Unravelling the dynamics of referral-to-treatment in the NHS

R. M. Wood [1, 2, 3]

[1] Modelling and Analytics, UK National Health Service (BNSSG CCG)
[2] School of Management, University of Bath
[3] Correspondence to South Plaza, Marlborough St, Bristol, BS1 3NX, United Kingdom

Abstract

Despite being the key measure of elective performance in the NHS, there is little awareness of the dynamics behind the 18-week referral-to-treatment (RTT) standard. In making the case that a better understanding can improve service planning, this paper introduces the main components governing the dynamics and explains – through a simple analogy – the interplay between them. In the second part of the paper, this thinking is extended to numerical results through a computer simulation model. Using publicly available national data for 2017-18, the model is used to investigate a range of scenarios where the effect of changes to referral rates and elective capacity is assessed for performance, waiting lists, and spend.

Keywords*: Service Planning, Elective Pathways, Referral-to-treatment, Stochastic Modelling, Simulation*

1. Introduction

In the National Health Service (NHS) of Great Britain (England, Scotland, Wales) waiting time performance for consultant-led elective care is measured by the duration of time from when a referral is received until the start of treatment. This is known as the referral-to-treatment (RTT) standard which specifies the constitutional requirement that at any point in time at least 92% of patients have been waiting under 18 weeks (The National Health Service Commissioning Board and Clinical Commissioning Groups Regulations, 2012).

RTT *performance* is simply calculated as the proportion of waiting patients that have been waiting under 18 weeks. To be included the rules state that a patient must be on an *open pathway*, defined by a referral having been received (a *clock start*) but no treatment (*clock stop*) having yet occurred (Dept. of Health, 2015). What constitutes a clock stop varies by specialty and patient condition, and could involve a consultation, procedure or perhaps no activity at all if the patient reneges from the waiting list. In 2017-18 there were 15.7m clock stops in England (NHS England, 2018a).

The purpose of this paper is to impart an understanding of the key dynamics underpinning an RTT pathway. This is approached by conveying the dynamical concepts involved in reliably answering questions typically faced by managers, such as: How would performance respond to a reduction in capacity or an increase in referrals? What additional capacity is required in order to hold or improve performance given increasing referrals or winter pressures? What would be the effect of cancelling

elective procedures in winter months? How much money could be saved by reducing RTT performance by a given percentage?

The ability to answer these questions reliably is crucial for effective service planning. If managers cannot credibly balance pathway performance and spend subject to targets on the former, constraints on the latter, and in respect of varying demand, then they cannot plan services that best meet the needs of their population.

The need for improved service planning has been recognised within the NHS (NHS England, 2014), where a shift from reactive to proactive planning is required; ensuring deteriorating performance or spiralling spend aren't confronted *after* they've arisen. Change is also required in the way in which service plans are arrived at. In the author's experience there is too often a disconnect in the annually produced Operational Plans where elective activity assumptions underpinning RTT performance trajectories are inconsistent with the performance assumptions underpinning the activity projections.

It is interesting as to the reasons why the NHS is challenged in these regards. One of the reasons that could be offered is a recognised deficit in the analytical skillset within health services that is undoubtedly required in making sense of RTT pathway dynamics (see Bardsley, 2016 in particular for the NHS but also supported by Slipicevic & Masic, 2012, Hemans-Henry et al, 2016, Wang et al, 2018).

It could also be as a result of the lack of available literature on the topic. There has been a huge volume of academic papers in health modelling, with investigators studying the demand and capacity related dynamics of individual services (Griffiths et al, 2006, Mayhew & Smith, 2008) as well as whole hospitals (Brailsford et al, 2004, Rashwan et al, 2015); however no paper could be found looking at the RTT pathway dynamics specifically, despite this target being announced in 2004. The only material addressing this to date is guidelines produced by an NHS regulator (NHS Improvement, 2017). While this resource provides some introduction to the key components of the RTT pathway it serves little in developing the dynamical concepts from these that can help answer the above questions. That is what this paper sets out to address.

2. Materials and methods

In understanding RTT dynamics it is appropriate to first start with outlining the components of the pathway. These are: clock starts, capacity, clock stops, performance and waiting list size. The first two of these represent the handles, or controls, of the pathway whereas the remainder are responses. That is, clock stops, performance and waiting list size (the *dependent* variables) are a *function* of clock starts and capacity (the *independent* variables).

