Submission No 20

INQUIRY INTO IMPACT OF AMBULANCE RAMPING AND ACCESS BLOCK ON THE OPERATION OF HOSPITAL EMERGENCY DEPARTMENTS IN NEW SOUTH WALES

Name: Dr Kendall Bein

Date Received: 10 September 2022

Dear Mr. Donnelly and Members of Portfolio Committee No.2 – Health,

Thank you for the opportunity to make a submission to the parliamentary inquiry into the impact of ambulance ramping and access block on the operation of hospital emergency departments in New South Wales.

I am an emergency physician working as a staff specialist in the Emergency Department (ED) of one of Sydney's inner-city tertiary referral hospitals and trauma centers. I have been an author on several journal articles on emergency department flow, analyzing over 10 million patient presentations to NSW EDs over the best part of a decade, and I continue to be actively involved in research on ED and hospital flow. I have been a part of the Australasian College for Emergency Medicine's working group on access measures. I have firsthand experience of some attempts at managing ambulance ramping including the use of "Transfer of Care" (TOC) beds. In this submission I speak only for myself. I do not claim to speak on behalf of any of the organisations I am a part of.

What's the take home thesis?

Ambulance ramping is the publicly visible tip of the iceberg to the broader issue of whole of hospital flow / whole of NSW healthcare flow – a complex, nuanced and seemingly intractable issue that is solvable but will require political will and significant paradigm shift to do so.

Many very smart people have worked to fix patient flow for many years. If there was a simple fix, or if one section could be fixed independently of the rest of patient flow, then it would already have been done. The complexity of the system, and the need to fix the system as a whole, makes fixing it difficult – seemingly impossible, and trying to fix a small subset doomed to failure, as any fix will depend on other sections of the system as a whole functioning well.

Poor flow in one area causes greater (more visible and functionally impactful) issues in the preceding areas, and thence into the areas that precede that, and so on. When you dam a river (obstruct flow) it is the floodplain upstream that floods (the visible effect) not the village downstream (the erector of the dam, and the cause of the obstruction). Playing "whack a mole" with the issue of ambulance ramping will not work. Ambulance ramping can't be fixed without addressing ED access block. ED access block can't be fixed without addressing ward occupancy, length of stay (LOS) and discharge. Shifting to the far broader and difficult focus of whole system flow will be required, and fixing it will require a systems based solution.

Kind regards

Dr Kendall Bein MB BS FACEM

Contents:

P1. Preamble and take home thesis
P2. (a) The causes of ambulance ramping, access block and emergency department delays
P4. (b) The effects that ambulance ramping and access block has on the ability and capacity of emergency
P6. (c) The impact that access to GPs and primary health care services has on emergency department
P7. (d) The impact that availability and access to aged care and disability services has on emergency

P8. (e) How ambulance ramping and access block impacts on patients, paramedics, emergency department and other hospital staff
P9. (f) The effectiveness of current measures being undertaken by NSW Health to address ambulance ramping, access block and emergency department delays
.. P9. TOC/NART
.. P12. ETP and other time based targets
P13. (g) Drawing on other Australian and overseas jurisdictions, possible strategies, initiatives and actions that NSW Health should consider to address the impact of ambulance ramping, access block and emergency department delays
.. P14. Human factors
.. .. P14. Dunbar's Number
.. .. P15. Engagement and responsibility
.. P16. Aspects of flow
.. .. P16. Increased Capacity slows flow
.. .. P18. Flow resources and the tragedy of the commons
.. .. P19. 8 Principals for managing a common
.. P21. Applying a commons solution to whole of hospital flow
.. .. P21. Prepare
.. .. P21. Commit
.. .. P21. Enact
.. .. P24. Re-measure
.. .. P24. Anticipating criticisms
P25. Appendix – calculating a bed base

(a) The causes of ambulance ramping, access block and emergency department delays

The cause of ambulance ramping is the lack of an available (empty) ED bed to unload into. The cause of the lack of available ED beds is access block – patients for whom the admission decision has been made, necessary early treatment done, and who are safe to be moved to the ward, filling those ED beds and not being placed on a ward. There is rarely if ever a time when there are more ambulances ramping than there are admitted patients awaiting a ward bed. If the admitted patients were placed on a ward, then the ambulances could unload.

The cause of ED access block is a lack of empty beds on the wards. The key word is empty. More beds may or may not be needed. An appropriate number of beds is calculatable (appendix). However, the current system of flow on the wards leads to a state of near 100% ward occupancy. Simply providing more ward beds would lead to the ward having more ***full*** ward beds and still 100% occupancy. Full beds do not allow the ED to unload to the ward.

Bein et al in "Does volume or occupancy influence emergency access block? A multivariate time series analysis from a single emergency department in Sydney, Australia during the COVID-19 pandemic" EMA 33(2) April 2021 https://doi.org/10.1111/1742-6723.13717 showed that elective surgery and hospital occupancy had significant effects for up to 2 days on ETP, while there were no significant lasting effects of either ED presentations or ambulance presentations on ETP. i.e. Hospital occupancy is the major determinant of ED access block.

Garling in "Final Report of the Special Commission of Inquiry Acute Care Services in NSW Public Hospitals" 2008 https://www.dpc.nsw.gov.au/publications/special-commissions-of-inquiry/special-commission-of-inquiry-into-acute-care-services-in-new-south-wales-public-hospitals/

makes pages of observations and recommendations that are as relevant to this inquiry now as they were when written – especially section 17 and 20. Following the chain of flow from ambulance to ED to ward, he notes:"

20.161 Access block and overcrowding in Emergency Departments do not only relate to the number of attendances. They are also caused by what is happening at 'the back end' of the hospital. Where patients are ready to be discharged home or into alternate levels of care, whether it be nursing home care, rehabilitation or other step-down care, but this is not yet possible for practical reasons, they occupy inpatient beds that are needed by those awaiting admission from the Emergency Department. It is also not an uncommon problem for patients who are deemed ready for discharge from the ICU to the ward to remain in ICU for a number of days due to the unavailability of an inpatient bed.

20.162 They illustrate that the pressures on Emergency Departments cannot properly be assessed in isolation from what is occurring in the rest of the hospital. From what I have observed, patient flow in the rest of the system determines patient flow in Emergency Departments to a large extent. There is interdependence in the patient journey between all of the services which a patient receives in the Emergency Department until the time of discharge and re-integration in their communities."

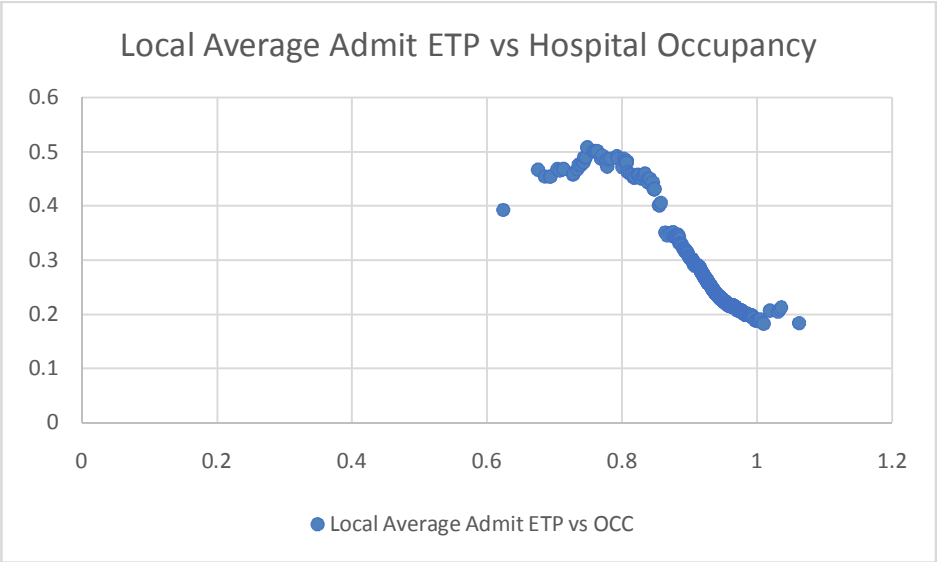
In other words, the wards suffer access block back to the community – RACF, placement for the homeless discharged from mental health, hospital outreach services for those patients requiring some degree of hospital level service but not the constant nursing/medical/allied health support of a hospital bed (i.e. could be at home).

