Coping but not Coping in Health and Social Care - masking the reality of running organisations well beyond safe design capacity

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

Fieldings, 12 Rivock Avenue, Steeton, Keighley, West Yorkshire, UK
01535 65816 eric.wolstenholme@symmetricsd.co.uk

Abstract

This paper develops a hypothesis that the 'normal' mode of operation for many organisations 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 demand, irrespective of their supply capability.

The irony is that such organisations can appear to cope at the strategic level. This is because operational managers employ a variety of well-intended, informal, survival techniques to meet performance targets. However, such practices can perpetually mask the underlying reality and have serious unintended consequences .

Evidence for the hypothesis has emerged from a number of studies carried out using system dynamics to identify and promote systemic practice in local health communities in the UK. If proved wholly or even partially correct there are some important messages in the paper for Health and Social Care management, the meaning of data and for modelling.

Introduction

System dynamics has been developed and successfully applied in a number of industries to 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 it was applied, 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 and encourage sustainable whole-system solutions rather than short–term 'fixes' around issues of:

- delayed hospital discharges

- variation and investment in new capacity
- elective wait times and increasing elective episodes
- community beds
- patient assessment efficiency
- reconfiguration of care pathways to better support service users

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 politically 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 future 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 sources of the maps are the mental models of the management teams from each agency involved, at an appropriate level of aggregation. The map is then transcribed into a computer-based model (a bespoke simulator), populated with the best data available and used to game play (simulate) different scenarios over time under different policies in each agency along the patient pathways. The idea is to create a learning environment in which the management teams can experiment with scenarios and policies risk free. 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 the management team that the model is capable of showing aggregate 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.

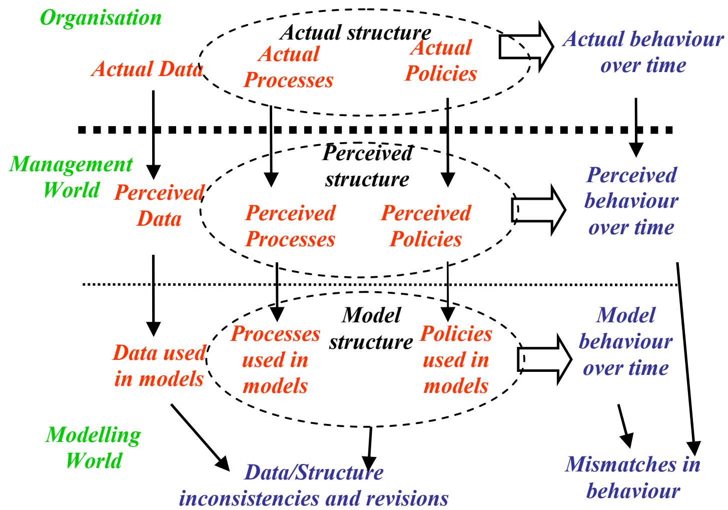

Figure 1 Inconsistencies and mismatches between actual, management and modelling perceptions of organisations

Figure 1 suggests that there are a number of 'worlds' in any organisation. In truth no one really knows the actual world situation. Managers, for example, work in a 'management world' where they have 'perceived' knowledge based on experience, intuition and interpretations of the data, processes and policies in use in an organisation. The operational core of health and social care organisations, often the clinicians and/or other practitioners, will work with their own set of assumption, beliefs and access to soft and hard data or evidence.

Increasingly (some of) this knowledge is gained from care pathway process mapping and is largely used for operational service improvement and management accountability purposes. Policy knowledge is often a mixture of strategic assumptions, policy guidelines and management rules and actions. However, analysis tends to be dominated by data. Process, policy and data and are seldom linked into a coherent whole.

In contrast, in the system dynamics modelling world an attempt is made to combine data, aggregated processes and policies in to an integrated whole, since all these ingredients are required to move the processes through simulated time.

