

La supply chain hospitalière face aux goulots d'étranglement : un cadre analytique intégratif pour les services d'urgences publics marocains

The hospital supply chain facing bottlenecks: an integrative analytical framework for Moroccan public emergency departments

Selma CHADLI, (Doctorante en Sciences de Gestion)

Laboratoire Modélisation Appliquée à l'Économie et à la Gestion (MEAGE)
Faculté des sciences juridiques, économiques et sociales d'Ain Sebâa,
Université Hassan II de Casablanca, Maroc

Kenza IDRISSE, (Docteure en Sciences de Gestion)

Laboratoire Mundiapolis de Recherche en Economie et Management,
Faculté de Management,
Université Mundiapolis CFC, Casablanca Finance City, Maroc

Anwar MEDDAOUI, (Docteur en Génie Industriel et Maintenance)

Laboratoire d'Ingénierie des Systèmes Intelligents, Industriels et Mécaniques
ENSAM
Université Hassan II de Casablanca, Maroc

Abdelhak YAACOUBI

Laboratoire de Modélisation Appliquée à l'économie et à la Gestion (MEAGE)
Faculté des sciences juridiques, économiques et sociales d'Ain Sebâa,
Université Hassan II de Casablanca, Maroc

Adresse de correspondance :	Faculté des sciences juridiques, économiques et sociales d'Ain Sebâa BP : 2634 •Route des Chaux et Ciments Beausite, Casablanca 20254 Téléphone : 05 22 34 34 82
Déclaration de divulgation :	Les auteurs n'ont pas connaissance de quelconque financement qui pourrait affecter l'objectivité de cette étude. Ils assument l'entière responsabilité de tout éventuel plagiat, de l'usage de l'intelligence artificielle dans la rédaction, ainsi que des résultats présentés dans cet article.
Conflit d'intérêts :	Les auteurs ne signalent aucun conflit d'intérêts.
Citer cet article	CHADLI, S., IDRISSE, K., MEDDAOUI, A., & YAACOUBI, A. (2026). La supply chain hospitalière face aux goulots d'étranglement : un cadre analytique intégratif pour les services d'urgences publics marocains. <i>International Journal of Accounting, Finance, Auditing, Management and Economics</i> , 6(7), 405–422. https://doi.org/10.5281/zenodo.20806221
Licence	Cet article est publié en open Access sous licence CC BY-NC-ND

La supply chain hospitalière face aux goulots d'étranglement : un cadre analytique intégratif pour les services d'urgences publics marocains

Résumé :

Contexte : Les services d'urgences hospitalières publics marocains sont confrontés à une tension croissante entre une demande de soins en forte augmentation, liée à l'extension de la couverture sanitaire universelle, et des processus organisationnels encore fondés sur la règle du premier arrivé, premier servi, indifférente à la gravité clinique des patients. Cette inadéquation génère des phénomènes de congestion structurels dont les causes relèvent de l'organisation interne des flux, et non d'un déficit de ressources.

Gap scientifique : Si la littérature internationale a largement documenté les mécanismes de saturation des urgences, les travaux spécifiques au contexte hospitalier public marocain restent rares et fragmentaires (El Fakir et al., 2016). Aucun cadre analytique intégratif n'articule explicitement, pour ce contexte, les théories managériales de focalisation, la modélisation mathématique des files d'attente, les outils de cartographie des processus et la perspective logistique de la supply chain hospitalière.

Objectif et cadre analytique : En réponse à ce déficit analytique, cet article propose un cadre intégratif mobilisant quatre corpus complémentaires : la Théorie des Contraintes (TOC), la théorie des files d'attente, les outils Lean-BPM (Value Stream Mapping et Business Process Management) et la perspective de la supply chain hospitalière.

Terrain et résultat principal : Ce cadre est mis à l'épreuve au Service des Urgences de l'Hôpital Mohammed Ben Abdallah de Mohammedia, sur la base de 323 jours d'observation continue et 85 486 admissions (juin 2024 - avril 2025). L'analyse empirique révèle un goulot structurel à l'étape des soins : 59,29 minutes d'attente pour 9,21 minutes de traitement effectif, soit un ratio de 6,4, signature mathématique de la contrainte dominante au sens de Goldratt et Cox (2004).

Contribution : L'article apporte une contribution théorique (modèle intégratif des quatre corpus), méthodologique (démarche reproductible en quatre temps : diagnostic processuel, identification des contraintes, modélisation de la variabilité, conception de l'état futur) et managériale (prescriptions opérationnelles à ressources constantes pour les urgences publiques marocaines).

Mots clés : Service des urgences, Théorie des Contraintes, Supply chain hospitalière, Lean-BPM, Goulots d'étranglement, Contexte marocain

Classification JEL : I10, I11, M11, L23

Type du papier : Recherche conceptuelle avec illustration empirique

Abstract:

Context: Moroccan public hospital emergency departments face growing tension between increasing healthcare demand and organizational processes still governed by first-come, first-served rules, indifferent to patient clinical severity. This mismatch generates structural congestion phenomena that are organizational in nature, not capacity.

Gap: Despite substantial international literature on emergency department crowding, research specific to the Moroccan public hospital context remains scarce and fragmented (El Fakir et al., 2016). No integrative analytical framework explicitly articulates, for this context, managerial focusing theories, mathematical queuing models, process mapping tools and the hospital supply chain perspective.

Objective and framework: This article proposes an integrative framework combining four complementary theoretical corpora: the Theory of Constraints (TOC), queuing theory, Lean-BPM tools (Value Stream Mapping and Business Process Management) and the hospital supply chain perspective.

Empirical application and main result: This framework is applied to the Emergency Department of Mohammed Ben Abdallah Hospital in Mohammedia, based on 323 consecutive days of observation and 85,486 admissions (June 2024-April 2025). The analysis reveals a structural bottleneck at the care stage: 59.29 minutes of waiting for 9.21 minutes of treatment, yielding a ratio of 6.4, the mathematical signature of the dominant system constraint.

Contribution: The article makes a theoretical contribution (an integrative model of four corpora), a methodological contribution (a reproducible four-step approach: process diagnosis, constraint identification, variability modeling, future state design) and a managerial contribution (operational prescriptions achievable at constant resources for Moroccan public emergency departments).

Keywords: Emergency department, Theory of Constraints, Hospital supply chain, Lean-BPM, Bottlenecks, Moroccan context

JEL Classification: I10, I11, M11, L23

Paper type: Conceptual Research with Empirical Illustration

1. Introduction

Il est, dans les couloirs des services d'urgences des hôpitaux publics marocains, une heure de la journée, souvent le lundi matin, où la salle d'attente ressemble à un carrefour sans signalisation. Des dizaines de patients, aux gravités cliniques radicalement différentes, s'y retrouvent confondus dans une même file, sous la seule autorité de l'ordre d'arrivée. Certains peuvent attendre. D'autres, cliniquement, ne le peuvent pas. Aucune règle ne les distingue. C'est de ce constat, trivial en apparence mais décisif en réalité, que naît la problématique du présent article. Le Maroc a consenti, depuis le début des années 2000, des efforts institutionnels considérables en faveur de l'accès aux soins : généralisation du Régime d'Assistance Médicale (RAMED), ambition de la Couverture Sanitaire Universelle inscrite dans les orientations royales, déploiement du réseau hospitalier public sur l'ensemble du territoire (MSPS, 2022 ; World Bank, 2022). Ces réformes ont porté leurs fruits : l'espérance de vie à la naissance dépasse désormais 77 ans, la mortalité infantile a été divisée par plus de cinq en quatre décennies, passant de 100 pour mille naissances vivantes dans les années 1980 à moins de 19 pour mille aujourd'hui (HCP, 2022 ; OMS, 2023). Ces acquis sont incontestables, et il serait injuste de ne pas les reconnaître.