Wards also suffer excess occupancy when their patients suffer overlong Length of Stay (LOS) or delays to the next phase of their care. Any time spent in a hospital bed while not progressing towards health is avoidable excess occupancy. Reducing hospital LOS where it does not compromise care, is functionally the same as creating empty beds, or reducing occupancy. E.g. If a patient waits one day less for their operation, they will get out of hospital one day earlier and will free up one hospital-bed-day for another patient who needs it more.

A quote often attributed to Paul Batalden and W. Edwards Deming is "Every system is perfectly designed to get the results it gets." Regardless of intent, best practice, or "best for patients", the results that the current system of flow on the wards "***gets***" is 100% occupancy, ED access block and Ambulance ramping. If the system achieved 85-90% occupancy at its equilibrium state, then there would be empty beds to unload Admitted patients from the ED, and hence empty ED beds to unload ambulances.

Garling (above) made the "Recommendation 125: NSW Health should commission a research project, the purpose of which is to establish what levels of risk and safety accompany varying levels of bed occupancy within a hospital facility, in order to determine a desirable bed occupancy level for NSW public hospitals." To my knowledge this hasn't been done. The following is unpublished pilot data / proof of concept for such a project.

As access block is linked to safety (section e) one way to do this would be to plot what Emergency Treatment Performance (ETP) (NSW's measure for access block) for admitted patients is for a given hospital occupancy level (± 3%). For one sample hospital it looks like this:



Here we see relatively stable ETP / ED access block up to a hospital occupancy of around 85% then a sharp decline. The implication is that a system that reached equilibrium occupancy around 85% would be running at both the best of making use of hospital beds while also optimising access block and patient safety. This is a single hospital's data, and complex systems have a way of producing unexpected effects, however it would seem this is a good and evidence based starting point.

Key in this is the concept that this is a system level problem. It will require system change to solve. The equilibrium state of the system must have empty beds. ***The 85-90% occupancy cannot be mandated as a KPI***. If it were, it will become a perverse incentive (Admitted patients would be access blocked in ED to preserve empty ward beds and the 85% KPI). It must be an emergent phenomenon from a system that prioritises and resources ward clinician engagement with flow, minimized wasted bed days, forward flow, discharge and outpatient resources that facilitate discharge. Such system changes will be discussed in section g.

(b) The effects that ambulance ramping and access block has on the ability and capacity of emergency departments to perform their function

The primary function of an ED is the resuscitation/stabilisation (if needed), assessment, and appropriate initiation of treatment for patients seeking medical help through the ED, then disposition to a suitable arena (hospital, home, other facility or outpatient service) and medical team (ICU, Operating theatre, ward team, primary care physician, etc) for their ongoing care. This patient care should be within an appropriate, caring, dignified and safe environment (or the best that can be reasonably provided). It should be provided at the quality achievable and expected by patients, the healthcare system and relevant bodies e.g. ACEM.

There are numerous secondary functions including: culture within the department and interactions with other departments within the hospital, ensuring the morale, safety and sustainable wellness of staff, training of the next generation of ED workers, disaster preparedness and occasional disaster management, and many, many more. Impact on secondary functions are better addressed in Section e.

Systems exist to ensure that ambulance ramping has reduced impact on resuscitation and stabilisation of patients. EDs have systems in place where critically unwell patients will bypass any queues or delays to reach appropriate resuscitation areas, and, with pre-arrival notification, bedspaces can be shuffled within even an overcrowded department to allow for critical interventions.

That said, there are certainly cases where the process of bed shuffling to create resuscitation space means that a patient who would optimally remain in a resus bed is stepped down earlier than ideal as the need of the incoming patient is greater. Again, when this happens, there always exists a patient who could be appropriately transferred to a ward (often mental health ward) or critical care space (ICU), or an admitted patient in an acute bed to be transferred to a ward, so a resus patient can be appropriately stepped down to an acute bed. i.e. the knock on effects of access block.

In non-rescucitation situations, ambulance ramping causes some limited impact on ED capacity and function. It creates a barrier to ED clinician's access to the patient, limits the care, dignity, safety and quality we can provide, prevents some aspects of assessment and management such as privacy required for history taking, ability to examine a patient, and imaging investigations. It is possible to medically manage a patient still on an ambulance trolley and certainly to provide urgent care, but not with the care, safety, dignity or quality that the patient should expect. Managing a patient on an ambulance trolley takes far greater time, effort and resources, denying those resources to other patients.

Furthermore, any corridor beds (TOC/NART/ambulance trolleys/patient beds placed in non-treatment areas) causes logistic and safety issues for all patients and staff within the department. Issues include:

- Lack of Privacy for history taking, examination, procedures, toileting, dressing, and even just the right to be somewhat separate from public view while unwell / in pain / nauseous / bleeding / not in their best state / not on their best behaviour.
- Difficulty for supportive family and friends to be nearby while in a public corridor.
- Lack of room and privacy for procedures
- Monitors not linked to central monitoring and recording
- Narrow uncomfortable bed with higher falls and overbalance risk
- Stimulating environment (a bad thing for agitated or delirious patient, or even just those feeling unwell and wishing for less hurly burly in their misery)
- Obstruction of a critical corridor and access to fire safety equipment
- Difficulty with accessing imaging
- Close quarters infection risk
- Patient safety, as the corridor they are placed in is a common flight path for escaping dangerous/mental health/drug affected patients
- Many resuscitation facilities that are immediately available at the bedside in planned acute monitored areas, such as wall oxygen, wall suction, and racks of common equipment such as oxygen delivery devices, gloves and vomit bags are not available in the ED corridor
- Increased risk of lost paperwork and other accompanying paraphernalia

By comparison the impact of Access block on ED capacity and function are far, ***FAR*** greater.

If EDs are access blocked then they are overcrowded. If EDs are overcrowded, then, by definition, they don't have the resources to perform their core functions. By definition, if EDs are access blocked, then they lack capacity and have reduced function.

An ED with access block has fewer beds available for patients who need them, fewer rooms in which to see patients with appropriate privacy, and fewer beds with increased level of care e.g. monitoring, negative pressure, meeting paediatric needs etc.

The time taken by clinicians to find physical space to see a patient increases dramatically, sometimes taking longer to find an appropriate space than it does to perform history, examination and investigations for that patient. This is not exaggeration or hyperbole.

The access blocked admitted patients become an increasing drain on ED resources - the time taken to care for them, the time taken to communicate with the team they are admitted under to discuss that management, the time taken to communicate with their friends and family – especially to apologise and explain why they remain access blocked in ED rather than receiving treatment in an appropriate ward with the subspecialty care available there, but that is not available in ED.

This extra resource cost for admitted access blocked patients is compounded when the patient's condition is worsened by being in ED. This particularly applies to patients with delirium or mental health issues whose issues worsen with prolonged stays in the overly stimulating environment of an overcrowded ED.

All these resource costs are taken from the normal function of the ED. In my hospital, there is a measurable decrease in function when more than a 20-25% of the average total daily admissions are admitted waiting for a ward bed. It is not uncommon to have double that number access blocked in ED, and on such days anywhere up to half a clinicians day can be spent caring for access blocked patients rather than caring for new presentations to ED.

(c) The impact that access to GPs and primary health care services has on emergency department presentations and delays

All evidence and expert opinion is that access to alternative care (primary health) is not influencing the burden of ED presentations. ED presentation numbers especially low acuity numbers are not a determining factor for ED delays, especially those of clinical significance.

