Balancing Supply and Demand in Health and Social Care Organisations

Professor Eric Wolstenholme, David Monk, Gill Smith and Douglas McKelvie

Abstract

This paper develops a hypothesis that the 'normal' mode of operation for many organisations today is well beyond their safe design capacity and that many health and social care organisations in the UK are in this position. This situation arises from having to cope with whatever demand arrives at their door irrespective of their supply capability. Risk levels can be high in these organisations and the consequences could be catastrophic.

The irony is that such organisations appear to cope at the strategic, 'whole-systems' level, where they can appear to be matching supply with demand and be in equilibrium. This is because operational managers employ a variety of well-intended, informal, survival policies to meet performance targets and avoid patient bottlenecks In fact, they are actively encouraged to find new ways of working and productivity gains to cope with more for less. However, such practices can perpetually mask the underlying reality and have severe unintended consequences. Organisations can become locked into a state of inefficiency, reduced patient safety and overspending. When additional funding is forthcoming it is immediately absorbed by the cumulated deficits, rather than being available for development and whole systems improvement.

Evidence for the hypothesis has emerged at many points along patient pathways in health and social care from a number of studies carried out using system dynamics simulation to identify and promote systemic practice in local health communities. The rigour involved in knowledge-capture and quantitative simulation model construction and running has identified mismatches between how managers claim their organisations work and the observed data and behaviour. The discrepancies can only be explained by surfacing informal coping strategies. Indeed, the data itself becomes questionable as it reflects more the actions of managers than the true characteristics of patients.

The result of capacity pressure can mean that managers are unable, physically and financially, to break out from a fire-fighting mode to implement better resource investment and development policies for systemic and sustainable improvement.

There are important messages in the paper for Health and Social Care management, the meaning of data and for modelling. The key message of the paper is that much-needed systemic solutions and whole system thinking can never be successfully implemented until organisations are allowed to articulate and dismantle their worst coping strategies and return to working within best practice capacities. This is the ultimate new way of working.

Introduction

System dynamics has been developed and successfully applied in a number of industries to identify help assist thinking and sustainable, counter-intuitive action in complex situations (Sterman 2003).

Recently the method has been extensively used by the authors in the field of health and social care. First, at a national level to influence government policy on reimbursement policy for delayed hospital discharges (Wolstenholme et al, 2004a) and more recently to assist local heath and social care communities in the UK to interpret and apply national policy frameworks for older people (Wolstenholme et al, 2004b and c).

System dynamics applications are currently underway by the authors in 10 health communities around the UK with the objectives of providing a visual and quantitative stimulus to strategic multi-agency planning. Specifically it is being used to identify encourage sustainable whole-system solutions rather than short–term 'fixes' around issues of:

- delayed hospital charges
- variation and investment in new capacity
- elective wait times and increasing elective episodes
- community beds
- patient assessment efficiency and times

This paper draws on experiences from current studies to create a hypothesis about how health and social care systems really operate to survive in a climate where they have to be seen to meet demand for their services, irrespective of their supply capabilities.

The paper will outline the process of application of system dynamics and explain more about the models that have been developed in Health and Social Care. It will then highlight some of the experiences that have given rise to the construction of the hypothesis and what the implications are for data, modelling and health and social care practice.

The process of applying system dynamics

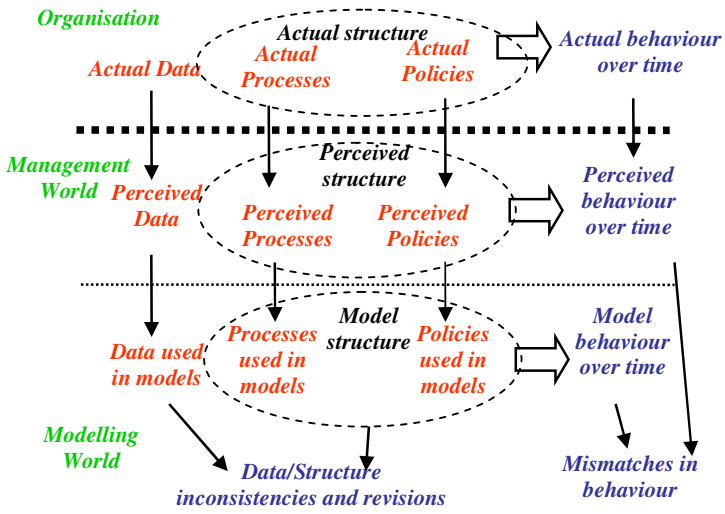
The process of system dynamics involves focussing on an issue of management concern and assembling data for variables associated with the concern, usually plotted over time, to centre thinking on both desirable and undesirable futures trends for the issue.

