Building the Foundation for Sophisticated Capacity Management in Hospitals

Learn:

1. Why current methods of allocating and managing capacity are simply inadequate to meet the growing complexity of the challenge.

2. What mathematical concepts underlie successful capacity management, and what healthcare can learn from the industries who have unlocked them.

3. How health systems are using tools to solve their mathematical challenges, and achieving stunning improvements in their capacity.
As the healthcare industry continues to recover from the impact of COVID-19, a transformation is taking shape with the use of AI and predictive analytics. Hospitals are turning to advanced technologies for fresh approaches to managing constrained healthcare assets like inpatient beds and operating rooms. Growth in patient volume will continue to rise due to the aging demographic and the higher rates of occurrence of chronic disease. Therefore, hospitals must strive to be more efficient with their resources than ever before. This means capacity management must become one of the most important areas of focus for health systems.

Sophisticated capacity management is an enormous operational and financial opportunity for every health system.

There are approximately 5,000 hospitals in the US, and a mid-sized hospital could have as much as $200–300 million in assets, excluding real estate.

This adds up to $1–1.5 trillion in health system assets across the country that are not being utilized as efficiently as possible.

Even a 10% improvement in the utilization of those assets would unlock $100–150 billion worth of value each year — or an average of $20–30 million per hospital per year.

Leaders have become acutely aware that the current methods of allocating and managing capacity are simply inadequate to meet the growing complexity of the challenge. The impact of the pandemic over the last 18 months has only amplified the urgency of adopting sophisticated capacity management methods as health systems were forced to overcome shortages in rapid-fire sequence. Initially, the shortages were in personal protective equipment (PPE), then ventilators, then intensive care unit (ICU) beds, and then regular inpatient beds. Building the foundation for sophisticated capacity management represents a massive opportunity for hospitals and health systems across the country.

Critical assets in hospitals, such as operating rooms (ORs), inpatient beds, and infusion chairs, are expensive, scarce, and currently not well optimized. The scheduling and allocation of these assets to patients, providers, and service lines are essentially the computerized version of the paper-and-pencil allocation methods that were in place decades ago — when there was plenty of capacity and operating efficiencies were not the most important priority for health system executives.

Operating rooms cost approximately $10–15 million per room, yet many ORs are idle during peak hours, while others will inevitably run late into the night, incurring unnecessary overtime costs. Surgeries are the financial engine for most hospitals since they generate half of the revenues and a majority of the earnings. Maximizing their utilization is certainly worth the investment in time and effort that will be required to make it happen in a scalable and sustainable manner.

In hospitals, inpatient beds are routinely at 90–95% occupancy levels — especially in urban centers. The demand for inpatient beds starts building early in the morning on most days while discharges only begin to occur in the afternoon and sometimes get delayed to the next day. This creates a relentless pressure on most hospitals, resulting in patients waiting for long periods of time before being placed in an inpatient unit or being “boarded” in the emergency department or the post-anesthesia care units (PACUs). It also causes hospitals to go on diversion where they cannot accept inbound transfers. For the average-size hospital, the financial impact of lost revenue from diversions and excess bed nights is often in the millions of dollars each year.
The demand for oncology services has been growing for the last several years. This has placed an enormous burden on infusion centers across the country. Infusion centers experience a daily “rush-hour” effect between the hours of 11am–2pm when all of the chairs in the center are occupied and nursing capacity is strained to near breaking point. Improving the utilization profile by smoothing out the peak hours and level-loading chair occupancies can improve the effective capacity of the infusion center by 10–15% while simultaneously reducing the wait time experienced by patients by 30–40% during peak hours.

These observations are consistent across many of the other expensive, constrained assets found through the health system, including imaging machines, radiation oncology, pharmacies, labs, emergency departments, specialty clinics, etc.

Sophisticated asset utilization methods that unlock capacity will deliver numerous important benefits to a constrained health system by:

1. Enabling more patients to be treated each day
2. Reducing wait times experienced by patients
3. Improving unit operating costs of the facility
4. Reducing capital expenditure needed to add new capacity
5. Improving staff satisfaction by eliminating the peaks and valleys in the workload
Solving the capacity management problem is complex and it cannot be solved by the EHR

Solving the capacity management problem in healthcare requires an appreciation of two core concepts:

1. **Matching**
   - The demand for each service with the supply of capacity (people, equipment, rooms) needed to provide that service.
   - Each service smoothly with the upstream and downstream services that are necessary in order to provide the appropriate care for each patient.