Richardson and Mountain in "Myths versus facts in emergency department overcrowding and hospital access block" Med J Aust 2009; 190 (7): 369-374. doi: 10.5694/j.1326-5377.2009.tb02451.x https://www.mja.com.au/journal/2009/190/7/myths-versus-facts-emergency-department-overcrowding-and-hospital-access-block point out:

"Importantly, there is no evidence for the often-proposed myth of "general-practice-type" or "inappropriate" patients leading to ED overcrowding. Discretionary presentations by patients with low-complexity conditions, who might reasonably be managed elsewhere, constitute an insignificant workload in most EDs. These patients are uncommon in major EDs, and the most frequent reason for them to attend an ED is because they were referred by a GP. They rarely require admission or even use of trolleys, they use minimal ED resources (less than 3% of all costs or resources in most EDs), are easy to deal with, and do not impose on the key functions of the ED (assessment of sick patients, complex treatments and resuscitation). They may attend an ED because no other options are available and, importantly, they often feel their

medical needs are urgent. They may spend a lot of time in waiting rooms, but this does not affect overall ED function. This myth is particularly problematic in that, if allowed to continue to be given credence, it continually diverts attention to solutions that cannot deal with the key issue causing dysfunction in the ED — that is, excessive numbers of admitted patients."

Access to GPs and primary health care services has little to no effect on emergency department presentations and delays.

I fear that this question is specifically aimed at the concept of "Patient Diversion".

The concept of "Patient Diversion" is fundamentally flawed. Patients who have any potential to be managed in an outpatient setting choose to present to ED because they feel their medical issue needs to be seen in a ED because:

- It is too complex for a primary care physician
- It is too urgent for a primary care physician
- Primary care physician is not available within an appropriate period of time
- Primary care physician is not available at a time convenient to ***me***

None of these issues will be solved by any "Patient Diversion" strategies. Patients choose where they go, and currently they go to the ED. "Patient Diversion" strategies won't change this because, at the end of the day, it is patients, not the system, who decide which service they present to.

More significantly, the framing is flawed. We should not seek "Patient Diversion". If we want patients to present to outpatient facilities other than Emergency Departments, then we should provide easily accessible outpatient services that are:

- Known to patients who need them
- Meet patient needs better than emergency departments
  (this should not be hard for any sub-specialty service)

We should pull (to a better outpatient service) not push (from ED) – ideally in a way that is cost (resource) efficient for the NSW Health care system, and in a way patients choose to use.

If viewed holistically, any resources that might be spent on "Patient Diversion" are better spent at the "back end" of the hospital in outpatient management of patients who are traditionally managed within the hospital system (e.g. through Hospital in the home, outpatient clinics, outreach services and step down units, Virtual medical care, RACF, etc.)

This would create empty ward beds, hence, unload EDs, hence, and give us room to perform our core function, thence possibly refer to these outpatient services (a virtuous cycle)

(d) The impact that availability and access to aged care and disability services has on emergency department presentations and delays

In my experience, access to aged care and disability services has little direct effect on ED presentations or delays.

There are very few presentations of the elderly or disabled to ED that would have gone to any existing outpatient service. Specifically, there are ED presentations that might be better served by such a service, but it is rarely lack of access that leads to ED presentation.

Acute on chronic illness presentations, Nursing Home referral presentations, Carer stress related presentation, DFQ presentations, new oligocopia presentations are coming to ED because the patient or their carer believes the patient needs hospital care and ED is the front door to the hospital. The problems the patient or carer bring to ED are issues that will always go through EDs, as the perceived need for medical intervention is never one that they feel can wait until later in the week. Planning through GP does not appear to saturate current access channels to the point they spill over to EDs.

On the other hand, lack of access to aged care and disability service has significant indirect impact on ED access block and function. This is a "back of hospital" flow issue. These services are frequently required for safe and timely discharge of ward patients. Delays to discharge means increased occupancy, fewer empty ward beds, increased ED access block and higher risks of ambulance ramping.

By contrast, lack of access to mental health, addiction and Drug and Alcohol services absolutely increases ED presentations and delays. Patients presenting with Mental health and Drug and Alcohol related problems are common, time consuming, frequently managed in ED less well than they would be by outpatient services, are often frequent presenters, and often would prefer to see a subspecialty service, or have even sought to do so, and only attended ED when they were unable to access such a service. Such patients suffer very lengthy stays in ED, contributing to overcrowding and access block, are often very resource and time intensive to care for. The ED is often a very bad place for such patients to be spending long periods as the loud, confusing, overcrowded, and over-stimulating environment is not good for their mental health.

(e) How ambulance ramping and access block impacts on patients, paramedics, emergency department and other hospital staff

Without a doubt the most troubling and tragic impact of ambulance ramping and access block is that it causes patient deaths.

The media is full of cases of patients who died waiting for an ambulance while ambulances waited to offload. And though it is easier to draw a direct line from this type of ambulance ramping causing death, the reality is that ED access block is just as responsible for patient deaths – it's just that the line is harder to draw. It is never as clear for any particular patient death whether it is the delays incurred in an overcrowded ED, or inevitable progression of the patient's disease that lead to their death. So to quantify the numbers of patient deaths attributable to access block, you need epidemiological studies looking at how many deaths occur when patients present to an ED when it is not access blocked, and how many more deaths occur when the same ED is access blocked.

Garling noted the fact that overcrowding is linked to patient deaths:

"20.87 The evidence from medical literature supports the conclusion that the most disturbing result of overcrowding in Emergency Departments is that it is associated with patient death. Where overcrowding compromises the provision of safe care, I regard it as dangerous and unacceptable."

More recently, Jones et al in "Emergency department crowding and mortality for patients presenting to emergency departments in New Zealand." EMA December 2020.

https://doi.org/10.1111/1742-6723.13699 notes that “Access block had the strongest association with 7-day mortality. That ED occupancy and the number of arrivals were not associated with increased mortality suggests that system issues related to long ED stays may be most important in the link between ED crowding and mortality.” Specifically, seven-day mortality is 10% higher for patients arriving at times when there was more than 10% hospital access block.

Dinh et al in “Predictors and in-hospital mortality associated with prolonged emergency department length of stay in New South Wales tertiary hospitals from 2017 to 2018” EMA 32(4) August 2020 https://doi.org/10.1111/1742-6723.13699 shows that the hazard ratio for 30-day all-cause mortality over time was 28% higher in those not meeting ETP benchmarks after adjusting for age, triage category, comorbidities, ICU and service-related group. Needless to say, over 3000 extra deaths each year in NSW due to ED overcrowding is not acceptable.

Ambulance ramping, access block and overcrowding are all linked to significant morbidity as well.

- Lack of Privacy for history taking, examination, procedures, toileting, dressing, and even just the right to be somewhat separate from others while unwell / in pain / nauseous / bleeding / not in their best state / not on their best behaviour.
- Difficulty for supportive family and friends to be nearby in an overcrowded ED
- Stimulating environment (a bad thing for agitated or delirious patient, or even just those feeling unwell and wishing for less hurly burly in their misery)
- Close quarters infection risk
- Patient safety, with increasing numbers and distress amongst co-located dangerous/mental health/drug affected patients who become more distressed with increased ED time from access block
- Longer delays to nursing care as staff are stretched thinner across more and more access blocked patients.
- Constantly interrupted sleep, spending 24+ hours in a chair not a bed, prolonged fasting for operations that cannot occur until a bed is allocated, etc.

For staff there is an understandable impact on their wellness, morale and mental fortitude. The increased crowding, noise, hurly-burly, infection risk and frustration of working in an overcrowded ED, not to mention feeling that the very environment you work in means you are letting your patients down, takes its toll. That fixing it is outside your control worsens the impact it has and breeds a sense of powerlessness, and that those with the power to influence it do not care for the wellness of you, your patients or your colleagues. Burnout is an ever present concern among any critical care staff, and access block is a large contributor in the modern ED.

(f) The effectiveness of current measures being undertaken by NSW Health to address ambulance ramping, access block and emergency department delays

TOC/NART

In 2015, the hospital I work at introduced “Transfer of Care” (TOC) beds where, in cases of ambulance ramping, 2-6 beds in ED corridors would be provided with allocated ED nurses, and availability of co-located monitoring for unloading Ambulances. This is similar to other “Nurse Ambulance Release Teams” (NART) trialed unsuccessfully at other hospitals in previous years, tragically resulting in at least 2 patient deaths.

Past president of ACEM, Assoc Prof. Dr Sally McCarthy said it was an inappropriate strategy to care for patients or manage delays in the system. "It's just replacing one dysfunctional system for managing the queues with another".