A map of patient pathways and the policies that make them work (referred to here as the process/policy structure of the organisation) is then created around the issue of concern. The source of the maps are the mental models of the management teams from each agency involved, at an appropriate level of aggregation. The map is populated with the best data available and simulated over future time under different policies in each agency along the patient pathways. The idea is to create a simulator on which the management teams can experiment with scenarios and policies in a risk free environment. The classical outcomes are improved understanding by the

management team of how agency plans and policies interact and commitment to more systemic polices, which benefit the whole patient pathway rather than any one agency.

Of course before using models for radical change ('what might be') it is necessary to establish a valid model of the current reality surrounding the issue ('what is'). The 'what is' phase of modelling is extremely important and should develop confidence in management team that the model is capable of showing behaviour over time consistent with their mental models of and data from the real system.

One of the major contributions of system dynamics arises from the quantitative rigour of the approach which is best explained by reference to Figure 1.



Inconsistencies and mismatches between actual, management and modelling perceptions of organisations

Figure 1 suggests that there are a number of 'worlds' in any organisation. In truth no one really knows the actual world situation. Managers work in a 'management world' where they have 'perceived' knowledge and interpretations of the data, processes and policies in use in an organisation.

This knowledge is largely in terms of data both for management and accountability purposes. However, increasingly some of the knowledge today is from perceptions of processes gained from process mapping. Policy knowledge is often a mixture of strategic ideals, policy guidelines and management rules and actions. However, analysis tends to be dominated by data and data, process and policy are seldom linked into a coherent whole.

In the system dynamics modelling world and attempt is made to combine data, aggregated processes and policies in to an integrated whole to generate simulated behaviour of the organisation over time.

This integration activity can expose serious inconsistencies between the perceived structure and the data of organisations. Further, running simulation models to create

behaviour over time exposes mismatches between the observed and simulated behaviour of the organisation Exploring the inconsistencies and mismatches can encourage a very open and rich dialogue between agencies as to how the organisation really works which challenges perceptions. It often turns out that numerous informal policies exist to keep the system functioning. This is an issue very relevant to the theme of this paper.

The models developed

A number of models have been developed in local health and social care situations, each of which have been tailored in detail to local circumstances and used for local issues. However, they all have generic underpinnings, which will be used to develop this paper. The generics can be considered as a template (referred to here as the whole system commissioning template (WSCT)) for the application of system dynamics in the analysis of patient pathways. An overview of the WSCT concerned with patient pathways between primary care, acute hospitals and post acute care is shown in Figure 2 and described below.