2.1 The bathtub analogy

In explaining to health managers the dynamics between these components the author has found useful the analogy of water flowing into and out of a bathtub. Here, the water represents the referrals which flow from the tap (clock starts); the water-level represents waiting list size; plughole diameter represents capacity, which determines the maximum clock stops per unit time; and the water temperature relates to waiting time, where water from the tap starts warm and cools over time. This is depicted in Figure 1.

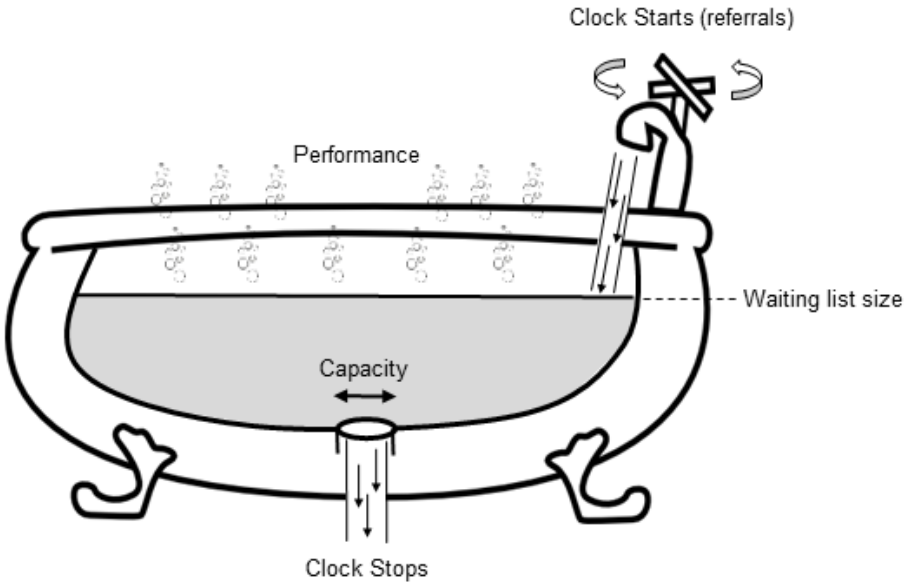

Figure 1 Bathtub representation of an RTT pathway

This analogy can be used to go some way in answering our above questions. For example, let's take the first one: How would performance respond to a reduction in capacity? One might first think that there would be a one-off step change in performance to a lower value. But consider a bathtub where the flow out is less than the flow in – what happens? The level of the water (waiting list) continues to increase and as it does the temperature (performance) continues to decrease as the new water added is having to wait increasingly longer before flowing down the plughole. Unless something is done this behaviour would continue, thus disproving the initial assessment that there would be some one-off adjustment.

The effects of mitigations can also be explored through the bathtub analogy. An intervention with the aim of negating these negative effects of constraining capacity may well be to reduce referrals too. One might think that provided referrals are reduced at the same time as capacity is constrained and by the same amount then the performance of the pathway would be unchanged. This is not the case. Here one is turning off the tap by the same amount as they are blocking the plughole and thus the system remains in balance and waiting list (water level) does not change. But the flow through the system has reduced and with it temperature (performance) since each drop of water added is now having to wait longer before passing down the plughole. The solution is to implement the referral reduction strategy before constraining capacity, so that the water can be brought down to a level which can support the same temperature (performance) as before with the new, reduced rate of flow. Understanding just how much lag is required is one of many dynamics that can be examined through the use of a simple mathematical model introduced in the next sub-section.

2.2. Simulation model

An understanding afforded by the bathtub analogy is useful in logically assessing what *kind of thing* will happen given a change in the dynamic of an RTT pathway, but a numerical solution is required in order to assess magnitudes and thus be useful in the quantification of service plans.

This is achieved through the use of a simple computer model. Modelling patient flow is established practice and has helped clinicians and service managers answer capacity-related questions along outpatient (Bowers et al, 2005), emergency (Gul & Guneri, 2015) and rehabilitation (Griffiths et al,

2013) pathways. However, there is currently no published method which can be used to help managers with the afore-mentioned types of question relating to the RTT pathway.