The Auditor-General recommended in 2013 that the practice of having paramedics stay with patients be phased out, but warned that hospitals should do this by improving patient flow through the hospital and not through additional release teams. "While these strategies free up ambulances, they do not reduce overcrowding in the hospital". "Performance may improve in the short term, but may not be sustainable."

The Ministry of Health admitted at the time that such strategies had been named in coronial enquiries as contributing factors to patient deaths.

Harriet Alexander "Pilot scheme to reduce ambulance delays implicated in past patient deaths." SMH September 17, 2015 https://www.smh.com.au/healthcare/pilot-scheme-to-reduce-ambulance-delays-implicated-in-past-patient-deaths-20150917-gjouoq.html

Julie Robotham Medical Editor "Casualties of a sick system" SMH January 10, 2009 https://www.smh.com.au/national/casualties-of-a-sick-system-20090110-gdt98i.html

5 years after the introduction of TOC beds we were in a position to examine and reflect upon the long-range systems effect that TOC beds have.

Do TOC/NART beds Work? (TL;DR: No)

By 2020, TOC beds had been used long enough for systems of patient flow to reach a stable equilibrium. The Covid-19 pandemic meant that from March 2020 onwards, TOC beds had to be eliminated for safe patient distancing. This allows us to compare TOC performance both with and without TOC Beds.

From a patient point of view, TOC beds increase discomfort, morbidity and mortality, but really make no meaningful difference to time spent on ambulance trolleys. Both 2019 and 2020 saw 88% of all ambulances offload within 30 minutes, with an average offload time of 18 minutes. What is most interesting is that from early December to late January when elective theatre is winding down, for both years, Ambulance offload KPI approaches 100%. It once again approached 100% in March/April 2020 when elective theatres shut for the start of the Covid pandemic and ward occupancy fell. The TOC KPI is intimately linked to ward occupancy. It is not linked to the presence of TOC beds.

That there is no meaningful change in offload times is to be expected. If action is taken to alleviate access block when ambulances are ramping, and levers to increase flow to wards are only pulled when the KPI is at risk, the equilibrium point will always lie just below the KPI. TOC bed occupancy is in itself not a trigger to action, so TOC beds make no difference to this equilibrium point. This will be true of most non-integrated, reactive systems.

In practice, the TOC beds fill with the first few ambulance arrivals to a fully access blocked ED. However nothing changes for the access block, and so as more ambulances arrive they become ramped in an access blocked department, now also suffering from 4-6 beds filled with access blocked patients who are obstructing a critical corridor.

Can TOC/NART beds work? (TL;DR: No)

When considered from the point of view of their role in ED and whole hospital patient flow, TOC beds fill the role of inferior acute beds.

With rubber walls and endlessly expandable number of chairs in the waiting room, rapid treatment or fast track area, no ambulance waits long to offload to an unmonitored chair. When ramped, ambulance offloads are inevitably waiting for monitored beds, or special requirement monitored beds (resus or isolation beds).

In comparison to acute monitored beds, TOC beds are inferior in all the ways ambulance trolleys are as listed in section b above.

They offer no benefit over Acute monitored beds. It could be argued that they could have a role in patient flow, if, when the first TOC bed was filled, it triggered protocol or immediate escalation to enact measures that rapidly moved already admitted patients who are boarding in the ED to ward beds, freeing spaces to unload first the TOC beds and then the other ambulance offloads that will inevitably be incoming. Such an argument would be false, but certainly warrants further examination.

The first issue with such a proposal would be the need for rapid escalation. Ad hoc decisions regarding changes in flow management do not appear to happen quickly. Protocolising an overcapacity plan might be a solution, however such solutions have their own issues.

Assuming an overcapacity protocol were to be designed, it would need to provide timely access to appropriate ward beds for admitted patients in the ED.

Timely access is an ambiguous phrase; however, we can attempt to put a more specific time frame on it. As it has been deemed that the vast majority of ambulances should be offloaded within 30 minutes, it seems reasonable that the same timeframe be applied to flow measures created such that the knock-on effects allow the timely offloading of ambulance trolleys. So, ward space for patients would need to be available within 30 minutes of the first TOC bed being filled. This would be difficult for an already-in-place protocol, and impossible for reactive solutions.

A 30 minute timeframe is not sufficient time to open a ward, call in extra staff, delay elective admissions or discharge existing patients. Any patient already ready to vacate a bed should already be in the discharge lounge, so mass mobilisation of ward staff to facilitate discharges cannot be relied upon to create beds.

Current disaster plans allow for moving admitted patients from the ED to ward corridors. A similar plan could be created as part of an overcapacity plan. Richard et al in “Patients Prefer Boarding in Inpatient Hallways: Correlation with the National Emergency Department Overcrowding Score” Emergency Medicine International · Dec 2011 https://doi.org/10.1155/2011/840459 and Viccellio et al in “The Association Between Transfer of Emergency Department Boarders to Inpatient Hallways and Mortality: A 4-Year Experience” Ann Emerg Med. 2009 Oct;54(4):487-91. https://doi.org/10.1016/j.annemergmed.2009.03.005 found that waiting in ward corridors for a bed was safe for patients, and acceptable to more patients than boarding in the ED, and that patients waiting on wards for a bed found a bed faster than boarding in ED – usually within a few hours.

It must be noted that any overcapacity plan reliant on TOC beds as a trigger could be enacted when the last **acute beds** are filled, or when more than 20% of the daily average admitted patients are access blocked in ED, removing the need for TOC/NART/ED corridor beds and their attendant morbidity/mortality.

However, implying that the choice is between TOC beds or ward corridor beds is failing to recognise that even needing to use TOC/NART beds or overcapacity plans means the hospital is already in a state of internal disaster that should rarely if ever occur – not a routine part of a system. Once again, "It's just replacing one dysfunctional system for managing the queues with another".

The hospital SHOULD have a protocolised overcapacity plan. It should have a protocolised plan or guideline for ALL foreseeable disasters. However as with any other disaster management plan, there need to be proactive systems that aim to ensure that it is rarely (or never) needed. The problem of whole hospital flow and especially ward flow must be solved, such that there are rarely any access blocked patients in the ED, and ambulance ramping does not occur. This means enacting system changes whereby there exist empty ward beds when the hospital is at equilibrium.

ETP and other Time based targets

Time based targets seem to be an appropriate KPI for Hospital flow, and, when followed, are capable of producing improvements in ED LOS and access block, until their value is eroded by gaming the system. To be clear, I agree with and wholeheartedly support time based targets, but feel they need to be the right targets.

Forero et al "Impact of the National Emergency Access Target policy on emergency departments' performance: A time-trend analysis for New South Wales, Australian Capital Territory and Queensland" EMA July 2018 31(2) https://doi.org/10.1111/1742-6723.13142 illustrates that introducing time targets meant that those time targets improved, but also found "Significant increases in short-stay admissions suggest a strategic change in ED process associated with NEAT implementation."

Hession et al "Gaming National Emergency Access Target performance using Emergency Treatment Performance definitions and emergency department short stay units" EMA Dec 2019 31(6) https://doi.org/10.1111/1742-6723.13295 shows one of the ways the target is gamed to produce improved numbers without improving the access block for those most vulnerable to its effects – the admitted, access blocked patients awaiting a ward bed.

Unfortunately, the problem is baked into the concept of KPIs. Goodhart's law states "When a measure becomes a target it ceases to be a good measure". The ETP KPI is ripe for gaming, and contains enough loopholes that it no longer reflects the goals for which it was introduced. The example of use of ED short stay units is illustrative. "Admitted patients" include those admitted to wards, and those admitted to short stay units. Patients admitted to short stay units make up 40-60% of all ED admissions, are under the control of ED physicians with an interest in flow and stay for short times. Ensuring near 100% of short stay admissions meet ETP KPIs is easy. If 100% of 40% of admissions meet ETP, then the remaining 60% of admissions – the ward admissions - can only meet ETP 20% of the time, and the "Admitted ETP" is still over 50%. If the goal of ETP was to get admitted patients to the wards within 4 hours, it can measure as 50% while only achieving 20%. Please note the nadir of the graph in section a.

