Nuvern International

Optimizing Healthcare Resource Allocation for Operational Efficiency and Cost Reduction Using Real-Time Analytics

Jeshwanth Reddy Machireddy[1]

[1]**Independent Researcher**

ABSTRACT

Healthcare systems are constantly struggling to balance managing limited resources to achieve high levels of operational effectiveness and controlling costs. This research proposes a theoretical model for real-time optimization of resource allocation across a network of regional health facilities using real-time analytics. It fills the gap between rigid resource planning and the need for agile, data-driven reactions and provides a systematic approach for regional resource coordination. The model integrates hospital operations' real-time data streams into a dynamic decision space. A mathematical formulation based on optimization is built to assign critical resources – medical staff, hospital beds, and critical equipment – in line with instantaneous demand and projected short-term need. The objective is to minimize overall operating costs while satisfying service demands and reacting to fluctuations in patient volume and acuity. The proposed framework is supported by a real-time analytics architecture that captures and processes real-time data on patient flow, resource usage, and system constraints. An optimization engine uses this data to update allocation decisions in real-time, enabling an adaptive and responsive approach to managing resources. The approach is notionally modest, with emphasis placed on structural integration of analytics and optimisation rather than exaggerated claims of performance. Potential uses are outlined to illustrate how the model can enhance efficiency, reduce waiting times, and restrict wasteful expenditure in theory. As a purely theoretical piece of work, this research highlights the opportunity and the limits of real-time data-driven optimization to control healthcare resources.

Keywords: analytics integration, healthcare resource optimization, hospital operations, real-time data, regional coordination, resource allocation, theoretical modeling

1 INTRODUCTION

Healthcare providers and administrators continually strive to balance limited resources with growing and random patient demands [1]. Operational efficiency in healthcare implies the effective utilization of resources such as hospital beds, medical staff, equipment, and supplies for delivering timely and quality care. Efficiency at a high level is of paramount significance in a period of constrained budgets and rising costs of care. At the same time, healthcare systems must maintain or improve quality of service so the patients receive appropriate care without unreasonable waits. Decisions on resource allocation – from staffing and bed assignments to equipment distribution – directly impact both operating costs and patient outcomes. Poor allocation can lead to such situations as under-staffed wards that risk compromising patient care, or, alternatively, over-provisioning that involves unnecessary operational costs.

Most of the healthcare allocation decisions are taken based on periodic planning and heuristics. For example, nurse staffing within hospitals is planned based on expected average patient volume, or the allocation of fixed numbers of beds to different departments based on historical trends. These traditional approaches, while intuitive, may be sluggish in adapting to real-time variability. Sudden surges in patient arrivals, unexpected equipment failures, or other types of disruptions can render a fixed plan obsolete. When such events occur, managers must make rapid adjustments, typically based on experience and manual reallocating. This ad hoc process can be variable and can fail to find truly optimal reallocations under pressure. Furthermore, localized decision-making at a single hospital or departmental level can fail to consider the broader regional context, where the overflow at one site could be met by the capacity at another site within the network. [2]

The development of health information technology has brought with it the possibility of more dynamic, data-driven management of resources. As electronic health records, real-time patient monitoring systems, and Internet-of-Things devices in hospitals become ubiquitous, there is now a

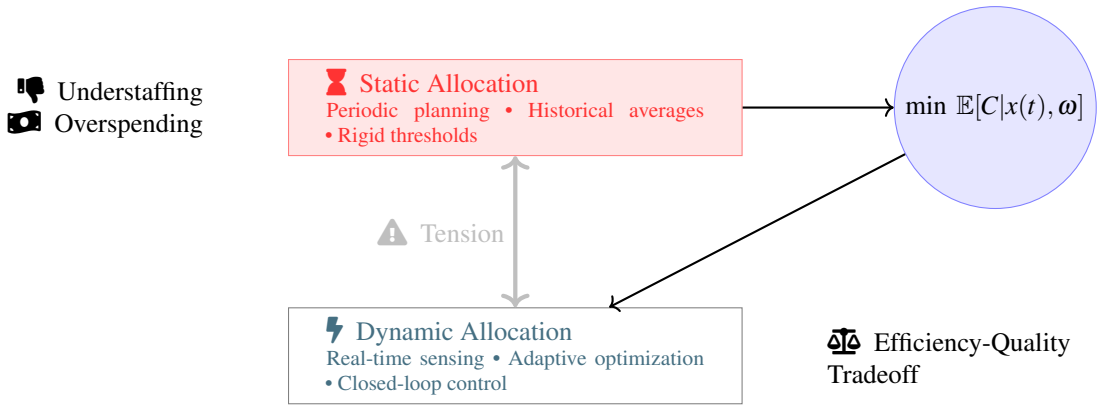

Figure 1. Duality of Healthcare Resource Allocation Paradigms

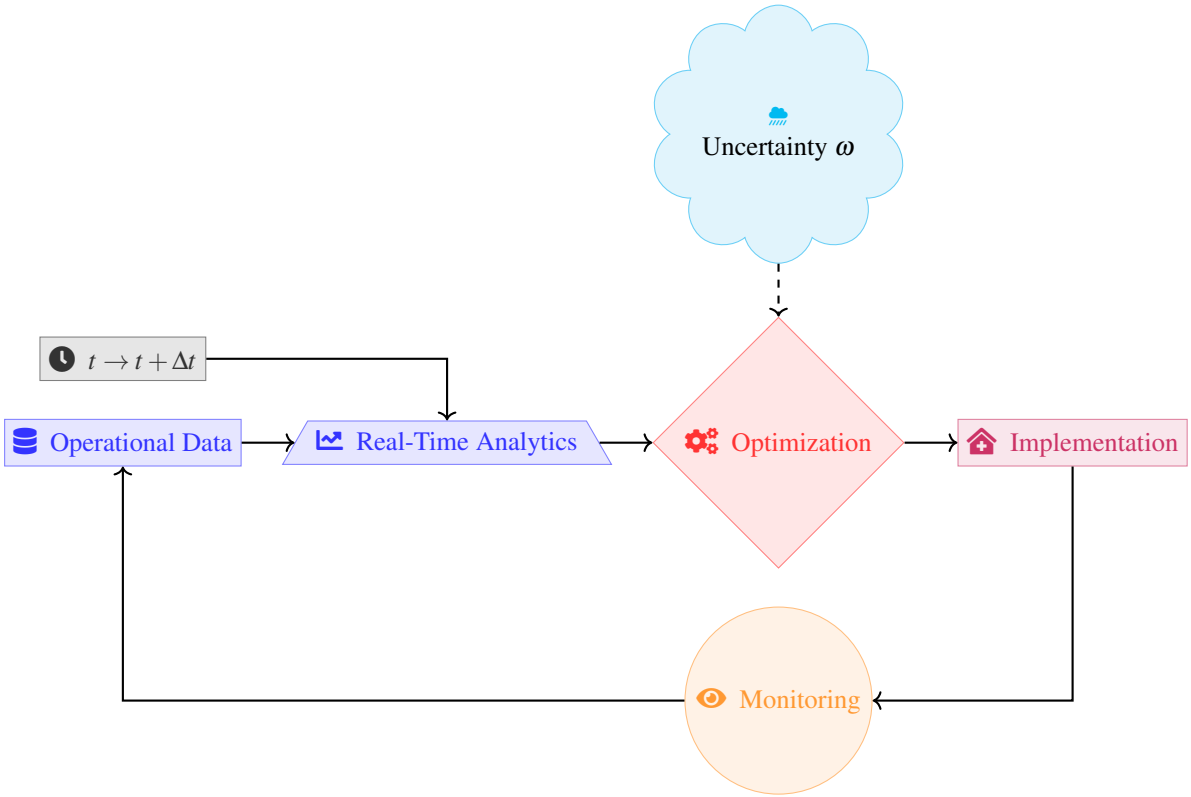

Figure 2. Closed-Loop Resource Optimization Ecosystem

Table 1. Examples of Healthcare Resource Allocation Decisions

Resource Type	Allocation Decision	Traditional Method	Potential Issues	Dynamic Alternative
Nursing Staff	Number per shift	Based on average demand	Under/overstaffing	Real-time workload adjustment
Hospital Beds	Departmental distribution	Historical patient trends	Misaligned capacity	Adaptive interdepartmental shifts
Equipment	Location assignment	Fixed room assignments	Idle or overused equipment	Sensor-based reallocation
Medical Supplies	Replenishment schedule	Periodic stock checks	Stockouts or surplus	IoT-driven auto-reordering
Specialists	On-call scheduling	Weekly plans	Delays or idle time	Demand-sensitive scheduling

tremendous amount of operational data being generated at all times. Patient admissions and discharges, current bed census, waiting room lengths, and even something like average treatment times can all be tracked in real time. Similarly, availability of supplies and personnel can be tracked across a network of facilities. If tapped effectively, this flow of data could enable health care systems to shift from pre-planned static resource allocation to a more responsive

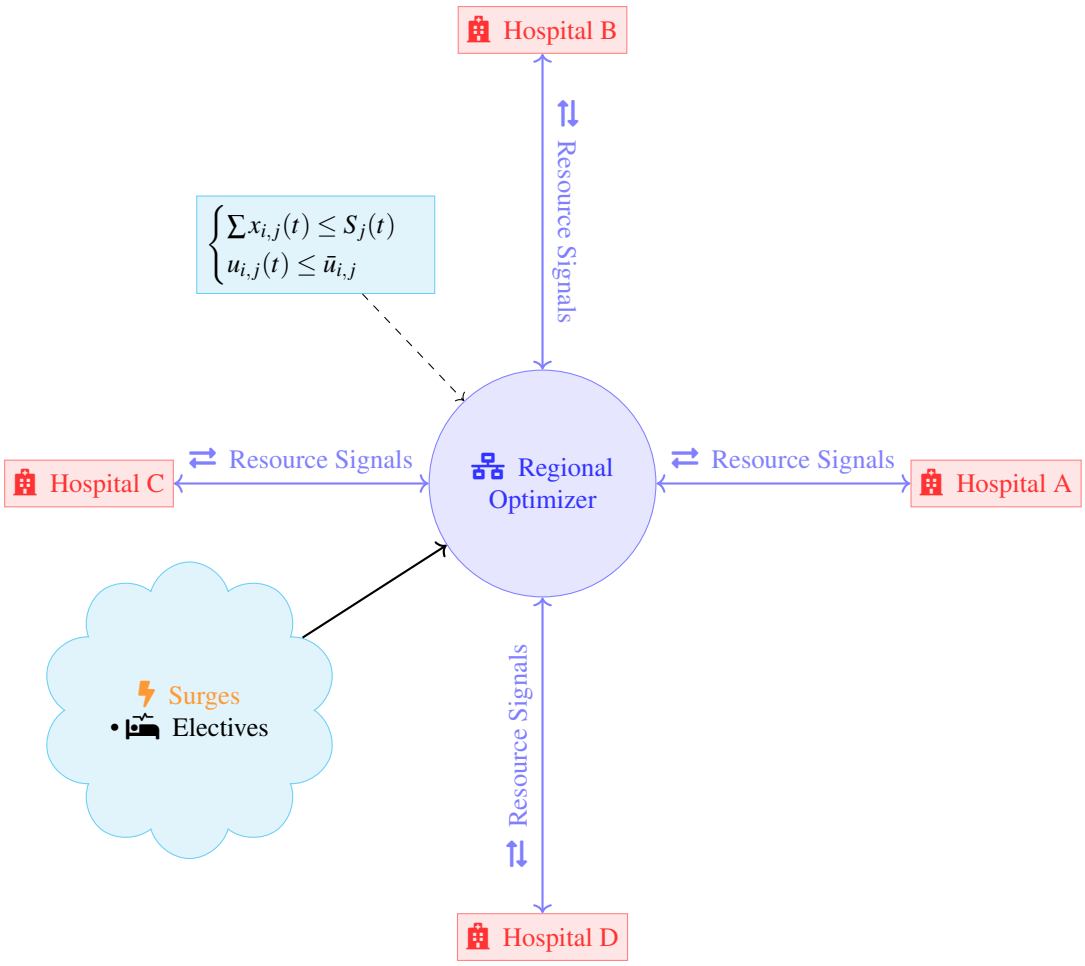

Figure 3. Networked Resource Allocation Under Stochastic Demand

operating mode in which decisions adapt as circumstances change.

Real-time analytics refers to the analysis of data as they are being received to gain instant insights and make timely decisions. In a healthcare operations setting, real-time analytics can be the nervous system of a resource management system, constantly sensing the state of the system and alerting nascent problems or opportunities. By integrating real-time analytics with an optimization model, there is the potential for creating a feedback loop: information on current conditions nourishes the model, the model suggests resource assignment changes, and the changes are made and then monitored. This type of closed-loop process would be a significant departure from traditional open-loop planning whereby decisions are made and only evaluated after the fact.

This article presents a theoretical framework that would seek to streamline healthcare resource planning at a regional level through real-time analysis. The framework would work to enhance operational effectiveness—by more closely aligning resources with patient demand as it arises—and to lower costs by preventing overstaffing or underuse of costly assets. Significantly, the approach is formulated to be adaptive and flexible rather than prescriptive; it provides a structured decision-making tool that can inform administrators, but it does not claim to substitute for human judgment or management [3]. The contributions of the work are conceptual, the authors propose a framework that blends continuous data monitoring and mathematical optimization methods within healthcare operations.

