

# Whole Systems in Health and Social Care

Introducing a commissioning template  
September 2005



## **1. Introduction**

This document describes a computer simulation model using System Dynamics, developed with the LGA and NHS Confederation and applied in a number of localities. The purpose of System Dynamics (SD) is to map flows (in this case of people accessing services) in a way that illustrates the effect of one part of the system on another, over time.

SD is used to focus on issues: this particular model was developed to support the concerns about delayed discharge. It can be used to explore aspects of demand, supply and service configuration.

## **2. Some of the Core Issues in Managing Whole Systems in Health and Social Care**

Health and social care provide a continuum of support for people in various stages of needing help. However, there is still a way to go before the aspirations of “putting the client/patient at the centre” are realised, since managing the whole system is complex. For instance, health operates in distinct sectors (primary care, secondary care in the community, secondary care in hospital, tertiary care). Meanwhile, social care is primarily concerned with providing packages of care in the community, and may be relatively unaware of the input provided by health colleagues to their clients. However, the recent focus on delayed hospital discharge has created significant joint work in this area. The interest in managing chronic disease may also spur closer collaboration on preventing hospital admission, but this work is less advanced.

The ultimate aspiration is to re-think the totality of the client/patient experience, including consideration of housing, jobs, healthy communities and social networks. Again, progress is patchy. Supported housing is becoming a significant resource in the delivery of social care, and public health is increasingly looking at lifestyle issues and the role of communities in addressing entrenched problems, such as drugs and alcohol.

Current issues in whole system commissioning include:

- Use of Hospitals. Are too many people being referred to hospital, or attending A&E? What would help? – pathways for chronic disease management; pathways for managing patients who need observation (or nursing care) rather than acute treatment; access to more extensive diagnostic facilities in the community; ways to manage a wider range of conditions through specialist GPs, rather than hospital-based consultants; pathways to divert mental health cases (including self-harmers).
- Shifting the Balance of Services. Are too many people in institutional forms of care? What would help? – more intensive forms of care in the home; support for carers; ways to manage risk in the community (eg telecare); more community-based treatments for health; more rehab services to ensure that patients have the best chance of recovery; more attention to the possibility of people improving to the point where they can “go home” again.
- Smoothing Throughput. Are services experiencing difficulties with variation in demand? What would help? – ways to buffer unexpected peaks and troughs of demand; ways to match capacity across the different elements of a care pathways; earlier warning of surges or trends upstream (eg changes in referral behaviour).
- Funding of Care. Are there perverse incentives to classify clients/patients in certain ways, in order to access services? What would help? – more flexible

ways to review continuing care needs; flexibility to give care **before** a client deteriorates to the upper bands of the Fair Access to Care matrix; more pooled budgets to encourage whole system outcomes (and overcome the tendency of individual agencies to guard their own resources).

### **3. Purpose of the Model**

The model is for commissioners in health and social care. Its purpose is to enable them to test their thinking, in order to develop more robust plans and strategies, and advance the whole system aims of the community.

The model provides:

- A **visual** means to analyse the complex processes involved in the patient journey from pre acute, through acute, to post acute sectors of the health and social care economy
- A **quantitative** analysis of flows through the economy, highlighting potential issues and opportunities for improvement
- The means to evaluate various **options for intervention**

It provides the ability to test strategies in the risk free environment of a simulator, before making investments and changes in the real world. The commissioning simulator is like a flight simulator: it enables people to “fly, crash and recover” in a safe environment, before implementing strategies for real.

Examples of strategies that can be tested include:

- Impact of a 12% increase in hospital beds
- Impact of a 10% increase in certain types of social care provision
- Impact of a 5% diversion of patients away from hospital

These are just a few of the strategies that can be tested with the model. The range is only limited by the combinations of parameters you choose.

### **4. Describing the Model**

The basic building blocks of an SD model (stocks, flows and converters) are described in the document “Introduction to the Use of Symmetric Templates”.

Figure 1 is a simplified diagram of the whole systems commissioning template, including only the main stocks and flows.