This is one example. There are many, many other ways ETP can and is exploited such that it no longer is an effective measure or target.

There is no good solution to Goodhart's law. The two things that come closest are:

- Not making the target or the measure a KPI, but allowing it to emerge as part of the system
- Make sure the target is actually the thing you wish to optimise. Setting it as a KPI will likely optimise it – though it still may have perverse incentives and unintended consequences.

If the goal is to get admitted patients to the ward, then set that as a target. As "admitted" can be gamed through short stay units, remove it from the equation by substituting it with "patients whose total hospital length of stay > 24 hours". As binary targets (e.g. 4 hours yes/no) can be gamed by prioritising patients who have not yet breached KPI, remove it from the equation by targeting an average. It won't be perfect, but the result is more likely to coincide with the actual goal.

For instance- Time target: Average ED LOS for all patients with [total hospital LOS (ED + Ward) > 24 hours] should be under 4 hours.

This is still not perfect – it can be gamed with transfers to other services – you would need to define a transfer with plans to be managed in an inpatient setting (even if transferred to an outpatient setting with plans to be admitted from there) as meeting the 24 hours requirement regardless of actual time spent in ED/hospital, and to an outpatient setting with plans for discharge as not. This is important. When a mental health patient remains trapped in ED for 20 hours before being transferred to a mental health facility with plans to continue their care there, they should count as 20 hours in ED for an admitted patient.

It does not give immediate feedback – you cannot calculate the time target in real time – it is only available a day later. This is good – it makes it harder to gameify and falsify, but harder to manage issues reactively – if managing reactively was at all desirable. That said, you could choose to manage reactively based on the consequences in real time – i.e. when the ED is overcrowded, access blocked or ambulances are ramping then react appropriately.

(g) Drawing on other Australian and overseas jurisdictions, possible strategies, initiatives and actions that NSW Health should consider to address the impact of ambulance ramping, access block and emergency department delays

Based on everything discussed above we find ourselves in a position where we need a systems based solution to the issue of whole of hospital flow – from Ambulance to ED to ward to discharge to the community. Failing to address any of these will mean access block somewhere in the system, and dooms any intervention to failure or undermining of intent by perverse incentives.

Any systems based solution needs to achieve:

- Minimal ambulance ramping
- Minimal ED access block
- Empty ward beds when they are needed (Ward occupancy of 85%-90% at equilibrium)
- All without compromising patient care

:- better than the current system does, and in a manner acceptable to the participants and funders of the system.

Let us begin by recognizing this is a large, complex system.

- Changes and problems in one part of the system will produce unanticipated problems in other parts
- One of the critical yet hardest to anticipate and manage parts of the system is human psychology
- Any incentive structure used to drive the system will be prone to Goodhart's law, so any target needs to be close enough to the desired goal that the goal is dragged along with the target as it is gamed, or, needs to be an untargeted emergent equilibrium point of the system, and we will need to deal with any perverse incentives and knock on effects generated along the way.
- Implementation will be difficult, with teething problems, and likely to make things worse – or harder before they get better.

Failure to understand or anticipate the influence of any part of a system will make it brittle, prone to failure, or subject to the influence of perverse incentives and unanticipated consequences and costs. As such, there are a few more factors that need discussion before we can construct a systems based solution.

Human factors

Dunbar's number

The concept behind Dunbar's number is that the human brain has a cognitive limit to the number of people they can keep in stable social relationships. Essentially when there are too many people in interlocking social relationships, then the human brain can't treat this larger number as it does the small number of close relationships it does have, can't navigate the larger social relationship web, and other factors like more restrictive rules, and enforced norms are needed to maintain a stable cohesive group.

Dunbar: "Coevolution of neocortical size, group size and language in humans." Behavioral and Brain Sciences, 16(4), https://doi.org/10.1017/S0140525X00032325

There remains controversy as to the exact size and range of Dunbar's number (most seem to agree that it is something like 100-200), or the range and quality of social dynamics that can be maintained with differing group size. Also, obviously, different people will have different group sizes they have the ability to, and feel comfortable navigating.

Prof Robin Dunbar: "Dunbar's number: why my theory that humans can only maintain 150 friendships has withstood 30 years of scrutiny" The Conversation May 13, 2021 https://theconversation.com/dunbars-number-why-my-theory-that-humans-can-only-maintain-150-friendships-has-withstood-30-years-of-scrutiny-160676

Thankfully those controversies don't apply to our discussion – only the general concept that human psychology can only deal on a close social level with a limited group – with a size in the ballpark of Dunbar's number. Outside of that group size our interactions are as with strangers, governed by social norms. It is not that we don't care about those outside the social group, just that that care is more abstract and distant. A caring and empathic person (who are hopefully

over-represented in healthcare) will still want the best for all people, however it is not the same as it is for those you have a social connection too – the limited group inside your Dunbar's number.

Engagement and responsibility

Currently those with engagement in and the responsibility and empowered to manage hospital flow (hospital executive) have only macro-economic controls to influence it. They can increase or decrease the supply of beds (surge beds) and resources (staffing numbers etc), and globally decrease demand (closing elective theatre and decreasing elective admissions), and apply political pressure to the coal face to prioritise discharging patients. They cannot directly discharge patients, affect length of stay or hospital occupancy, or micromanage obstacles to flow. Any change they make will take significant time to take any effect on forward flow.

Currently those with these microeconomic controls – discharge, recognition and navigation of obstacles to flow, redistribution to outpatient settings in a manner that is safe and acceptable to patients, engagement of outpatient flow resources such as Hospital in the home, clinics or virtual care, and with the ear of- and relationship with- patients and carers – are the ward teams. Currently they do not have engagement with or responsibility for hospital flow.

This is not a criticism of the ward teams – it is an inevitability of the system's interplay with human factors. The ward team's primary responsibility is to their patients not to hospital flow.

E.g. if Mrs Smith is capable of being safely discharged, however doing so is not as convenient for her – her children can't pick her up until tomorrow or better still the weekend, when one of them can stay overnight with her and help her settle back in at home, then it seems reasonable to delay her discharge until everything is optimal. Even if they were aware of Mrs Jones, access blocked in ED (which they are not) – she is a patient of a different team who they have no responsibility for – she is outside their Dunbar number of social responsibility, the hospital has hundred of beds – not just the bed occupied by Mrs Smith, and finding a bed is someone else's responsibility. They are still good people who, if they were aware of her, would wish Mrs Jones the very best (they would wish that of any stranger) and would not do anything to worsen her plight, but they don't see that anything they are doing is worsening her situation, and even if this were pointed out, they don't feel they should compromise Mrs Smith's best possible outcome.

There are many other aspects of ward flow that are better micro-economically managed by the ward team than macro-economically by the hospital as a whole – awareness of subtleties of care, individual needs of their patients, subspecialty level understanding of the patient's illness and patterns of recovery, and ward level organisational factors. They have firsthand knowledge of obstacles to flow which lengthen hospital LOS, increase hospital occupancy and contribute to access block so can respond faster and with greater focus.

For ward flow to work, those with the levers to make it work – the ward teams – need to be engaged in flow in a way that has them take into account both the convenience of those patients ready for discharge and those patients suffering the morbidity and mortality attached to access block, and choose to prioritise one patient's morbidity and mortality over another's comfort and convenience unless in the most extenuating circumstance. They need to see obstacles to flow as more than an inconvenience that they must patiently wait to clear, but as a threat to the morbidity and mortality of patients they care about, prompting them to vocally advocate for flow, seek new systems of care that promote flow, and plan in advance to minimize the impact of

such obstacles. Accounting for human psychology, this will probably mean keeping both the access blocked and the patients who would be inconvenienced but without compromise of care - within the team's Dunbar number sized circle of social responsibility.

Aspects of Flow

Increased capacity slows flow

It seems counter-intuitive that **increased capacity worsens flow**, however this is what we find occurs in real life. This confusing effect is further compounded by the fact that as a hospital increases its scale and scope of service, it will require increased capacity, asks for extra beds, is given extra beds, but flow worsens.