The remainder of this paper is organized as follows. Section 2 covers background concepts and the context of healthcare resource allocation problems and opportunities for real-time data exploitation. Section 3 presents the formal model formulation, decision variables, constraints, and objective function that define the optimization problem. Section 4 describes the real-time analytics architecture supporting the model, detailing how the data is collected and processed in real time. Section 5 describes the optimization logic and solution approach to deploying the model in a live setting. Section 6 provides a discussion on potential theoretical applications of the model, illustrating how it can be applied in various scenarios and what benefits can be achieved. Section 7 outlines this theoretical approach's

assumptions and limitations. Finally, Section 8 concludes the paper with a summary of findings and reflections on future research directions.

2 BACKGROUND

Effective allocation of healthcare resources has been a subject of interest to operational management of hospitals and health systems for many years. Conventionally, the allocation of resources is conducted according to a combination of statistical forecasting, heuristic guidelines, and administrative experience. For example, hospital managers use historical data of patient admissions to predict staffing needs for a given month or season. Regulatory policy and budgets also influence such plans – limiting the number of personnel that can be rostered or the number of beds that can be operated. The allocation plans so produced are typically static over the planning horizon (e.g., fixed staffing rosters or pre-defined bed allocations by department). While such approaches can be calibrated to represent average conditions, they typically do not possess the responsiveness to respond to real-time variability. Classical resource planning essentially offers a required baseline but is not necessarily the best recipe for success when dealing with the underlying uncertainty of healthcare demand.

Numerous techniques have been given by operations research and management science to enhance resource planning in healthcare facilities. Mathematical models have been developed for issues like nurse scheduling, operating room scheduling, ambulance location planning, and medical supply inventory control. Linear programming, integer programming, and simulation solutions have been shown to reduce costs and improve service levels in some applications. Many of these models, however, assume input parameters that are stationary or updated infrequently [4]. For instance, a scheduling model might assume a stationary forecast of patient volume for the day, or an inventory model might use average consumption rates. In reality, patient arrivals can unexpectedly surge (e.g., due to accidents or disease outbreaks), and resource availability can change (staff can call in sick, equipment can break down). Static assumption-based models can become inferior as soon as actual circumstances differ from expectations.

Emerging technologies are gradually enabling more dynamic approaches to resource management. Real-time data streams from healthcare operations are becoming ever more available through the information technology infrastructure of modern hospitals. Emergency departments often have real-time feeds of key metrics such as waiting room patients and treatment times. Hospital-wide dashboards can display current bed utilization by unit, and staffing software can show which nurses or physicians are signed on at any particular time. Also, regional health systems with multiple facilities can share certain information centrally – for example, a regional command center might monitor ambulance status, intensive care bed availability, or operating room schedules across the network in real time. All of this information is an opportunity for data-driven decision support systems that update their recommendations in real time as new information arises.

Real-time analytics in healthcare operations is the analysis and processing of these continuous streams of data to produce actionable insights in real time. Hospitals have started utilizing real-time analytics primarily for monitoring over the last few years. A few examples include patient flow dashboards that trigger alerts to managers when wait times exceed thresholds, or predictive algorithms that warn of an imminent bed shortage based on the current admission rates [5]. Such systems fall short of decision-making or action. The leap from insight to decision usually relies on human judgment: the manager gets the alert and then, say, decides to call in extra staff or open an overflow ward. While human expertise is vital, especially in complex and sensitive environments like healthcare, there is a interest in more automated or systematically optimized decision-making processes. The idea is not to substitute human decision-makers, but to furnish them with model-based suggestions considering a larger number of alternatives and information than any human could conceivably combine in a moment of crisis.

Thus, there is a gap between what can be done with real-time monitoring and the application of formal optimization methods in the decision instant. Bridging this gap entails combining ideas from the two disciplines: continuous data analysis in order to keep the system knowledge up to date, and optimization techniques to compute good decisions based on the current state (and perhaps anticipated future states). This is conceptually the same as what is performed in other industries by automated control systems or real-time supply chain management, though healthcare presents unique challenges. The system is stochastic and multi-dimensional – many kinds of constraints and resources, and decision consequences can be life-critical. Further, any automated recommendation system in healthcare must be transparent and tunable since clinicians and administrators must trust and check the reasonableness of the resource shifts.

In a regional healthcare system, the complexity is even more. Each hospital or clinic in the network will likely have its own priorities and constraints, yet the allocation of resources can often be optimized by an approach that looks across the system. For instance, if a hospital is confronted with a sudden surge of emergency patients, a regional view might suggest diverting some ambulances to other facilities in the area that have capacity, or redirecting roving available personnel or equipment from one facility to another in an effort to cope with the surge [6]. In the absence of a coordinated policy, each hospital can only operate within its remit – with the potential result of one hospital being overwhelmed while another has available capacity. One model spanning multiple facilities can, in theory, calculate

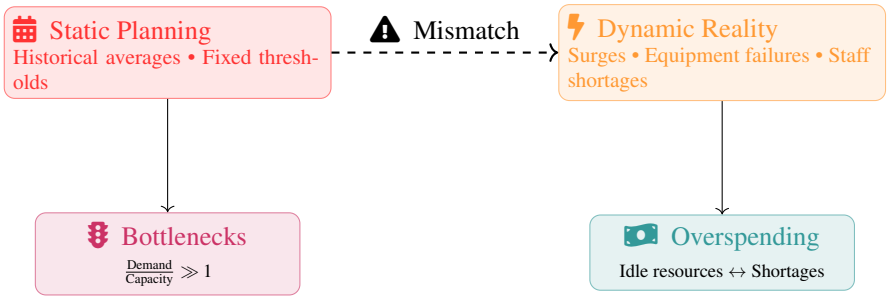

Figure 4. Static Planning vs. Dynamic Reality Mismatch

Table 2. Comparison of Traditional vs. Real-Time Healthcare Resource Allocation

Aspect	Traditional Approach	Real-Time/Modern Approach
Data Source	Historical records, periodic forecasts	Live operational data streams
Update Frequency	Monthly or seasonal	Continuous / real-time
Responsiveness	Limited to anticipated conditions	Adaptive to current system status
Decision Support	Heuristics and manual oversight	Algorithmic recommendations and alerts
Scalability	Department- or hospital-level	Network-/system-level integration

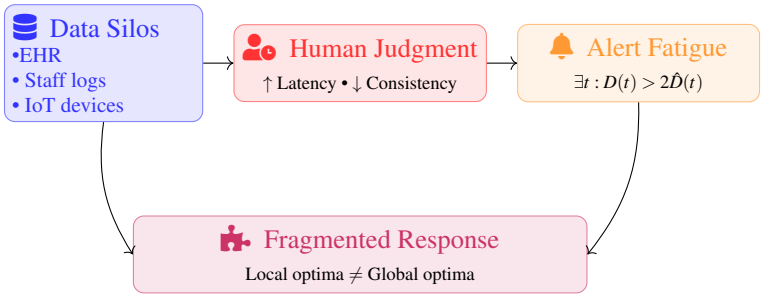

Figure 5. Human-Centric Decision-Making issues

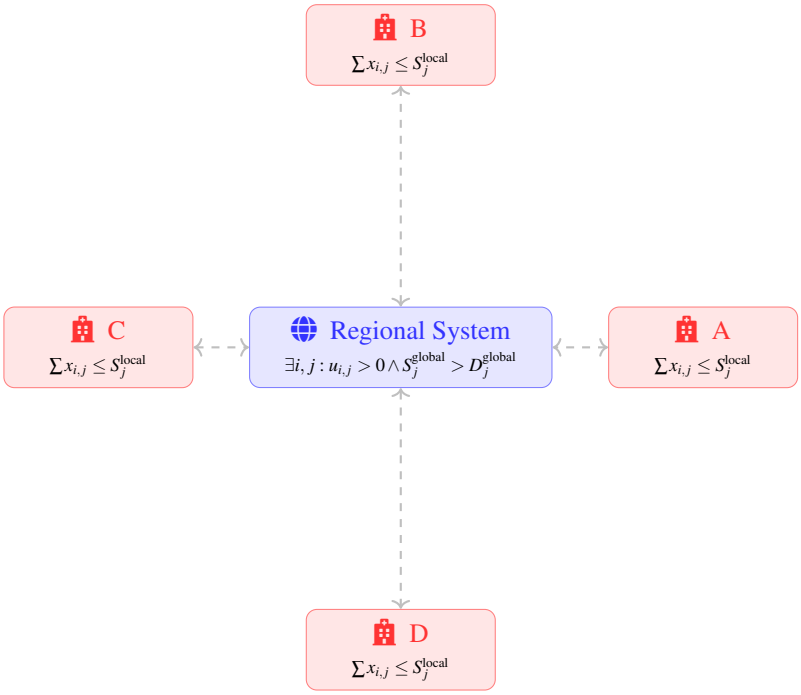

Figure 6. Fragmented Resource Allocation in Regional Networks

allocations that bring more effective use of the total regional resources, balancing load and sharing resources where appropriate. Yet to achieve this in practice, one requires not only data from each facility, but also a decision engine that can handle the scale and urgency of multi-facility optimization.

The research context for this research, therefore, is the confluence of two trends: increasing availability of real-time operational data in healthcare, and development of algorithms and computing power that can quickly solve optimization problems of complexity. The model suggested herein takes advantage of these advances by speculating on how they may be combined. Unlike retrospective review or re-planning on some periodic schedule, this model suggests that data and decisions cycle constantly. The issue addressed is not in an individual resource allocation problem (e.g., single nurse staffing or single inventory), but rather in creating a generalizable framework for the general resource allocation problem at the regional system level. This scope is intentional – it is meant to outline a high-level architecture that could be instantiated in many different particular use cases, all held together by the common theme of employing real-time data to make operations decisions.

3 MODEL FORMULATION

The foundation of the suggested framework is a decision-making mathematical program at discrete decision epochs

Table 3. Key Variables in the Resource Allocation Model

Symbol	Description	Units / Type
$D_{i,j}(t)$	Demand for resource j at facility i at time t	Required quantity (input)
$x_{i,j}(t)$	Amount of resource j allocated to facility i at time t	Decision variable (continuous or integer)
$u_{i,j}(t)$	Unmet demand for resource j at facility i at time t	Slack variable (non-negative)
$S_j(t)$	Total available supply of resource j in the system at time t	Resource pool (input parameter)
C_j	Cost per unit of using resource j per interval	Cost coefficient
P_j	Penalty per unit of unmet demand for resource j	Penalty coefficient

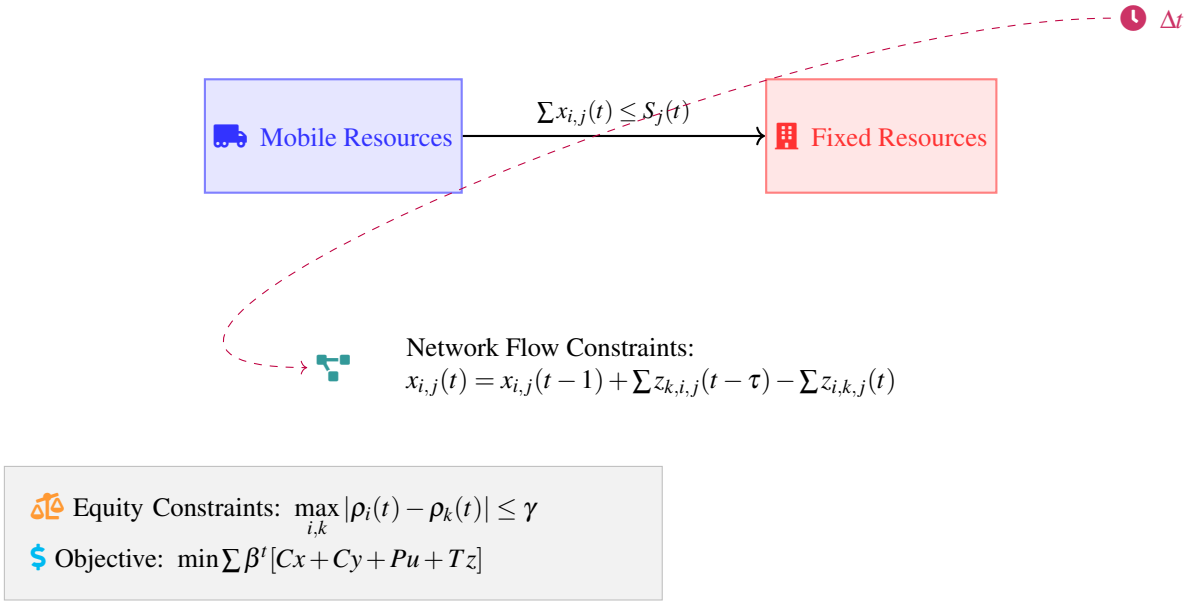

Figure 7. Network Flow Architecture with Temporal Coupling

Table 4. Summary of Optimization Constraints

Constraint	Formulation	Interpretation
Demand coverage	$x_{i,j}(t) + u_{i,j}(t) \geq D_{i,j}(t)$	Total allocation + slack must meet or exceed demand
Resource availability	$\sum_i x_{i,j}(t) \leq S_j(t)$	Total assigned resources cannot exceed system availability
Non-negativity	$x_{i,j}(t) \geq 0,\ u_{i,j}(t) \geq 0$	No negative allocations or unmet demand allowed

for the allocation of heterogeneous resources across a geographically dispersed healthcare network. Although the previous discussion assumed an hourly rolling horizon with linear cost and single-interval optimization, field deployment often demands a richer formulation that can capture multi-period dynamics, indivisibility constraints, equity requirements, and robustness to forecast error. This longer chapter therefore outlines the decision variables, constraints, and objective terms at more detail, explains temporal coupling among intervals, and outlines canonical extensions—e.g., integer constraints, fairness metrics, and probabilistic guarantees—that preserve tractability while enhancing realism.