Pourtant, à l'intérieur des hôpitaux publics, cette transformation de l'accès aux soins se répercute sur des structures organisationnelles qui n'ont pas évolué au même rythme. Selon le Livre Blanc 2020 de la Société Marocaine de Médecine d'Urgence, plus de six millions de patients se rendent annuellement aux urgences hospitalières publiques, avec une progression de près de 10 % par an (SMMU, 2020). Cette pression quantitative s'accompagne d'une inadéquation qualitative documentée : le Ministre de la Santé déclarait devant la Chambre des Conseillers que 80 % des prestations dispensées aux urgences hospitalières relevaient de soins non urgents (CESE, 2023), traduisant l'ampleur du report de la demande primaire sur ces services en l'absence de mécanisme de différenciation à l'entrée.

La gestion des flux patients y reste largement fondée sur des logiques administratives héritées, où chaque unité fonctionne selon sa propre rationalité et où le parcours du patient n'est pas pensé comme un système à optimiser globalement (Marty, 2003). Les services d'urgences en offrent l'illustration la plus saisissante : accueillant sans sélection préalable une demande composite (urgences vitales, pathologies chroniques en décompensation et consultations relevant ordinairement de la médecine de ville), ils fonctionnent dans la plupart des établissements sans aucun mécanisme formel de hiérarchisation clinique (El Fakir et al., 2016). Or, la littérature internationale est formelle sur ce point. Asplin et al. (2003), dans un modèle analytique devenu central dans le domaine, ont montré que les facteurs de congestion des urgences se distribuent entre les flux entrants, les processus de passage et les capacités d'aval, et que les leviers les plus accessibles se situent précisément dans l'organisation interne des processus, qui relève de décisions managériales et non d'investissements capacitaires. Hoot et Aronsky (2008), après une revue systématique portant sur 93 études, confirment que la saturation résiste aux tentatives d'amélioration qui n'en traitent pas les causes structurelles. Morley et al. (2018), dans une revue systématique portant sur 102 études publiées entre 2000 et 2018, identifient une inadéquation persistante entre les causes de la congestion, majoritairement liées au volume et à la nature des arrivées, et les solutions proposées, essentiellement orientées vers l'organisation interne du flux. La littérature converge vers une hypothèse forte : dans de nombreux contextes hospitaliers, et particulièrement dans ceux à ressources contraintes, c'est davantage la règle d'ordonnancement des flux que le nombre de ressources disponibles qui détermine la performance du service (Goldratt & Cox, 2004 ; Morley et al., 2018). Cette hypothèse guide notre démarche, sans prétendre à une démonstration causale universelle.

C'est dans cet espace, entre le diagnostic bien documenté de la littérature internationale et le vide analytique qui persiste pour le contexte marocain, relevé notamment par El Fakir et al. (2016), que s'inscrit le présent article. Notre objectif est de proposer un cadre analytique intégratif mobilisant quatre corpus théoriques complémentaires : la Théorie des Contraintes (TOC), la théorie des files d'attente, les outils Lean-BPM et la perspective de la supply chain hospitalière. Ce cadre vise à fournir aux gestionnaires et aux chercheurs marocains les outils conceptuels nécessaires pour identifier, modéliser et résoudre les goulots d'étranglement dans leurs services d'urgences, dans le respect des contraintes budgétaires structurelles du secteur public.

La question centrale qui oriente cette recherche est la suivante : dans quelle mesure l'articulation de la Théorie des Contraintes, de la théorie des files d'attente, des outils Lean-BPM et de la perspective de la supply chain hospitalière permet-elle de construire un cadre analytique cohérent et opérationnellement pertinent pour la gestion des flux patients dans les services d'urgences publics marocains ?

Cette question se décline en trois sous-questions complémentaires. La première est théorique : quels sont les fondements et les apports spécifiques de chacun des quatre corpus, et selon quelle logique articulatoire forment-ils un cadre cohérent ? La deuxième est analytique : dans quelle mesure ce cadre permet-il de diagnostiquer et de modéliser les goulots d'étranglement observés dans un service d'urgences public marocain ? La troisième est prescriptive : quelles recommandations opérationnelles ce diagnostic autorise-t-il pour améliorer la performance à ressources strictement constantes ?

Sur le plan des apports, cet article contribue à trois niveaux distincts. La contribution théorique réside dans la proposition d'un modèle intégratif articulant quatre corpus, appliqué explicitement au contexte hospitalier public marocain. La contribution méthodologique tient à la formalisation d'une démarche reproductible en quatre temps dont les étapes, les outils et les livrables sont clairement définis. La contribution managériale, enfin, consiste en des prescriptions opérationnelles permettant d'améliorer la performance des urgences à ressources budgétaires constantes, enjeu central dans le secteur public marocain.

Pour répondre à ces questions, l'article procède en deux temps : une revue critique et sélective de la littérature organisée autour des quatre cadres mobilisés (section 2) ; puis une illustration empirique adossée à une étude de cas unique, fondée sur une observation continue de 20 mois et des mesures chronométriques de processus dans un service d'urgences marocain (section 3), suivie d'une discussion des résultats et limites (section 4).

2. Revue de la littérature

2.1. Le service des urgences : un système organisationnel ouvert et non linéaire

Appréhender le service des urgences par le seul prisme des activités cliniques qui s'y déroulent, c'est manquer l'essentiel de sa complexité organisationnelle. Pour en saisir la logique profonde, et comprendre d'où viennent ses dysfonctionnements, il faut d'abord le considérer pour ce qu'il est : un système ouvert, dynamique et non linéaire, dont le comportement ne peut être déduit de la somme de ses composantes prises isolément.

La Théorie Générale des Systèmes, formalisée par von Bertalanffy (1968) puis adaptée aux organisations humaines par Le Moigne (1977), fournit ici un cadre heuristique indispensable. Von Bertalanffy montrait que c'est l'organisation des relations entre les éléments d'un système, et non les éléments eux-mêmes, qui génère les propriétés observables de l'ensemble. Transposée au service des urgences, cette perspective permet de comprendre pourquoi optimiser localement chaque poste de travail ne produit pas nécessairement une amélioration globale : les interactions entre sous-systèmes médical, administratif et logistique peuvent annuler, voire inverser, des gains obtenus en un point du processus.

Le service des urgences entretient des échanges permanents de matière (patients, médicaments, équipements), d'énergie (ressources humaines, compétences) et d'information (données cliniques, décisions thérapeutiques) avec un environnement institutionnel complexe : services hospitaliers d'aval, partenaires préhospitaliers, structures médico-sociales (Asplin et al., 2003 ; Marty, 2003). Sa dynamique est structurellement marquée par une variabilité des flux entrants qui confère au système un comportement non linéaire : une légère variation dans les conditions d'entrée peut engendrer des perturbations considérables dans l'ensemble du processus (Plsek & Greenhalgh, 2001).