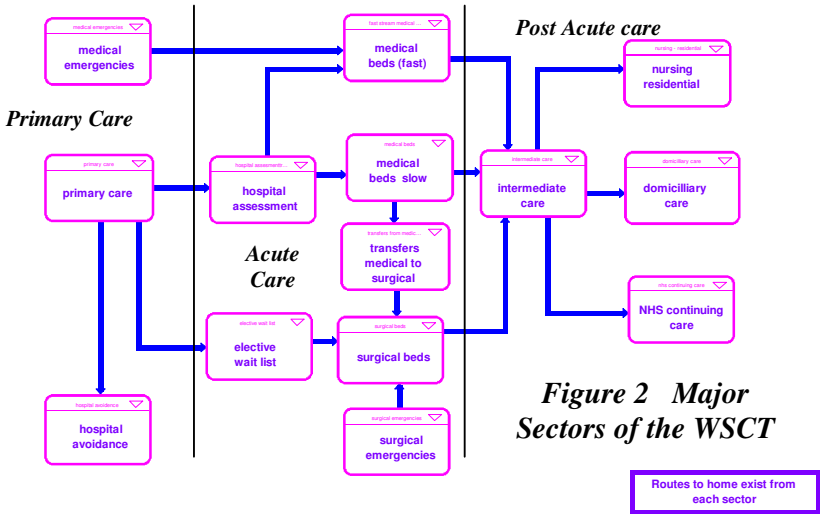


Figure 2 Major Sectors of the WSCT

In the WSCT the flow of patients at an aggregate level from primary care into acute hospitals is classified as being via two major routes - the elective route (mainly surgical) and the non-elective (mainly medical emergencies). Medical patients are classified as 'fast' or 'slow'. Fast refers to the simpler cases, who will stay in hospital a relatively short time and require minimal past acute care. Some of these may be day case patients. 'Slow' refers to more complex cases, who stay in hospital a relatively long time and require significant post acute care.

The flow of patients from hospital to post acute care is classified as being via post acute intermediate care to the post acute services of nursing/residential, domiciliary care and NHS continuing care. The WSCT has always allowed for some systemic solutions, such as the emerging routes of hospital avoidance via facilities such as pre-acute intermediate care. It has also always incorporated some informal coping policies, such as and the use of 'outliers'. This is the term given to medical patients using surgical beds who are transferred when medical beds become fully occupied.

Data to populate such a model is essentially 'flow' data. This consists of:

1. current and forecast demand,
2. the proportions of patients flowing down each pathway,
3. the capacities of each sector
4. the average lengths of stay in each service on the pathway, usually broken down into treatment, assessment and waiting for discharge components

The models would typically be run over 3 years on a daily basis and used to examine polices such as inter-agency capacity planning and hospital avoidance.

Elements of the 'what is' analysis in the community studies based on the WSCT

Having mapped the formal process/policy structure for the organisation and agreed data with the management teams from each agency, models were run under capacity constraints to identify the effects of policies on the location and extent of bottlenecks occurring along the patient pathways (for example in, accident and emergency, elective surgery and delayed hospital discharges) and on other performance measures (for example, number of elective acute episodes).

It was frequently found in the studies that there were indeed inconsistencies between the structure and data claimed for the organisation and between the simulated behaviour from the models and the perceived behaviour of the real organisation.

These inconsistencies and mismatches are explored in the next sections of the paper.

Description of an important model structure - capacity

Figure 3 shows details of how capacity was represented in the model for each agency.

This process/policy structure shown in Figure 3 can be considered generic for any service between a service purchaser and service deliverer of health and social care, and was agreed as being applicable by each agency along the patient pathways of Figure 2.