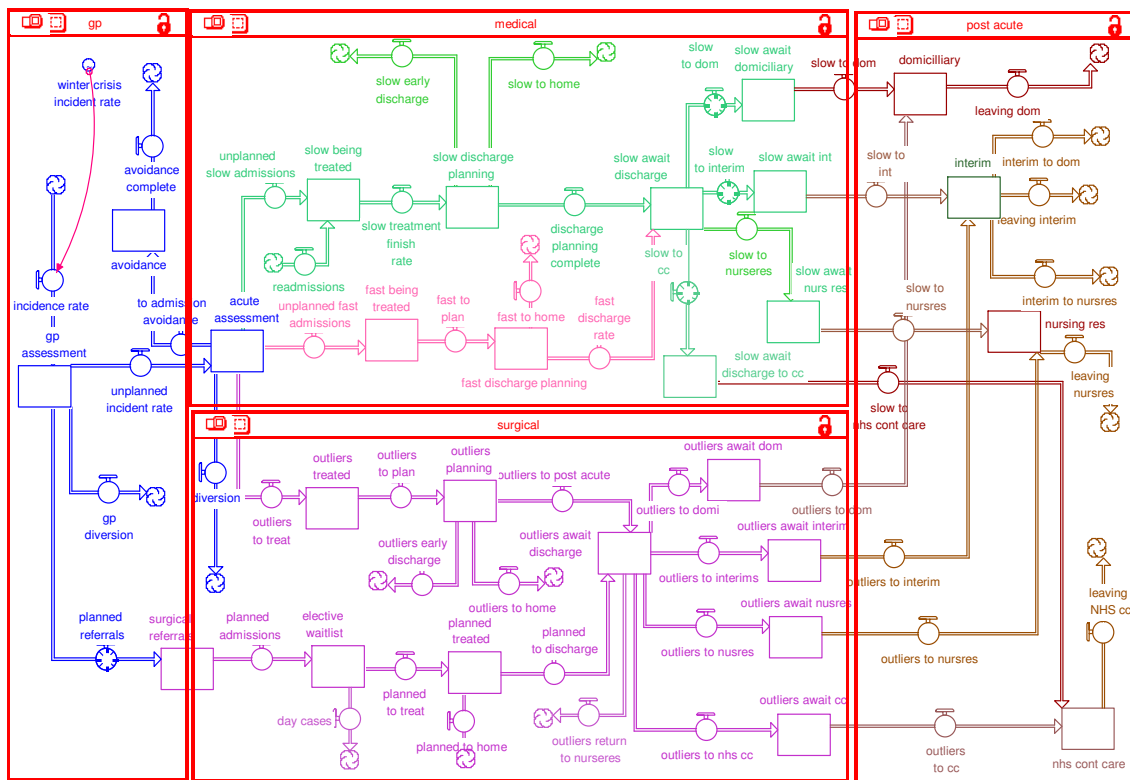
The pathways are as follows:

- People come through their GP and are directed to the elective waitlist (via a consultant – not shown) or considered for unplanned admission. The model does not consider people going to the GP for minor conditions or turning up in A&E as “walking wounded”: it only deals with potential admissions.
- The GP can choose to maintain the patient at home. People who reach A&E can be dealt with in 3 ways: diverted using new investment, diverted by redeploying community resources or admitted. Diversion using new capacity has a delay for the new service to come on stream.

## Introducing a Whole Systems Commissioning Template

- Unplanned admissions are considered as “fast” (staying in hospital a few days) or “slow” (in-patient time several weeks). Where there are insufficient medical beds, people will “overspill” from medical to elective beds (medical outliers). In response to pressure in A&E, people can also be discharged early from the slow medical stream, but this will increase the rate of re-admissions.
- Elective patients are treated as day cases or admissions (currently 50:50 split). No constraints on capacity for day surgery have currently been included. Capacity constraints on elective admissions, however, mean that some get blocked, so the proportion of elective admissions can be less than 50%.
- For all 3 streams, there is an option to go home with no package, or to have post-acute care. There are capacity constraints on all forms of post-acute care.