C'est précisément cette non-linéarité qui rend inefficaces les approches gestionnaires fondées sur des équilibres stables ou des prévisions linéaires. Hoot et Aronsky (2008) identifient trois catégories de facteurs de congestion (entrants, de passage et de sortie) et établissent que les leviers les plus accessibles sont ceux qui agissent sur l'organisation interne des processus, laquelle relève de décisions managériales. Lynn et Kellermann (1991) le formulent de manière encore plus tranchante : les urgences ne sont pas saturées parce qu'elles manquent de ressources ; elles sont saturées parce qu'elles sont organisées de manière inadaptée à la diversité de leur demande. C'est cette inadaptation organisationnelle que les cadres théoriques suivants permettent de diagnostiquer et de corriger.

2.2. La Théorie des Contraintes : de l'identification du goulot à son traitement

Parmi les approches théoriques mobilisées dans le champ du management hospitalier, la Théorie des Contraintes (Theory of Constraints, TOC) occupe une position singulière : elle est à la fois la plus simple dans son principe et la plus radicale dans ses implications managériales. Développée par Eliyahu Goldratt à partir de son roman *The Goal* (Goldratt & Cox, 2004), elle repose sur un postulat dont la portée n'est souvent appréciée à sa juste valeur qu'a posteriori, une fois mis à l'épreuve du terrain : la performance globale d'un système n'est pas déterminée par l'efficacité moyenne de ses composantes, mais par celle de son maillon le plus faible, la contrainte. Il s'ensuit que l'effort d'amélioration, aussi bien intentionné soit-il, ne produit aucun effet sur la performance globale s'il ne cible pas la contrainte dominante.

L'opérationnalisation de ce principe passe par un processus de focalisation itératif en cinq étapes : identifier la contrainte ; l'exploiter au maximum de la capacité existante, sans engager de ressources supplémentaires ; subordonner l'ensemble du système à son rythme ; élever la contrainte lorsque les mesures d'exploitation ont été épuisées ; revenir enfin à la première étape pour identifier la nouvelle contrainte (Goldratt, 1990). Ce cycle n'est pas seulement un outil d'amélioration continue : c'est une discipline de pensée qui oblige à distinguer ce qui importe vraiment, la contrainte, de ce qui attire l'attention, les performances locales.

La transposition de la TOC aux services hospitaliers, initiée par Kershaw (2000) et prolongée par une littérature empirique aujourd'hui substantielle, a confirmé la fertilité de cette transposition. Bhattacharya et al. (2010) ont documenté des réductions significatives des temps d'attente dans des urgences britanniques, sans augmentation des effectifs. Rohleder et al. (2011) rapportent une diminution de 30 % des délais d'attente dans un service canadien, à ressources strictement constantes, par la seule réorganisation du processus autour de la contrainte identifiée. Litvak et Bisognano (2011) ont montré que la gestion de la variabilité des flux, levier accessible à coût quasi nul, produit des gains comparables à ceux obtenus par des investissements capacitaires. Ronen et al. (2018) ont par la suite formalisé ces principes dans un manuel dédié à l'amélioration hospitalière combinant TOC et Lean, et montrent que l'identification du maillon contrainte constitue, dans des environnements à ressources limitées, le préalable incontournable à toute démarche d'amélioration durable.

Dans les hôpitaux publics marocains, la pertinence de la TOC tient précisément aux conditions dans lesquelles elle génère le plus de valeur : celles de la contrainte budgétaire et de l'abondance de ressources organisationnelles inexploitées. Le Tableau 1 propose une taxonomie des

contraintes spécifiques à ce contexte, élaborée à partir de la littérature et de nos observations de terrain.

Tableau 1 : Taxonomie des contraintes dans les services d'urgences publics marocains¹

Type de contrainte	Nature	Exemples concrets	Conséquence sur les flux	Levier TOC	Références
Physique	Ressources matérielles	Boxes d'examen, lits d'observation, équipements d'imagerie	Saturation rapide lors des pics d'affluence saisonniers	Exploiter la capacité existante avant tout investissement	Asplin et al. (2003) ; Hoot & Aronsky (2008)
Humaine	Ressources en personnel	Effectif médical/paramédical insuffisant en heures de pointe	Allongement des délais de première évaluation clinique	Réorganiser les rotations en fonction des pics identifiés	Visser & Beech (2005) ; OMS (2023)
Organisationnelle	Règles et processus internes	Absence de protocole de triage, règle FIFO sans différenciation clinique	Désordre dans la priorisation ; patients graves en attente derrière des cas bénins	Modifier la règle : du FIFO vers l'ordonnancement clinique	Goldratt (1990) ; Lynn & Kellermann (1991) ; El Fakir et al. (2016)
Politique interne	Procédures institutionnelles	Paiement exigé avant toute prise en charge clinique	Retard clinique introduit avant même le premier contact médical	Décision managériale et arbitrage institutionnel	Visser & Beech (2005) ; MSPS (2022)
Aval hospitalier	Saturation des services d'aval	Indisponibilité des lits d'hospitalisation, délais de transfert prolongés	Blocage des sorties du SU, accumulation des patients traités	Coordination à l'échelle hospitalière, gestion anticipée des lits	Asplin et al. (2003) ; Broyles & Cochran (2011)
Informationnelle	Qualité et circulation de l'information	Absence de SIH intégré, dossiers papier, traçabilité incomplète	Retards décisionnels, répétitions d'exams, erreurs d'orientation	Outils simples et partagés ; tableau de triage lisible par tous	Dumas et al. (2013) ; El Fakir et al. (2016)

Source : élaboration des auteurs.

Ce que ce tableau révèle, et que la pratique confirme, c'est que la contrainte dominante dans les SU marocains est le plus souvent de nature organisationnelle (une règle absente, un protocole inadapté, une procédure mal séquencée) et non de nature physique ou humaine. Identifier cette contrainte, l'exploiter, puis subordonner le reste du système à son rythme constitue le fil directeur opérationnel du modèle intégratif que nous proposons.

2.3. La théorie des files d'attente : fondement mathématique de la gestion des priorités

Si la TOC donne le cadre stratégique du diagnostic, la théorie des files d'attente en fournit la grammaire mathématique. Développée par Erlang dans les années 1910 pour les systèmes téléphoniques et formalisée par Kendall (1953) dans sa notation canonique A/B/c, cette théorie modélise les phénomènes d'attente qui surgissent lorsque la demande excède momentanément la capacité de service, situation structurelle dans tout service d'urgences.

Le paramètre central de ces modèles est le facteur d'utilisation $\rho = \lambda / (c \times \mu)$, où λ désigne le taux d'arrivée et μ le taux de service. Une propriété fondamentale, souvent sous-estimée dans la gestion hospitalière, est la non-linéarité de la relation entre utilisation et temps d'attente : pour des valeurs de ρ dépassant 0,7 à 0,8, les temps d'attente croissent de manière exponentielle, non linéaire (Green, 2006). Un service fonctionnant à 85 % de sa capacité théorique peut, du seul fait de la variabilité stochastique des flux, connaître des phénomènes de congestion intenses. La loi de Little (1961), selon laquelle $L = \lambda W$ (le nombre moyen d'entités dans un système est le produit du taux d'arrivée par le temps moyen de séjour), établit quant à elle la relation fondamentale qui relie flux, capacité et délais d'attente, et constitue un outil de diagnostic rapide accessible sans données complexes.

