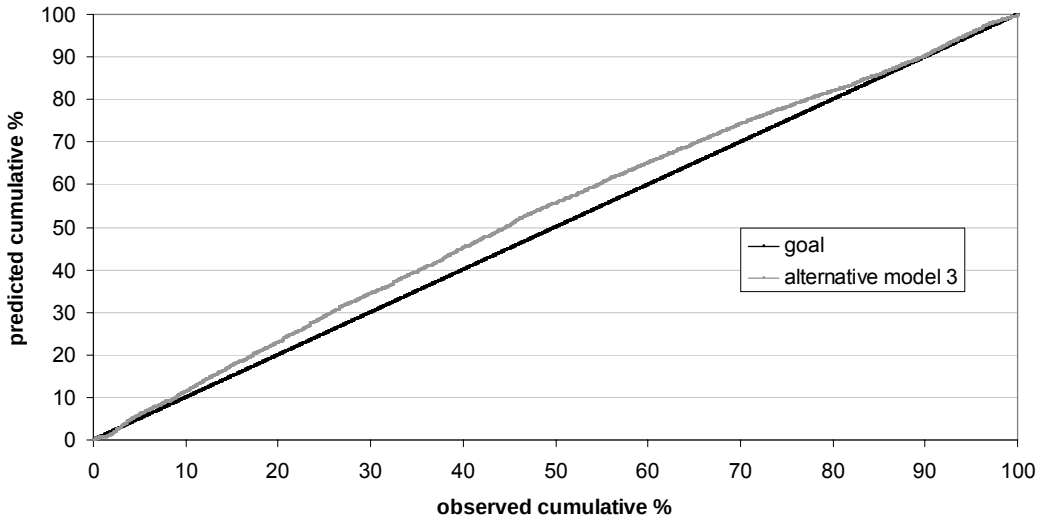

Figure 20: Comparison of the quality of fit generated by alternative model 3 using May 2002 data. Perfect agreement would be on the diagonal line.

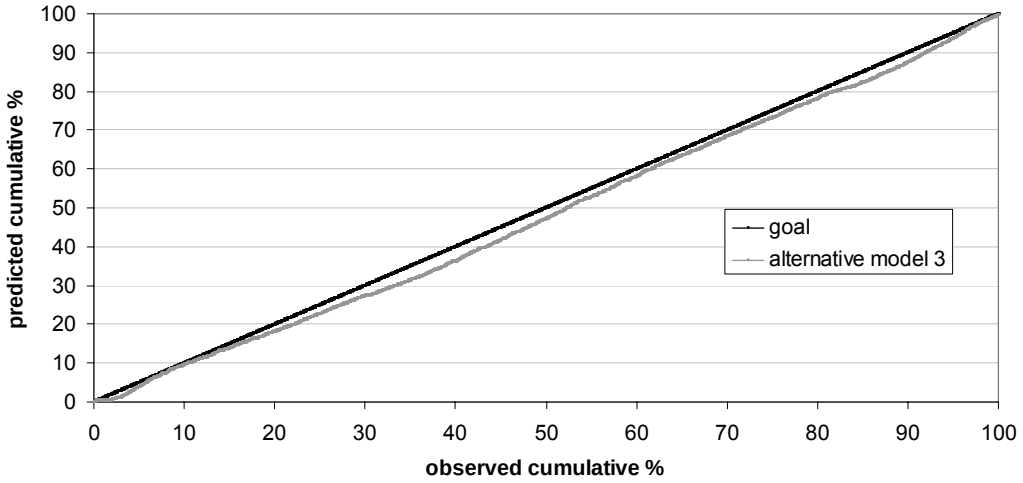

Figure 21: Comparison of the quality of fit generated by alternative model 3 using June 2002 data. Perfect agreement would be on the diagonal line.