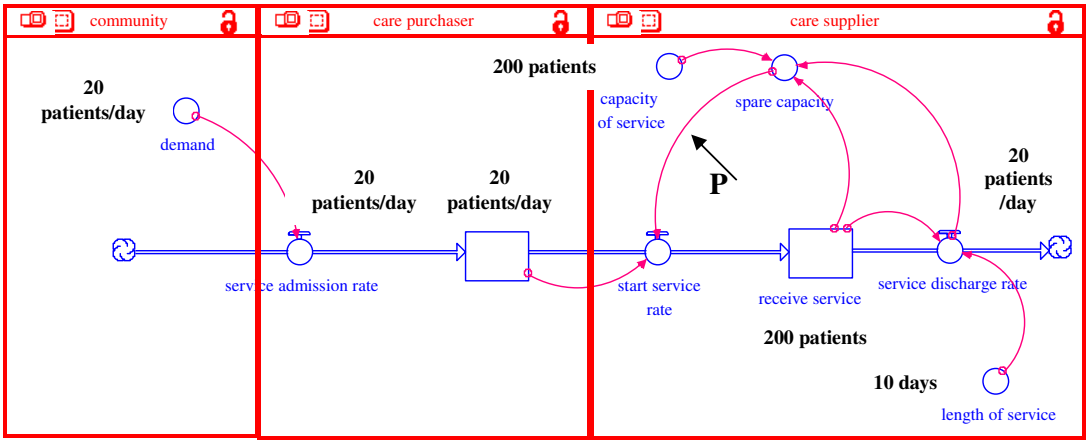


Figure 3. Formal process/policy structure used in the WSCT

Figure 3 is presented in stock-flow terms as used by system dynamics models in general and the 'ithink' software (in which the models were constructed), in particular. Processes are represented as 'pipes' along which resources flow from infinite sources and to infinite sinks in 'clouds' outside the model boundary. In the case of Figure 2 the resource flowing is patients. The 'valves' on the pipes represent flow variables or 'rates', which are driven either by outside factors such as demand or by internal factors such as management policies. The 'boxes' on the pipes represent 'stocks' or accumulations of the resource. The single lines represent information feedback by which the states of the stocks are used to determine the policies controlling the rates.

So in Figure 3 the assumptions are that external demand determines the service admission rate per day, which cumulates in the stock named 'await service', from where service starts. The service start rate is determined by the spare capacity of the service, allowing for there being people waiting. Spare capacity is the difference between service capacity and the number receiving service, plus replacement for those leaving. Service finish rate is determined as those patients receiving service divided by an average, pre-defined length of stay.

There is effectively only one formal policy link (P) in Figure 3 and the system is assumed to work by a balancing feedback process using spare service capacity to control service start rates assuming a given average length of stay.

This structure for representing capacity was employed within the WSCT and the studies at numerous points. For example:

1. the acceptance of patients into elective surgery from a wait list,
2. the acceptance of patients into post acute services from a wait discharge hospital stock.
3. the admission of patients into acute medical services from an accident and emergency stock

Examination of data

Figure 3 also shows a set of data items for demonstration purposes, consistent with the process/policy structure described.

So, for example, if a steady stream of 20 people are admitted per day to the service, 50 await the service, the service is full (200 people receiving service and service capacity 200) and the length of service is 10 days, then the discharge rate from the service MUST be 20 people per day and 20 people per day can start the service. Such data consistency between the average length of service, service capacity and numbers awaiting the service is essential for the system to be in true equilibrium.

Any variations in the admission rate will then be absorbed by the await service stock, which will show a fluctuating level of bottleneck as it absorbs the difference between demand and supply.

Examination of data/structure inconsistencies

What was found in practice was a mismatch between the data collected and the process/policy structure in Figure 3. For example, using the previous data, it was found that the data collected for the average length of service could be say 20 days, but that the system was still in equilibrium and no one waited for service. However, this is not theoretically possible under the structure of Figure 3. If the average length of service is 20 days then, with a stock of 200 receiving service 10 are discharged per day and 10 will start the service. If 20 are admitted per day but only 10 start the service, then the await service stock must rise by 10 per day and there would be a severe wait problem after only a few days.

Examination of behaviour mismatches

In some cases data inconsistencies could be rationalise to some extent and the models moved into running mode. However, at this stage there would typically be greater accumulations and bottlenecks in the model output than in the real organisation, which seemed to be more or less in equilibrium with supply matching demand.

Hence discussions took place about what process/policy structure must really exist to allow these inconsistencies and mismatches to exist.

A search for alternative structure

It is generally well accepted in all situations along the patient pathway that only very limited waiting is acceptable from a patient need point of view. In terms of emergency services no waiting at all is permitted. Hence the formal capacity policy of Figure 3 often has to be overridden in practice. This results in capacity being exceeded and it is at this point that informal policies come into play. Some of these policies, such as diversion to other services are well known and were built into the WSCT template at an early stage. Others are less frequently voiced. All have unintended consequences.