This effect is probably best seen in an analogous situation of traffic flow. There are many parallels between road traffic flow and hospital patient flow, and lessons can be learned from road traffic that can be applied to patient flow

These parallels include:

- Fixed resource capacity, and variable inputs and outputs dependant on flow and external factors (number of road lanes and length, drivers choice of when to drive and access to alternate transport for road traffic; fixed bed space, ability for management to control elective admissions, and ability for teams to discharge early or late for hospital patients)
- Flow worsened by start - stop progress (traffic lights in road traffic; delays to radiology, OT, or being seen late on daily ward round/outlier status for hospital patients)
- Exponential worsening of flow with congestion (due to lane changing and reduced safe driving speeds in traffic; and by later ward round, divided duties of ward teams, delays to access flow resources for hospital patients)
- Limitless demand from service users that easily exceeds any practical resource provision (extra road lanes are filled as soon as finished. There is always more demand for hospital beds for elective admissions)

The following four case studies of road traffic flow are useful as they show what many would consider counterintuitive results, but the flow principals behind the results are both predictable and significant in managing both road traffic flow and hospital patient flow.

In the early 2000's, the I-10 Katy freeway in Texas, USA was labeled by the American Highway Users Alliance as the second worst bottleneck in the nation. In 2005 travelling from Taylor to Pine Oak (the currently measured ends of the highway) took 52 minutes. So plans were made to expand the highway from 6-8 lanes to 26 lanes. Doing so has markedly worsened flow on the highway. Highway expansion finished in 2010, and in 2011, it took 47 minutes to travel its length. By 2014, it took 70 minutes.

This is the result of increasing capacity without approaching the issue of congestion. Once the same congestion recurs in an area with greater capacity, the flow is inevitably worse. In the hospital system, there is often a need for increased bed numbers – but this should be a **scaling** effort to match increased inflow, and new or changing patient demographics. Appendix shows an approach to calculating bed numbers to suit the scale of service

Increasing bed numbers to attempt to fix flow is counterproductive. Flow does not require more beds, it requires less congestion – more free flow resources (free beds, empty radiology slots, more porters, ward rounds that don't lose efficiency by travel to outliers.) Without fixing this congestion issue, any new beds created to fix flow will be rapidly filled, and efficiency of flow resource use will fall still further.

In 2000, Minnesota Department of Transport tested the effectiveness of their ramp meters (which limit number of cars per minute entering a stretch of freeway) by switching off the ramp meters for 8 weeks. They found that highway capacity **decreased** by 9%, Travel time **increased** by 22%, traffic speed **decreased** by 7%, and Accidents **increased** by 27% (Levinson et al. "Ramp meters on trial: Evidence from the Twin Cities metering holiday" Transportation Research Part A: Policy and Practice Volume 40(10) Dec 2006) https://doi.org/doi:10.1016/j.tra.2004.12.004.) There were some exceptions to the improved efficiency with ramp meters turned on including some short trips and Off Peak time periods.

The implication this has for hospital patient flow is again the lesson of the marked efficiency gains of lower congestion, though this may not be as great in cases where flow efficiency is influenced less by congestion – i.e. Day-stay cases.

A similar effect was seen in Stockholm Sweden in 2006 when they introduced a 1-2 euro charge for driving in the centre of the city. Numbers of cars in the inner city dropped by 20%, and because of the exponential effect of congestion on flow, a small reduction results in a markedly greater improvement on travel times and safety – with queue times decreasing by a third in the morning peak hour, and by half in the afternoon. Travel time also became far more predictable for commuters – the reduced congestion also reducing variability of delays.

The lesson for hospital flow here regards the exponential nature of congestion. The impact of the thousandth car on a road is far greater slowing of far more cars than the impact of the first or the tenth. Traffic Jams occur with increasing frequency as traffic congestion crosses critical thresholds. The same is true of hospital flow. Congestion exponentially slows flow. A small reduction in congestion will result in a larger improvement in flow and a net gain in the number of patients completing their journey to health per day. The difficulty for those charged with managing that flow is to create that reduced congestion and prevent it from re-filling to the previous level. Stockholm created it by changing incentives – placing a trivial charge on entering the city. The same is not directly translatable to the hospital system.

In 2008, researches at Department of Complex Systems Science, Nagoya University Japan demonstrated that traffic jams occur spontaneously, without any causative bottlenecks – once traffic reaches a certain density (congestion). Starting with 22 cars driving on a 230m circumference track at 30 km/h, traffic jams spontaneously formed and continued to propagate throughout the traffic, eventually stopping forward movement within the jam.

Yuki Sugiyama et al "Traffic jams without bottlenecks—experimental evidence for the physical mechanism of the formation of a jam" New J. Phys.10 March 2008. https://doi.org/10.1088/1367-2630/10/3/033001 https://www.youtube.com/watch?v=Suugn-p5C1M

The hospital flow implications of this finding are that clearly defined bottlenecks are not needed to disrupt smooth flow. Once there is sufficient congestion, flow will be disrupted spontaneously. In NSW's hospital system, 99% congestion is clearly sufficient congestion to

cause such spontaneous disruptions of flow. Focus solely on fixing bottlenecks will not solve the flow issue. Focus on congestion on wards needs to be addressed simultaneously.

Looking at the findings and failings of traffic analysis shows us that the solution to our problem of access block and its impact on flow is not just more beds. It is improved flow through decreased congestion. Any bed increase needs to be only to offset/match increases in demand (i.e. scaling), and these increases must be separate to, and preceded by creation of systems that will reduce occupancy and maintain that reduced occupancy in the long run.

The concept of flow resources and the Tragedy of the Commons

Within the hospital system exist a large number of resources that strongly influence flow, are limited in supply but are in high demand by many parts of the hospital. These flow resources include hospital beds (especially beds on a team's home ward, and specialist beds such as ICU, respiratory, monitored and isolated beds), operating theatre availability, clinic time, space and financing, procedure suite availability, medical, nursing and allied health staffing, availability of radiology, and discharge resources. As mentioned above, it is not just the availability of these flow resources, but how long they will be occupied by the patients of individual departments as they journey from illness to health that will affect flow.

In 1833, William Forster Lloyd described a situation where individual decision-makers seek to use a shared resource, where demand for that resource exceeds its supply, where overutilisation leads to benefit of the individual but is harmful to the common good. This is a good parallel to the situation in a hospital where individual departments seek to use shared flow resources .

This situation is known to economic and game theory literature as "The Tragedy of the Commons". The commonly described model is of a pasture (the common) shared by a group of farmers who all have a right to graze their cattle on it. Grazing more cattle will degrade the common and lower the production for all the farmers, but the benefit of adding one more cow for any individual farmer is greater than that farmer's loss in production. So all farmers have the economic drive to overgraze even though all of them suffer in the end.

Economics views this situation in terms of externalities. Externalities are the costs or benefits that affect a party who did not choose to incur that cost or benefit. In the classic Tragedy of the Commons problem, it would be the decreased production of all cattle on the common due to overgrazing. In the hospital flow model, it is the slowed flow (out of ED, and hospital wide) due to utilisation of flow resources in excess of their optimal use.

Until recently, economics has recommended one of two solutions to the problem.

The Governmental Solution: a governing authority takes responsibility for the common and regulates how it can be utilised and by whom. The correct system of regulation will allow the common to be utilised at its optimum rate without risk of being degraded by overuse. This is what we see in NSW hospitals.

The Non-Governmental solution – ie Privatisation: The common is divided into privately owned parcels amongst those with a claim to it. It ceases to be a common, and all externalities and benefits are borne solely by the owner of the parcel.

The hospital system utilises the governmental solution with the executive regulating how flow resources are allocated. As discussed, these resources are not being optimally utilised implying the current regulations need to be revised by those with the responsibility and authority to do so.

Game theory literature includes tens of thousands of papers on the Tragedy of the Commons. Its approach and solutions differ greatly from those found by economics. Even the most simplistic game theory models establish that a common won't be overgrazed and destroyed, but rather will reach a suboptimal equilibrium at the point where all but the most efficient farmers have had their marginal utility reduced to zero by overgrazing. From there it seeks to find game play strategies for the players of the game.

This is the situation (eroded marginal utility) that we find in NSW hospitals, with near 100% ward occupancy, inefficient ward flow, access block and ambulance ramping.