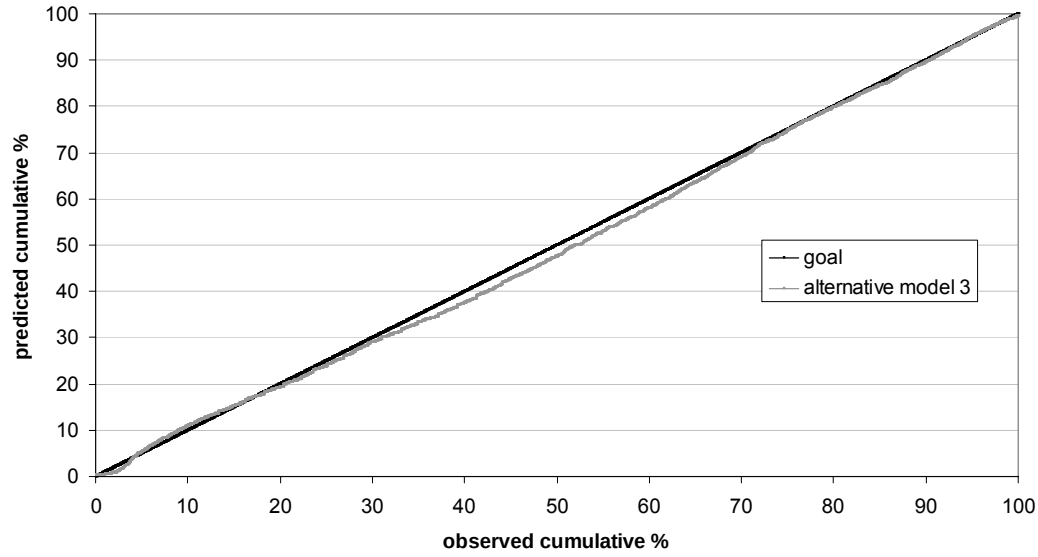

Figure 22: Comparison of the quality of fit generated by alternative model 3 using July 2002 data. Perfect agreement would be on the diagonal line.

Although the model does not quite reflect the observed rates each month, it can be seen that the actual shape of the expected curve compared to the observed curve is very similar. Typically in A&E departments the achievement of a consistent service will depend on numerous factors, for example staff availability or the random occurrence of major incidents. Despite random variations of this nature it is interesting to see that the model continues to provide reasonable predictions of monthly observations based on a single set of model parameters.

The average time that each patient spent in A&E was 4.62 hours in May, 3.98 hours in June and 4.2 hours in July. The calibrated model using all the data has an expected mean time of 4.3 hours and so we would expect the model to predict patients will be processed faster in May than they were. This can be seen to be the case by looking at Figure 20 which shows the alternative model line to be above the goal line. Similarly, we would expect the model to underestimate the time that patients were processed in June and once again this is what occurred, as Figure 21 shows the alternative model to be below the goal line. As July also had a mean time of less than 4.3 hours we would expect the alternative line to be below the goal line but to a lesser extent than June. Again, Figure 22 shows this to be true.

Re-designating patients: Alternative model 4

Many of the reasons for looking at the process times of patients is for the hospital to meet stringent A&E targets. One possible solution to this problem is by reconsidering the point at which a patient can be considered to have been discharged 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'. 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 four hours. 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.

This is currently a two stage process and 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 reclassified treatment. The model is therefore changed so that the 'long treatment' path becomes a one-process model where the time spent in the process is equal to half of that spent in the long treatment path of alternative model 3. This gives us model 4 which is represented in Figure 23.

Issues arising in relation to the 4-hour national target

By the end of 2003 A&E departments were expected to achieve a standard of 90% completion within 4 hours. Since then the standard has been further tightened to 98%, a difference of 8 percentage points. Figure 24 is a partial ready-reckoner, based on alternative model 3 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 hours. This equates to an average completion time on the vertical axis of 1.75 hours (1:45 hour: minutes).