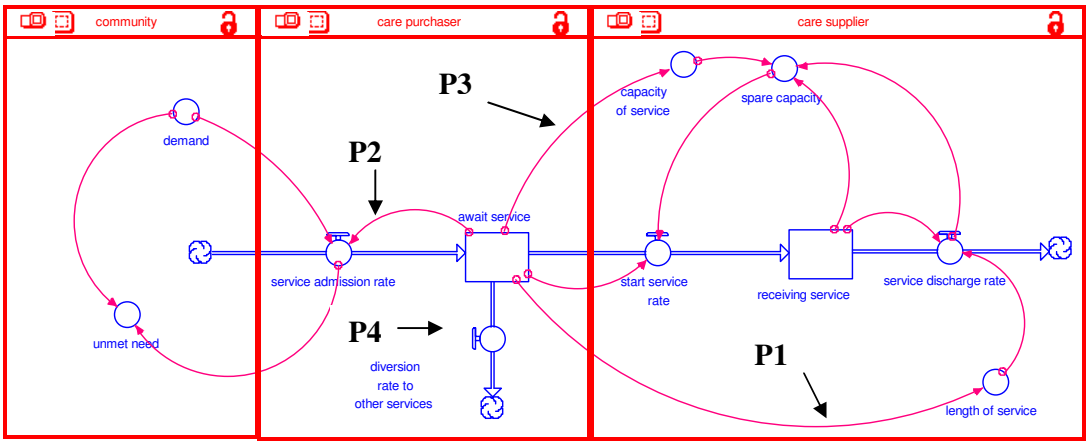


Figure 4. Formal plus informal process/policy structure that emerged in the community based system dynamics studies

Fig 4 shows the process/policy structure that emerged from numerous discussions during the exploration of the inconsistencies and mismatches. Figure 4 identifies four

informal coping policies (P1- P4), acting in addition to the formal capacity policy of Figure 3. Of course, not all of these policies apply at every point along the patient pathways. A description of the polices and example of where each occur are given below:

1. **Length of Service Policy (P1).** Length of service emerged as being a managerial policy, rather than a constant based on patient need and condition as shown in Figure 3. Length of service was described as a way of generating 'elastic' capacity, and was varied where ever changes occurred in the numbers of people awaiting the service. If demand was high, the length of stay would be reduced and if demand was low the length of stay would be increased. Examples of this are the practice in acute hospitals of discharging patients early if admission pressure is high and the rationing of the length of domiciliary service in social services to expedite discharges from were hospital.

2. **Service admission rate (P2).** Service admission rate emerged from discussions as a second managerial control variable. If patients had to wait for a given service then it was likely that the admission threshold is changed. An example of this is the behavioural response of some GPs would be to vary admissions to elective surgery in line with stabilising wait times.

3. **Capacity of Service (P3).** Direct additions to service emerged as a third control variable where this was easily arranged. An example of this is the spot purchase of domiciliary care in social services, again in response to delayed hospital discharges.

4. **Overspill (P4).** Service overspill emerged as a for the control variable. In cases where no waits at all were possible patients would be moved into other services. An example of this is the practice in acute hospitals of moving emergency medical patients to surgical beds when demand is high. People moved in this way are sometimes referred to as 'outliers'.

The coping mechanisms described are all well-intended policies aimed at keeping system performance within bounds and supply within capacity. It is not claimed here that any of these polices were unknown to management in each local area of application. What the studies described here did was to demonstrate the cumulative effect of such practices on the global behaviour of a patient pathway across multiple agencies.

The policies represent ways of making it appear at the strategic level that the organisation is coping and mask the fact that the organisations are working beyond their design capacity. A little over capacity is perhaps good for motivation, but operating well in excess of capacity for prolonged periods and institutionalising the coping policies can have serious consequences. Some of these are described below.

Implications of the findings

There are serious implications in these findings for the operation of health and social care agencies, the real meaning of data and the process of system dynamics modelling

Unintended consequences of coping strategies for health and social care

Changes in gate keeping thresholds hold back demand, but this becomes absorbed by stocks outside the health and social care system and cumulative unmet need pushes responsibilities back on families, charities and communities.

Reducing lengths of stays in acute hospitals create more incomplete episodes of care and readmissions. Institutionalising the practice of outliers results in numerous disruptive bed shifts for patients and inefficiencies for hospital consultants and occupational therapists who often waste time and money locating their patients.

Rationing home help hours can result in patient dissatisfaction and later increases in higher cost interventions. Buying external capacity in social services usually means buying at a premium rate, which leads to cost escalation.

Even if cost is the sole consequence the service suffers since, when extra money is forthcoming it merely goes to pay off cumulated deficits rather than better ways of working.

The problem with all these informal policies is that they are fixes to cope with working beyond the design capacity of the organisation.