Game theory offers several solutions to the tragedy of the commons with mutual cooperation between players becoming the optimal strategic option, and the common maintained at optimal utilisation. (Florian K. Diekert "The Tragedy of the Commons from a Game-Theoretic Perspective" Sustainability 2012(4) https://doi.org/ 10.3390/su4081776), and that the best solutions are available for a common managed by small groups of players with long term goals (Stewart, A. et al. "Small groups and long memories promote cooperation. Sci Rep 6, 26889 (2016)"; https://doi.org/10.1038/srep26889)

In 2009, Elinor Ostrom became the first, and one of only two women to win the Nobel Prize for Economics "for her analysis of economic governance, especially the commons". She analysed real world solutions to the Tragedy of the commons, and the conditions required for such solutions. She identified real word examples of a common administered by a community without overuse of its resources, and without governmental or privatisation solutions, and the commonalities that allowed this to happen.

8 Principles for Managing a Common

- Define clear group boundaries.
- Match rules governing use of common goods to local needs and conditions.
- Ensure that those affected by the rules can participate in modifying the rules.
- Make sure the rule-making rights of community members are respected by outside authorities.
- Develop a system, carried out by community members, for monitoring members' behaviour.
- Use graduated sanctions for rule violators.
- Provide accessible, low-cost means for dispute resolution.
- Build responsibility for governing the common resource in nested tiers from the lowest level up to the entire interconnected system.

These principals could be instituted as a method for managing the Tragedy of the Commons in hospital flow, by giving individual departments responsibility and authority to administer (as a department) their own, predefined, fixed bed base, with predefined penalty for times when they over-ran this bed base.

Clear group boundaries are defined as the department. As the department, its staff and patients falls well within the Dunbar number of social responsibility, it means the flow of all incoming patients will be taken into account as well as the convenience of those being discharged and the issues facing those with obstacles to their ongoing flow.

The rules governing the common, participation by all in the creation, monitoring, enforcement and low cost resolution of those rules, along with responsibility in nested tiers can be met by incorporating discussion of flow and its impact on the department into departmental morbidity and mortality meetings. The graded sanctioning of individual violators can also be discussed in a collegiate and respectful manner as other M+M sanctions are.

The final principal: “Make sure the rule-making rights of community members are respected by outside authorities” needs close examination. Handing each individual department authority and responsibility to administer its own bed base includes their responsibility to place and incoming admissions from ICU or ED, or transfers from other hospitals, and finally elective admissions etc. The hospital should take a very hands off approach to the departments administration of the bed base unless its access block is taking flow resources from other departments – eg access block to ED, ICU.

In particular, it would be very tempting when one department cannot place ED / ICU patients, for the hospital to step in and place them as outliers in empty beds within another department’s bedbase. This should not ever be done. Doing so will automatically and permanently cause the system to fail. If a department is given authority and responsibility for its bedbase, and through good stewardship has less than 100% occupancy, allowing it to manage its own inflow, then taking that bedspace sends the message that they do not have authority or stewardship, and that good behavior will be penalised by confiscation of resources. Any long range plans to manage its flow will be thwarted by the hospital in favour of departments who fail to manage their flow. It will encourage maintaining 100% occupancy in order to prevent theft of its resources, and encourage bad behavior.

That said, it will be up to hospital management to negotiate and enforce penalties for exceeding bed base – as measured by access block to ICU, ED or planned transfers in. This should be negotiated at the same time as the bed base is granted to the department, in concert with all parties, be graded, and be tailored to the individual department. Examples might include enforcing extra ward rounds aimed at discharge, curtailing or cancellation of elective admissions, provision of reports as to what went wrong and plans to mitigate in the future, transfer of care to other teams with ongoing input through consultation only. The hospital may wish to have a few reserve surge beds to manage the (hopefully rare) over-running of bed base to minimize the impact on patients.

The methods the departments use to achieve flow for their granted bedbase will be up to them. They may be advised by the hospital, the executive should seek to help them remove any obstacles to flow and be responsive to requests for aid, especially in any extenuating circumstance, but not try to micromanage how they achieve flow. That empowerment is part of the engagement we seek. In contrast, they should consider appropriate rewards – e.g. in times of hospital-wide theatre shutdown for reasons of fixing access blocks, departments who maintain good flow should be able to keep using their standard theatre times.

Applying a Commons solutions to Whole hospital flow

Each hospital will need to tailor their own variant of a commons solution to their own hospital and to departments within the hospital. There will not be a one-size-fits-all answer, and the following pages should not be seen as a standard blueprint. What follows is what it might look like to have ward teams engage with whole hospital flow by accepting stewardship of their own bed base. Again this is not a detailed prescription – but a picture of what the process might look like:

Preparation

The first stage of such a wide ranging systems change will be to get realistic measures of both the issue and the resources available and needed. Determining what size bed base each subspecialty will need, and what resourcing, in and outpatient services etc are needed are outside the scope of this paper, and will need to be determined as a collaborative discussion with departments and administration.

There exist many ways of measuring access block, delays in management and especially what is causing extended ward length of stays such as Functional Resonance Analysis Method (FRAM), and commissioning such a study may prove necessary, however a simpler version may be enough.

Ask teams to create an anticipated time line (from ED arrival to eventual discharge with all expected time-points for resource use) for the 4-5 bread and butter admission pathologies their subspecialty see (describe work as imagined), then track a few patients of each type through their admission to find how the actual ward flow differs from this idealised path (describe work as done)

This will identify the roadblocks that exist in current systems (and possibly methods to unblock them), give teams interest, buy-in and appreciation for their own nuanced systems of flow, and a sense of ownership (and being heard) during the process. Calculate the bed base of all the subspecialties (likely on a per season basis – see Appendix)

Figure out how to cope with surge management including winter surge and need for seasonal variation in bedbase (especially for respiratory and geriatrics). Ensure there are enough beds in the hospital to cope with handing the bed base to subspecialties. If there are insufficient beds, either magic some up, or relook at elective admissions, or acceptable hospital LOS. Hospital LOS may need to be curtailed through some of the alternate outpatient pathways described earlier in this essay.

Commit

One of the core principals of Elinor Ostrom's effective management of a common is "Make sure the rule-making rights of community members are respected by outside authorities"

In the next few pages we will be describing one example of a process of handing ward teams stewardship of their bed base. The hospital will need to abide by the contract that that team gets to manage the bed base. If the team is doing well, and has some empty beds, it will be tempting to use those empty beds to short term solve flow issues in the rest of the hospital by filling them with access blocked patients. This is guaranteed to make the handing over of stewardship fail. Handing a team their bed base includes their right to keep beds fallow

awaiting future admissions, or to negotiate with other teams to board their excess patients in return for future promise of the same, or to manage other flow resources.

Similarly, once a ward team is handed a bed base, unless pre-negotiated, it is not OK to reduce it because they frequently have empty beds, or because other teams are performing badly and need more beds to compensate for poor flow. This is punishing good behaviour. This is why it is critical to calculate bed base for ALL ward teams in the preparation process to ensure the right bed base is handed to each team. Revision later will be exponentially difficult. Pre-negotiating seasonal variation in bedbase is perfectly reasonable (including graduated increase/decrease over the course of a season if needed).

Finally, there needs to be commitment to a principle of empowerment not prescription. The executive's main role in flow would shift from attempting to manage flow directly to

- Provision of information and resources to ward teams to measure and enable ward flow
- Mediate and troubleshoot issues between teams and with surge management
- Measure and allocate resource needs to teams
- Manage poor ward flow through discussions with AMOs and HODs aimed at problem solving, re-assessment of needs, advice as needed, and in very rare cases where patient safety is threatened or in pre-negotiated actions for breach of bed base, taking over management/curtailing activity as a circuit breaker

The primary flow role of hospital management should be to enable and empower the ward teams to control their own flow, rather than to control it for them.