At the highest level, let the planning horizon be discretized into a finite set $\mathscr{T} = \{t_0, t_1, \ldots, t_H\}$ of decision epochs, each of duration Δt (for example, fifteen minutes or one hour). The healthcare network comprises a set $\mathscr{I}$ of facilities and a set $\mathscr{J}$ of resource categories. For each resource $j \in \mathscr{J}$, the system distinguishes between a pool of mobile units—which can in principle be redeployed between facilities—and a pool of fixed units that cannot leave their home facility. Denote by m_j and $f_{i,j}$ the cardinalities of these two pools, respectively. Time-indexed decision variables include the allocation of mobile resources $x_{i,j}(t)$, the utilization of fixed resources $y_{i,j}(t)$, and shortfall variables $u_{i,j}(t)$ that capture demand not satisfied during interval t. For resources representing clinical staff, $x_{i,j}(t)$ and $y_{i,j}(t)$ can be interpreted as staffed person-hours, whereas for equipment categories they represent unit-time capacity slices.

Demand for each resource is expressed through a de-

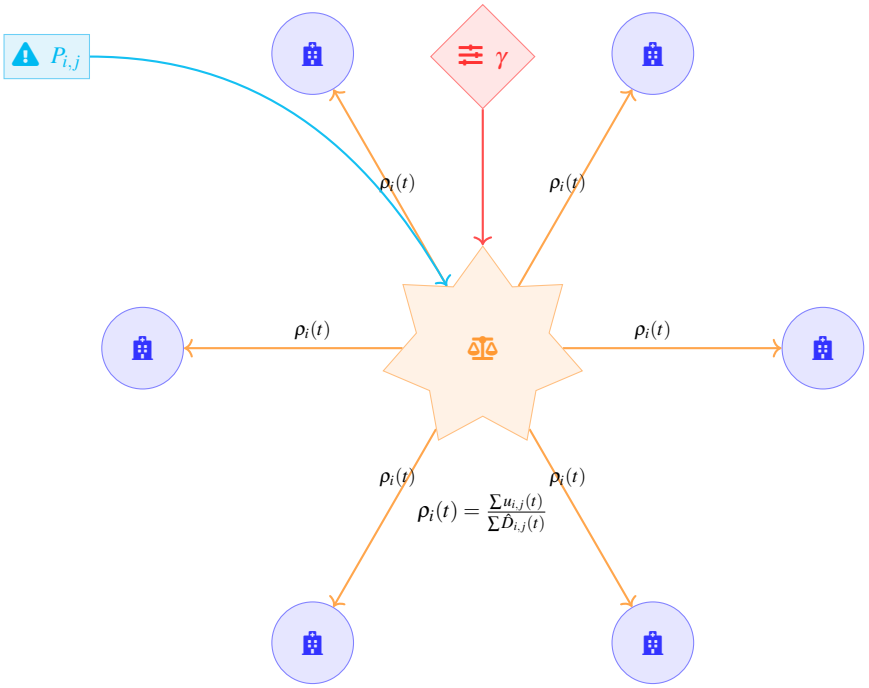

Figure 8. Equity Constraint Enforcement Mechanism

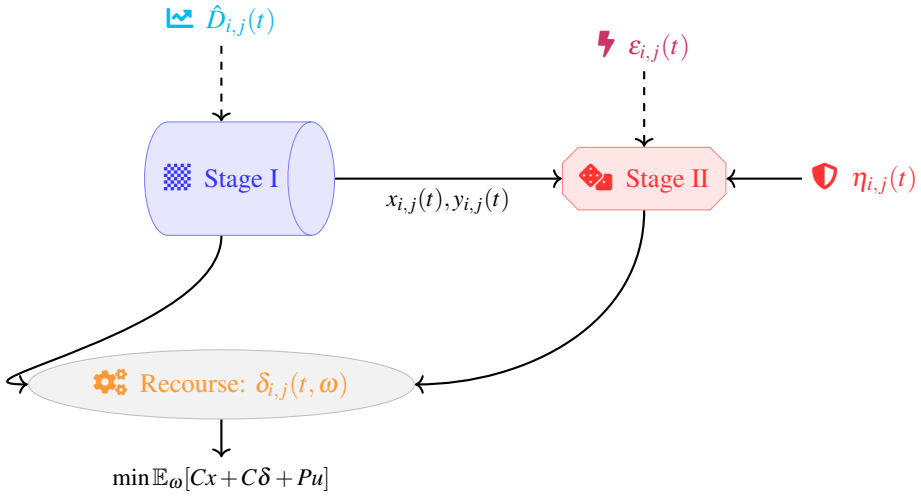

Figure 9. Two-Stage Stochastic Optimization Framework

terministic forecast $\hat{D}_{i,j}(t)$ and an error term $\varepsilon_{i,j}(t)$ such that the realized demand is $D_{i,j}(t) = \hat{D}_{i,j}(t) + \varepsilon_{i,j}(t)$. While the subsequent optimization employs $\hat{D}_{i,j}(t)$ as a point estimate, the design anticipates forecast error by embedding protective capacity and by permitting rapid re-optimization when observed deviations exceed an alert threshold. This approach functions analogously to a first-order robust control law.

Because many resource categories are inherently indivisible—one cannot allocate half a ventilator or fraction of a physician beyond granularity constraints—the model treats the allocation variables for those categories as integers. Let $\mathscr{J}_{\text{int}} \subseteq \mathscr{J}$ denote the subset requiring integral decisions. Then

$$x_{i,j}(t) \in \mathbb{Z}_{\geq 0} \quad \text{and} \quad y_{i,j}(t) \in \mathbb{Z}_{\geq 0} \qquad \forall j \in \mathscr{J}_{\text{int}},\, i \in \mathscr{I},\, t \in \mathscr{T}.$$

For divisible resources (such as aggregate nursing labor measured in decimal hours), the non-negativity domain $\mathbb{R}_{\geq 0}$ suffices.

Resource conservation constraints govern the total supply available for allocation. Let $S_j(t)$ represent mobile supply and $\bar{f}_{i,j}$ denote the baseline quantity of fixed resource j

at facility i. Then

$$\sum_{i\in\mathscr{I}} x_{i,j}(t) \le S_j(t), \qquad 0 \le y_{i,j}(t) \le \bar{f}_{i,j} \qquad \forall j \in \mathscr{J}, t \in \mathscr{T}.$$

Demand satisfaction is modeled through a coverage constraint that couples supply to needs:

$$x_{i,j}(t) + y_{i,j}(t) + u_{i,j}(t) \ge \hat{D}_{i,j}(t) \qquad \forall i \in \mathscr{I},\ j \in \mathscr{J},\ t \in \mathscr{T}.$$

Here $u_{i,j}(t)$ represents deferred or unmet demand—which triggers penalties in the objective—and is bounded above by a policy-chosen maximum $\bar{u}_{i,j}$ to reflect ethical limits on service denial.

Temporal coupling enters via continuity and ramp-rate constraints. For mobile resources, redeployment between epochs consumes transfer time $\tau^j_{i\to k}$ and may incur a transit cost. To express this, define binary motion indicators $z_{i,k,j}(t)$ that equal 1 if a unit of resource j is scheduled to transfer from facility i at time t to facility k and becomes available at $t + \tau^j_{i\to k}$. The conservation of flow for each mobile unit $r = 1, \dots, m_j$ can then be enforced through a network-flow substructure:

$$x_{i,j}(t) = x_{i,j}(t-1) + \sum_{k\in\mathscr{I}} z_{k,i,j}\left(t - \tau^j_{k\to i}\right) - \sum_{k\in\mathscr{I}} z_{i,k,j}(t) \quad \forall i, j, t > t_0.$$

This expression guarantees that units neither disappear nor duplicate across the horizon, while also encoding transfer delays. If transit times are negligible relative to Δt, the formulation simplifies to a first-difference limit:

$$|x_{i,j}(t) - x_{i,j}(t-1)| \le \Delta_{i,j},$$

with $\Delta_{i,j}$ representing a ramp-rate bound chosen to limit staffing shocks.

Equity protections can be embedded by introducing Gini-like dispersion measures or by constraining the maximum relative shortfall across facilities. One pragmatic metric is the ratio of unmet demand to total demand at each facility for high-acuity resources:

$$\rho_i(t) = \frac{\sum_{j\in\mathscr{J}_{\text{crit}}} u_{i,j}(t)}{\sum_{j\in\mathscr{J}_{\text{crit}}} \hat{D}_{i,j}(t)}.$$

A fairness constraint then limits disparity: [7]

$$\rho_i(t) - \rho_k(t) \le \gamma \qquad \forall i, k \in \mathscr{I},\ t \in \mathscr{T},$$

with γ defining the tolerated inequity band. This linear formulation avoids nonlinear inequality indices yet captures the spirit of proportional service access.

The multi-period objective must balance operating costs, transfer costs, and service penalties. Let $C_{i,j}$ denote per-unit operating cost, $T_{i,k,j}$ the cost of transferring one unit of resource j from facility i to k, and $P_{i,j}$ the penalty for unit shortfall. A discounted horizon cost is

$$\min \sum_{t\in\mathscr{T}} \beta^{t-t_0} \left[\sum_{i,j} \left(C_{i,j}x_{i,j}(t) + C_{i,j}y_{i,j}(t) + P_{i,j}u_{i,j}(t) \right) + \sum_{i,k,j} T_{i,k,j} z_{i,k,j}(t) \right]$$

where $0 < \beta \le 1$ applies temporal discounting; choosing $\beta < 1$ prioritizes near-term performance.

Stochastic extensions can be accommodated by treating $\varepsilon_{i,j}(t)$ as a random variable with known distribution and embedding recourse decisions. A two-stage formulation defines first-stage allocations $x_{i,j}(t)$ based on the forecast and second-stage adjustments $\delta_{i,j}(t,\omega)$ contingent on realization $\omega \in \Omega$:

$$x_{i,j}(t) + \delta_{i,j}(t,\omega) + u_{i,j}(t,\omega) \ge \hat{D}_{i,j}(t) + \varepsilon_{i,j}(t,\omega),$$

with scenario-weighted cost

$$\min \sum_{t\in\mathscr{T}} \sum_{i,j} \left(C_{i,j}x_{i,j}(t) + \mathbb{E}_\omega \left[C_{i,j}\delta_{i,j}(t,\omega) + P_{i,j}u_{i,j}(t,\omega) \right] \right).$$

Because enumerating Ω can be computationally prohibitive, sample average approximation or robust-optimization surrogates provide tractable alternatives. In a robust variant, the error term resides within an uncertainty set $\mathscr{U}_{i,j}(t) = \{\varepsilon : |\varepsilon| \le \eta_{i,j}(t)\}$ and the constraint tightens to guard against worst-case deviations:

$$x_{i,j}(t) + y_{i,j}(t) \ge \hat{D}_{i,j}(t) + \eta_{i,j}(t).$$

Parameter $\eta_{i,j}(t)$ acts as a demand buffer that is calibrated via historical volatility or a service-level target.

Hierarchical decomposition supports practical solvability. A master problem allocates aggregate supply to facilities, while subproblems refine intra-facility assignment across departments. The master's decision variables can be aggregated totals $X_i(t) = \sum_j x_{i,j}(t)$, yielding a smaller linear program. Benders cuts or dual prices then communicate marginal valuations back to subproblems, iterating until convergence. Similarly, column-generation techniques decompose by resource type: each resource class solves a knapsack-like subproblem generating columns (allocation patterns), which a restricted master problem selects to minimize cost under facility demand cover constraints. [8]

Dual analysis yields managerial insight. Let $\lambda_j(t)$ be the dual multipliers on the system-wide supply constraints. Economically, $\lambda_j(t)$ approximates the shadow price of augmenting supply of resource j during interval t. A high $\lambda_j(t)$ signals scarcity and can inform strategic investments such as recruiting additional staff or acquiring equipment. Likewise, duals on equity constraints quantify the opportunity cost of fairness; if the lagrangean multiplier on $\rho_i(t) - \rho_k(t) \le \gamma$ is large, relaxing equity by a small ε might deliver substantial

cost savings—a trade-off boards may deliberate in policy sessions.

Computational testing—beyond the scope of this purely theoretical exposition—would typically scale solution time as a function of problem size, verify dual-based sensitivity analysis, and test robustness margins. Even so, the structural elements presented herein already indicate paths for integrating actual real-time healthcare operations data into a mathematically sound but operationally advanced allocation engine balancing efficiency, equity, and pragmatism in the face of uncertainly.