This integration activity can expose serious inconsistencies between the perceived processes and data of the organisation. Further, running simulation models to create behaviour over time can expose mismatches between the observed and simulated behaviour of the organisation. Exploring the inconsistencies and mismatches challenges perceptions and can encourage a very open and rich dialogue between agencies as to how the organisation really works. 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 share a number of common structural features, which will be used to develop this paper. These features when modelled together can be used to create 'a template' (referred to here as the whole system commissioning template (WSCT)) for the application of system dynamics learning and 'what if' analysis of patient pathways. An overview of the WSCT concerned with patient pathways for older people between primary care, acute hospitals and post acute care is shown in Figure 2 and described below.

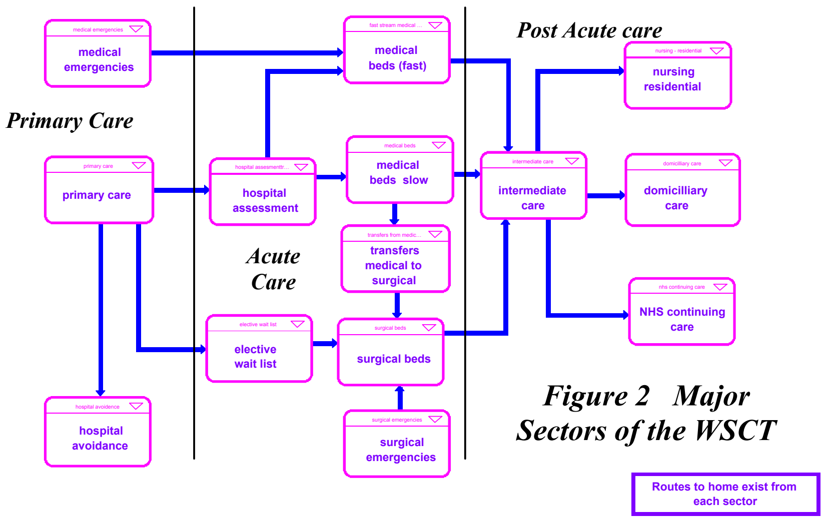

Figure 2 Major Sectors of the WSCT

In the WSCT the flow of patients at an aggregate level from primary care into acute hospitals is classified (and simplified) 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 post acute care. 'Slow' refers to more complex cases, who stay in hospital a relatively long time and require significant post acute care. Surgical patients are classified as urgent, non-urgent or day cases.

Patients flow from hospital in to a range of intermediate care services and onward to nursing/residential, domiciliary care and NHS continuing care.

The WSCT always incorporated some informal coping policies, such as early discharge from hospital when demand was high and the use of 'outliers'. The latter 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 and shaped the process/policy structure for the organisation and agreed data with the management teams from each agency, models are run under capacity constraints to identify the location and extent of bottlenecks occurring along the patient pathways. For example, in accident and emergency, elective surgery and hospital discharge.