The computer model used here is based on stochastically simulating clock starts (referrals) and clock stops each day within the simulated period, where capacity dictates the maximum number of clock stops that can occur each day. At the end of each simulated day the clock stops, performance, and waiting list size are recorded within the computer program, where performance is simply the proportion of those waiting who have been waiting under 18 weeks. Spend can be calculated from the clock stops using an average measure of cost based on activity per clock stop.

These measures can then be used to readily assess the cost-benefit impacts of various "what if" scenarios, which is possible through the flexibility provided by the simulation model. This is explored in the next section using some real-life data from the NHS.

3. Results

The effect of a number of hypothetical *what if* scenarios on pathway dynamics is now assessed. The aim is to show, using the model introduced in Section 2.2, how clock stops, performance, and waiting list size respond to various changes in the handles/controls on the pathway – clock starts and capacity (see the outset to Section 2 for an introduction of the five pathway components).

The hypothetical scenarios considered here are based on the questions posed in Section 1 which, in the author's experience, represent a range of typical issues faced by healthcare managers when planning elective pathways. These involve changes precipitated by factors both under and outside of the control of managers, such as increasing capacity or increasing clock starts respectively (noting the latter could arguably be under the control of managers through their influence on GP referral policy).

In illustrating the effect of these scenarios on the pathway dynamics, the model is calibrated based on real-life national data for RTT pathways in England for the financial year 2017-18 (NHS England, 2018a). Note that the model can also be applied to pathways at locality and/or specialty-level – this is discussed more in Section 4.

Some points should first be noted in interpreting results for the ten hypothetical what if scenarios presented in the rest of this section. Each scenario starts from a baseline starting point at time zero. At this point England average performance and total waiting list size are fixed at their values at the end of 2017-18 (87.2% and 3,843,182 respectively) and the system is assumed in balance where clock start and capacity rates are set equal to the England average daily clock stops in 2017-18 (43,149). For each scenario, 365 days are simulated from time zero based on adjusted versions of clock starts and capacity in line with the scenario in question. With an average cost per clock stop of £2,933 (from Programme Budgeting data and NHS England, 2018a) a baseline elective spend of £46.3b can be calculated.

Faced with multi-billion pound funding shortfalls (NHS England, 2014) a reasonable management consideration would be to reduce spend through constraining capacity. However, if this is attempted without also reducing demand (i.e. clock starts) then there could be drastic effects on waiting times (noting mechanics exemplified through bathtub analogy in Section 2.1). Cutting capacity by 10% (see **Scenario 1** in Figures 2 and 3 and Table 1) would save £4.6b in one year but lead to a projected 41% increase in waiting lists (95% confidence intervals 33% to 49%) and a reduction in RTT performance from 87.2% to 78.2% (75.9% to 80.9%). If the cut to capacity is made gradually over the course of the year (**Scenario 2**) then the negative effects on waiting times are less pronounced but at the expense of a (£1.1b) lesser saving.

If it is possible to reduce demand as well as capacity (**Scenario 3**) then the negative effects on performance are reduced. However, they are not totally mitigated for the reasons explained in Section

2.1: in order to ensure no ultimate drop in performance then some lag on constraining capacity, and foregoing of saving, is required (**Scenario 4**). If the ending of the lag were to be extended then this would lead to a lower water level (waiting list) which can support a higher temperature (performance) but at the cost of more water passing having passed through (i.e. more clock stops / spend).

Due to multi-year contracts and economic viability of provider organisations, constraining capacity may actually not be as easy as it sounds. If capacity (beds, consultants, theatres) cannot be readily taken out but efforts to reduce referrals by 10% are successful (**Scenario 5**) then performance would increase to 97.6% (93.1% to 100%) and waiting lists would fall by 31% (23% to 39%). However, it is important to note that under this scenario there would be no saving – reducing referrals would not be reducing spend.

Up to now there has been an assumption that the system is in balance and that deviations from this are based on interventions at our discretion. However, as is widely known, there is an ageing population with increasing comorbidities which, unchecked, would likely push up referrals. If this occurs then the system would be out of balance and waiting times would worsen and worsen. Increasing capacity (**Scenario 6**) would mitigate this with a similar, but opposite, dynamic to that explained in Scenarios 3 and 4 (that is, a lag is necessary on increasing capacity to bring the water level up to a higher level that is required to support the same temperature given the higher through-flow).