4 REAL-TIME ANALYTICS ARCHITECTURE

The real-time analytics architecture, a layered ecosystem that ingests heterogeneous data streams, converts raw observations into consistent decision variables, and supplies the optimization engine with an ever-refreshed representation of the healthcare network, forms the operational heartbeat of the proposed optimization framework. The architecture must satisfy stringent requirements on latency, throughput, fault tolerance, security, and interpretability while being sufficiently modular to accommodate evolving data sources and analytical models. As a result, its architecture draws from principles of distributed systems engineering but is informed by the idiosyncrasies of clinical operations, regulatory oversight, and ethical stewardship.

At its edge, the architecture engages with heterogeneous data producers that are scattered across the regional healthcare setting. Each facility exposes device telemetry, transactional logs, and contextual metadata that collectively characterize real-time operational state. Admit-discharge-transfer messages, nurse call-bell presses, point-of-care device usage counters, laboratory order milestones, and supply chain bar-coding events are typical examples. The majority of these sources post events asynchronously on milliseconds-to-seconds timing, while others—such as periodic bed census snapshots—arrive on slower, batched rhythms [9]. The ingestion layer therefore employs a high-throughput message broker, abstracting protocol heterogeneity through the use of adapters that translate vendor-specific standards (e.g., Health Level Seven, Digital Imaging and Communications in Medicine, or proprietary real-time location system packets) into a canonical internal schema. Each incoming record is time-stamped at the ingestion gateway on a monotonic clock synchronised via Network Time Protocol stratum services to minimize skews that would otherwise cause causal ordering inaccuracies downstream.

A stream processing fabric immediately downstream manages low-latency transforms to cleanse, normalise, and enrich the data. Data cleansing deals with syntactic anomalies—such as incorrectly formed identifiers or unit mismatches—via rule-based validators and light-weight probabilistic corrections. Normalisation translates synonymous codes into a shared vocabulary through a mapping dictionary dynamically versioned in a manner that allows evolving clinical terminologies to be updated without downtime. Enrichment introduces computed features such as derived acuity scores, rolling utilisation ratios, and interarrival rate estimates. To maintain exactly-once semantics, each transformation step is wrapped in idempotent operators committing transactional offsets only after stateful operations succeed, thus immunizing the pipeline from partial failures or replay storms.

A persistent mutable state store underpins the stream processor that retains the most recent value of each operational metric for each facility–resource pair. Conceptually, this store is a key–value map indexed by the composite key (i, j) for facility i and resource type j, with each value a high-granularity time series of the most recent N observations, where N is tuned to balance memory footprint and modelling fidelity. This persistence tier provides support for windowed aggregations—e.g., computing a five-minute moving average of emergency department arrivals—while providing a single source of truth for downstream consumers. To support sub-second read/write performance under concurrent workloads, the store is sharded across a distributed in-memory database cluster with automatic partition rebalancing and replica placement optimised for both latency and resilience. Crash consistency is ensured by a write-ahead log which replicates delta records to remote object storage, enabling time-travel queries and disaster recovery without synchronous I/O bottlenecks. [10]

On top of the streaming substrate core is a model orchestration layer that manages online inference, predictive analytics, and feature engineering. Here, microservices encapsulate models for short-horizon demand forecasting, patient acuity progression, staff availability prediction, and equipment failure likelihood. Each microservice exposes gRPC endpoints that accept feature tensors and produce predictive scores or probability distributions. Model weights and hyperparameters are managed via a versioned registry that supports blue–cyan rollouts, with effortless promotion of new model versions once they have demonstrated superior back-testing performance. The orchestration fabric includes a feature store that materialises real-time and offline-computed features for inference, providing the model with consistent semantics between retraining and serving. Capture feature lineage metadata that records provenance, allowing auditors to trace decision inputs back to raw sensor or transactional events.

A unique aspect of the architecture is the latency-aware scheduler that orchestrates predictive inference calls according to service-level objectives. For example, surge prediction for emergency departments must complete in hundreds of milliseconds because its outputs directly influence the next optimisation step. Conversely, though, elective surgery slot demand forecasts for the next day tolerate multi-second latencies. The scheduler exploits this asymmetry by mapping tight-latency workloads to high core frequency and memory bandwidth nodes and relegating less time-sensitive

Table 5. Key Components of the Real-Time Analytics Architecture

Layer	Function	Examples / Tools
Data Acquisition	Collect raw data from multiple systems	EHRs, IoT sensors, staffing systems, telemetry
Data Processing	Clean, transform, and unify data streams	Timestamp alignment, error filtering, in-memory updates
Analytics	Predict demand and resource availability	ML forecasts, trend analysis, KPI monitoring
Optimization Engine	Solve allocation problem per interval	LP/IP solver, heuristics, warm-starts
Decision Execution	Implement or recommend actions	Dashboard, alerts, system triggers (e.g., staff redeployment)

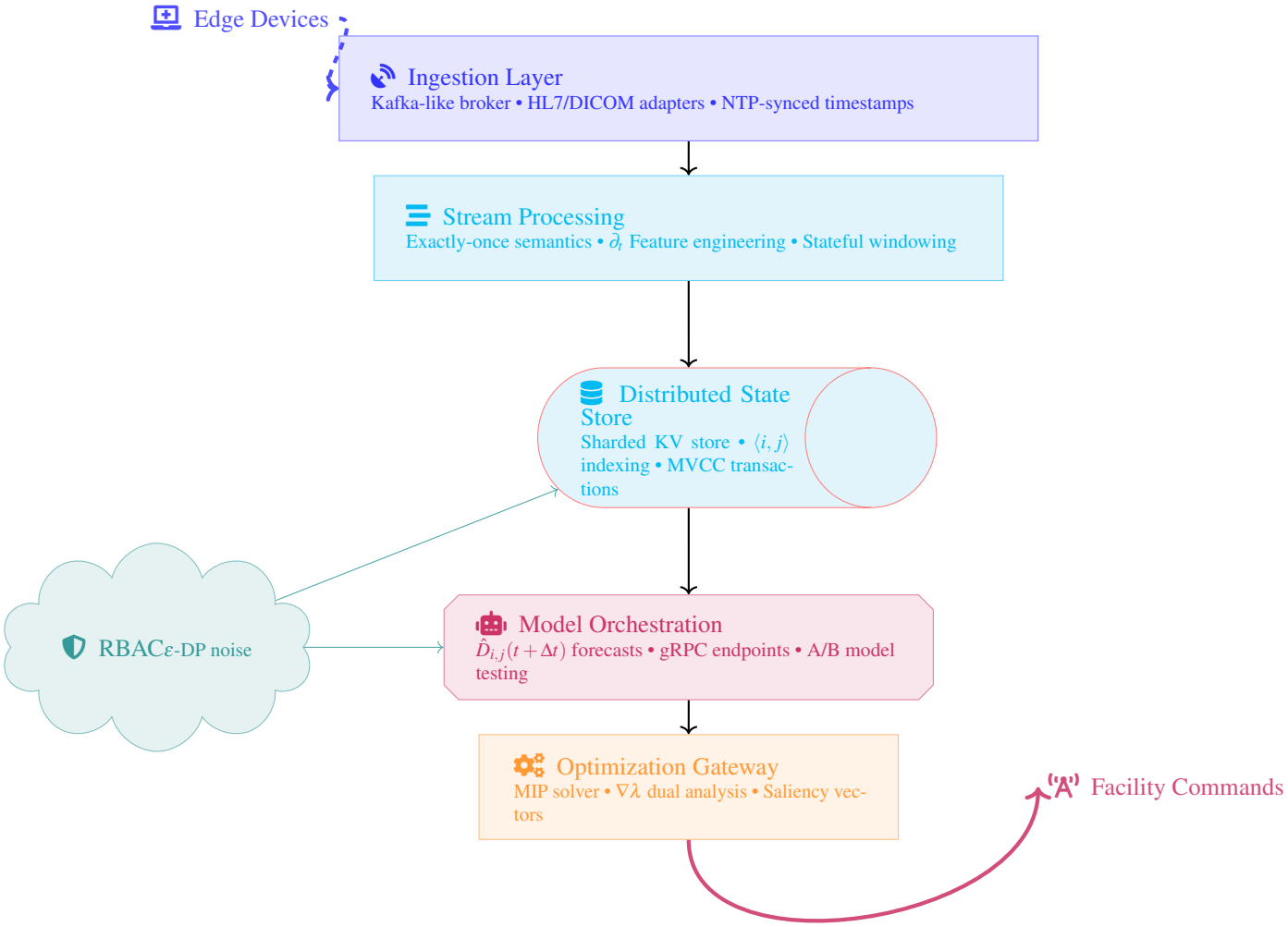

Figure 10. Multi-Layer Analytics Architecture with Stateful Processing

Table 6. Examples of Real-Time Data and Their Role in Decision Support

Data Type	Source System	Role in Optimization Model
Patient Admissions / Discharges	EHR / ADT Systems	Determines $D_{i,j}(t)$ for nurses, beds, equipment
Staff Clock-In/Clock-Out	Badge systems, HR platforms	Contributes to $S_j(t)$ for staffing categories
Equipment Usage Logs	IoT telemetry, device logs	Tracks real-time availability of critical resources
Wait Times / Queue Lengths	ED dashboards, triage systems	Proxy for unmet demand or hidden surges
Predictive KPIs	ML or statistical forecasting models	Informs future $D_{i,j}(t+h)$ for rolling-horizon planning

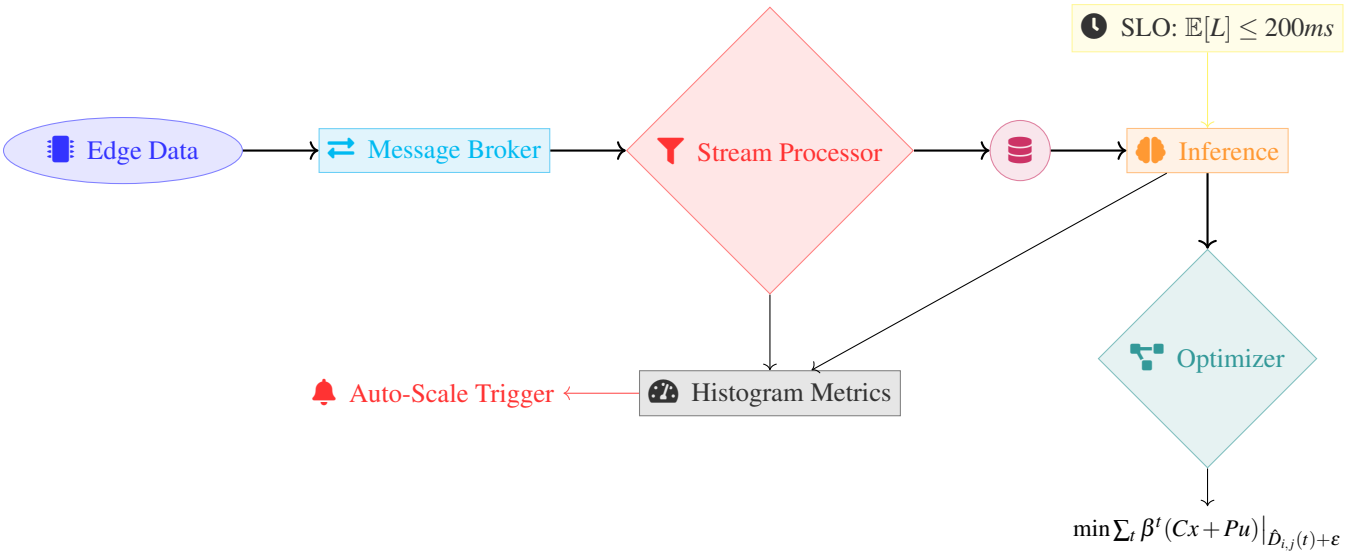

Figure 11. Real-Time Pipeline with Latency-Aware Scheduling

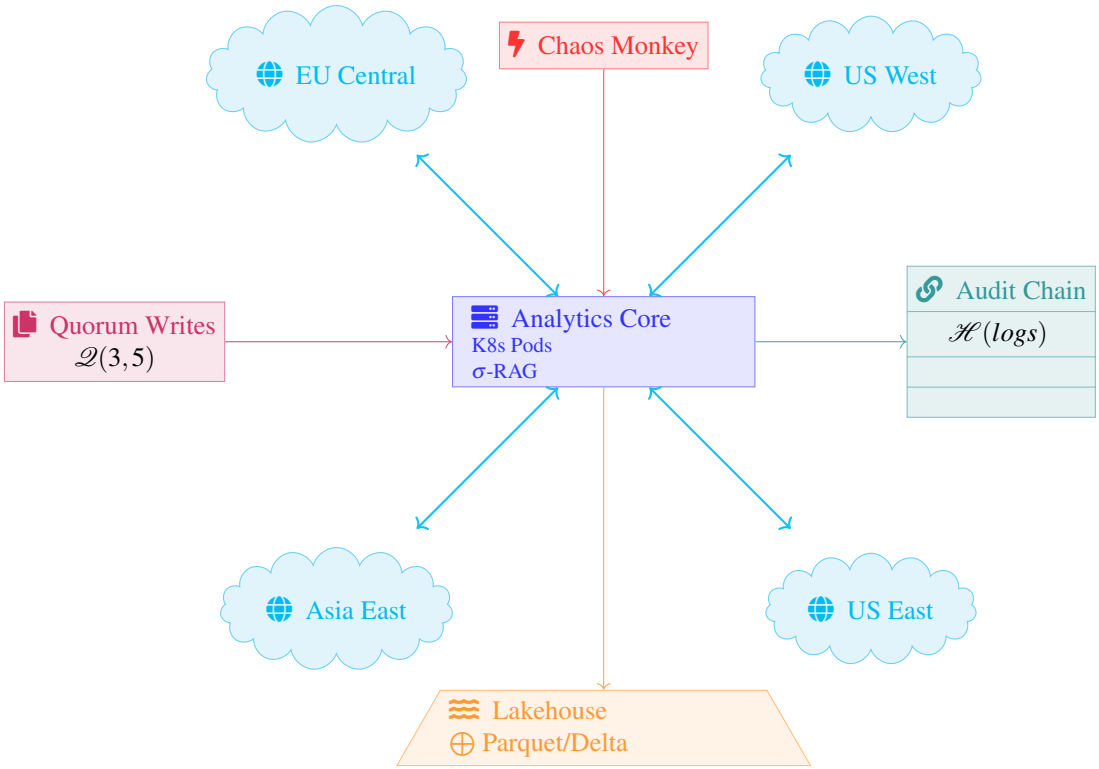

Figure 12. Geo-Distributed High-Availability Architecture

workloads to cost-efficient spot instances. This elasticity is controlled via a queue depth and response-time percentile feedback signal and scales computational resources linearly with workload burstiness without human intervention.