L'apport le plus significatif de la théorie des files d'attente pour la gestion des urgences réside dans les modèles à files à priorités. Contrairement aux modèles M/M/c standards qui traitent

¹ Les colonnes 'Type de contrainte', 'Nature' et 'Levier TOC' procèdent de la synthèse bibliographique (Goldratt, 1990 ; Asplin et al., 2003 ; Visser & Beech, 2005). Les colonnes 'Exemples concrets' et 'Conséquence sur les flux' intègrent les observations de terrain conduites au SU de l'Hôpital M. Ben Abdallah de Mohammedia (2024-2025).

tous les clients identiquement, ces modèles distinguent plusieurs classes de patients selon leur niveau de gravité clinique. Cox et Smith (1961) ont établi mathématiquement qu'une politique de priorisation réduit significativement les temps d'attente des patients les plus graves sans allonger de manière disproportionnée l'attente des patients moins urgents, à condition que le mécanisme de priorisation soit mis en œuvre dès l'entrée dans le système. Cette démonstration mathématique fournit le fondement quantitatif des systèmes de triage formalisés et permet de répondre, données à l'appui, à la résistance institutionnelle que suscite souvent leur introduction.

2.4. Lean-BPM : rendre les processus visibles pour les rendre transformables

Diagnostiquer un goulot d'étranglement est une chose. Le rendre visible, à tous les acteurs du service et pas seulement aux gestionnaires, en est une autre, sans laquelle aucune transformation organisationnelle durable ne peut être conduite. C'est précisément la fonction des outils Lean-BPM, dont l'articulation constitue l'un des apports méthodologiques les plus robustes de la littérature en management hospitalier des deux dernières décennies.

La Value Stream Mapping (VSM), héritée du Lean Manufacturing (Womack & Jones, 1996) et adaptée aux soins de santé par Mazzocato et al. (2010), est une cartographie schématique du flux d'un patient à travers l'ensemble des étapes de sa prise en charge. Sa force tient à une propriété simple mais décisive : elle rend visible, d'un seul coup d'œil, la proportion du temps de séjour effectivement consacrée à des activités à valeur ajoutée clinique, par opposition aux temps d'attente et aux activités sans valeur directe pour le patient. Dans les services d'urgences, cette proportion s'avère systématiquement défavorable : entre 15 et 25 % du temps de séjour correspond à du soin effectif, le reste étant absorbé par des attentes inter-étapes et des activités sans valeur clinique directe (Mazzocato et al., 2010 ; Radnor et al., 2012). La VSM en rend compte de manière immédiatement communicable aux équipes soignantes.

Le Business Process Management (BPM), via sa notation standard BPMN, complète la VSM en offrant une modélisation rigoureuse des logiques décisionnelles, des conditions de branchement et des responsabilités entre acteurs (Dumas et al., 2013). L'organisation par couloirs de nage (swimlanes) permet de visualiser explicitement les interfaces entre sous-systèmes et d'identifier les points de rupture de coordination, souvent invisibles dans la gestion courante mais particulièrement coûteux en termes de délais. VSM et BPMN ne sont pas substituables : le VSM quantifie les gaspillages, le BPMN modélise les décisions. Leur combinaison dans une démarche Lean-BPM intégrée produit un outil de diagnostic et de transformation d'une efficacité remarquable (Radnor et al., 2012).

2.5. La supply chain hospitalière : le service des urgences comme nœud critique

La quatrième perspective mobilisée dans notre cadre intégratif, celle de la supply chain hospitalière, apporte une dimension que les trois précédentes ne couvrent pas entièrement : la vision réseau, systémique, dans laquelle le service des urgences n'est pas une unité isolée mais un nœud dont les performances sont co-déterminées par celles de ses partenaires en amont et en aval.

La supply chain hospitalière peut être définie comme l'ensemble coordonné des flux physiques, informationnels et financiers qui permettent la production et la livraison des soins, depuis le premier contact du patient avec le système de santé jusqu'à sa sortie définitive de l'hôpital (de Vries & Huijsman, 2011 ; Rivard-Royer et al., 2002). Dans cette perspective, le service des urgences occupe une position stratégique particulière : il est à la fois un point d'entrée majeur dans le système et un nœud d'articulation entre les soins préhospitaliers et les unités d'hospitalisation, dont la saturation se propage mécaniquement en amont et en aval (Visser & Beech, 2005 ; Broyles & Cochran, 2011).

La recherche marocaine commence à investir ce champ : Ziat et al. (2019), dans un travail pionnier sur la supply chain hospitalière au Maroc, proposent un cadre générique pour en

décrire la complexité selon une approche holiste et inductive ; Haboubi et al. (2024) étendent cette perspective à la supply chain hospitalière 4.0, intégrant les apports de l'intelligence artificielle et des systèmes d'information dans la gestion des flux de soins. Cette inscription dans une logique de supply chain présente trois avantages analytiques pour notre cadre. Elle permet d'abord d'identifier les facteurs de performance qui dépassent le périmètre direct du service, notamment la gestion des lits d'aval. Elle permet ensuite d'adapter les outils de gestion des flux logistiques industriels, dont la TOC et la VSM sont précisément des exemples canoniques, à la réalité du flux patient. Elle permet enfin de repenser le triage non comme un simple geste clinique mais comme un mécanisme de régulation logistique : allouer les ressources critiques en priorité aux patients dont l'état le justifie, c'est optimiser l'ensemble de la chaîne de production de soins.

2.6. Limites et tensions critiques des approches mobilisées

La rigueur scientifique impose, avant de proposer une synthèse intégrative, de signaler les tensions et limites que la littérature identifie dans chacun des corpus mobilisés. Trois débats méritent d'être explicitement posés.

La première tension oppose deux conceptions de la congestion. Les tenants d'une lecture processuelle (Goldratt & Cox, 2004 ; Hoot & Aronsky, 2008) soutiennent que les inefficiences organisationnelles constituent la contrainte dominante, remédiable sans investissement capacitaire. À l'inverse, Morley et al. (2018) montrent que la grande majorité des solutions documentées, centrées sur l'organisation interne du flux, ne permettent pas de traiter les causes profondes de la saturation, qui tiennent au volume et à la nature des arrivées, ainsi qu'à l'insuffisance des ressources d'aval. La vérité opérationnelle est probablement entre les deux : l'approche processuelle est nécessaire mais non suffisante, et ses gains atteignent un plafond au-delà duquel seul un renforcement capacitaire permet de progresser.

La deuxième tension concerne la transférabilité du Lean dans le secteur public hospitalier. Radnor et al. (2012) ont documenté les conditions sous lesquelles le Lean échoue dans les organisations publiques : absence de culture de la performance, résistances institutionnelles, faible informatisation des processus. Ronen et al. (2018) soulignent que la définition même du maillon contrainte et du throughput doit être adaptée à une logique de soin non marchande. Pour les urgences marocaines, ces mises en garde invitent à ne pas transposer mécaniquement des protocoles conçus pour des systèmes dotés de ressources et d'outils de gestion très différents.