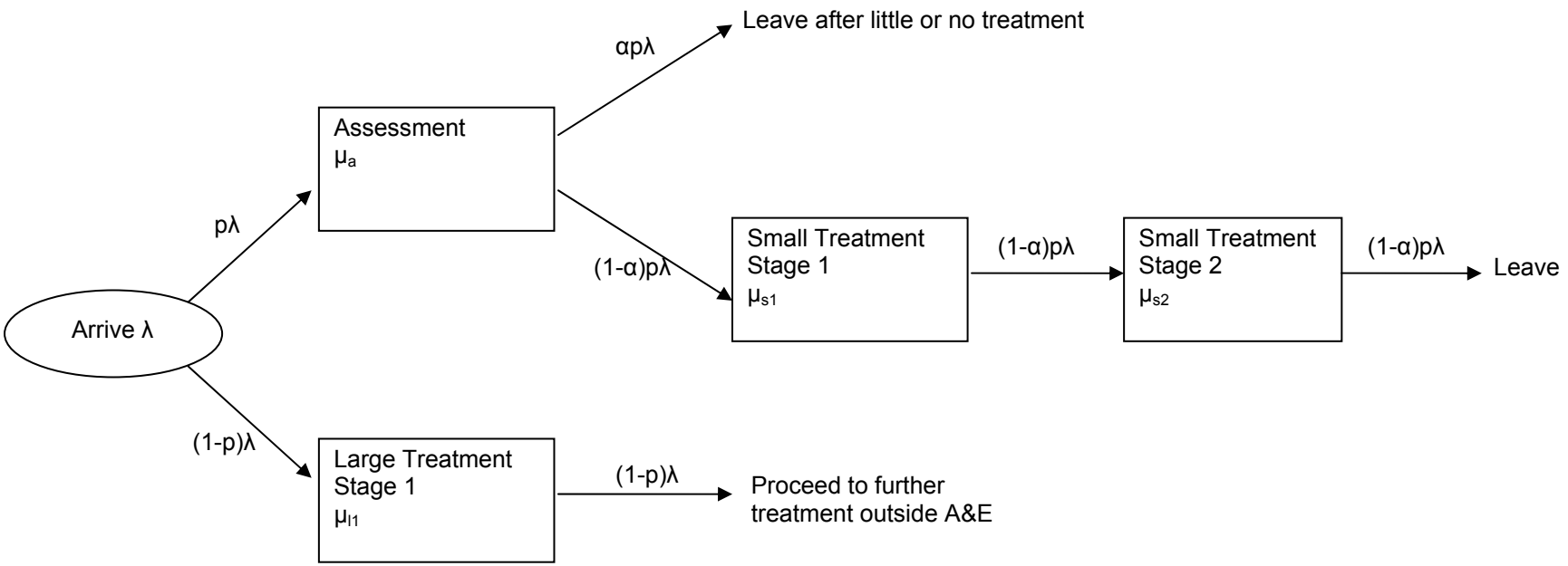

Figure 23: Diagram of alternative model 4

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

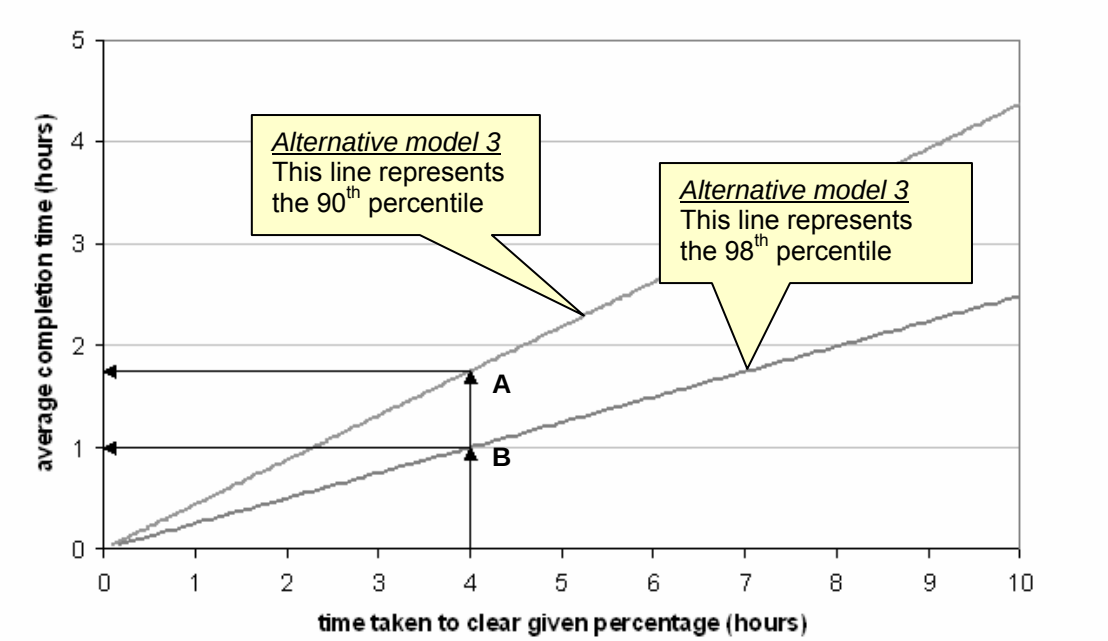

Figure 24: Alternative model 3showing how the average completion time changes when moving from the 90th to the 98th percentile represented by points A and B

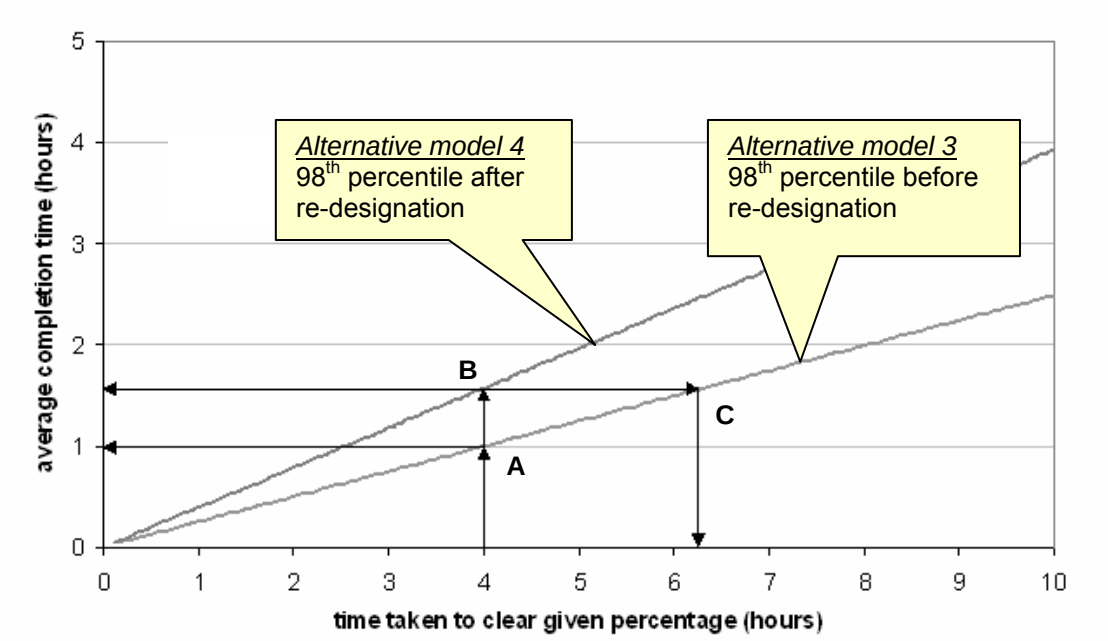

Figure 25: A comparison of the time taken to complete the A&E treatment of 98% of patients based on alternative models 3 and 4

Let us now consider alternative model 4 in which the long treatment path has been truncated after stage 1 following re-designation. (Figures 13 and 23 refer). Figure 25, by comparing the 98th percentiles in alternative model 3 and model 4, shows the difference this makes to the achievement of the 4 hour 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 Figure 13 i.e. 0.99 hours.