2. **Supply**
   - Availability of the right staff member with the right skills
   - Availability of the necessary equipment (robot, pump, imaging machine, etc.)
   - Availability of the right type of room (OR, procedure, examination room, etc.)

3. **Demand**
   - Number of incoming patients seeking treatment
   - Type of treatment needed by each patient
   - Arrival time (including delays) of incoming patients
   - Expected length of treatment for each patient

Meanwhile, the supply of capacity is interconnected, constrained, and hard to predict. To deliver any healthcare service, the equipment, the staff, and the facilities must be simultaneously available at that exact moment in time. There are a finite number of rooms, staff, and pieces of equipment. Finally, staff members can call in sick, equipment can break down, and rooms can be unavailable at short notice. Hence, the supply signal is also highly uncertain.
It takes sophisticated algorithms to dynamically find the equilibrium between demand and supply signals every day, throughout the day — particularly since they are both highly uncertain signals. First, the demand signal must be accurately predicted in terms of the number of patients, the type of treatment they require, the duration of the treatment, and the probability they will be early or late. Next, the availability of the resources needed for a specific service must be determined using constraint-based optimization algorithms.

Health systems have largely ignored the complexity of the problem of matching the supply and demand signals throughout the day. Instead, they have relied on simple calendaring methods or manual allocation techniques. These are simply inadequate. As a result, patients and staff encounter bottlenecks at various points in the day for virtually every service process. The cycle repeats every day without fail.

### Linking

Each patient experiences a sequence of individual services as part of a single encounter with the health system. For example, they may get their blood drawn in the lab, see their provider, and then get a procedure done. Alternatively, they may enter the system through the emergency department, undergo a series of diagnostic tests, and then get admitted into an inpatient unit for treatment.

Each of these independent services is like a node in an interconnected network. Each patient travels through a unique set of nodes to receive their required treatment. Optimizing the flow across an interconnected network of nodes involves the science of topological network analysis, Directed Acyclic Graph (DAG) theory, and other esoteric mathematical methods. Again, health systems have largely ignored this complexity. For the most part, patient itineraries are created by “inspection” — the scheduler simply schedules the individual services as if they were totally independent of each other and prints out the schedule for the patient. As a result of ignoring the connected interdependency between the individual services, patients are often stuck in waiting rooms between the various services that must be completed as part of the same encounter.

### The EHR cannot solve this problem

The EHR was built as a repository of every patient encounter — it does an excellent job of maintaining the database and providing all of the necessary patient data and billing information required by the health system.

EHRs allocate resources on a first come, first served basis. They do not provide any of the sophisticated mathematical methods like prediction, simulation, optimization, AI, or machine learning that are required to dynamically address the supply-demand balance throughout each day. As a result, they are incapable of providing the kind of prescriptive guidance that is required for the frontline to make dozens of decisions each day based on a nuanced understanding of the ever changing dynamics of the supply-demand balance.
Proven approaches to solving the capacity management problem exist all around us

Package delivery companies like FedEx and UPS have encountered the problems of matching and linking for decades. They have built sophisticated models and operating practices that allow them to thrive despite the inherent complexity of their business models. Specifically, it is impossible for them to predict, on any given day, the millions of people all across the country who will decide on that very day to ship a package from Point A to Point B. Meanwhile, their “supply” of resources — aircraft, trucks, trailers, vans — are constrained and captive within a limited geographical radius. They have to build robust demand forecasting models and design their network of hubs to be flexible and resilient to a volatile demand signal. Once a package delivery van has been loaded, the driver must make as many as 100 stops during the day in order to drop off and pick up packages. The route they follow is an optimized traversal sequence across the stops they make that must be dynamically rebalanced based on the specific addresses of the packages that happen to be on the van that day. Yet, they are able to fulfill their delivery time obligation 99.9% of the time day after day.