Enact: An Example of what subspecialty team controlled bed base might look like

A subspecialty surgical team is handed its bed base of 20 beds with the expectation that on average 17 will be in use, and 3 will be fallow awaiting the predictable influx of new admissions from various sources (ED, ICU, elective, interhospital transfer)

They are required to accept into those beds any new patients from any of those sources. If at 10 am on any day there are access blocked patients (ED admitted pts with EDLOS > 8 hours, ICU patients deemed ready for the ward and notified to the team, etc) then this will be escalated to HOD or their nominated representative (such as the on call AMO on weekends), and if unresolved by midday, the problem is escalated to hospital management with expectation that team's elective admissions will be curtailed until the access block is resolved. Theatre is not cancelled, but only useable for progressing flow of already admitted patients. Fallow theatre time becomes an emergency list for that team to promote its flow, then for the hospital. Elective admissions for that day may need to be discharged and rescheduled.

Teams are welcome to negotiate boarding of patients within other subspecialty teams bed bases, however for these outliers, further rules exist

They (and boarders in ED) should be seen first on ward rounds.

When a bed is cleared within that team's bed base, it is first given to ICU patients awaiting a ward bed in that team's bed base, then to outliers within other team's bed base, then ED, patients requiring admission from clinics, interhospital transfers, etc and then rooms and elective admissions.

Hospital management will leave the subspecialty with responsibility and authority for stewardship of their bed base. It won't be used by the hospital to meet flow shortfalls by other teams (see Commit above), and while they are managing flow within their bed base, they will not be subject to flow measures applied to the hospital under flow management of the hospital such as theatre cancellation. (i.e. they get to keep operating because they are the good team who is handling their flow.) From time to time, there will be some public health or NSW health exceptions to this hands off approach such as Covid shutdowns, unavoidable ward closures, disasters, etc, and the bed base will still be subject to other hospital wide flow measures such as overcapacity protocols, disaster management, and hospital wide processes such as holiday theatre reductions etc will still apply.

They may wish to engage in a more team based care model, whereby a consultant in discussion with a patient's own AMO can facilitate that patient's progress/discharge without the AMO being present.

Begin with those most likely to succeed

Pilot the program with the subspecialty teams most likely to succeed and most keen to engage. In my hospital, there has been a string of recent theatre shutdowns and corresponding concerns by some surgical subspecialties that despite their good discharge performance, they are unable to admit elective patients or operate. There are some who will be very keen to take control of their bed base in return for the control it will give them.

Pour extra resources (though not extra bed base) into helping them succeed. There will be teething problems. Help them figure these out. The solutions they come up with will likely be able to be adapted and employed by the teams who follow them. Any extra resources distributed at this early stage will be high yield.

Engage the larger middle

Following the success of the pilot subspecialties, extend the process to other subspecialties now keen to engage – especially those in control of specialised beds (Cardiology, neurology and respiratory with their monitored beds, ASU and Bipap beds). Finally extend to all the subspecialties.

Help the stragglers

Some will find the process harder. Geriatrics face a very specific form of discharge block – the need for RACF beds for many of their discharges that are not available as soon as the patient is fit for discharge. Others have a very small bed base (e.g. endocrine) and so one or two patients constitute a significant surge. It is probably best to roll these in with linked subspecialties (in the endocrine example, renal, and they can engage with the flow portion of the renal M+M.)

Again, it is likely that extra resources will need to be allocated to these stragglers, however these will be appropriate, targeted, and high yield – far higher yield than generic untargeted flow resources being vaguely directed at ward flow. It is likely that some teams will need a personalised variation on the stewardship contract to manage their flow, and possibly a disproportionate bed base to handle their specialised circumstance. These management decisions are well within the supervisory, resource allocation and advisory role hospital management might take.

Attempt to anticipate and solve implementation issues

There will be teething problems with any new system, and a large change in whole hospital systems of flow will be no exception. While it will never be possible to pre-anticipate all issues, some can be predicted and discussed pre-emptively during the planning phase.

There will be resistance to change – especially from those who currently benefit most from longer ward length of stays and poor flow, and from those who do not feel that change is needed at all.

During the initial phases of introduction, flow will likely suffer before improving. Making more flow resources available, and introducing new systems when activity is lower will increase the chance of successful introduction.

Some things we try will not work. Blame for not succeeding at first, will not lead to the feeling that teams can innovate and try new things. Any discussion of teething issues will need to be framed carefully and without blame or accusation.

Leadership will be key. Each department will need to designate a leader in flow who will have to become familiar with aspects of flow, and nuances of how it applies to their department's specific situation. They will need to be a liaison between the department, department head, medical demand unit and executive. Especially when systems are being set up they will probably need an FTE fraction assigned for dealing with these issues.

Re-measure, compare systems between teams, cross-pollinate and innovate

As with any quality improvement process, key to achieving sustainable improvement is to complete the cycle by measuring the results of change and giving feedback to the next round of quality improvement. Although what works for one team may not work well for another, allowing comparison of methods and results is likely to give other teams ideas for further improvement, and give executive ideas where best to spend resources to further improve flow.

Anticipating Criticisms

The most common criticism aimed at this plan is "What happens to patients if a team has run out of beds and can't place them?" The first response is that a situation where the patients of one subspecialty team are access blocked is vastly preferable to the situation we currently face where the patients of all the hospital teams are access blocked to the point of ambulance ramping. Whats more the team at the heart of the problem will actually have incentive to fix the problem rather than attribute it to being unfixable and "someone else's problem".

However the situation actually bears further examination. The reason for the access block can only be one of a few things:

There has been a surge of activity overfilling the appropriately prepared 15% of fallow beds kept for ideal occupancy. Surges are short lived and will quickly be able to be brought under control. Appeals from HOD to hospital executive for brief extra resources for unanticipated surge management are not unreasonable.

The team has not been managing flow. Note that the bed base has been created to meet expected demands at 85% bed occupancy. This means that the team has over-admitted elective patients or has excess length of stay 15% above what was negotiated when bed base

was handed over. The team needs to curtail elective admissions and discharge patients taking overlong stays in hospital. In short, the team needs to engage in flow management.

The demand has grown. Through time, emergence of new diseases or treatments, or something else, the overall demand for the teams service has increased. This will be identifiable by increasing numbers of admissions from ED, increasing requests for elective admissions, and maintenance of LOS for bread and butter presentations. In this case, this is a scale of service issue, and will require discussions between HOD and executive making a business case for increasing the size and bed base of the department, along with an even larger increase in other flow resources, as, **increased capacity slows flow**, and so will require proportionately greater amount of non-bed flow resources (medical, nursing and allied health staff, discharge resources, clinics, access to radiology, procedures, theatres and clinics etc) than would be expected to maintain the same LOS.

Appendix – calculating a bed base.

Deciding on a bed base for a subspecialty is outside the scope of this inquiry, and will always be a more complex decision for the hospital executive. Hence it has been placed in an appendix rather than in the body of the text. This is more to illustrate that it is possible and a knowable quantity.

As many illnesses are seasonal, I would recommend calculating a bedbase for each season for each subspecialty. E.g. Respiratory teams will likely need larger bed bases in winter. It is not unreasonable to have a bedbase with planned slow increase/decrease over the course of a season (eg respiratory in autumn to peak in winter)

Have each subspecialty nominate its “bread and butter” most common admissions, up to 2/3 to ¾ of their total admission numbers.

Through consensus of expert opinion, trial patient tracking, average of previous years, or mean of peer hospitals, decide between subspecialty and hospital executive what an acceptable average LOS for each of those illnesses is. For the remaining 1/3 to ¼, use the average LOS of all the remaining unaccounted illnesses based on past experience.

For each of these illnesses (and as a single entity, the catch all rest of demand)

Patient-bed days required for that illness for the season

= Bedbase for that illness x days in the season x ideal occupancy(85%)

= LOS for that illness x number of that type of presentation per that season

So bedbase for each common illness can be calculated as

Bedbase = LOS for illness x No of that type of presentation per that season

---

days in the season x ideal occupancy(85%)

Add these together for each of the common illnesses, and for the left over catch all set for the total seasonal bedbase

This should give a starting point for hospital executive and subspecialty department to discuss what their required bedbase should be for each season.

Again it should be noted that this needs to be tailored to the needs of the hospital, and the quirks of the department – e.g. endocrine with a small inpatient numbers but potential for high surge needs, and largely consultative service, or Aged care's potential to have their discharge access block due to RACF access block outside their control.