As a result of re-designation, we find that the average can be allowed to increase to 1.57 hours, whilst still achieving the target of 98% in 4 hours (point B, Figure 25). The time taken to treat 98% of patients under the original classification (alternative model 3) with this average completion time would be 6.31 hours (6:19 hours: minutes) which is, of course well, outside the target (point C, Figure 25). 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 new model reduces the chance of a particularly long treatment as only one process has to be passed through now rather than two.

Table 2 provides a comparison of both alternative model 3 and alternative model 4 for average completion times of 1 to 12 hours based on the 90th, 95th, 98th and 99th percentiles. As is 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 hours, 99% of patients would be completed in 15.65 hours (15:39 hours: minutes) with re-designation and 23.69 hours (23:41 hours: minutes) without re-designation.

Average Time	90%		95%		98%		99%	
	Mod 3	Mod 4	Mod 3	Mod 4	Mod 3	Mod 4	Mod 3	Mod 4
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

Table 2: The time taken to clear a specified percentage of patients given the average time for alternative model 3

Conclusions

We have shown that work flows in A&E departments can be represented as a queuing process that can be modelled accurately on the basis of monthly work flow data and is quite general. The components or stages in the model replicate the situation on the ground quite accurately, although this was not the main purpose of the model which

as previously noted was to replicate the overall process by accurately predicting completion times.

A key result is the ability to relate Government targets which are expressed as the percentage of cases completed in a given time from the time of arrival to discharge to the underlying average time in the system (or alternatively the percentage of cases outstanding in the system for any given average completion time). Furthermore different versions of the model showed that all were generally robust at high clearance percentiles giving fairly comparable results.

The stated aim of the NHS Plan (2000) was to complete the treatment of 75% of patients in 1 hour with an eventual aim of clearing all patients (100%) inside 4 hours from the time of arrival to discharge. Alternative model 3 in this paper shows that the original 75% standard would translate into an average completion time of 0.78 hours (or 47.0 minutes). By using alternative model 4, which re-designates some patients, this underlying average completion time is increased to 1.16 hours (or 69.8 minutes). (Note this is not the average time of alternative model 4, rather it is the average time for alternative model 3 before the long treatment is truncated).

Today the target is to clear 98% of patients in 4 hours. Meeting this target using the re-classification of patients under model 4 is the equivalent of only 47.5% patient completion using alternative model 3. Thus the possibility of re-designating some patients is clearly a significant aid towards achieving the 98% target where re-designation has occurred. Indeed it seems highly improbable that the original target could ever have been achieved without changing the basis for counting patients through the system.

Based on our analysis we have shown that some improvements in completion times since the NHS plan can be attributed to genuine efficiency improvements but also to re-designation. This is not necessarily a criticism of re-designation if it results in patients being cared for in more appropriate surroundings rather than in the same setting but with a ‘different label’. However, in comparing the ‘old’ with the ‘new’ is not comparing ‘like’ with ‘like’ and so true improvements are somewhat less than headline figures might suggest.

In view of the large differences in average completion time that result 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 it was introduced. For one thing it raises concerns about the credibility of some reported performances. The original 75% in four hours equating to a completion average of 47 minutes seems wildly optimistic in retrospect (which may be why it was abandoned), whereas a 1-hour average borders on the impossible (alternative model 3).

Even with re-designation (alternative model 4) an average completion time of 1.5 hours is extremely stretching. We conclude therefore that 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

hours in the system[4]). We conclude 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. 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 four hours (say 90%), but in a walk in or urgent care centre 95% or higher (but never more than 98%).

In conclusion 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.

References

L. Mayhew and E. Carney Jones (2003) Evaluating a New Approach for Improving Care in an Accident and Emergency Department. Cass Business School (60pp).

Mayhew L. (1987) Resource Inputs and Performance Outputs in Social Security. *Jrnl Operational Research Society*, 38(10), 913-928.

Brailsford, S.C., V.A. Lattimer, P. Tamaras, J.C. Turnbull (2004) Emergency and on demand health care: modelling a large and complex system. Journal of the Operational Research Society 55, 34-42.

Gorunescu, F., S.I. McClean, P.H Millard (2002) A queuing model for bed occupancy management and planning of hospital. Journal of the Operational Research Society 53, 19-24.

Ross, S. (1997) Introduction to Probability Models Sixth Edition. Academic Press, San Diego US.