**Fig 1: Stocks and Flows in the Commissioning Template (converters not shown)**



The model uses the following concepts:

### Demand

- This is driven by an incidence of people going to their GP (potential admissions – less serious cases are not considered). The average value of the incidence rate is 100/day, but the actual rate is determined by a graph which shows 3 peaks over the 3 years that the model runs.

### Constraints

- People can only flow through the various routes in the model where there is sufficient capacity. The general equation for allowing people “through” is to take the minimum of demand at that point or spare capacity. The general equation for deciding spare

## Introducing a Whole Systems Commissioning Template

capacity is to take current capacity less capacity in use. This format is used for all resource types (eg medical/surgical beds, nursing/residential care etc)

- It is useful to show a graph of capacity against usage to test the points in the 3-year run where capacity constraints come into effect. This will be determined partly by capacity levels and partly by the surges in demand given by the peaks in the incidence graph. At these points, the model will be particularly sensitive to small changes in demand.

### Sources of capacity

- Capacity for the main stocks (beds, post acute care etc) can be “developed” (shown by an inward flow) and lost (shown by an outward flow). In addition, the capacity can be “occupied” and unavailable for further use until “freed up”
- Where capacity is increased, we can either:
  - Show this as an instant change or
  - Show a delay while the capacity is developed

In some cases, we do the former (eg allowing the user to select an increase in beds) and in other cases, we do the latter (eg allowing the user to select the development of a “new” resource for diverting people from acute admission, with an inbuilt delay). We also show “background” changes in capacity due to market forces. There is a 10%/year gradual decrease in available nursing/residential capacity and a corresponding 10% growth in the domiciliary market

- We also provide for a “redeployment” of capacity in one part of the model. Domiciliary and nursing/residential care (which are generally used in a post acute situation) can be used for pre-acute diversion (see “diversion to existing capacity”). In this case, the capacity available for post acute is reduced by the amount redeployed to pre acute. However, an efficiency saving of 50% has been introduced so that twice as many people can be treated for the same capacity in a pre-acute situation as are treated for this capacity in a post-acute situation.

### Routes

People can be “directed” down various routes at a branching point:

- A percentage can be sent one way or another (eg where 70% GP referrals go down the elective route)
- One way can be invoked until full and then another brought into play (eg where medical admissions are taken until there is no capacity and they are then switched to surgical beds, as medical outliers)

### Apportioning access to capacity

Sometimes capacity is apportioned according to “rules”:

- Outliers are given a proportion of post acute care (nursing/residential, domiciliary etc) depending on the percentage of the total demand that outliers represent (outlier flow to nursing/residential as percentage of outlier flow + other discharges to nursing/residential)

### Recalculating percentages

Where percentages going to different routes are expressed for the entire flow, they are recalculated for the subset (those requiring post acute care).

Example: the proportion of fast medical to interim care  
User inputs 6% (of fast medical discharge) which is adjusted to 60% (of post acute flow), given that 90% fast medical go home

### Initial values

If the model “started from scratch”, it would take some time for the inputs to create a realistic “steady state”. Hence the model is populated with initial values.

- Initial values for stocks – these are derived by calculating what the normal daily in-flow would be based on all upstream values, and multiplying this by the normal length of stay within that stock; the initial values for some “waiting” stocks are set at zero; and for some services that are assumed to be always under pressure, the initial value is set as a (high) percentage of total capacity

### Measuring wait times

Wait times for selected parts of the process (eg for elective admission, for discharge to nursing/residential) are provided by setting a “clock” on entry to each process and “stopping the clock” when the exit point is reached.

However, we discount the situation where a person goes in and out of a stock in the same day (eg patients waiting discharge who go out same day are not counted as having waited even one day). This adjustment is made in the “perceived admissions” part of the model.

## **5. Running the Demonstration Model**

The template is designed with 6 fixed runs, to enable people to go through a structured demonstration, and a seventh option which allows parameters to be changed under user control.

The suggested sequence for demonstrating the model is as follows:

- Familiarise with the model

The model will open at the user interface level.