La troisième tension touche aux limites des modèles de files d'attente en environnement non stationnaire. Si la loi de Little (1961) et les modèles M/M/c fournissent des estimations solides pour des systèmes stables, leur applicabilité est remise en question dans des services soumis à des pics hebdomadaires et saisonniers marqués. Comme nous le verrons dans la section empirique, les données de Mohammedia documentent précisément cette non-stationnarité : l'amplitude de 318 patients entre le jour le plus calme et le plus chargé invalide l'hypothèse de flux stationnaire sur laquelle reposent les modèles analytiques classiques, ce qui justifie le recours à la modélisation ARIMA pour caractériser les dynamiques temporelles du flux.

2.7. Vers un modèle analytique intégratif : synthèse et articulation

Ces tensions reconnues, et les ajustements qu'elles imposent intégrés, l'examen des quatre corpus théoriques développés dans les sections précédentes révèle une convergence vers un même diagnostic opérationnel : la performance d'un service des urgences dépend avant tout de l'organisation des processus qui régissent l'utilisation des ressources disponibles. Chaque corpus apporte, sur cette réalité, un éclairage irréductible à celui des autres.

D'un point de vue épistémologique, ces quatre corpus ne se situent pas au même niveau d'abstraction et n'exercent pas la même fonction dans le cadre intégratif. La Théorie des

Contraintes joue le rôle de cadre structurant : c'est elle qui définit l'objet de l'analyse (la contrainte dominante) et qui oriente les choix d'action. La théorie des files d'attente joue un rôle analytique : elle quantifie mathématiquement les phénomènes identifiés. Les outils Lean-BPM jouent un rôle opérationnel : ils rendent les processus visibles, mesurables et transformables. La supply chain hospitalière, enfin, joue un rôle contextuel : elle inscrit le service dans un réseau dont les dynamiques conditionnent et sont conditionnées par les performances du service. Cette hiérarchisation épistémologique est fondamentale, car elle détermine la séquence logique du modèle intégratif : on ne peut opérationnaliser que ce qu'on a d'abord identifié et quantifié. Elle s'inscrit par ailleurs dans la logique du cycle PDCA (Plan-Do-Check-Act ; Deming, 1986), qui constitue le cadre d'amélioration continue gouvernant la démarche de recherche-action : la TOC et les files d'attente nourrissent la phase Plan (diagnostic et modélisation), le Lean-BPM soutient la phase Do (conception et mise en oeuvre), et la supply chain oriente la phase Check et Act (évaluation des impacts sur l'ensemble du réseau hospitalier).

Le Tableau 2 ci-dessous propose une synthèse des quatre cadres, organisant leurs fondements, leurs outils principaux, leurs applications spécifiques aux urgences et leur contribution respective au modèle intégratif que nous proposons.

Tableau 2 : Synthèse comparative des quatre cadres théoriques mobilisés

Cadre théorique	Fondement	Outil principal	Application aux urgences	Apport au modèle intégratif
Théorie des Contraintes (TOC)	Optimisation systémique (Goldratt & Cox, 2004)	Processus de focalisation en 5 étapes	Identification du goulot dominant ; subordination des ressources	Ciblage de l'action sur la contrainte sans investissement additionnel
Théorie des files d'attente	Modélisation stochastique (Kendall, 1953)	Modèles M/M/c, loi de Little, files à priorités	Quantification des temps d'attente selon la variabilité et la priorité clinique	Fondement mathématique de la politique de triage
Lean-BPM (VSM + BPMN)	Réingénierie processuelle (Womack & Jones, 1996)	Cartographies AS-IS et TO-BE	Visualisation des flux, mesure des gaspillages, identification des ruptures	Visibilité opérationnelle et co-construction des processus cibles
Supply chain hospitalière	Systémique et logistique (de Vries & Huijsman, 2011)	Analyse des flux patients, amont-aval	Positionnement stratégique du SU dans la chaîne de soins globale	Intégration réseau et prise en compte des contraintes d'aval

Source : élaboration des auteurs, d'après Goldratt & Cox (2004), Kendall (1953), Mazzocato et al. (2010) et de Vries & Huijsman (2011)

Ce que ce tableau ne montre pas encore, c'est la manière dont ces quatre cadres s'articulent dans une démarche opérationnelle cohérente. C'est l'objet de la Figure 1, qui formalise le modèle intégratif proposé comme une démarche en quatre temps : un diagnostic systémique et processuel (BPMN AS-IS + VSM), une identification et une classification des contraintes (taxonomie TOC), une modélisation quantitative des effets de la variabilité et de la priorisation (files à priorités), et enfin la conception de l'état futur optimisé (BPMN TO-BE + VSM TO-BE).

La propriété la plus importante de ce modèle, celle qui lui confère une pertinence particulière dans le contexte des hôpitaux publics marocains, est qu'il génère des améliorations de performance sans investissement additionnel. Les quatre cadres mobilisés visent tous, à leur façon, à changer une règle plutôt qu'à acheter une ressource. C'est précisément ce dont les services d'urgences marocains ont besoin.

Figure 1 : Le modèle analytique intégratif proposé pour les services d'urgences publics marocains

THÉORIE DES CONTRAINTES (TOC) Identifier · Exploiter · Subordonner · Élever	THÉORIE DES FILES D'ATTENTE Modéliser · Quantifier · Prioriser	LEAN-BPM (VSM + BPMN) Cartographier · Visualiser · Optimiser	SUPPLY CHAIN HOSPITALIÈRE Intégrer · Coordonner · Piloter
▼ DÉMARCHE EN QUATRE TEMPS ▼			
① DIAGNOSTIC Systémique & processuel (BPMN AS-IS + VSM)	② IDENTIFICATION des contraintes dominantes (taxonomie TOC)	③ MODÉLISATION des effets de la variabilité (files à priorités)	④ CONCEPTION de l'état futur optimisé (BPMN TO- BE + VSM)
RÉSULTAT : Amélioration de la performance organisationnelle SANS investissement additionnel			

Source : élaboration des auteurs

3. Application illustrative au contexte marocain

3.1. Terrain et démarche méthodologique

Le cadre intégratif développé dans la section précédente n'est pas une construction purement abstraite : il a été élaboré en dialogue constant avec un terrain empirique précis : le Service des Urgences de l'Hôpital Mohammed Ben Abdallah de Mohammedia, établissement public de référence de la région Casablanca-Settat. L'accès à ce terrain a été obtenu dans le cadre d'une démarche de recherche-action (Kemmis & McTaggart, 1988), organisée autour du cycle itératif planification, action, observation et réflexion, et conduite en étroite collaboration avec le chef du service et les équipes soignantes.

Les données mobilisées dans cette section sont issues d'un dispositif d'observation continue de 323 jours consécutifs (juin 2024 - avril 2025), articulé autour de quatre instruments complémentaires. Le premier est l'observation directe et continue, conduite sous ses formes participante et non participante, impliquant une présence physique régulière au sein du service pendant la totalité de la période d'investigation. Le deuxième est le chronométrage direct : chaque étape du parcours patient (accueil/enregistrement, paiement des frais, temps d'attente en salle, consultation médicale, soins) a fait l'objet d'une mesure chronométrique individuelle, permettant de construire une base de 85 486 parcours documentés. Le troisième est l'analyse systématique des registres administratifs d'admission, fournissant les données catégorielles (motif, sexe, âge, heure) pour chaque admission. Le quatrième est un ensemble d'entretiens semi-directifs conduits avec les professionnels du service (médecins urgentistes, infirmiers, agents d'accueil, agents de caisse, cadre de santé, direction médicale), permettant de confronter les mesures objectives aux représentations des acteurs et de co-valider le diagnostic. L'unité d'analyse retenue est le passage du patient individuel, défini comme l'intervalle entre l'enregistrement à l'accueil et la sortie du service. La triangulation de ces quatre sources assure la validité interne des résultats présentés (Yin, 2018). Les indicateurs sont mesurés sur la totalité de la période, non sur des échantillons, garantissant l'exhaustivité et la représentativité.