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

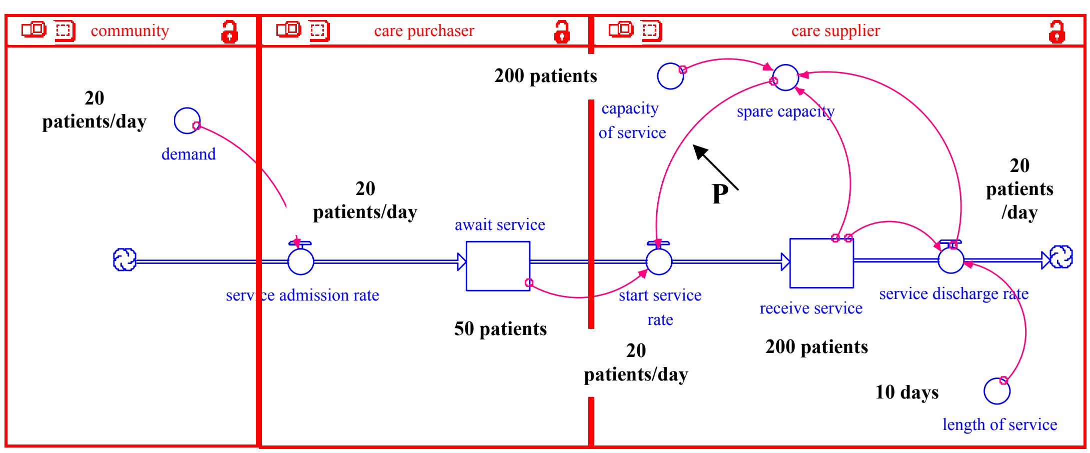

Figure 3. Formal process/policy structure used in the WSCT

Figure 3 is presented in stock-flow terms, the common language of 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 represented as 'clouds' outside the model boundary. In the case of Figure 3 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 and/or professional practice. 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 in policy implementation.

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 being discharged from the service. The discharge rate is determined as those patients receiving service divided by an average, pre-defined length of service.

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 has been found (to date) in practice is 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 might 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 process/policy 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 rationalised to some extent and the models moved into running mode. However, at this stage there would typically still 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. In terms of emergency services, very little waiting at all is permitted. In such cases the formal capacity constraint of Figure 3 has to be overridden. 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.

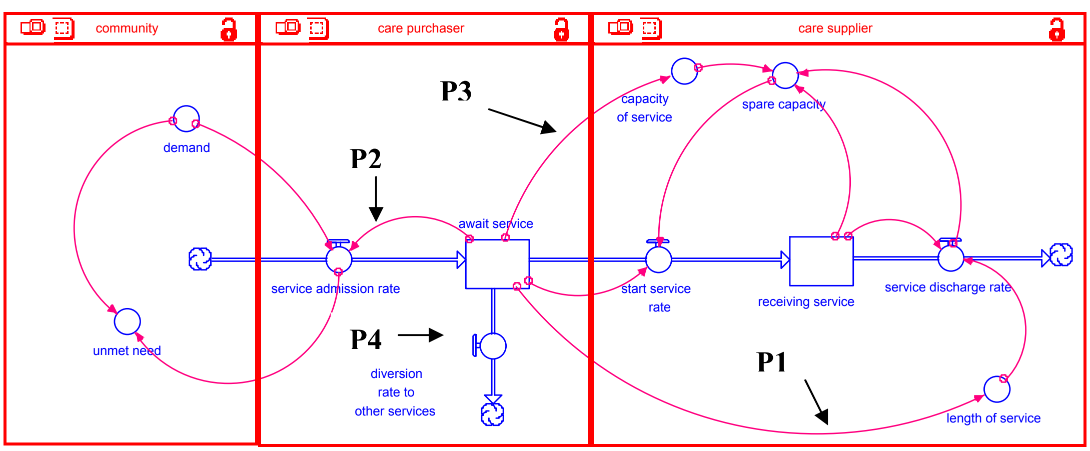

Figure 4. Formal plus informal process/policy structure that emerged in the community based system dynamics studies

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 examples of where each occur are given below:

1. **Length of Service Policy (P1).** Whilst it might be expected that length of service is a constant based on patient need and condition, it consistently emerged as being a managerial policy as shown in Figure 3. Length of service was described as a way of generating 'elastic' capacity, which could be varied wherever 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 hospital into the service. It is also evident in the transfer of patients to 'leave' beds in acute psychiatry.

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

3. **Capacity of Service (P3).** Direct additions to service capacities emerged as a third control variable, if this could be easily arranged. Two examples of this are the spot purchase of domiciliary care in social services, again in response to delayed hospital discharges and

4. **Overspill (P4).** Service overspill emerged as a fourth 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 referred to as 'outliers'. This practice also surfaced in the use of post-acute intermediate care by hospital consultants. Any spare places in intermediate care, whether appropriate or not, would be immediately filled to relieve hospital pressures.

It is important to emphasise that coping mechanisms described are all well-intended and the best actions that can be achieved in the interests of patients. They are all policies aimed at minimising the extent by which capacity is exceeded. Neither is it claimed here that any of these polices are unknown.