Service managers may be familiar with seasonal trends in referral rates, where colder weather can act as a catalyst to pre-existing conditions. Without appropriate capacity management this can mean fluctuating RTT performance (**Scenario 7**), noting the lag in clock start (referral) and performance peaks and troughs. Another seasonal effect could be the management response to surges in non-elective demand over winter months, where the cancelling of elective procedures has recently been debated. With procedures making up an estimated 20% of clock stops (NHS England, 2018a), the effect of such a reduction in capacity for just two months (**Scenario 8**) would bring about deterioration in performance to 84.6% (83.1% to 86.2%). Note that the comparability of this drop (2.6 percentage point expectation) to that observed in the winter of 2017-18 where elective procedures were widely cancelled (2.3 percentage points) lends support to the validity of the model in making such predictions.

Scenarios targeting specific performance, waiting list, or spend can also be investigated using the model. For example, it can be calculated that achieving the constitutional target of 92% (**Scenario 9**) would require an investment of £1.5b. Note that this is a one-off in-year investment required to bring the water level (waiting list) down to 3,339,752 – the number required to support such a performance (water temperature). Note that the investment required would rise to £1.7b if regulatory assumptions (NHS England and NHS Improvement, 2018) of a 0.8% increase in GP referrals (clock starts) are factored in (**Scenario 10**).