Les indicateurs temporels utilisés dans cet article sont définis opérationnellement comme suit. Le lead time total correspond à l'intervalle, exprimé en minutes, entre le début de l'enregistrement administratif à l'accueil et la sortie effective du patient (retour à domicile ou début du transfert vers une unité d'hospitalisation). Le temps à valeur ajoutée (VA) désigne la durée effectivement consacrée à des actes cliniques ou diagnostiques en contact direct avec le patient. Le temps sans valeur ajoutée (NVA) comprend l'ensemble des temps d'attente inter-étapes et des activités administratives ne produisant pas de bénéfice clinique direct. Le ratio attente/traitement rapporte la durée d'attente observée à l'étape des soins au temps de traitement effectif à cette même étape. Les cas aberrants (valeurs supérieures au 99^e centile) ont été identifiés et vérifiés sur pièce avant inclusion dans les statistiques descriptives.

Sur le plan éthique, cette recherche a bénéficié d'une autorisation formelle de la Direction de l'Hôpital Mohammed Ben Abdallah de Mohammedia et de l'accord du chef du service des urgences. Aucune donnée nominative relative aux patients n'a été collectée ou traitée : les

observations portent exclusivement sur des flux agrégés, des temps de cycle et des statistiques d'admission. L'anonymat des professionnels de santé ayant participé aux entretiens a été rigoureusement préservé. Ces travaux s'inscrivent dans le cadre d'une convention de recherche-action (Kemmis & McTaggart, 1988) entre l'équipe de recherche et l'établissement. Les données du modèle MILP d'ordonnancement et les projections de l'état futur après introduction du triage ne sont pas mobilisées dans cet article, qui se limite à la documentation empirique de l'état actuel (AS-IS).

Tableau 3 : Matrice méthodologique d'articulation des quatre cadres théoriques

Cadre	Objectif	Données requises	Outil mobilisé	Livrable attendu
TOC	Identifier la contrainte dominante et focaliser l'action	Temps de cycle observés, files d'attente, capacités	Processus de focalisation en 5 étapes (Goldratt, 1990)	Taxonomie des contraintes et identification de la contrainte dominante
Files d'attente	Quantifier l'impact de la variabilité et de la priorisation	Taux d'arrivée, taux de service, nombre de serveurs	Modèles M/M/c, loi de Little ($L=xW$), files à priorités	Estimation des temps d'attente selon le niveau de priorité clinique
Lean-BPM (VSM + BPMN)	Cartographier et concevoir l'état futur	Chronométrages directs, flux et ressources par étape	Value Stream Mapping et BPMN AS-IS et TO-BE	Cartographies processus, quantification des gaspillages, état futur
Supply chain hospitalière	Inscrire le SU dans le réseau hospitalier	Données d'aval (lits, taux d'occupation, transferts)	Analyse des flux entrants et sortants, coordination réseau	Identification des contraintes d'aval et scénarios de coordination

Source : élaboration des auteurs, d'après Goldratt & Cox (2004), Kendall (1953), Mazzocato et al. (2010) et de Vries & Huijsman (2011).

3.2. Portrait statistique du flux entrant

Sur la période d'observation, le service a enregistré 85 486 admissions, soit une fréquentation journalière moyenne de 265,5 patients. Ces chiffres ne constituent pas, en soi, une anomalie pour un établissement de référence régionale. Ce qui frappe davantage, c'est l'amplitude des variations : entre le jour le plus calme (98 admissions) et le jour le plus chargé (416 admissions en plein été), la variation atteint 318 patients, soit plus du triple du plancher minimal. Une modélisation ARIMA(1,0,1) validée sur l'ensemble de la série révèle une forte persistance temporelle des niveaux d'activité ($AR(1) = 0,962$) et confirme un effet hebdomadaire hautement significatif : le lundi concentre en moyenne 62 patients supplémentaires par rapport au dimanche ($p < 0,0001$), traduisant un report de la demande accumulée pendant le week-end.

L'analyse désagrégée du flux révèle plusieurs structures importantes. Sur le plan saisonnier, la fréquentation est significativement plus élevée en juillet-août (moyenne : 307 patients/jour, maximum 416 le 29 juillet 2024) et en décembre-janvier (291 patients/jour, avec un pic de 411 patients le 16 décembre 2024), tandis que le printemps et l'automne offrent une charge relativement stable autour de 255 patients/jour. Sur le plan hebdomadaire, la dynamique est marquée : le lundi (+62 patients vs dimanche), le mardi (+38), le mercredi (+31), le jeudi (+30) et le vendredi (+21) concentrent le report de la demande accumulée sur le week-end. Ces deux effets, saisonnier et hebdomadaire, sont structurellement prévisibles et constituent précisément les fenêtres de planification que le modèle ARIMA permet de formaliser et d'anticiper.

Sur le plan démographique, la décomposition du flux entrant révèle la présence de trois sous-populations aux profils cliniques et aux besoins organisationnels distincts. Les hommes représentent 43,9 % des admissions (moyenne : 116,5 patients/jour, écart-type 23,1, maximum 187), les femmes 38,1 % (101,1 patients/jour, écart-type 18,8, maximum 170) et les enfants 18,0 % (47,9 patients/jour, écart-type 12,2, maximum 101). Les enfants présentent un profil de variabilité particulier : leur fréquentation connaît des pics prononcés durant les vacances scolaires (juillet 2024 : 101 enfants en un jour, décembre 2024 : 92 enfants), probablement liés aux infections virales saisonnières et aux accidents domestiques ou sportifs. Ces pics

pédiatriques exigent une anticipation spécifique des ressources, que la règle FIFO est structurellement incapable de garantir.

L'analyse par catégorie clinique constitue l'apport le plus décisif pour comprendre pourquoi l'organisation actuelle génère structurellement des iniquités dans la prise en charge. Les maladies générales (MLD) constituent la catégorie dominante avec 49 411 admissions (56,5% du total), suivies des traumatismes et imprudences (IMP : 22 034, 25,2%), des agressions (AGR : 7 319, 8,4%) et des accidents de la voie publique (AVP : 6 131, 7,0%). Les intoxications (398 cas), les brûlures (140 cas) et les noyades (42 cas) représentent ensemble moins d'1% des admissions en volume, mais concentrent une part disproportionnée des cas vitaux nécessitant une intervention immédiate. Cette hétérogénéité clinique n'est pas homogène dans le temps : les MLD atteignent leurs pics en hiver (pathologies respiratoires, gripes), tandis que les cas d'imprudence et d'AVP s'intensifient en été (activités extérieures, vacances scolaires). Ainsi, non seulement le volume varie selon les jours et les saisons, mais la composition clinique du flux entrant varie qualitativement, rendant d'autant plus nécessaire un mécanisme de différenciation à l'entrée que la règle FIFO ignore totalement.