Amidst predictive services is a metrics monitoring system that instruments every component of the pipeline [11]. Each microservice exports structured logs and latency histograms to an observability stack of time-series databases and distributed tracing systems. Alert rules specify thresholds on key indicators—message lag, model response time, and feature drift. When a threshold is breached, an incident automation service triggers remediation workflows that can include auto-provisioning another streaming node, restarting a failed container, or rolling back to a prior stable model version. These self-healing capabilities are essential in maintaining continuous uptime during network partitions,

hardware failures, or unexpected workload spikes.

Sitting atop the analytics fabric is the data transformation gateway that brokers between the real-time metrics universe and the optimisation engine. The gateway periodically snapshots the state store, extracting the most recent demand forecasts $\hat{D}_{i,j}(t)$, supply availability $S_j(t)$, and auxiliary covariates—e.g., predicted staff clock-in adherence or predicted discharge probability. The gateway rearranges these inputs into the solver’s expected data structures and invokes the optimisation service through an asynchronous call. For consistency, every snapshot is logically isolated using a multi-version concurrency-control transaction: the optimiser works on a view that is stable while it solves, even while new data still arrives in the stream processor.

Solver runtime can be containerized within the same Kubernetes cluster as the analytics services or offloaded to an HPC farm according to problem size. In either situation, after a allocation decision has been received, the gateway commits the solution back to the state store and publishes it to a command topic on the message broker. Facilities are subscribed to specific command channels according to their identifier so that only relevant instructions—e.g., redeploying two respiratory therapists or relocating a mobile ultrasound—arrive at local dispatch consoles. Commands include effective timestamps, validity windows, and rationale annotations so that human supervisors can cross-check allocation decisions against real-time dashboards before authorising execution.

In the interest of explainability, the gateway enriches each decision record with a saliency vector [12]. This vector quantifies the marginal contribution of significant input features—such as ED arrival volume or ICU bed occupancy—on the final assignment for each resource. Saliency calculation uses dual variable analysis for linear constraints and Shapley value approximation for nonlinear elements of the model. By offering these explanations alongside numeric suggestions, the system gives clinicians and administrators an interpretable narrative that inspires trust and makes override simpler when domain instinct identifies a nuance the algorithm misses.

Privacy, security, and compliance concerns are embedded in each layer. Data ingress encrypts payloads through transport-layer security and employs per-facility authentication tokens confirmed by a mutual TLS handshake. In the analytics fabric, role-based access controls restrict operators to least-privilege interactions, and differential privacy noise can be injected into metric queries backing research dashboards without violating patient confidentiality. Audit trails capture immutable ledgers of data transformations, model invocations, and decision outputs, hashed and anchored to a permissioned blockchain that delivers cryptographic integrity without suffering the throughput limitations of public chains. Vulnerabilities are always addressed by ongoing penetration testing and static analysis, and container images are supply-chain attested to ensure provenance.

The architecture facilitates horizontal expansion not only in scale but also in feature space. Edge-analytics extensions allow latency-sensitive inference—e.g., ventilator alarm triage—to be run on ward-level gateways, with reduced round-trip latency to the central cluster. Federated learning can be achieved where the law, such as that which protects substance-abuse treatment records, requires that certain parameters never leave the originating facility in their unprocessed form. In this case, learners on premises compute gradient updates which are centrally summed, updating a global model without divulging protected datasets [13]. Further, the message broker’s topic hierarchy for the subject can incorporate external data feeds—weather alerts, traffic trends, or infectious-disease surveillance reports—that enrich demand forecasting models with exogenous signals.

Cost efficiency and scalability also benefit from adopting a lakehouse strategy to the storage of historical data. Raw event streams are compressed and written column-wise to object storage, partitioned by facility and event date. Analytical questions for model retraining, root cause analysis, or retrospective audits can leverage distributed SQL engines that read directly from the lakehouse without disrupting hot path operations. A metadata catalog records schema evolution so that historic records can still be read even as source systems migrate to new coding conventions.

High availability is achieved through multi-zone deployment across geographically distributed data centres, each of which can handle ingestion, transformation, and optimisation tasks in full. Failover replication employs quorum writes and asynchronous geo-replication for the hot path so that regional disasters or connectivity loss degrade capacity gracefully, not catastrophically. Regular chaos-engineering exercises inject controlled failures to verify that redundancy mechanisms meet recovery time goals, while service-level metrics provide quantitative proof of resilience.

As a final component, a simulation sandbox interfaces with the production analytics pipeline so that scenarios can be tested without endangering live operations. Synthetic event generators replay historical traces or simulate extreme-demand scenarios so that administrators can observe model responses, adjust penalty parameters, and test ramp-rate constraints before pushing policy changes live. The sandbox uses the same code base as production, ensuring fidelity, but routes decisions to a shadow dashboard. Simulation findings guide parameter refinement, predict future bottlenecks, and support training exercises for human operators who must interpret and sometimes override algorithmic recommendations [14].

Briefly, the real-time analytics infrastructure functions as a digitally synchronized nervous system for the regional healthcare network. It ingests fine-grained operational events, distills them into high-value decision variables, enables predictive models to forecast near-future states, and provides optimisation-derived recommendations at a

rhythm that matches clinical workflows. Each layer—ingestion, stream processing, state storage, model orchestration, optimisation gateway, and command dissemination—provides expert capabilities while working together under an overall design driven by latency, reliability, security, and interpretability requirements. By marrying modern data-engineering abstractions to domain-specific safeguards, the architecture provides the robust computational substrate required for adaptive, cost-aware, and patient-centered resource management.

5 OPTIMIZATION LOGIC

The optimization reasoning is the operational element of the manner in which the model continuously comes to inform decisions. Effectively, it constitutes a control feedback loop in the operations management scenario. At each decision interval (e.g., each hour or each time a significant event like a surge occurs), the system gathers the latest information (through the architecture of Section 4) and builds the optimization problem as described in Section 3. It then solves this problem to determine recommended resource allocations $x_{i,j}(t)$ for the next interval. After these suggestions have been implemented, the system idles for the subsequent epoch (or trigger) when there is new data ready, and this cycle is then repeated. The process is the same as using a rolling horizon or model predictive control method wherein each solve is given current state data and potentially forecasts, does the initial segment of the solution, and readjusts subsequently.

One of the key aspects of the optimization logic is that the solution needs to be able to be attained in a timely manner. The theoretical model can be very large in terms of variables (especially if there are many facilities and types of resources), and if expanded to include binary or integer decisions (e.g., whether or not to hire an extra staff member, which can be represented as a yes/no decision), then the problem is computationally intensive. To address this, the justification can employ certain techniques [15]. In the first instance, whenever solving the model occurs at time t, solution from the previous time step $(t-1)$ could be employed as a good start or a good warm start. The majority of optimization solvers allow starting using an earlier solution, and the same can help speed up convergence quite easily if the system's state has not changed much from one time step. This means that if demand and resources at time t are similar to $t-1$, the new optimum is likely to be a slight adjustment of the one before, and a warm-started optimizer will find it in no time.

Second, the optimization smarts can decompose or simplify the problem when the situation calls for it. If solving the entire problem takes too long, a decomposition approach could be employed. For example, the assignment over each type of resource j can be calculated in parallel across different threads or machines, since linking between other resource types in the objective function is only via possibly common facilities (if P_j penalties make each resource effectively meet its own requirement). There would be some coordination if cross-resource constraints are present (not something our basic formulation includes, but an extension could, e.g., a total limit on personnel of various types). Or alternatively, a top-down approach can be used: first, one decides on a top-level distribution of overall capacity over facilities (e.g., with disregard of all resources in bulk or for the most critical resource constraints), and then, in a follow-up stage, further distributions are calculated within each facility or across each type of resource. Hierarchical optimization can reduce dimensionality level by level and can mirror how decisions are being made in the world (e.g., first decide how many patients each hospital will see, and then decide how to allocate staff within each hospital).

Another technique of the optimization reasoning is to employ constraints or penalties that promote long-term stability of decisions. Because the model is solved repeatedly, there is a risk of oscillations or wild oscillations – e.g., at time t the solution will allocate more staff to Hospital A and less to Hospital B, but at time $t+1$ this is reversed, and this seesawing might be destabilizing. To prevent this, the model can have a smoothing term or constraint. One way is to add a term to the objective that punishes reallocation changes, e.g.,

$sum_{i,j} H_j |x_{i,j}(t) - x_{i,j}(t-1)|$ with some penalty weight H_j that makes high frequency large reallocations less appealing unless due to large changes in demand. Another way is to impose ramping constraints, for example: [16]

$$|x_{i,j}(t) - x_{i,j}(t-1)| \leq \Delta_{i,j},$$

limiting the extent to which a decision needs to change within an interval by some value

$Delta_{i,j}$. These mechanisms ensure the logic of optimization produces recommendations not merely optimal in cost for the moment but with knowledge of implementation realizability and temporal consistency. In real implementation, this increases the likelihood that administrators and personnel will be able to implement the recommendations (since they avoid whiplash resulting from constant variations) and imposes confidence in the stability of the system.

The optimization logic also addresses data and prediction uncertainty. Real-time analytics provides the current best estimate of and near-future state, but even so, between the decision points unpredictable things can happen (for example, a catastrophe can cause a surprise surge in demand). To address this, the logic can add some robustness. A way to do this is to employ conservative demand estimates: for example, utilize an inflated $D_{i,j}(t)$ or a high percentile of forecasted demand distribution when missing demand is very expensive, leaving room for a buffer. This is similar to saving some capacity in case of surprise. Another approach is to solve not for one case but to consider a small set of

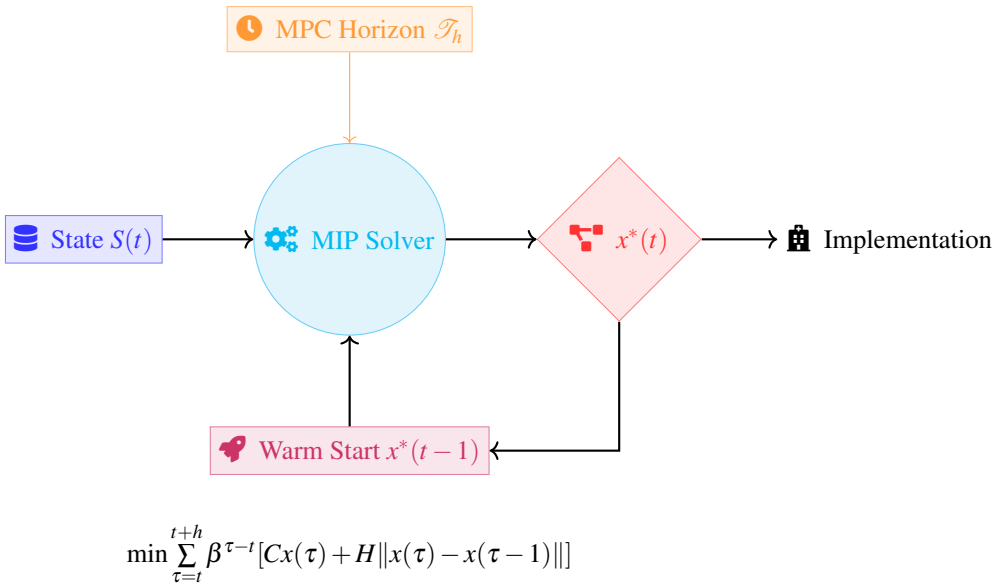

Figure 13. Model Predictive Control Loop with Warm-Start Injection

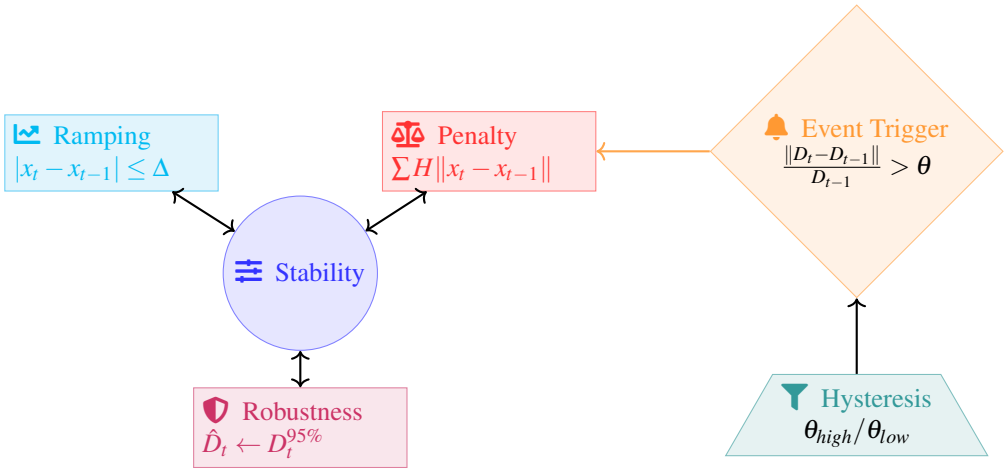

Figure 14. Stability Preservation Mechanisms with Adaptive Triggers

cases (e.g., normal and worst-case demand) and make a decision that is well enough across cases (a bounded form of robust optimization). But scenario-based optimization may be more computationally demanding, so there is a balance. Generally, the simpler and more pragmatic solution is to rely on the ability to re-solve quickly: if something unexpected happens, the next iteration will catch up with it and react. The secret is that iterations are frequent enough so that the system never gets behind reality too far.