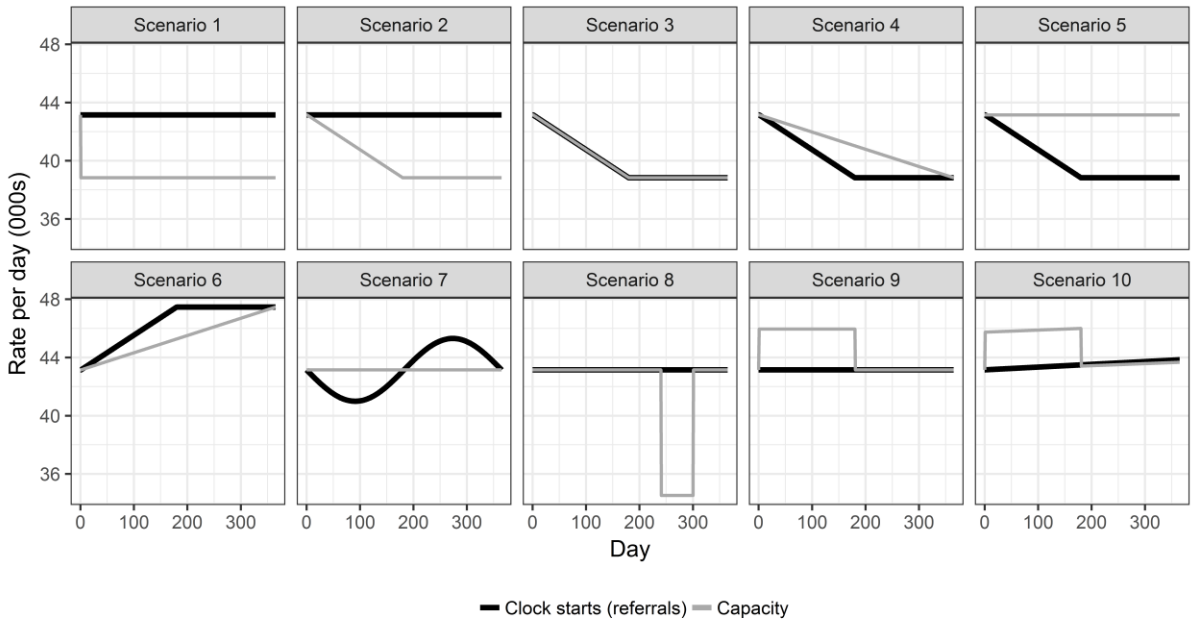

Figure 2 Pathway inputs (capacity and clock starts) for scenarios considered

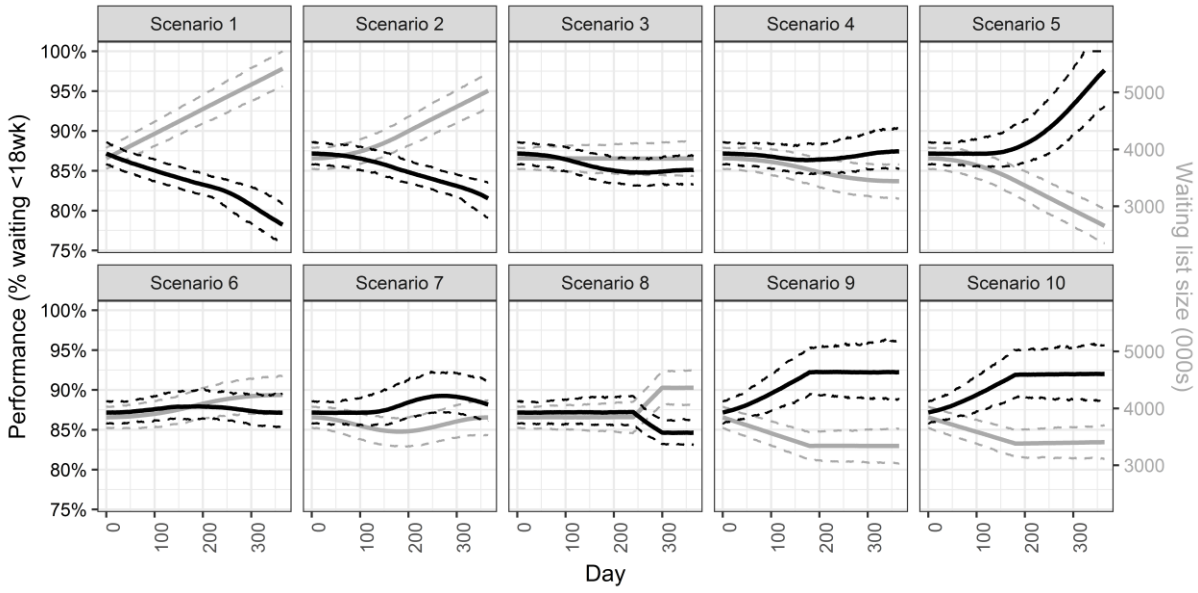

Figure 3 Projected outputs (performance and waiting list) for scenarios considered (with 95% confidence intervals)

Table 1 Projected outputs (clock stops and spend) for scenarios considered

Scenario	Clock stops (000s)	Elective spend (£b)
Baseline	15,792	46.3
1	14,218	41.7
2	14,606	42.8
3	14,606	42.8
4	15,005	44.0
5	15,792	46.3
6	16,580	48.6
7	15,793	46.3
8	15,274	44.8
9	16,297	47.8
10	16,353	48.0

4. Discussion

In conveying an understanding of the key dynamics underpinning an RTT pathway this paper addresses a gap in the literature where, to date, there has been no published account of how the various components of the pathway interact. The bathtub analogy introduced in Section 2.1 provides a fitting abstract which can be used to explain the dynamics to health managers in relatable terms. And the model of Section 2.2 builds on this, providing numerical assessments of interest to elective service planning.

Some of these assessments have been illustrated in Section 3 where a number of scenarios have been evaluated at national level. The model can also be applied at local level. At the author's organisation it has been used as part of annual Operational Planning rounds in order to understand the relationship between incremental spend and RTT performance (busting the myth that there is a linear relationship where £*x* buys a *y* percentage point improvement in RTT). The model has also been used at specialty level to understand the effect of *QIPP* saving strategies on specific services (such as how much money can be released by constraining MSK capacity in line with planned referral reduction without deteriorating performance). Evaluation could also go down to sub-specialty or individual consultant level but, as outlined in regulatory guidance (NHS Improvement, 2017), this should be assessed with caution owing to low data volumes.