It is interesting that some of these coping policies are now sometimes quoted as formal policies and it is likely that more will be institutionalised as organisations are urged to 'realise more for less'. Finding new ways of working is very necessary for the future of health and social care in a world of limited resources and sustainable solutions are emerging from the studies described. However, it is suggested from the evidence here that a necessary precursor to the implementation of these would be to allow organisations to surface from the burden of working beyond design capacity. Only then can sustainable policies be designed.

The meaning of data during periods of coping policies

There is a tendency in management to believe that more data is better and millions of pounds is spent annually in health and social care to increase the quantity, quality and usability of data. Further data is the evidence usually used in statistical analysis for the purpose of organisational analysis and change management. Data seems to have the magical property of **appearing** to be absolute and solely a characteristic of the entities measured. So, for example, every medical condition can be shown to have an average treatment length of stay with a given standard deviation.

However, as shown here data can much more often reflect the management actions undertaken during its period of collection than the characteristics of the entities measured. So data collected on lengths of service during periods of applying coping policies, reflects nothing more than management overload and bears no mathematical

relationship to the numbers of patients in the system, the service capacities or, indeed, the characteristics of the patients.

As an absolute minimum it is fundamental to know what processes and policies where in place during a given period of data collection for the data to have any meaning.

Lessons for system dynamics modelling

When trying to validate a system dynamics model it is essential to know the process/policy structure that exists at both the formal and informal levels. The examples of the type given in this text should be incorporated into the process of application so that everyone is aware of the need to openly discuss the formal and informal policies in place and to surface these at a very early stage of enquiry.

If there are informal policies in place they must be incorporated into the model together with the data that reflects them and the consequences they cause. These behavioural feedback effects, for example, varying treatment lengths of stay around nominal averages, are vital to establishing a valid 'what is' model of system behaviour. Interestingly, in system dynamics modelling it is often said that feedback is difficult to find. This is perhaps because we do not probe enough beneath the surface of linear processes and formal descriptions of system polices.

Moreover, the first stage of the 'what might be' analysis in system dynamics should be to expose the real unmasked behaviour of the system when coping policies are withdrawn. Only then is it sensible to try to demonstrate the effects of systemic policies to really redesign the system and counter the exposed behaviour.

The modelling insight here is that:

one of the main objective of a system dynamics study should be to identify where systems are deviating from best practice and to demonstrate firstly the merits of a return to best practice

To achieve the aim we require new data for the 'what might be' phase of system dynamics studies such as best practice capacities and length of stays. ***Not past data associated with past practice, which we wish to replace.***

It is somewhat ironic that system dynamics models are sometimes criticised as being invalid, because they cannot reproduce past data, when actually they can be demonstrating that it is the data that is invalid.

Conclusions

This paper has attempted to create a hypothesis to explain patterns of discrepancies between the way organisations are described to work and their observed behaviour that has emerged from studies of applying system dynamics in health and social care.

Analysis of mismatches has surfaced cases where many informal policies seem to dominate behaviour. These policies are a result of coping with demand well outside the design capacities of the organisations and mask sever side effects, which mitigate

against the successful implementation of sustainable policies for real systemic improvement.

Removing such coping strategies and returning to best practice is suggested to be a major first step in creating sustainable change. This is the ultimate new way of working to get more from less.

References

Sterman, J, (2003). Business Dynamics –Systems Thinking and modelling for a complex world. Irwin, McGraw-Hill, Boston.

Wolstenholme, E. F., Monk, D., Smith, G. and McKelvie, D. (2004a). "Using System Dynamics to Influence and Interpret Health and Social Care Policy in the UK." Proceedings of the 2004 System Dynamics Conference, Oxford, England.

Wolstenholme, E. F., Monk, D., Smith, G. and McKelvie, D. (2004b). "Using System Dynamics in Modelling Health and Social Care Commissioning in the UK." Proceedings of the 2004 System Dynamics Conference, Oxford, England.

Wolstenholme, E. F., Monk, D., Smith, G. and McKelvie, D. (2004c)**.** "Using System Dynamics in Modelling Mental Health Issues in the UK." Proceedings of the 2004 System Dynamics Conference, Oxford, England.