Appendix

Derivation of a Three Stage Hypo-Exponential Probability Distribution

Assume we have three exponential distributions each with a different parameter.

$$X_1 \sim \exp(\lambda_1) \quad X_1 \sim \exp(\lambda_2) \quad X_1 \sim \exp(\lambda_3)$$

Then we are able to derive the probability function of $X_1 + X_2$ as shown in Ross (1997):

[4] That this possibility occurs is easily detected because the completion time distribution would be truncated after 4 hours i.e. there would be a cliff-edge effect in a typical graph showing the distribution of completion times.

$$
\begin{aligned}
f_{X_1+X_2}(t) &= \int_0^t f_{X_1}(s)\, f_{X_2}(t-s)ds \\
&= \int_0^t \lambda_1 \exp[-\lambda_1 s]\mu_2 \exp[-\lambda_2 (t-s)]ds \\
&= \lambda_1\lambda_2 \exp[-\lambda_2 t]\int_0^t \exp[-(\lambda_1-\lambda_2)]ds \\
&= \frac{\lambda_1}{\lambda_1-\lambda_2}\mu_2 \exp[-\lambda_2 t](1-\exp[-(\lambda_1-\lambda_2)t]) \\
&= \frac{\lambda_1}{\lambda_1-\lambda_2}\lambda_2 \exp[-\lambda_2 t] + \frac{\lambda_2}{\lambda_2-\lambda_1}\lambda_1 \exp[-\lambda_1 t]
\end{aligned}
$$

We are then able to repeat this process to find the combination of all three distributions.

$$
\begin{aligned}
f_{X_1+X_2+X_3}(t) &= \int_0^t f_{X_1+X_2}(s) f_{X_3}(t-s)ds \\
&= \int_0^t \left\{ \frac{\lambda_1}{\lambda_1-\lambda_2}\lambda_2 e^{-\lambda_2 s} + \frac{\lambda_2}{\lambda_2-\lambda_1}\lambda_1 e^{-\lambda_1 s} \right\} \lambda_3 e^{-\lambda_3 (t-s)} ds \\
&= \frac{\lambda_1\lambda_2\lambda_3}{\lambda_1-\lambda_2} e^{-\lambda_3 t}\int_0^t e^{-(\lambda_2-\lambda_3)s} ds + \frac{\lambda_1\lambda_2\lambda_3}{\lambda_2-\lambda_1} e^{-\lambda_3 t}\int_0^t e^{-(\lambda_1-\lambda_3)s} ds \\
&= \frac{\lambda_1\lambda_2\lambda_3 e^{-\lambda_3 t}}{(\lambda_1-\lambda_2)(\lambda_2-\lambda_3)}\left(1-e^{-(\lambda_2-\lambda_3)t}\right) + \frac{\lambda_1\lambda_2\lambda_3 e^{-\lambda_3 t}}{(\lambda_2-\lambda_1)(\lambda_1-\lambda_3)}\left(1-e^{-(\lambda_1-\lambda_3)t}\right) \\
&= \frac{\lambda_1\lambda_2\lambda_3 e^{-\lambda_3 t}}{(\lambda_1-\lambda_2)(\lambda_2-\lambda_3)} - \frac{\lambda_1\lambda_2\lambda_3 e^{-\lambda_2 t}}{(\lambda_1-\lambda_2)(\lambda_2-\lambda_3)} + \frac{\lambda_1\lambda_2\lambda_3 e^{-\lambda_3 t}}{(\lambda_2-\lambda_1)(\lambda_1-\lambda_3)} - \frac{\lambda_1\lambda_2\lambda_3 e^{-\lambda_1 t}}{(\lambda_2-\lambda_1)(\lambda_1-\lambda_3)} \\
&= \frac{\lambda_1\lambda_2\lambda_3 e^{-\lambda_1 t}}{(\lambda_2-\lambda_1)(\lambda_3-\lambda_1)} + \frac{\lambda_1\lambda_2\lambda_3 e^{-\lambda_2 t}}{(\lambda_1-\lambda_2)(\lambda_3-\lambda_2)} + \frac{\lambda_1\lambda_2\lambda_3 e^{-\lambda_3 t}}{(\lambda_1-\lambda_3)(\lambda_2-\lambda_3)}
\end{aligned}
$$

Simple integration then allows us to derive the probability distribution function of the combined distribution.