Ride-sharing services like Uber and Lyft cannot be certain of the demand on any given day since people decide on the spur of the moment that they would like a ride from Point A to Point B. Yet, they must be able to provide the car and the driver to any street corner of any major city in the world at any time of the day within a short time interval (e.g., 10 minutes). They cannot simply look at a calendar and the available roster of drivers on duty in order to make assignments. Instead, they build sophisticated demand forecasting models based on time of day, day of week, seasonality, weather, zip code, events, etc. and then marry it with a deep understanding of the driving patterns of their drivers (who aren’t even employees who can be forced to work at particular times). Importantly, when the supply-demand equilibrium is off balance, they actively intervene in real-time by offering incentives to bring more drivers on the roads, surge pricing to defer demand by 30–45 minutes, or issue automated text messages to reposition drivers closer to where the current demand is located. Major airlines face a similar challenge in balancing supply and demand at each of their major hubs. They have to execute the turnaround of hundreds of aircraft upon their arrival at a gate and push it back out for another flight within 45 minutes. They do this safely and correctly thousands of times each day (Delta Airlines has approximately 5,000 daily flights). Turning around an aircraft requires stringing together dozens of independent services (baggage handling, cabin cleaning, crew changeover, passenger boarding, catering, etc.). They have built sophisticated models to match the demand for each service (e.g., baggage handling) with the availability of the supply, which includes equipment (tows, tugs, ramps, etc.) and personnel within each section of the terminal. This matching has to take place in a tightly defined zone since it’s difficult to reposition equipment across the vast expanse of a major airport, which can easily be 5–6 miles from one end to the other. Furthermore, they have mastered linking services in the right order to maximize efficiency. For example, the cabin can only be cleaned after incoming passengers have left the plane but it must be cleaned before new passengers board the plane.

The success of this approach has been remarkable. The number of “aircraft movements” (takeoff or landing) at a major airport over the last 20 years has increased by a factor of 5–10x even though the number of runways has largely remained the same since it takes 25–30 years to get a new runway operational in a major city like Atlanta or New York.
Applying these approaches to healthcare

The approaches described above can readily be applied to healthcare. The trick is to focus on one asset at a time and understand the nuance of the supply-demand dynamics at a very detailed level to build the right optimization algorithms that can generate prescriptive insights on a real-time basis. Consider focusing on these three specific assets: operating rooms, inpatient beds, and infusion chairs as a way of getting started.

Operating rooms

As especially high-value assets, ORs must be staffed and scheduled appropriately to ensure that they are being utilized to their full potential. The demand for OR time depends on many factors, such as the number of patients seen in the clinic over the prior weeks, the percent of those patients that needed surgery, the subset of those patients who could accommodate a specific day for getting their surgery done, and the length of their case in comparison to other cases that have already been booked by that surgeon on that day. Each of those factors in isolation is hard to estimate — predicting the combination of all of those factors is virtually impossible. On the other hand, the supply of OR time is allocated to surgeons and service lines as fixed “blocks” — e.g., all day on Mondays and Wednesdays. A fixed and static allocation of blocks could never be expected to match a volatile and unpredictable demand. This is why there is so much turmoil in the scheduling of operating rooms every single day. Approximately 30% of the cases on any given day are being performed by someone other than the original owner of the block that was scheduled for that day. A large number of add-on and emergent cases are not really add-on cases, they’re just squeezed into the schedule to take advantage of open time in the OR.

There is a better way

Predictive algorithms can proactively anticipate the surgeons and service lines that are unlikely to fully utilize a block. This knowledge can make blocks available to other surgeons who might be able to better use the operating room through a “liquid marketplace” that facilitates the easy exchange of OR time between surgeons. Mining the historical patterns of the actual usage of block time by surgeons and service lines can unearth patterns in contiguous time left unused, total abandonment of blocks, or a consistent pattern of repeatedly releasing a specific block. Intelligent algorithms can prescribe specific opportunities to “collect” and reallocate these blocks without adversely impacting the practices of the surgeon or service line whose blocks are being repurposed.