La lecture clinique du flux révèle par ailleurs une réalité organisationnellement décisive : 70 % des patients présentent des pathologies non urgentes, tandis que 30 % relèvent d'une intervention rapide ou prioritaire, proportions établies par relecture systématique des dossiers selon les critères de l'Emergency Severity Index adapté au contexte local (données primaires, Hôpital Mohammed Ben Abdallah, 2024-2025). Ces deux populations entrent dans le même couloir, franchissent le même guichet administratif, s'acquittent des mêmes formalités financières et s'assoient dans la même salle d'attente. Aucun dispositif ne les distingue à l'entrée. C'est cette indifférence organisationnelle qui génère la contrainte dominante identifiée par l'application de la TOC.

3.3. Diagnostic VSM AS-IS : le coût organisationnel du goulot

La cartographie VSM du processus actuel, conduite sur la base des chronométrages directs et de l'analyse des registres, révèle un tableau saisissant. Sur les 96,38 minutes que passe en moyenne un patient dans ce service, 17,93 minutes seulement correspondent à des activités à valeur ajoutée clinique, soit 18,6 % du temps de séjour total. Les 78,45 minutes restantes (81,4 %) sont du temps d'attente pure, pendant lequel le patient occupe l'espace, génère de l'anxiété et, pour une fraction d'entre eux, se dégrade cliniquement sans qu'aucun mécanisme ne le détecte.

Le goulot principal se situe à l'étape des soins : un patient y attend en moyenne 59,29 minutes avant que l'on touche à son cas, pour un traitement qui durera 9,21 minutes. Le ratio de 6,4 qui en résulte n'est pas une anomalie statistique : c'est, au sens précis de Goldratt et Cox (2004), la signature mathématique d'un goulot d'étranglement : une ressource dont la capacité de traitement est structurellement inférieure à la demande qui lui est adressée, non pas parce que les médecins manquent, mais parce que la règle FIFO (premier arrivé, premier servi) adresse à cette ressource critique des patients bénins avec la même priorité que les patients graves.

Le Tableau 4, le tableau 5 et les Figures 2 et 3 ci-dessous présentent les indicateurs AS-IS qui sont mesurés sur le terrain, et qui détaille la répartition par catégorie clinique.

Ce que ces données illustrent, au-delà de leur valeur descriptive, c'est la validation empirique du cadre intégratif proposé en section 2. La TOC identifie l'étape des soins comme contrainte dominante (ratio 6,4). La théorie des files d'attente prédit que ce ratio résulte de la conjonction entre une forte variabilité des flux entrants (amplitude de 318 patients entre le minimum et le maximum journalier) et une règle de service FIFO insensible à la gravité clinique. Le BPMN et le VSM rend visible le coût organisationnel de cette règle : 81,4 % du temps de séjour perdu. La perspective de la supply chain révèle enfin que ce goulot est amplifié par l'absence de

différenciation à l'entrée, qui mélange dans le même flux des patients dont les besoins de ressources sont radicalement différents.

Tableau 4 : Indicateurs clés de performance du SU de l'Hôpital Mohammed Ben Abdallah de Mohammedia (AS-IS)

Indicateur	Valeur mesurée	Interprétation diagnostique
Période d'observation	Juin 2024 – Avril 2025 (323 jours consécutifs)	Continuité ininterrompue, base de données exceptionnelle
Volume total d'admissions	85 486 admissions	Représentativité statistique incontestable
Fréquentation journalière moyenne	265,5 patients/jour	Min : 98 / Max : 416 (pic estival), soit une amplitude de 318 patients
Profil clinique du flux entrant	70% cas non urgents / 30% cas graves	Coexistence dans un flux unique sans différenciation à l'entrée
Effet hebdomadaire (ARIMA)	AR(1) = 0,962 / +62 patients le lundi vs dimanche	Fort effet de report de la demande, saisonnalité hebdomadaire avérée
Lead time moyen mesuré (AS-IS)	96,38 minutes par patient	Durée totale de séjour (toutes étapes incluses)
Temps à valeur ajoutée (VA)	17,93 min (18,6% du lead time)	Fraction correspondant effectivement à du soin clinique
Temps sans valeur ajoutée (NVA)	78,45 min (81,4% du lead time)	Temps d'attente pure et activités sans valeur clinique directe
Attente à l'étape des soins (goulot)	59,29 minutes	Goulot principal identifié, soit 9,21 min de traitement effectif
Ratio d'attente / traitement	6,4 : signature mathématique du goulot	Un patient attend 6,4 fois le temps que durera son soin

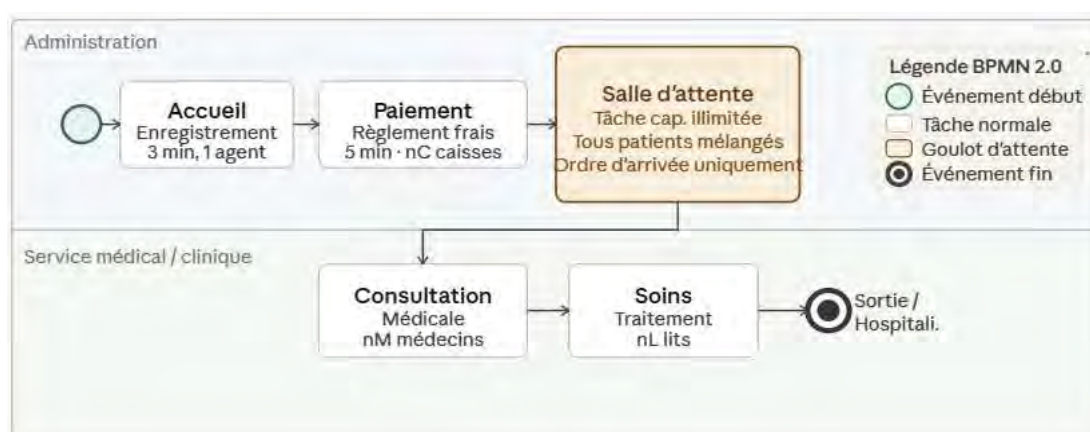
Source : données primaires collectées par les auteurs (juin 2024 – avril 2025)

Tableau 5 : Répartition des admissions par catégorie clinique, SU Hôpital M. Ben Abdallah de Mohammedia (juin 2024 - avril 2025)

Catégorie clinique	Total admissions	% du flux	Profil de risque
Maladies générales (MLD)	49 411	56,5 %	Pics en hiver (infections respiratoires, grippe)
Traumatismes / imprudences (IMP)	22 034	25,2 %	Pics en été (activités extérieures, vacances)
Agressions (AGR)	7 319	8,4 %	Distribution régulière, prédominance masculine
Accidents voie publique (AVP)	6 131	7,0 %	Pics estivaux et jours fériés (x3 les weekends)
Intoxications (INT)	398	0,5 %	Faible volume mais gravité maximale : prise en charge immédiate obligatoire
Brûlures	140	0,2 %	Cas nécessitant une orientation spécialisée rapide
Noyades	42	< 0,1 %	Caractère saisonnier (été), pronostic vital engagé

Source : données primaires collectées par les auteurs. Analyse par catégorie clinique sur 323 jours d'observation.

Figure 2 représente la cartographie BPMN du processus actuel, permettant de visualiser la structure organisationnelle qui génère les indicateurs présentés ci-dessus.