$$
\begin{aligned}
F_{X_1+X_2+X_4}(t) &= 1-\int_t^\infty f_{X_1+X_2+X_4}(s)ds \\
&= 1-\int_t^\infty \frac{\lambda_1\lambda_2\lambda_3 e^{-\lambda s}}{(\lambda_2-\lambda_1)(\lambda_3-\lambda_1)} + \frac{\lambda_1\lambda_2\lambda_3 e^{-\lambda_2 s}}{(\lambda_1-\lambda_2)(\lambda_3-\lambda_2)} + \frac{\lambda_1\lambda_2\lambda_3 e^{-\lambda_3 s}}{(\lambda_1-\lambda_3)(\lambda_2-\lambda_3)} ds \\
&= 1-\left[\frac{-\lambda_2\lambda_3 e^{-\lambda_1 s}}{(\lambda_2-\lambda_1)(\lambda_3-\lambda_1)} - \frac{\lambda_1\lambda_3 e^{-\lambda_2 s}}{(\lambda_1-\lambda_2)(\lambda_3-\lambda_2)} - \frac{\lambda_1\lambda_2 e^{-\lambda_3 s}}{(\lambda_1-\lambda_3)(\lambda_2-\lambda_3)} \right]_t^\infty \\
&= 1-\left(\frac{\lambda_2\lambda_3 e^{-\lambda_1 t}}{(\lambda_2-\lambda_1)(\lambda_3-\lambda_1)} + \frac{\lambda_1\lambda_3 e^{-\lambda_2 t}}{(\lambda_1-\lambda_2)(\lambda_3-\lambda_2)} + \frac{\lambda_1\lambda_2 e^{-\lambda_3 t}}{(\lambda_1-\lambda_3)(\lambda_2-\lambda_3)} \right)
\end{aligned}
$$

For our small treatment path (see Figure 13) the following are the values of λ_i's.

$$\lambda_1 = \mu_a - p\lambda$$
$$\lambda_2 = \mu_{s1} - (1-\alpha)p\lambda$$
$$\lambda_3 = \mu_{s2} - (1-\alpha)p\lambda$$

There are two things to note. Firstly, as patients are assessed between those needing treatment and those who leave with little or no treatment we are assuming that this occurs constantly at the assumed proportion so that patients entering the small treatment path are entering at an exponential rate still. Secondly, when it comes to selecting the process times there is no overall time constraint on the variables this time i.e. we do not select the required average waiting time and force the parameters to give us this value as we did with alternative model 2.

Cass Business School
City of London

FACULTY OF ACTUARIAL SCIENCE AND INSURANCE

Actuarial Research Papers since 2001

Report Number	Date	Publication Title	Author
135.	February 2001.	On the Forecasting of Mortality Reduction Factors. ISBN 1 901615 56 1	Steven Haberman Arthur E. Renshaw
136.	February 2001.	Multiple State Models, Simulation and Insurer Insolvency. ISBN 1 901615 57 X	Steve Haberman Zoltan Butt Ben Rickayzen
137.	September 2001	A Cash-Flow Approach to Pension Funding. ISBN 1 901615 58 8	M. Zaki Khorasanee
138.	November 2001	Addendum to "Analytic and Bootstrap Estimates of Prediction Errors in Claims Reserving". ISBN 1 901615 59 6	Peter D. England
139.	November 2001	A Bayesian Generalised Linear Model for the Bornhuetter-Ferguson Method of Claims Reserving. ISBN 1 901615 62 6	Richard J. Verrall
140.	January 2002	Lee-Carter Mortality Forecasting, a Parallel GLM Approach, England and Wales Mortality Projections. ISBN 1 901615 63 4	Arthur E.Renshaw Steven Haberman.
141.	January 2002	Valuation of Guaranteed Annuity Conversion Options. ISBN 1 901615 64 2	Laura Ballotta Steven Haberman
142.	April 2002	Application of Frailty-Based Mortality Models to Insurance Data. ISBN 1 901615 65 0	Zoltan Butt Steven Haberman
143.	Available 2003	Optimal Premium Pricing in Motor Insurance: A Discrete Approximation.	Russell J. Gerrard Celia Glass
144.	December 2002	The Neighbourhood Health Economy. A Systematic Approach to the Examination of Health and Social Risks at Neighbourhood Level. ISBN 1 901615 66 9	Les Mayhew
145.	January 2003	The Fair Valuation Problem of Guaranteed Annuity Options : The Stochastic Mortality Environment Case. ISBN 1 901615 67 7	Laura Ballotta Steven Haberman
146.	February 2003	Modelling and Valuation of Guarantees in With-Profit and Unitised With-Profit Life Insurance Contracts. ISBN 1 901615 68 5	Steven Haberman Laura Ballotta Nan Want
147.	March 2003.	Optimal Retention Levels, Given the Joint Survival of Cedent and Reinsurer. ISBN 1 901615 69 3	Z. G. Ignatov Z.G., V.Kaishev R.S. Krachunov
148.	March 2003.	Efficient Asset Valuation Methods for Pension Plans. ISBN 1 901615707	M. Iqbal Owadally
149.	March 2003	Pension Funding and the Actuarial Assumption Concerning Investment Returns. ISBN 1 901615 71 5	M. Iqbal Owadally