These methods have delivered an enormous impact at over 40 major health systems spanning 200 facilities and over 2,000 operating rooms, as illustrated below.
A large healthcare network of 11 surgical centers that performs 164,000 surgeries each year in 139 ORs across 8 states, needed to improve OR access for its 1,400 surgeons. Almost half of all surgeons were independent, and many of those were “splitters” who could perform cases at any health system. Novant needed to retain these “splitters” to maintain a strong capacity during the onset of COVID-19, as thousands of surgeries needed to be rescheduled.

iQueue for ORs helped create a streamlined, transparent platform, where surgeons could easily view available OR time, trust the accuracy of the information, and confidently and proactively access the rooms they needed. From May 2020 to 2021 — throughout the height of COVID-19 surges — Novant saw these results:

- 8% Increase in Number of Cases
- 12.7% Increase in volume from case surgeon splitters
- 8% Increase in OR minutes
- 5–6% Increase in Prime Time and Staffed Room Utilization respectively

Rush wanted to recruit new surgeons, but 100% of its OR time was blocked. As a small academic system, Rush had little available time to offer new hires and could not successfully take on additional surgeons. At the same time, they were looking to improve the experience of the surgical staff. In order to make a change, they needed to have a deep analysis of their utilization data. With tools like iQueue for OR to help surgeons and leadership see actual day-to-day time usage and room availability, plus make future predictions, Rush surgeons were empowered to:

- Make informed choices and optimize individual utilization, leading to better capacity management overall
- Release block time earlier in advance, allowing more operations to be scheduled
- Decrease unused time blocks overall by 34%
Inpatient beds

Most hospitals — especially in urban centers — are extremely tight on inpatient bed capacity. It’s not unusual to see occupancy levels in the 90–95% range (even exceeding 100% for several hours during the day) on a consistent basis. As with most assets in healthcare, the demand pattern for inpatient beds is volatile and hard to predict. Meanwhile, the availability of beds depends on the pace at which patients can be discharged. It’s hard to increase the supply of beds in the short-term since beds need to be licensed, and more importantly staffed, in order to be available for patients.

A particularly challenging aspect for inpatient bed units is that the arrival pattern for a bed starts to ramp up early in the morning as patients come out of surgery and often need to be placed in an inpatient unit to recover. Meanwhile, the departure pattern of discharges only picks up steam in the afternoon. This leads to a consistent window of several hours each day where shortages are acutely felt — resulting in patients being “boarded” in the ED or the PACU or, even worse, forcing the hospital to go on diversion and decline inbound transfers from external sources.

Most capacity management teams are forced to rely on simplistic dashboards, Excel spreadsheets, and manual census counting and reporting that is aggregated from all of the units. Again, there is a better way. It’s possible to build accurate predictive models for each hour of the day for each unit in the hospital based on historical patterns of admits and discharges coupled with real-time intelligence for what is actually happening at each moment throughout the day. Armed with an accurate prediction, staff can focus on accelerating the discharge of patients from specific units or making intelligent choices regarding the placement of incoming patients to a bed in a particular unit.

These methods are demonstrating powerful results in 4,000 beds at 15 different hospitals. One example is illustrated below.

CASE STUDY

An integrated healthcare delivery system serving Colorado, southern Wyoming, and areas of Nebraska, UCHealth ran lengthy morning bed meetings off a manual spreadsheet, during which nursing and medical leadership reported reactively on fast-changing existent data. With iQueue for Inpatient Beds, they’ve seen impressive results and now spend much less time on reporting out and reviewing data, while making proactive, effective decisions.

37% reduction in time to complete ICU transfers
(Drastically increasing total number of requests and overall share volume)

8% decrease in Opportunity Days
(difference between Med/Surg Length of Stay & CMS Length of Stay)

4% decrease in time to admit
(despite 18% increase in COVID-19 census)

90% improved confidence in critical capacity decisions
(compared to only 50% previously)
Infusion centers

Virtually every infusion center in the country experiences a midday “peak” between the hours of 11am–2pm when every chair is occupied and several patients are camped out waiting to begin their infusion. Just as rush hour traffic adversely impacts the commute for everybody, the peak hours at an infusion center impact both patients (who are forced to wait to begin treatment) and nurses (who are forced to scramble to try and keep things on track).