Frequency and triggers for re-optimization are part of the logic design. While a standard schedule (e.g., hourly) is conceptually easy, the system may also offer event-based triggers [17]. For example, if a hospital emergency room suddenly sees 10 more patients in 15 minutes (a departure from the norm), the system may trigger an immediate optimization out-of-cycle in reaction to the surprise change, rather than waiting until the next hour. This requires the architecture to support on-demand solving of the solver and the logic for what constitutes a trigger threshold. Establishing equilibrium for this is paramount to avoid over-computation or infinitesimal constant tweaking for every slight change. The logic can use hysteresis or threshold rules (e.g., do not start a new solve unless a metric changes by more than X

Optimization integration is another nuance. If the analytics layer foresees an escalation in demand two or three intervals in advance, then the optimization logic can either adjust by effectively solving a multi-interval lookahead (as detailed, akin to model predictive control) or by pushing the current demand upward by anticipation. Which to use may depend upon how crucial lead time preparation is. For instance, if the employment of additional workers has a lead

time (it will take time for someone to arrive), then the model can potentially have to make decisions one interval ahead: in essence, the optimisation at time t might include some decisions which are meant to meet demand at $t+1$. This can be formulated by adding $D_{i,j}(t+1)$ to the optimisation currently, with some weighting or by simply lengthening the model's time horizon. The logic of optimization can thus potentially have a lookahead horizon parameter and weighing of future intervals' goals, which can be tuned empirically.

Algorithmically, if the underlying problem is linear (or linear with few integer variables), then modern solvers can solve relatively large instances in seconds, especially with warm starts and assuming that the structure of the problem is utilized. If the issue becomes sophisticated (e.g., a big integer problem with lots of binary variables), then heuristic approaches could be employed. For instance, one may apply a greedy approach to allocate resources to the most urgent shortfalls first, or apply metaheuristics (such as genetic algorithms or simulated annealing) that can be executed within a bounded time budget to offer a good solution although not necessarily optimal [18]. The theoretical model itself does not need a specific method of finding a solution, but the logic layer is responsible for determining appropriate methods of finding a solution within real-time operation's time constraints.

Briefly, the optimization logic governs the data-driven reoptimization loop so that the solution process is efficient, stable, and responsive to change. It balances between optimality and pragmatism by introducing computation shortcuts, decision smoothing, and robustness. The logic transforms the static mathematical model into a dynamic process that makes the system more efficient and cost-effective on an on-going basis.

6 APPLICATIONS

An application could be dealing with an unplanned surge in ED demand within a single hospital in a regional health network. Consider Hospital A receives an unplanned surge of patients (for example, due to a multi-car accident on a busy night). Historically, Hospital A would address this surge by pushing its own limits—having on-duty staff treat more patients and diverting some ambulances to other hospitals once capacity is obviously strained. Under the new model, though, the reaction would be more proactive and streamlined. Real-time data from Hospital A's ED (rise in patient arrivals, queue length, severity mix) would be automatically analyzed. The system can recognize that unless action is taken, estimated waiting times will exceed acceptable standards and some patients will depart without treatment (a loss to patient care as well as to hospital funds) [19]. The optimization program can recommend redirecting resources such as having some on-call physicians or nurses call in (if present), or even temporarily transferring a crew from Hospital B's less busy urgent care center to Hospital A's ED. It can also recommend diverting ahead of time a percentage of incoming ambulances to Hospital B and Hospital C, before Hospital A is completely filled, to redistribute the load. In doing so, the model incurs costs for these activities: calling in extra staff carries an overtime expense (reflected in the cost term C_j for staff at Hospital A), and relocating patients to a rather more distant hospital might incur a cost in terms of patient inconvenience or ambulance fuel, etc. By comparing those costs with the penalty of idle demand (patients not being treated on time), the model determines an allocation that might, for instance, assign two extra nurses and redirect 20% of ambulances to other facilities for the next 2 hours. This theoretically maintains waiting times in line at Hospital A (increasing efficiency and quality of care) at lower expense than reactive overstaffing or uncontrolled diversion after the fact. Additionally, the area as a whole handles the surge efficiently, using available capacity at Hospital B and C that otherwise would have been unused. As far as cost savings are concerned, avoiding unnecessary overcrowding at Hospital A will prevent expensive downstream consequences (like patients needing more intensive care due to delays, or the hospital needing to open overflow space temporarily at tremendous financial expense). The situation shows how coordination enabled in real-time using the model would be able to achieve efficiency (fast service for emergency cases) and cost control (preventing unnecessary activation of resources and sharing load regionally).

Another possible use is in the management of elective procedures and bed capacity within a region. Let us assume two hospitals, each with a fixed number of operating rooms (ORs) and post-operative beds. On a given day, Hospital X had a light elective schedule (some surgeons were out or cases had been delayed), while Hospital Y is nearly booked solid for surgeries and its recovery beds are occupied. Independent management of Hospital Y can experience overcapacity – being forced to cancel a surgery at the last moment or retain patients longer in the recovery room due to a lack of ward beds, while Hospital X is operating below capacity (inefficient use of resources). Under region-wide optimization planning, the system would be notified of this discrepancy in advance using real-time schedule and bed use data [20]. It may suggest moving one or two elective cases from Hospital Y to Hospital X (if both or other similarly qualified surgeons are present at X). By doing so, the model reduces the workload at Hospital Y (preventing costly overtime for staff and potential use of pricey contingency strategies like opening up another area for recovery) and optimizes utilization at Hospital X (more effectively utilizing its staffed OR time that would otherwise go underutilized). The expenses that would be factored in the model would be any transfer bother or logistical cost of the patient, and perhaps a small penalty if something is done at a non-preferred facility. These may be offset, however, by the expense benefits of not over-allocating Hospital Y's

Table 7. Scenarios Demonstrating Real-Time Optimization Benefits

Scenario	Challenge	Traditional Response	Optimized Response
ED Surge at Hospital A	Unexpected patient influx	Staff overstretch, reactive diversion	Preemptive staff reallocation, ambulance rerouting
Elective Surgery Overload at Hospital Y	OR and bed congestion	Last-minute cancellations or overtime	Case redistribution to Hospital X with spare capacity
MRI Overload at Hospital M	Imbalanced imaging demand	Patient backlog, deferred scans	Shift patients or technicians, balance load with Hospital N

capacity and avoiding having to cancel procedures (which can incur money penalties or lost revenue). In real time, they are able to make decisions a day or a few hours ahead of time as the schedules get sorted out, illustrating how the model is not only reacting to the crises of the moment but also optimizing operations on a day-of basis. This leads to total improved efficiency (all available ORs and staff are utilized productively) and cost savings (less overtime, no temporary staffing agency nurses, etc., and better patient throughput so no surgical backlog).

A third example setting could be involved with the assignment of specialized equipment and accompanying technicians in a region. Take a fleet of expensive-to-run and expensive-to-staff medical imaging devices (MRI devices). Perhaps Hospital M and Hospital N both possess a single MRI scanner, but on any given day most of the demand for MRI scans is coming to Hospital M (due, say, to an influx of trauma patients who need to be imaged), and Hospital N's machine is idle for hours. The traditional method would be that each hospital book and uses its MRI separately, and when Hospital M is delayed, patients wait longer or some non-emergency scans are rescheduled to another day. With an optimization system in real time, one could control these assets in total. If the real-time data indicates heavy usage at M and light usage at N, the system can suggest shifting some outpatients or non-emergency scans from M to N (if patients are agreeable and it's feasible transport-wise) or even sending a mobile MRI unit (if available) to M [21]. It could also look to switch an MRI tech's N to M shift for the day to enhance throughput on M's machine (since often the scan volume is limited by available machines as well as techs). The model would weigh the travel or coordination expense against the worth of reducing the queue at M. The effect could be that a predetermined number of patients are requested to report to N for their scan during the afternoon, keeping both machines comparatively busy and avoiding a six-hour wait at M. The gain in efficiency comes from balancing patient service time and machine utilization, and the cost reduction comes from putting existing N capacity to use rather than putting M on overtime or considering purchasing another machine for M down the road. Although this scenario is operationally complex (since it involves patient voluntary cooperation and coordination), it is hypothetical to prove that the above kind of resource-sharing justification can be determined by the model in a systematic way rather than relying on ad-hoc judgment.

These examples illustrate how the real-time analytics and optimization model can be applied to every feature of healthcare operations. In all cases, the theme is realignment of resources according to circumstances to avoid waste: whether that is a surplus of emergency units, inappropriateness of elective surgical cases, or unused utilization of expensive equipment. By doing that, the system tries to reconcile service quality (throughput, wait times) with expenses (labor, overtime, utilization of capital). Each situation also underscores the value of the input data and constraints – the model's suggestions are only as good as the data it is given (e.g., knowing that Hospital X has available OR time or that patients can be transferred to Hospital N's MRI is valuable information). Thus, while these examples show potential benefits, they also reflect the importance of strong data sharing and operational flexibility in the health network.

7 LIMITATIONS

The model presented here is intentionally theoretical, and its capacity to generate run-time value is contingent on a number of assumptions that must be thoroughly tested [22]. Perhaps most critically, the model takes as given the availability of high-fidelity, low-latency data streams across all facilities within the regional network. Real-world hospital information systems also are heterogeneous and include multiple discrete vendor platforms, custom interfaces, and legacy databases that do not necessarily interoperate. Even where technical integration succeeds, semantic variability—e.g., in non-standard coding of categories of diagnosis, staffing jobs, or types of beds—can propagate embedded errors to optimization inputs. A mistake in defining an intermediate-care bed as an intensive-care bed, say, would cause the solver to overestimate critical capacity, potentially delaying escalation procedures. These errors might be latent and not reveal themselves except under data-stress conditions, and thus could be difficult to identify with standard data-quality testing. Its success therefore depends on continuous monitoring of data governance through automated validation testing and periodic domain-expert inspection sessions, cross-checking algorithmic quantification with frontline indicators.

Latency of data is a second, related limitation. The optimization cycle should be able to handle near-real-time snapshots of state, but some operational parameters—e.g., actual start and end of surgery or lab turn-around times—won't be available until several years after the fact. Decades or even only two decades of minutes' latencies may render the decision horizon nonexistent with little or no time for corrective intervention prior to its shifting once more. For example, if the post-anesthesia care unit nurse redeployment

Table 8. Example Cost-Efficiency Tradeoffs in Real-Time Decisions

Use Case	Cost Factors	Efficiency Gains	Key Model Features Used
ED Surge Management	Overtime pay, patient transport	Lower wait times, reduced crowding	C_j, P_j, real-time $D_{i,j}(t)$
Surgery Load Balancing	Surgeon reassignment, patient preference cost	Higher OR utilization, fewer cancellations	Forecasting, inter-facility capacity view
MRI Scheduling	Technician shift change, patient transfer	Reduced queue, balanced machine use	Resource pooling, load shifting

by the model is contingent on an estimated end-of-case time that subsequently materializes thirty minutes later, the unit might be unstaffed at the clinically most difficult time. Buffering techniques—like conservative lag compensation or time-stamped confidence intervals—are half the solution but sacrifice the high-granularity responsiveness that makes real-time optimization valuable in the first place. This latency tolerance for decision agility trade-off is inescapable and cannot be completely avoided [23]. Privacy and security requirements are another source of complication.