150.	Available August 2004	Finite time Ruin Probabilities for Continuous Claims Severities	D. Dimitrova Z. Ignatov V. Kaishev
151.	August 2004	Application of Stochastic Methods in the Valuation of Social Security Pension Schemes. ISBN 1 901615 72 3	Subramaniam Iyer
152.	October 2003.	Guarantees in with-profit and Unitized with profit Life Insurance Contracts; Fair Valuation Problem in Presence of the Default Option[1]. ISBN 1-901615-73-1	Laura Ballotta Steven Haberman Nan Wang
153.	December 2003	Lee-Carter Mortality Forecasting Incorporating Bivariate Time Series. ISBN 1-901615-75-8	Arthur E. Renshaw Steven Haberman
154.	March 2004.	Operational Risk with Bayesian Networks Modelling. ISBN 1-901615-76-6	Robert G. Cowell Yuen Y, Khuen Richard J. Verrall
155.	March 2004.	The Income Drawdown Option: Quadratic Loss. ISBN 1 901615 7 4	Russell Gerrard Steven Haberman Bjorn Hojgarrd Elena Vigna
156.	April 2004	An International Comparison of Long-Term Care Arrangements. An Investigation into the Equity, Efficiency and sustainability of the Long-Term Care Systems in Germany, Japan, Sweden, the United Kingdom and the United States. ISBN 1 901615 78 2	Martin Karlsson Les Mayhew Robert Plumb Ben D. Rickayzen
157.	June 2004	Alternative Framework for the Fair Valuation of Participating Life Insurance Contracts. ISBN 1 901615-79-0	Laura Ballotta
158.	July 2004.	An Asset Allocation Strategy for a Risk Reserve considering both Risk and Profit. ISBN 1 901615-80-4	Nan Wang
159.	December 2004	Upper and Lower Bounds of Present Value Distributions of Life Insurance Contracts with Disability Related Benefits. ISBN 1 901615-83-9	Jaap Spreeuw
160.	January 2005	Mortality Reduction Factors Incorporating Cohort Effects. ISBN 1 90161584 7	Arthur E. Renshaw Steven Haberman
161.	February 2005	The Management of De-Cumulation Risks in a Defined Contribution Environment. ISBN 1 901615 85 5.	Russell J. Gerrard Steven Haberman Elena Vigna
162.	May 2005	The IASB Insurance Project for Life Insurance Contracts: Impart on Reserving Methods and Solvency Requirements. ISBN 1-901615 86 3.	Laura Ballotta Giorgia Esposito Steven Haberman
163.	September 2005	Asymptotic and Numerical Analysis of the Optimal Investment Strategy for an Insurer. ISBN 1-901615-88-X	Paul Emms Steven Haberman
164.	October 2005.	Modelling the Joint Distribution of Competing Risks Survival Times using Copula Functions. I SBN 1-901615-89-8	Vladimir Kaishev Dimitrina S, Dimitrova Steven Haberman
165.	November 2005.	Excess of Loss Reinsurance Under Joint Survival Optimality. ISBN1-901615-90-1	Vladimir K. Kaishev Dimitrina S. Dimitrova
166.	November 2005.	Lee-Carter Goes Risk-Neutral. An Application to the Italian Annuity Market. ISBN 1-901615-91-X	Enrico Biffis Michel Denuit

167.	November 2005	Lee-Carter Mortality Forecasting: Application to the Italian Population. ISBN 1-901615-93-6	Steven Haberman Maria Russolillo
168.	February 2006	The Probationary Period as a Screening Device: Competitive Markets. ISBN 1-901615-95-2	Jaap Spreeuw Martin Karlsson
169.	February 2006	Types of Dependence and Time-dependent Association between Two Lifetimes in Single Parameter Copula Models. ISBN 1-901615-96-0	Jaap Spreeuw
170.	April 2006	Modelling Stochastic Bivariate Mortality ISBN 1-901615-97-9	Elisa Luciano Jaap Spreeuw Elena Vigna.
171.	February 2006	Optimal Strategies for Pricing General Insurance. ISBN 1901615-98-7	Paul Emms Steve Haberman Irene Savoulli
172.	February 2006	Dynamic Pricing of General Insurance in a Competitive Market. ISBN1-901615-99-5	Paul Emms
173.	February 2006	Pricing General Insurance with Constraints. ISBN 1-905752-00-8	Paul Emms
174.	May 2006	Investigating the Market Potential for Customised Long Term Care Insurance Products. ISBN 1-905752-01-6	Martin Karlsson Les Mayhew Ben Rickayzen
175.	December 2006	Pricing and Capital Requirements for With Profit Contracts: Modelling Considerations. ISBN 1-905752-04-0	Laura Ballotta
176.	December 2006	Modelling the Fair Value of Annuities Contracts: The Impact of Interest Rate Risk and Mortality Risk. ISBN 1-905752-05-9	Laura Ballotta Giorgia Esposito Steven Haberman
177.	December 2006	Using Queuing Theory to Analyse Completion Times in Accident and Emergency Departments in the Light of the Government 4-hour Target. ISBN 978-1-905752-06-5	Les Mayhew David Smith