First look at the underlying model. Click down-arrow at top left of screen. Select any part of the model and use the +/- tabs (bottom left) to zoom in or out, and the navigation arrows to move around.

Go back to the interface layer (click the up-arrow at top left of screen). Explore the graph pads to see the range of data that can be displayed (go from one “sheet” to the other by clicking the turned-up corner at bottom left of the graph pad). Note that a whole set of graphs relating to post-acute and intermediate care are to be found on a different screen (button called “more graphs tables”).

Look at the range of data that can be changed in run 7, by clicking on any of the “data” buttons (acute settings, discharge rates etc).

Check how to run the model. It runs for 3 years, in days. When you press “run”, the simulation will go for the first “year” and then stop – press “run” again, and once more when it pauses after the 2<sup>nd</sup> year.

If you want to clear the graphs, press “restore graphs and tables”. This is necessary from time to time, since the “single element” ones are comparative graphs which redraw on the same space and can clutter your view).

- The Fixed Runs
  - **Run 1: Establishing Base Performance**

The model has been set initially with some data which allows the economy to “manage”, despite a number of inherent issues.

Press “run” and interpret what is going on.

Best graphs to look at are:

- Left hand No 3 (elective wait time)
- Right hand No 1 (delayed transfers)
- Also look at left hand No 2. This shows the seasonal variations in demand. When demand is highest, medical beds-in (line 1) meets medical capacity (line 2) and the overspill causes medical outliers (line 5). Note that surgical beds are always full (lines 3 and 4).
- Also look at right hand No 2. This shows the cumulative elective operations, which are currently constrained by bed-blocking and medical outliers.

### Interpretation

The situation in run 1 is unsustainable. Delayed discharges are tolerable (£3.5m over 3 years: see display under left hand graph pad), but elective wait times are way over target

#### • **Run 2: Intervene in Acute Sector**

This scenario illustrates a strategy of adding medical and surgical beds (12% are added automatically on day 180 of the simulation).

Press “run” and interpret what is going on.

Best graphs to look at are:

- Left hand No 3 (elective wait time)
- Right hand No 1 (delayed transfers)
- Left hand No 2 (capacity and usage of hospital beds)
- Right hand No 2. (cumulative elective operations)

### Interpretation

Elective wait time is better, but not ideal (LH No 3). Delayed transfers have soared (RH No 1): note that the display under the left hand graph (“fund generated by reimbursement”) is over £8.5m for the 3 years.

The additional hospital capacity can be seen as a step-up of line 1 in the LH No 2 graph: this has reduced the outliers (line 5 on this graph). Making better use of hospital beds has improved elective productivity (RH No 2).

So the hospital has benefited from this intervention, but things have got worse in social care. This is an example of unilateral action – the hospital has acted to improve its performance, but it is not getting the full benefit (since the extra beds mean even more demand on post-acute care, and more delayed discharges). Meanwhile, the SSD is struggling to cope. This underlines a fundamental concept in SD: capacities need to be balance across the whole patient pathway, or increases in one part will simply cause bottlenecks elsewhere.

#### • **Run 3: Intervene in Post-acute Sector**

This illustrates an alternative strategy of adding 10% post-acute care (domiciliary care, continuing care and interim care) on day 90 of the simulation. Note that care

## Introducing a Whole Systems Commissioning Template

homes are not increased – in fact, the model has a 3% per annum shrinkage in available care home places, to reflect the current market.

Press “run” and interpret what is going on.

Best graphs to look at are:

- Left hand No 3 (elective wait time)
- Right hand No 1 (delayed transfers)
- Left hand No 2 (capacity and usage of hospital beds)
- Right hand No 2. (cumulative elective operations)

### Interpretation

Elective wait time is good (LH No 3). Delayed transfers are very good (RH No 1) and the reimbursement fund is under £0.5m for the 3 years.

In addition, outliers are minimal (LH No 2) and elective productivity is good (RH No 2).

This is an example of a win-win situation: both the hospital and the SSD have benefited by balancing capacity across the process, so that the SSD is able to cope with discharges.