In order to operationalise any changes in capacity there must be increases or decreases to activity provided. The amount and type of activity to put in or take out can be coarsely calculated as the simple product of change in clock stops (Table 1) and average activity per clock stop, which, based on 2017-18 data (NHS England, 2018b), is 0.08 admissions, 0.44 daycase procedures, and 1.17 outpatient appointments. Recall from Section 1 that some clock stops may have "*no activity at all if the patient reneges from the waiting list*". Whilst the incidence of this outcome is relatively low (at the author's organisation < 10%), the separating out of activity and non-activity based clocks stops could be facilitated by a convenient extension to the bathtub analogy where a leak or over-spilling water represents the reneging of patients whose waiting time has simply got too long (and the patient has either died, significantly deteriorated, or opted for private treatment). Note that there is an increased likelihood of this happening in scenarios where waiting lists are larger – this would reduce average activity/cost per clock stop (quantifying by how much could be the topic of further work).

This paper argues that effective operational planning of RTT pathways requires a solid grasp of the underlying dynamics. Two reasons are offered in Section 1 as to why the NHS has not made more use of analytical approaches in understanding and planning RTT pathways. The first – a gap in the literature – should start to be addressed by this paper, as well as the associated model script, coded in

open source software *R*, and available from the author. The second – insufficient analytical thinking on the ground – requires more of a cultural shift. There are seeds of hope here where recently there have been initiatives in the NHS such as a specialist Health Analytics graduate training programme and the setting up of an *R* community for analysts to share scripts and model code.

References

Bardsley, M. (2016). Understanding analytical capability in health care. Do we have more data than insight. *The Health Foundation*. https://www.health.org.uk/publication/understanding-analytical-capability-health-care

Brailsford, S.C., Lattimer, V.A., Tarnaras, P., Turnbull, J.C. (2004). Emergency and on-demand health care: modelling a large complex system. *Journal of the Operational Research Society*. 55 (1).

Bowers, J., Lyons, B., Mould, G., Symonds, T. (2005). Modelling outpatient capacity for a diagnosis and treatment centre. *Health care management science*. 8 (3).

Department of Health (2015). Referral to treatment consultant-led waiting times – Rules Suite. https://www.gov.uk/government/publications/right-to-start-consultant-led-treatment-within-18-weeks

Griffiths, J.D., Price-Lloyd, N., Smithies, M., Williams, J. (2006). A queueing model of activities in an intensive care unit. *IMA Journal of Management Mathematics*. 17 (3).

Griffiths, J. D., Williams, J. E., & Wood, R. M. (2013). Modelling activities at a neurological rehabilitation unit. *European Journal of Operational Research*. 226 (2).

Gul, M., & Guneri, A. F. (2015). A comprehensive review of emergency department simulation applications for normal and disaster conditions. *Computers & Industrial Engineering*. 83.

Hemans-Henry, C., Blake, J., Parton, H., Koppaka, R., Greene, C. M. (2016). Preparing Master of Public Health graduates to work in local health departments. *Journal of Public Health Management and Practice*. 22(2).

Mayhew L., Smith, D. (2008). Using queuing theory to analyse the government's 4-H completion time target in accident and emergency departments. *Health Care Management Science*. 11 (1).

National Health Service England (2014). Five year forward view. https://www.england.nhs.uk/wp-content/uploads/2014/10/5yfv-web.pdf

National Health Service Improvement (2017). Referral to treatment pathways: A guide to managing efficient elective care. https://improvement.nhs.uk/resources/elective-care-guide/

National Health Service England (2018a). Consultant-led Referral to Treatment Waiting Times Data 2017-18. https://www.england.nhs.uk/statistics/statistical-work-areas/rtt-waiting-times/rtt-data-2017-18/

National Health Service England (2018b). Monthly Hospital Activity Data. https://www.england.nhs.uk/statistics/statistical-work-areas/hospital-activity/monthly-hospital-activity/mar-data/

National Health Service England and National Health Service Improvement (2018). Refreshing NHS Plans for 2018/19. https://www.england.nhs.uk/wp-content/uploads/2018/02/planning-guidance-18-19.pdf

Rashwan, W., Abo-Hamad, W., Arisha, A. (2015). A system dynamics view of the acute bed blockage problem in the Irish healthcare system. *European Journal of Operational Research*. 247 (1).

Slipicevic, O., Masic, I. (2012). Management knowledge and skills required in the health care system of the federation Bosnia and Herzegovina. *Materia socio-medica*. 24 (2).

Wang, Y., Kung, L., Byrd, T.A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organisations. *Technological forecasting and social change*. 126.