Statistical Research Papers

Report Number	Date	Publication Title	Author
1.	December 1995.	Some Results on the Derivatives of Matrix Functions. ISBN 1 874 770 83 2	P. Sebastiani
2.	March 1996	Coherent Criteria for Optimal Experimental Design. ISBN 1 874 770 86 7	A.P. Dawid P. Sebastiani
3.	March 1996	Maximum Entropy Sampling and Optimal Bayesian Experimental Design. ISBN 1 874 770 87 5	P. Sebastiani H.P. Wynn
4.	May 1996	A Note on D-optimal Designs for a Logistic Regression Model. ISBN 1 874 770 92 1	P. Sebastiani R. Settimi
5.	August 1996	First-order Optimal Designs for Non Linear Models. ISBN 1 874 770 95 6	P. Sebastiani R. Settimi
6.	September 1996	A Business Process Approach to Maintenance: Measurement, Decision and Control. ISBN 1 874 770 96 4	Martin J. Newby

7.	September 1996.	Moments and Generating Functions for the Absorption Distribution and its Negative Binomial Analogue. ISBN 1 874 770 97 2	Martin J. Newby
8.	November 1996.	Mixture Reduction via Predictive Scores. ISBN 1 874 770 98 0	Robert G. Cowell.
9.	March 1997.	Robust Parameter Learning in Bayesian Networks with Missing Data. ISBN 1 901615 00 6	P.Sebastiani M. Ramoni
10.	March 1997.	Guidelines for Corrective Replacement Based on Low Stochastic Structure Assumptions. ISBN 1 901615 01 4.	M.J. Newby F.P.A. Coolen
11.	March 1997	Approximations for the Absorption Distribution and its Negative Binomial Analogue. ISBN 1 901615 02 2	Martin J. Newby
12.	June 1997	The Use of Exogenous Knowledge to Learn Bayesian Networks from Incomplete Databases. ISBN 1 901615 10 3	M. Ramoni P. Sebastiani
13.	June 1997	Learning Bayesian Networks from Incomplete Databases. ISBN 1 901615 11 1	M. Ramoni P.Sebastiani
14.	June 1997	Risk Based Optimal Designs. ISBN 1 901615 13 8	P.Sebastiani H.P. Wynn
15.	June 1997.	Sampling without Replacement in Junction Trees. ISBN 1 901615 14 6	Robert G. Cowell
16.	July 1997	Optimal Overhaul Intervals with Imperfect Inspection and Repair. ISBN 1 901615 15 4	Richard A. Dagg Martin J. Newby
17.	October 1997	Bayesian Experimental Design and Shannon Information. ISBN 1 901615 17 0	P. Sebastiani. H.P. Wynn
18.	November 1997.	A Characterisation of Phase Type Distributions. ISBN 1 901615 18 9	Linda C. Wolstenholme
19.	December 1997	A Comparison of Models for Probability of Detection (POD) Curves. ISBN 1 901615 21 9	Wolstenholme L.C
20.	February 1999.	Parameter Learning from Incomplete Data Using Maximum Entropy I: Principles. ISBN 1 901615 37 5	Robert G. Cowell
21.	November 1999	Parameter Learning from Incomplete Data Using Maximum Entropy II: Application to Bayesian Networks. ISBN 1 901615 40 5	Robert G. Cowell
22.	March 2001	FINEX : Forensic Identification by Network Expert Systems. ISBN 1 901615 60X	Robert G.Cowell
23.	March 2001.	Wren Learning Bayesian Networks from Data, using Conditional Independence Tests is Equivalant to a Scoring Metric ISBN 1 901615 61 8	Robert G Cowell
24.	August 2004	Automatic, Computer Aided Geometric Design of Free-Knot, Regression Splines. ISBN 1-901615-81-2	Vladimir K Kaishev, Dimitrina S.Dimitrova, Steven Haberman Richard J. Verrall
25.	December 2004	Identification and Separation of DNA Mixtures Using Peak Area Information. ISBN 1-901615-82-0	R.G.Cowell S.L.Lauritzen J Mortera,
26.	November 2005.	The Quest for a Donor : Probability Based Methods Offer Help. ISBN 1-90161592-8	P.F.Mostad T. Egeland., R.G. Cowell V. Bosnes Ø. Braaten

27.	February 2006	Identification and Separation of DNA Mixtures Using Peak Area Information. (Updated Version of Research Report Number 25). ISBN 1-901615-94-4	R.G.Cowell S.L.Lauritzen J Mortera,
28.	October 2006	Geometrically Designed, Variable Knot Regression Splines : Asymptotics and Inference. ISBN 1-905752-02-4	Vladimir K Kaishev, Dimitrina S.Dimitrova, Steven Haberman Richard J. Verrall
29.	October 2006	Geometrically Designed, Variable Knot Regression Splines : Variation Diminishing Optimality of Knots. ISBN 1-905752-03-2	Vladimir K Kaishev, Dimitrina S.Dimitrova, Steven Haberman Richard J. Verrall

Papers can be downloaded from

http://www.cass.city.ac.uk/arc/actuarialreports.html

Faculty of Actuarial Science and Insurance

Actuarial Research Club

The support of the corporate members

- CGNU Assurance
- English Matthews Brockman
- Government Actuary's Department

is gratefully acknowledged.

Cass Business School
106 Bunhill Row,
London EC1Y 8TZ.

ISBN 978-1-905752-06-5