Continuous data feeds may transmit covered health information across network infrastructures that range from highly robust to decidedly less robust. Even with encryption in transit and at rest, attack by an adversary like ransomware or efforts at data exfiltration may have the potential to disrupt input signals or taint the integrity of optimization results. Regulationally, all data sharing between sites must comply with jurisdictional privacy regulations and institutional review processes. In order to that end, however, some potentially useful data fields might be missing or hidden, limiting model information. Moreover, the privacy mechanism overhead of technology such as differential privacy perturbations or secure multiparty computation introduces delay and reduces numerical accuracy, both detrimental to solution quality. Balancing the strong privacy protections against optimization timeliness requirements is a precarious matter beyond that purely technical. Coming to the model structure, linear penalty and cost are used for convenience by the model, but the majority of real cost functions are threshold or nonlinear.

In scheduling staff, marginal labor costs can increase tremendously beyond scheduled time due to overtime penalties and loss of productivity due to fatigue. For capital, maintenance cycles and economies of scale introduce nonlinearities in the cost function that linear approximations will not capture. Omitting these nonlinearities could lead to allocations that are biased towards less expensive theoretical solutions, which when realized, possess stealth or lag costs. The incorporation of piecewise-linear approximations or nonlinear programming techniques may introduce realism but may make run times unacceptably lengthy if problem size increases with dozens of facilities and resources [24]. Thus, any run deployed will strike a balance between fidelity and solvability and will employ hybrid methods such as hierarchical decomposition or heuristic post-processing in order to be capable of modeling nonlinear effects as an approximation of a predominantly linear core model. Another set of assumptions deals with the fluidity of resource transfer.

The model theory employs personnel and equipment in the blink of an eye from location to location without consideration for travel time, turn-around procedures, and institutional hurdles like licensure reciprocity to create state-to-state practicing physicians. In practice, however, short physical distances between locations can translate into significant transfer latency in the form of traffic congestion, shift relays, or waiting to sterilize equipment and rescribe the inventory lists. These dynamic costs make proposed allocations arrive too late to satisfy the demand they were intended to, creating churn as the next optimization cycle overcorrects in the opposite direction. Modeling resources with imbedded travel-time constraints or buffer stockages would alleviate this issue somewhat but at the expense of model simplicity. Human factors add complexity to a simplified mathematical solution.

The system takes for granted that staff will embrace convenient reassignment orders, yet psychological and contractual considerations influence personnel willingness to embrace. Repeated intra-shift changes can result in disruptions and fatigue or decrease job satisfaction, thereby contributing to turnover or absenteeism. Physicians would also resist transfer of elective cases if they perceive reputational or financial losses are incurred. These response patterns might be difficult to measure but are tough determinants of real system performance. Qualitative interaction with nurse unions, hospital employee committees, and patient advocacy groups must therefore supplement algorithm deployment, including joint governance institutions and feedback arrangements for policy reformulation [25]. Organizational incentives form a second constraint. The cost drivers in the model are one regional view, but in reality each facility will have separate budgeting needs and reimbursement systems.

A high-end-margin community hospital will not willingly forfeit high-revenue elective work to a tertiary center even as the regional optimizer suggests this to work through up-bedding. Incentives could be tied to shared revenue deals or performance metrics linked to pooled outcomes, strategies that entail executive-level coordination and, if so mandated, legislative ratification. Without such coordination, local opt-outs may jeopardize global optimality of the model, replacing decision-making with politically brokered or ad-hoc alternatives. Adherence to rules also places limitations within which the model can operate. Staff requirements by professional accrediting organizations, licensure occupancy maxima, or special trauma designations may exclude straight cost-based calculation.

For instance, an activity that nominally is saving money

could inadvertently violate a mandated nurse-to-patient ratio in a critical care unit. While no matter how elegantly such constraints are programmed to appear, policy space is heterogeneous and non-stationary: policy rules update, at some rate which sometimes can be highly rapid, based on public health emergencies or changes to policy. Consequently, the set of constraint must therefore refresh continually with new rules without negatively affecting solver efficiency—a far too frequent un-noticed burden upon system administrators. Concerns with equity add other constraints on naked optimisation. Straightforward cost minimisation can generate allocations that systematically benefit high-density urban centers, inadvertently reinforcing rural or disadvantaged group inequalities. [26]

Including fairness constraints—e.g., guaranteeing minimum levels of service or equating travel burdens—also introduces multi-objective trade-offs and the potential for conflict between efficiency and equity. While such extensions are computationally tractable, the weightings involved must be open to political and ethical rather than computational engineering judgment. Algorithmic explainability and transparency are also limitations. Clinicians and administrators will accept decisions better if they can see why; but even the tractability of a linear model with thousands of variables may preclude intuitive explanation.

If front-line workers cannot see how changes in their local measures influence system-wide recommendations, they may not believe the output or challenge agendas. Produce easily interpretable surrogate models, visualization dashboards, or sensitivity analyses that reveal key drivers of allocation changes is therefore useful, but such products carry development overhead and occasionally flatten nuances in the underlying optimization heuristics. Computational robustness is also an operational limitation. Even on high-grade hardware and with the most modern solvers, unforeseen spikes in problem size or data aberrations can cause solver failure or unsustainable runtimes.

allback heuristics like, for example, rule-based allocation or proportional allocation must be invoked in order to help provide for continuity of operation, but the latter will materially yield different outcomes and lose trust if invoked on a regular basis. Predefining switching thresholds to automatic heuristic modes and clearly specified communication protocols to alert human operators are requirements. Legal liability is also a constraint. When an algorithmic recommendation leads to adverse clinical outcomes—e.g., a patient deteriorates after delayed transfer—the issue of fault can become an issue. [27]

Developers can be subject to product-liability suits, hospitals can be held liable for malpractice, and insurers can deny coverage. Such legal risks might lead to conservative parameter settings that minimize efficiency gains through playing safe with over-allocation. Organizations' risk-management departments will therefore need to be involved in governance and planning of the optimization platform, and matching documentation of decision rules will need to be kept for forensic analysis in case of adverse outcomes. These disclaimers suggest that the design outlined above, as theoretically elegant as it is, is not plug-and-play or one-size-fits-all. Real-world operating efficacy will entail iterative tweaking, policy adjustment, stakeholder engagement, and ongoing surveillance.

The utility of the model also therefore depends on not only its mathematical elegance but the socio-technical context in which it exists. Future work will continue hybrid robustness techniques blending rapid reoptimization and probabilistic safeguarding, human-scale escalation triggers, and reward-compatible systems that harmonize private facility objectives with local needs. Only by surmounting such multi-dimensional constraints will the full promise of real-time analytics-based resource optimization be maximally realized in the healthcare environment.

8 CONCLUSION

The theoretical framework described throughout this paper delineates a structured approach to dynamically optimizing health resource allocation throughout regional networks through the integration of real-time analytics with mathematical optimization. Leveraging prior research based on discrete, static planning horizons, the model makes the argument that data-driven feedback mechanisms can systematically balance operational effectiveness with cost control, even in complex and stochastic clinical environments. The proposal rests on three interdependent pillars: a precise optimization formulation that balances the cost of allocations against penalties for unmet demand, an analytics architecture that can ingest and transform continuous streams of data into actionable parameters, and an iterative decision logic that rescales recommendations as conditions evolve. Within this structure, every cycle of data acquisition essentially equates to an empirical snapshot of the regional ecosystem, and every solve-implement cycle constitutes a methodical attempt to re-configure scarce resources for alignment with current clinical priorities [28]. In theory, this iterative cycle between sensing and deciding can minimize the inertia that has historically infested hospital scheduling, inventory provisioning, and inter-facility coordination. But the very same properties that promise responsiveness create questions, for the pursuit of near-continual reoptimization inexorably runs up against sociotechnical, ethical, and computational issues outside the purely mathematical space of the model.

In terms of systems thinking, the most important contribution of the model is its potential to render clear the often inscrutable trade-offs among cost, quality, and timeliness of care. By converting resources and demand into numerical terms and assigning calibrated cost coefficients and penalty functions to allocation choices, the framework makes what are frequently intuitive or tacit managerial judgments explicit. Transparency can underlie governance processes

that must entail clear expression of priorities, particularly when various stakeholders have conflicting goals. For instance, a chief financial officer would prioritise marginal cost savings above everything else, and clinical leadership, reducing patient wait times or specialist staff availability to cover emergent cases. The model's objective function allows these competing priorities to be measured, weighed, and revisited in a formalized discussion, thereby fostering a data-driven decision culture without requiring a particular normative methodology. Secondly, since the optimization is solved multiple times, changing strategic priority can be incorporated gradually—penalty parameters may be raised or lowered to align with changing policy priorities—without across-the-board redesign. Essentially, the model provides a controllable mathematical board upon which regulators may trace shifting institutional desires with methodological constancy.

A second key consequence follows from the model's management of uncertainty. Whereas the formulation itself remains determinstic at each instant of decision-making, uncertainty is addressed implicitly through the short cadence of reoptimization and, where desired, scenario-sensitive adjustment of input parameters. The approach acknowledges that healthcare operations themselves too seldom present the luxury of perfect foresight, particularly with episodic bursts of demand or supply interruptions [29]. Instead of inserting a difficult stochastic program that could be infeasible to solve within operational time constraints, the model uses regular feedback to estimate robustness. It implicitly relies on the assumption that the next decision period will occur soon enough to rectify any forecast error that is discovered during the current interval. This nimble philosophy takes conceptual ideas from model predictive control in engineering and algorithmic trading strategy in finance, both of which substitute rapid reoptimization with full probabilistic hedging. The tradeoff, naturally, is that purely disastrous events—those that inundate the system faster than the feedback loop can respond—are still an ongoing risk. Accordingly, contingency planning outside the optimization horizon may still be required, e.g. in the guise of strategic buffers or emergency measures overriding normal cost-driven recommendations.

From an implementation point of view, arguably the biggest challenge is data governance and interoperability. The proposed analytics pipeline presumes some degree of digital maturity and harmonization between sites that is not yet universal. For example, real-time streams from electronic health records, device telemetry, and workforce management systems must be synchronized into a common schema with low latency. Such integration usually requires heavy investment in middleware, data standards, and information-exchange policies, all of which require organizational negotiation as well as technological deployment. Also, precise data and provenance are critical; one faulty feed—say, a distorted staffing number due to a clock-out glitch—could propagate into misallocations with tangible clinical impacts. These practical realities do not diminish the theoretical importance of the model, but they do set a realistic constraint on short-term implementability, especially in disjointed health markets where digital platforms and incentive regimes are heterogeneous [30]. A sound strategy of implementation could then begin with low-scope pilots in integrated delivery systems or between geographically close hospitals already collaborating in data-sharing agreements and gradually expand as technical legitimacy and organizational comfort increase.

Another dimension, however, requiring careful attention is the relation between professional discretion and algorithmic recommendation. Medicine is an area where decision-making carries extremely heavy ethical implications, and clinicians would naturally anticipate discretion to conform to the idiosyncratic traits of individual patients. The model's distributional choices are largely focused on macro-level distribution of resources—bed allocation, equipment placement, staffing complements—rather than micro-level clinical habits. Yet frontline professionals will necessarily experience the downstream effects of these choices in day-to-day workflow. If, for example, the system frequently redeploys cohort nurses or redeploys ventilators among intensive care units, subsequent workflow adaptations could impact line-of-sight supervision, continuity of care, and staff morale. Designers of systems thus need to integrate explainability mechanisms through which stakeholders can question the justification of particular recommendations. Such explainability can reduce top-down control perceptions and enable educated acceptance, thus lowering the chance of resistance or policy non-adherence. In effect, then, the optimization engine has to be a decision support colleague, not a black-box despot, upholding the integrity that mathematical efficacy is a means of enhancing, not degrading, the human values informing healthcare delivery.

Computation itself is both an enabler and a constraint. While today's solvers are amazingly quick, it is still potentially non-trivial to determine an optimal setpoint for a large mixed-integer program with dozens of facility and resource types within strict real-time constraints, particularly when integer limits, fairness staffing rules, or nonlinear cost functions are in play. The model's default linear approximation is consequently a trade-off between solvability and accuracy, capturing first-order cost dynamics while enabling almost instantaneous solutions [31]. But. This. simplification overlooks second-order effects such as schedule fatigue, learning curves, or nonlinear overtime premiums. Our research agenda thus. involves examining decomposition strategies, surrogate models, and heuristic approximations that might enhance model fidelity without diminishing tractability. Column generation or Benders decomposition for example might decompose the problem by. resource type or geographic cluster, facilitating parallel solution streams that converge to a master allocation plan.

Consequently, machine-learned surrogate functions might be able to estimate solution corrections based on historical demand trends while reserving full optimization for big jumps in system state. Hybrid approaches might be able to find an operational compromise between precision and speed and retain the real-time responsiveness that composes the framework's motivating force.

Budgetary considerations extend beyond the proximal operating cost coefficients of the model. Implementing the analytics-optimization loop means incurring initial and ongoing capital investments for data infrastructure, software licensing, and change management. Estimates of return on investment thus should equate initial and continued expense against anticipated savings in operations and quality gains. Because budgets for healthcare tend to be siloed by department or by facility, cross-department or cross-facility savings might prove politically infeasible to reinvest. For example, a local choice which redistributes elective patients from an overbooked urban hospital to a suburban affiliate can reduce total overtime cost, but the suburban facility can take on added workload without attendant budget recognition [32]. Aligning monetary incentives with system-wide optimization is thus critical; shared savings programs, centralized funding for overflow costs, or performance bonuses tied to aggregate efficiency metrics could be the approach to align motivations. The theoretical model will necessarily be based on a single objective function, but without coordinated fiscal systems, operating decisions are likely to revert to parochial considerations despite sophisticated analytics.