### • **Run 4: Combine the interventions in Runs 2 and 3**

This assumes that the benefits seen in runs 2 and 3 will be combined by adding both 12% hospital beds AND 10% post-acute capacity.

Press “run” and interpret what is going on.

Best graphs to look at are:

- Left hand No 3 (elective wait time)
- Right hand No 1 (delayed transfers)
- Left hand No 2 (capacity and usage of hospital beds)
- Right hand No 2. (cumulative elective operations)

### Interpretation

Elective wait time is good (LH No 3). Delayed transfers are not good (RH No 1) – and the reimbursement cost is back to nearly £5.5m for the 3 years.

Outliers are still an issue (LH No 2).

Elective productivity is slightly better than with just additional post-acute care (RH No 2).

The combined strategy is expensive but has disappointing results: the extra beds have negated the benefits of the extra post-acute care. Capacity needs re-balancing.

### • **Run 5: Divert from Hospital**

This looks at the relative merit of a smaller investment (5% reduction in admissions – for instance, by developing diversion services).

Press “run” and interpret what is going on.

Best graphs to look at are:

- Left hand No 3 (elective wait time)
- Right hand No 1 (delayed transfers)
- Left hand No 2 (capacity and usage of hospital beds)

## Introducing a Whole Systems Commissioning Template

- Right hand No 2. (cumulative elective operations)

### Interpretation

Elective wait time is quite good (LH No 3). Delayed transfers are similar to base case (RH No 1) –reimbursement cost is £3.5m for the 3 years.

Outliers are improved (LH No 2).

Elective productivity is slightly better than base (RH No 2)

This strategy is promising (and cheaper), but there are still delayed transfers....

### • **Run 6: Divert from Hospital AND increase Post-acute Care**

This takes the previous run and adds domiciliary care. The strategy is to keep some patients out of hospital and help others at the discharge point.

Press “run” and interpret what is going on.

Best graphs to look at are:

- Left hand No 3 (elective wait time)
- Right hand No 1 (delayed transfers)
- Left hand No 2 (capacity and usage of hospital beds)
- Right hand No 2. (cumulative elective operations)

### Interpretation

Elective wait time is good (LH No 3). Delayed transfers are quite good (RH No 1) – reimbursement cost is £1.9m for the 3 years.

Outliers are much improved (LH No 2).

Elective productivity is good (RH No 2).

This strategy is the best yet. But there are many other variables we could experiment with.

For instance, what if we could:

- Change the proportion of “fast” and “slow” cases coming into hospital? (as might happen under a regime of chronic care in the community)
- Reduce lengths of stay in the hospital? (this effectively adds capacity)
- Increase the rehab rate from interim care?

### • **Run 7: Applying your own Changes to the Variables**

When the “run 7” button is clicked, you can re-run the simulation as often as you like, changing parameters on the data screens. You can make the changes at the “pause points” (after each “year”) or before the run.

Note that the data inputs from all of the pre-loaded runs are still switched on, and you will first need to un-set many of the slider controls. Use the navigation buttons called “Acute Settings”, “Post Acute 1” and “Post Acute 2” to view the data, and go through each slider setting.

You will probably find that the “best” solution (which benefits both the hospital and the SSD) involves a mix of strategies.

## 6. General Lessons from the Template

## Introducing a Whole Systems Commissioning Template

- Deciding what is a “good result” may not be straightforward. Some service configurations may create a smooth flow (no bottlenecks) but at the expense of overall throughput
- Capacities have to be tuned to demand or there will be capacity constraints (blockages) or unused capacity (surplus)
- Capacities have to be *balanced* across the supply chain or there will be bottlenecks
- Capacities will need a gearing based on how long they are in use. For a flow through 2 operations where the time spent in the second is twice as long, the capacity at the second point has to be twice as large in order to balance the flow
- Where there is variable demand, there needs to be a response which:
  - Is flexible enough to cope with the variation or
  - Invokes a resource which can buffer the “peaks and troughs”
- There are several possible responses to bottlenecks:
  - Add more capacity (but beware unbalancing capacities downstream of the bottleneck)
  - Change the process (eg split the flow and run some in parallel)
  - Change the dynamics of the process (eg speed up one part by decreasing lengths of stay)