However, the studies were able to surface and demonstrate the vast extent and cumulative effect of such practices on the global behaviour of patient pathways across multiple agencies.

It is also important to recognise that 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 inevitably becomes absorbed by stocks outside the health and social care system. As cumulative unmet need increases, responsibilities are pushed back on families, charities and communities. Ultimately, there is a kick-back on services, with a higher proportion of people entering as emergencies.

Reducing lengths of stays in acute hospitals creates more incomplete episodes of care and readmissions. Institutionalising the practice of outliers results in numerous disruptive bed shifts for patients and inefficiencies for medical staff who often waste time and money locating and assessing their patients.

Using spare places in intermediate care to relieve hospital pressures results in ‘mixed use’ facilities and it becomes impossible to determine true intermediate care capacity. The overall consequence of reducing lengths of stay is to produce an increase in the ‘revolving door phenomena’, whereby a small population of people recycle continuously through hospital and become a significant problem.

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. Such cost increases distort the benefits of later investments, when extra money merely goes to pay off cumulated deficits, rather than create better ways of working.

The problem with all these informal policies is that they are fixes to achieve imposed performance targets whilst 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 review 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 claimed to have an average treatment length with a given standard deviation. However, it is suggested here that data may well more often reflect the management actions undertaken during its period of collection, rather 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. This cause and effect dilemma is captured in Figure 5.

Figure 5: Does data create action – OR – does action create data??

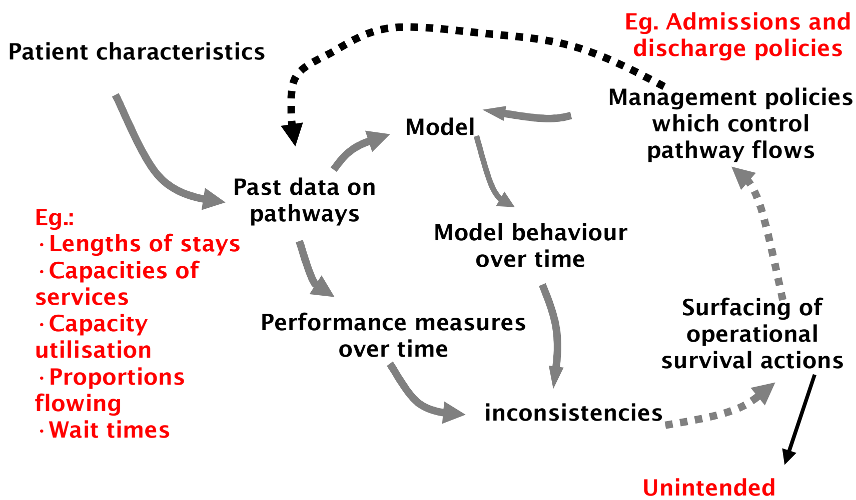

Figure 5 is an extension of Figure 1 and shows that data can lead to action, but that action can also lead to data. As an absolute minimum it is fundamental to know what processes and policies were in place during a given period of data collection for the data to have any meaning and to be used for decision making.

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. 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 (espoused) and informal (in-action) policies in place and to surface these at a very early stage of enquiry. System dynamics is one of the few ways to uncover inconsistencies and neglected questions.

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. The industry tends to prefer spreadsheets – these ignore feedback and don't kick back and show the real legacy of planning.

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 objectives 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 also 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. The caption on Figure 5 could just as easily read: **" does data validate a model – OR – does a model validate data??"**

Beyond Fire Fighting – the creation of truly systemic action

Once some of the most undesirable coping policies have been removed it is possible to really focus on the merits of balanced interventions to improve the whole of the patient pathways. System dynamics is then applicable in its traditional role of testing new ideas and assessing the gains and penalties of innovative new ways of working.

Conclusions

This paper has attempted to create an emerging hypothesis to explain patterns of discrepancies between the way organisations are described to work and their observed behaviour emerging 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 severe 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. It is suggested that his 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.