Moving on to patient outcomes, the model's current design monitors service performance mostly through unmet demand penalties—a proxy for wait time reduction and capacity adequacy. However, healthcare quality consists of a set of elements including continuity, safety, and patient experience. Subsequent implementations would use soft constraints or multi-objective formulations capturing these dimensions. Care continuity, e.g., would be captured through penalizing very frequent patient transfer or staffing rotations disrupting provider-patient continuities. Safety considerations—e.g., nurse to patient minimum ratios—can be built in as hard constraints. Patient experience measures, while harder to quantify real-time, can be approximated through real-time sentiment or satisfaction scores collected using digital check-in and surveys. Balancing these non-monetary goals against cost factors would involve either scalarization methods or Pareto frontier analyses to uncover trade-off possibilities to decision-makers. The theoretical work here therefore lays a foundation but leaves scope intentionally for more qualitative quality-of-care improvement.

Ethical stewardship also extends more generally to issues of equity. Local optimization that does nothing more than minimize cost risks reallocating resources in a way that exacerbates inequities, favoring high-density cities over low-income or rural segments whose demand arrives in the form of smaller but necessary [33]. Inclusion of equity-sensitive language—such as minimum levels of service or fairness constraints that ensure proportional access to resources across geographic or demographic communities—can avoid this consequence. Secondly, the governance process needs to include a diverse range of stakeholders, including citizens, in order to decide on suitable equity thresholds and scrutinize algorithmic results. Frequent audit procedures are suggested, searching for systematic bias or drift. While these variables add complexity, they also support the legitimacy of the system so that computational efficiency does not take precedence over societal commitments to fair and equitable care.

Keeping these different issues in mind, the path from theory to practice is likely to be a proto-centric chain. Early pilots would adopt a small set of decisions—e.g., rescheduling nurse shifts within a single integrated health system—validating the real-time analytics pipeline and solver performance under controlled load. What is discovered through such pilots can be cycled back into model calibration, user interface design, and change-management planning. Future growth can include other resources, apply to decisions across facilities, or introduce predictive analytics with longer lookahead horizons. Throughout, key performance indicators need to be monitored closely to gauge actual gains against baseline operations. Ideally, such evaluations would include not just cost and throughput metrics but also clinical outcome proxies and staff engagement surveys. Successful iterations across time can foster increased adoption, leading to regional deployments that realize the complete potential of the framework. Regardless, continuous tuning is expected; healthcare systems remain in development, and the model must be kept responsive to new care pathways, technological innovations, and policy direction. [34]

REFERENCES

[1] S. Shojaee, N. Hajizadeh, H. Najafimehr, *et al.*, "Bayesian adjustment for trend of colorectal cancer incidence in misclassified registering across iranian provinces.," *PloS one*, vol. 13, no. 12, e0199273–, Dec. 13, 2018. DOI: 10.1371/journal.pone.0199273.

[2] G. Wester, L. Rand, C. Y. Lu, and M. Sheehan, "The ethics of grandfather clauses in healthcare resource allocation.," *Bioethics*, vol. 35, no. 2, pp. 151–160, Oct. 11, 2020. DOI: 10.1111/bioe.12815.

[3] K. Kovacic, S. Matta, K. Kovacic, C. M. Calkins, K. Yan, and M. R. Sood, "Healthcare utilization and comorbidities associated with anorectal malformations in the united states.," *The Journal of pediatrics*, vol. 194, pp. 142–146, Dec. 1, 2017. DOI: 10.1016/j.jpeds.2017.10.010.

[4] K. Bond, L. Sandman, and E. Gustavsson, "137:oral-taking a chance on health: The lottery principle, healthcare resource allocation, and orphan drugs," in *Abstracts*, BMJ Publishing Group Ltd, Apr. 28, 2022, A21.2–A21. DOI: 10.1136/bmjgh-2022-isph.58.

[5] Y. Shen and Z. Sun, "Estimating the spatial correlation and convergence of china's healthcare resources allocation: Evidence from the yangtze river delta region.," *Archives of public health = Archives belges de sante publique*, vol. 80, no. 1, pp. 207–, Sep. 14, 2022. DOI: 10.1186/s13690-022-00958-4.

[6] K. Selvarajah, P. M. Zadeh, Z. Kobti, K. A. Pfaff, and M. Kargar, "A palliative care simulator and visualization framework," in Germany: Springer Singapore, Jun. 6, 2019, pp. 317–327. DOI: 10.1007/978-981-13-8566-7_31.

[7] A. Baghbanian, G. Torkfar, and Y. Baghbanian, "Decision-making in australia's healthcare system and insights from complex adaptive systems theory," *Journal of Health Scope*, vol. 1, no. 1, pp. 29–38, May 15, 2012. DOI: 10.5812/jhs.4623.

[8] A. Briggs, "Rational allocation of resources available for healthcare: Understanding cost effectiveness analyst.," *Journal of the International AIDS Society*, vol. 17, no. 4, pp. 19 492–19 492, Nov. 2, 2014. DOI: 10.7448/ias.17.4.19492.

[9] A. Morton, "Aversion to health inequalities in healthcare prioritisation: A multicriteria optimisation perspective.," *Journal of health economics*, vol. 36, pp. 164–173, Apr. 13, 2014. DOI: 10.1016/j.jhealeco.2014.04.005.

[10] K. Sharkey and L. Gillam, "Should patients with self-inflicted illness receive lower priority in access to healthcare resources? mapping out the debate," *Journal of medical ethics*, vol. 36, no. 11, pp. 661–665, Sep. 3, 2010. DOI: 10.1136/jme.2009.032102.

[11] B. Yuan, J. Li, Z. Wang, and L. Wu, "Household registration system, migration, and inequity in healthcare access.," *Healthcare (Basel, Switzerland)*, vol. 7, no. 2, pp. 61–, Apr. 11, 2019. DOI: 10.3390/healthcare7020061.

[12] D. G. Smithard and J. Haslam, "Covid-19 pandemic healthcare resource allocation, age and frailty.," *The New bioethics : a multidisciplinary journal of biotechnology and the body*, vol. 27, no. 2, pp. 127–132, Apr. 3, 2021. DOI: 10.1080/20502877.2021.1917101.

[13] J. Yue, Q. Fu, Y. Zhou, *et al.*, "Evaluating the accessibility to healthcare facilities under the chinese hierarchical diagnosis and treatment system.," *Geospatial health*, vol. 16, no. 2, Nov. 3, 2021. DOI: 10.4081/gh.2021.995.

[14] K. K. Venkatesh and N. Kumarasamy, "The number of hiv-infected individuals decreases by half in india: Impact on the global epidemiology of hiv/aids," *Future HIV Therapy*, vol. 2, no. 5, pp. 395–398, Sep. 19, 2008. DOI: 10.2217/17469600.2.5.395.

[15] K. Fushimi, H. Hashimoto, Y. Imanaka, *et al.*, "Functional mapping of hospitals by diagnosis-dominant case-mix analysis," *BMC health services research*, vol. 7, no. 1, pp. 50–50, Apr. 10, 2007. DOI: 10.1186/1472-6963-7-50.

[16] C. D. Costanzo, "Healthcare resource allocation and priority-setting. a european challenge.," *European journal of health law*, vol. 27, no. 2, pp. 93–114, Mar. 2, 2020. DOI: 10.1163/15718093-12271448.

[17] G. Badano, "Substance in bureaucratic procedures for healthcare resource allocation: A reply to smith," *Journal of medical ethics*, vol. 45, no. 1, pp. 75–76, Jul. 26, 2018. DOI: 10.1136/medethics-2018-104932.

[18] R. Cookson, C. McCabe, and A. Tsuchiya, "Public healthcare resource allocation and the rule of rescue," *Journal of medical ethics*, vol. 34, no. 7, pp. 540–544, Jun. 30, 2008. DOI: 10.1136/jme.2007.021790.

[19] J. Y. C. Yip, "Healthcare resource allocation in the covid-19 pandemic: Ethical considerations from the perspective of distributive justice within public health," *Public health in practice (Oxford, England)*, vol. 2, pp. 100 111–, Mar. 28, 2021. DOI: 10.1016/j.puhip.2021.100111.

[20] M. Schlander, "Hta agencies need evidence-informed deliberative processes comment on "use of evidence-informed deliberative processes by health technology assessment agencies around the globe".," *International journal of health policy and management*, vol. 10, no. 3, pp. 158–161, Mar. 1, 2021. DOI: 10.34172/ijhpm.2020.22.

[21] C. Munthe, D. Fumagalli, and E. Malmqvist, "Sustainability principle for the ethics of healthcare resource allocation.," *Journal of medical ethics*, vol. 47, no. 2, pp. 90–97, Nov. 5, 2020. DOI: 10.1136/medethics-2020-106644.

[22] B. Ip, L. Au, A. Chan, *et al.*, "Express: Evolving ischemic stroke subtypes in 15 years: A hospital-based observational study," *International journal of stroke : official journal of the International Stroke Society*, vol. 17, no. 4, pp. 17 474 930 211 005 953–

17 474 930 211 005 953, Apr. 7, 2021. DOI: 10.1177/17474930211005953.

[23] J. Walter, "The evolving science of disorders of consciousness calls for an inclusive framework for healthcare resource allocation.," *AJOB neuroscience*, vol. 12, no. 2-3, pp. 151–153, May 7, 2021. DOI: 10.1080/21507740.2021.1904035.

[24] A. A. Jaffer, P. J. Karanicolas, L. E. Davis, *et al.*, "The impact of tranexamic acid on administration of red blood cell transfusions for resection of colorectal liver metastases.," *HPB : the official journal of the International Hepato Pancreato Biliary Association*, vol. 23, no. 2, pp. 245–252, Jul. 5, 2020. DOI: 10.1016/j.hpb.2020.06.004.

[25] S. Sinclair, "How to avoid unfair discrimination against disabled patients in healthcare resource allocation," *Journal of medical ethics*, vol. 38, no. 3, pp. 158–162, Dec. 3, 2011. DOI: 10.1136/medethics-2011-100093.

[26] N. Smith, C. Mitton, A. Davidson, and I. Williams, "A politics of priority setting: Ideas, interests and institutions in healthcare resource allocation," *Public Policy and Administration*, vol. 29, no. 4, pp. 331–347, Apr. 23, 2014. DOI: 10.1177/0952076714529141.

[27] R. C. Cope, J. V. Ross, M. Chilver, N. Stocks, and L. Mitchell, "Characterising seasonal influenza epidemiology using primary care surveillance data.," *PLoS computational biology*, vol. 14, no. 8, e1006377–, Aug. 16, 2018. DOI: 10.1371/journal.pcbi.1006377.

[28] K. Syrett, "Healthcare resource allocation in the english courts: A systems theory perspective," *Northern Ireland Legal Quarterly*, vol. 70, no. 1, pp. 111–129, Mar. 8, 2019. DOI: 10.53386/nilq.v70i1.235.

[29] F. Tian and J. Pan, "Hospital bed supply and inequality as determinants of maternal mortality in china between 2004 and 2016," *International journal for equity in health*, vol. 20, no. 1, pp. 51–51, Jan. 30, 2021. DOI: 10.1186/s12939-021-01391-9.

[30] H. Haghparast-Bidgoli, A. A. Kiadaliri, and J. Skordis-Worrall, "Do economic evaluation studies inform effective healthcare resource allocation in iran? a critical review of the literature," *Cost effectiveness and resource allocation : C/E*, vol. 12, no. 1, pp. 15–15, Jul. 11, 2014. DOI: 10.1186/1478-7547-12-15.

[31] M. Sarkies, L. M. Robins, M. Jepson, *et al.*, "Effectiveness of knowledge brokering and recommendation dissemination for influencing healthcare resource allocation decisions: A cluster randomised controlled implementation trial," *PLoS medicine*, vol. 18, no. 10, pp. 1–23, Oct. 22, 2021. DOI: 10.1371/journal.pmed.1003833.

[32] P. Mitchell, "What is enough? sufficiency, justice, and health: Carina fourie and annette rid, eds, 2017, oxford university press (new york, ny, 978-0-19-938526-3, 336 pp.)," *Journal of Bioethical Inquiry*, vol. 16, no. 3, pp. 473–475, Aug. 21, 2019. DOI: 10.1007/s11673-019-09936-y.

[33] I. N. Olver, "Application of a framework to guide policy on healthcare resource allocation decisions to data linkage projects," *International Journal of Population Data Science*, vol. 5, no. 5, Dec. 7, 2020. DOI: 10.23889/ijpds.v5i5.1527.

[34] S. Leng and A. Yener, *Handbook of Large-Scale Distributed Computing in Smart Healthcare - Resource Allocation in Body Area Networks for Energy Harvesting Healthcare Monitoring*. Springer International Publishing, Aug. 8, 2017. DOI: 10.1007/978-3-319-58280-1_20.