Adding capacity is rarely the only (or even the best) alternative

- Interventions which route the flow away from the main process are useful since they cut out areas later in the flow where problems may arise. Hence interventions **early** in the supply chain are likely to be more cost-effective than those **later** in the supply chain
- Interventions are more effective if they take place **before** a problem has manifested itself (ie responding in anticipation of demand is better than after the event)
- “Re-cycling” (where part of the flow goes back round a loop) may only involve small percentages of the flow but can create significant impacts (particularly if the “backflow” interferes with normal flow and/or occurs in a portion of the supply chain which is particularly sensitive to volume change). An example is readmission (and particularly the increased rate of readmission following early discharge)

Other loops can involve external organisations. For instance, an undue rate of surgical admission from nursing homes might raise the question of whether more simple surgery can be done within the nursing home

- Time is an essential consideration in balancing flows: delays in one part of the process (particularly the sort of significant delays while new capacity comes on stream) can create blockages
- The interaction of flows, capacity constraints and delays can create a complex process, but this may be overlaid by unexpected behaviour due to local reactions to the conditions (unintended consequences). For instance, using early discharge to relieve pressure in A&E increases the level of readmission. A particular problem is created where **rules** set up to deal with expected situations actually **undermine** appropriate responses to these conditions. Hence it is important to review **policy** and the degree of operational freedom.

## Introducing a Whole Systems Commissioning Template

The behaviour caused by the interaction of **structure** (the underlying rules and process logic) and **data** (the numbers) can also be studied by looking at **causal loop diagrams** (see “Introduction to the Use of Symmetric Templates” for further discussion on the components of SD) .

One important spin-off benefit of the template is that it illustrates the areas of the model most sensitive to change. As well as offering a clue as to where intervention may be most effective, these sensitive areas show where monitoring of data is particularly important. **These should be incorporated as Performance Indicators and monitored for early signs of issues.**

### Examples

- Early discharges/day
- Medical outliers/day
- Numbers in acute assessment/day
- Readmissions/week
- Average length of stay for each flow
- Utilisation (eg of medical/surgical beds, nursing/residential places)

The model also illustrates the sort of **results** which enable organisations to demonstrate the progress they are making.

### Examples

- Elective wait list
- Elective wait times
- Number of operations performed
- Delayed transfers
- Reimbursement total

## **7. Using the Model in your Local Context**

Symmetric are aware that every locality has subtle differences from those illustrated in the demonstration model. The model is totally flexible and can be altered to reflect your local context. In order to do this, you will need to:

- Check the structure of the model
- Obtain the data for your locality
- Decide the scenarios to run

## **8. Symmetric Assistance**

Symmetric can provide various levels of support for organisations using the whole systems Commissioning Template, including:

- Workshops for key managers to explore the concepts of whole systems and appraise their applicability to the local economy
- Assistance to project teams in understanding the template and disseminating basic insights for discussion in the organisation
- Mentoring of the project team, with a view to testing concepts in the template against local data

## Introducing a Whole Systems Commissioning Template

- Training and support to the project team for amending the template to meet local conditions (roles may vary – we can undertake the customisation, or train and assist the project team to do so)
- Support for adopting the template (standard or customised) and embedding it in an organisation's Performance Framework, as a strategic commissioning tool
- Support for the change management necessary to adopt new ways of thinking, in order to optimise results from adopting the template
- Support for development of a full Performance Management Framework, incorporating the template

### **Symmetric SD Limited**

Symmetric SD is a management consulting organisation focussed on health and social care environments. Symmetric takes a whole system approach to its work and has at its core the use of the System Dynamics methodology.

#### Contact Details

David Monk

P: 01273 564 560

M: 07866 707 195

E: [david.monk@symmetricsd.co.uk](mailto:david.monk@symmetricsd.co.uk)