The demand signal for infusion on any given weekday depends on the oncologists who are at their clinics that day, the number of patients they see, the percent of patients who need infusion, the mix of treatment lengths, and patient arrival patterns. Meanwhile, in order to deliver an infusion treatment, the chair, the nurse, the pump, and the drugs must all be ready and available for the infusion to start. As in prior examples, attempting to match a volatile, unpredictable demand signal with a constrained, interconnected, volatile supply availability is a hard optimization problem to solve. Yet, most cancer centers attempt to schedule infusion treatments by simply looking at the calendar and making a determination. This is a direct result of using a simple EHR grid-based calendar to schedule a service that is inherently too volatile to be scheduled in such a manner.

There is a better way of optimizing infusion scheduling to create a level-loaded schedule that ramps up smoothly, keeps chair occupancy nearly flat throughout the day, and then ramps down smoothly. First, it requires building a sophisticated demand forecasting model for each day of the week to accurately predict the number of patient treatments and the mix of treatment lengths. Next, it requires understanding the long list of supply constraints, which include operational hours, number of chairs, configuration of chairs into pods, pharmacy operations, nurse allocation approaches (e.g., primary nursing), nurse workloads, etc. Balancing the demand and supply signals throughout the day requires the use of constraint-based optimization algorithms that sift through the trillions and trillions of possible template configurations and offer up a scheduling template that is unique for each hour of each day. Once these templates are baked into the EHR scheduling program, schedulers can confidently offer patients an appointment slot knowing that they are contributing toward creating the flattest possible chair occupancy profile for that given day despite the inherent uncertainties of the demand-supply balance.

These methods have delivered a 10–15% increase in treatment capacity while lowering the wait time experienced by patients during the peak hours by 30–40%. iQueue for Infusion Centers has been deployed at 400 infusion centers that collectively operate 8,500 chairs belonging to over 90 different health systems. One case example is The University of Kansas Cancer Center.

CASE STUDY

The University of Kansas Cancer Center

UKCC comprises 13 locations across Missouri and Kansas. At those locations, 12,000 unique patients are seen each year making up over 100,000 treatment appointments. The team at UKCC turned to iQueue for Infusion Centers after experiencing common pain points, including volume peaks from 10am–2pm, patients stressed by delays, nurses missing lunches, and physicians frustrated with add-on variability. After implementing iQueue, UKCC can make operational improvements and create better outcomes for its patients and staff, including proactive schedule management.

UKCC results

- 32% lower average chair wait time
- 16% lower average drug wait time during peak hours
- 19% increase in volumes
- 17% increase in patient hours

Additionally, UKCC improved clinician satisfaction since nurses can now take lunches, offered more time for leadership to support the frontline, improved patient and staff safety measures, and increased overall capacity.
Applying these sophisticated methods have delivered massive improvements in operational performance in several other industries for decades, and have the potential to unlock enormous value for health systems. In addition to improving the patient experience, sophisticated capacity management can reduce staff burnout while adding tens of millions of dollars in annual improvement to the bottom line.

The ultimate vision is a world where every asset within the health system is optimized — from ORs, infusion centers, and inpatient units to emergency departments, pharmacies, labs, imaging, and more. In this scenario, every leader in every care setting has the tools they need to address complex problems and thus increase the capacity of all assets across the entire system.

This future state cannot be achieved through spending and building alone, nor solely through process improvement initiatives. It requires a concerted effort to fully optimize the assets that already exist. To approach the level of efficiency already demonstrated in asset-reliant industries like transportation and logistics, health systems must apply the same sophisticated level of math to their daily operational workflows. They must offer tools that calculate and solve the capacity management problems their leaders, providers, and staff already engage with daily, helping them strike the ideal balance between under- and overutilization across the entire system. Supporting thousands of decisions made by the frontline each day with accurate, prescriptive recommendations will eventually lead to organizational transformation and a fuller, more efficient use of related assets.

Now is the time for leaders to embrace a fundamental shift in capacity management as critical to the success of their healthcare organizations, to achieve better healthcare across the spectrum.