Source : élaboration des auteurs. Notation BPMN 2.0.

Figure 3 : Répartition du temps de séjour AS-IS (Value Stream Mapping simplifié)

LEAD TIME TOTAL MESURÉ : 96,38 MINUTES PAR PATIENT (AS-IS)	
VA : 17,93 min (18,6% du séjour) ✓ Soin effectif	NVA : 78,45 min (81,4% du séjour) Attente pure, dont 59,29 min au goulot des soins
18,6%	81,4% >> GOULOT PRINCIPAL : ratio attente/traitement = 6,4
<i>Un patient attend en moyenne 59,29 minutes pour recevoir 9,21 minutes de traitement effectif</i>	

Source : élaboration des auteurs à partir des données chronométriques collectées sur le terrain

4. Discussion

Les résultats présentés dans la section précédente invitent à plusieurs niveaux de réflexion, qui excèdent la seule portée descriptive du cas étudié.

L'apport de chacun des quatre cadres théoriques dans l'interprétation des résultats empiriques peut être explicitée point par point. La TOC identifie l'étape des soins comme contrainte dominante (ratio 6,4) et prescrit de l'exploiter en priorité avant toute réflexion capacitaire. La théorie des files d'attente explique mathématiquement pourquoi ce ratio est aussi élevé : le facteur d'utilisation ρ de l'étape des soins, combiné à la forte variabilité des arrivées (amplitude de 318 patients entre jour minimum et maximum), maintient le système structurellement au-dessus du seuil d'emballlement exponentiel des attentes. Les outils Lean-BPM rendent ce diagnostic visible : la VSM révèle que 81,4 % du temps de séjour est du gaspillage organisationnel, non du soin ; le BPMN montre que l'absence de triage à l'entrée est la cause racine structurelle de ce gaspillage. Enfin, la perspective de la supply chain hospitalière révèle que la contrainte ne peut être pleinement résolue au niveau du seul service des urgences : la résorption du goulot à l'étape des soins dépend aussi de la disponibilité des lits d'aval, laquelle relève d'une coordination à l'échelle du réseau hospitalier.

En premier lieu, le ratio d'attente/traitement de 6,4 observé à l'étape des soins trouve un écho direct dans la littérature internationale. Rohleder et al. (2011) rapportent des ratios comparables dans des urgences nord-américaines ; les études de Mazzocato et al. (2010) sur le Lean Healthcare documentent systématiquement des proportions de temps à valeur ajoutée inférieures à 25 %. Ce que le cas de Mohammedia apporte de spécifique, c'est la preuve empirique que ces ordres de grandeur ne sont pas propres aux systèmes de santé à hauts revenus : ils caractérisent aussi les services d'urgences publics marocains, avec la même intensité, et pour les mêmes raisons structurelles.

En second lieu, la convergence des quatre cadres théoriques vers un diagnostic commun, à savoir l'absence de mécanisme de différenciation clinique à l'entrée comme source première du goulot, confirme la pertinence empirique du modèle intégratif proposé. Chaque cadre, mobilisé isolément, aurait conduit au même constat par un chemin différent. Leur articulation en un modèle séquentiel en quatre temps offre quelque chose de plus : une démarche reproductible, dont les étapes sont clairement définies et dont les outputs (cartographies BPMN, VSM, taxonomie de contraintes, modèles de files à priorités) constituent autant de livrables opérationnels pour les gestionnaires hospitaliers.

Enfin, et c'est peut-être l'enseignement le plus important pour la politique de santé au Maroc, le goulot documenté dans ce service résulte d'une règle, la règle FIFO, et non d'une insuffisance de ressources. Cette distinction n'est pas rhétorique : elle a des implications budgétaires directes. Là où une approche capacitaire chercherait à recruter des médecins supplémentaires ou à étendre les locaux, une approche par la TOC et le Lean-BPM cherche à modifier l'ordonnancement du flux existant. Les travaux de Litvak et Bisognano (2011) ont montré que cette approche est non seulement moins coûteuse, mais souvent plus efficace, y compris dans des systèmes opérant sous forte contrainte budgétaire, comme c'est le cas dans les hôpitaux publics marocains.

Il convient de reconnaître explicitement les limites épistémiques de la validation proposée dans cet article. Le cas présenté constitue une mise à l'épreuve initiale du cadre intégratif, non une démonstration causale de sa supériorité sur des approches mono-cadre. L'établissement d'une telle preuve requerrait une comparaison systématique avec d'autres services utilisant des approches alternatives, ou une évaluation avant-après de l'introduction des interventions prescrites, ce que le design de la présente recherche n'a pas encore rendu possible. Ce que le cas de Mohammedia démontre, c'est la cohérence interne du cadre, la plausibilité empirique de ses diagnostics et la fécondité analytique de son articulation. C'est une contribution à l'étayage progressif d'une hypothèse managériale, non une preuve définitive.

Il convient également de souligner la dimension méthodologique de la présente contribution. La démarche de recherche-action adoptée, combinant observation continue sur 323 jours, triangulation chronométrique et documentaire et co-construction du diagnostic avec les équipes soignantes, représente en elle-même un apport à la littérature sur la logistique hospitalière en contexte africain. Kemmis et McTaggart (1988) rappellent que la validité d'une recherche-action ne réside pas seulement dans ses résultats, mais dans la qualité de l'implication des acteurs dans le processus de construction de la connaissance. À cet égard, le fait que le diagnostic empirique présenté en section 3 ait été co-validé avec le chef du service renforce considérablement la crédibilité des conclusions avancées.

5. Conclusion

Cet article est parti d'un constat simple, mais dont les implications organisationnelles sont profondes : dans les services d'urgences publics marocains, la règle qui gouverne l'ordre de traitement des patients ignore la gravité clinique de ceux-ci. Il a cherché à montrer, à travers une revue critique de la littérature et une illustration empirique inédite, que cette règle constitue une hypothèse fortement plausible de contrainte dominante du système, et que sa modification représente, selon les conclusions convergentes des quatre corpus mobilisés, le levier d'amélioration le plus accessible dans le contexte des urgences publiques marocaines.

Pour fonder analytiquement ce diagnostic et en dégager des prescriptions opérationnelles, nous avons proposé un cadre intégratif articulant quatre corpus théoriques complémentaires : la Théorie des Contraintes, la théorie des files d'attente, les outils Lean-BPM et la perspective de la supply chain hospitalière. Chacun de ces cadres apporte, sur la réalité des urgences, un éclairage irréductible à celui des autres ; leur articulation dans une démarche séquentielle en quatre temps (diagnostic, identification des contraintes, modélisation quantitative, conception de l'état futur) produit un modèle qui va au-delà de la simple juxtaposition de théories.

Sur le plan théorique, la contribution centrale réside dans la proposition d'un modèle intégratif cohérent articulant quatre corpus complémentaires appliqués au contexte hospitalier public marocain, domaine encore peu couvert par la littérature internationale. Sur le plan méthodologique, cet article formalise une démarche reproductible en quatre temps dont les étapes, les outils mobilisés et les livrables attendus sont clairement définis, offrant un protocole applicable à d'autres services d'urgences aux caractéristiques comparables. Sur le plan managérial, les prescriptions opérationnelles formulées, à savoir l'introduction d'un protocole de triage formalisé, la différenciation des circuits de prise en charge selon la gravité clinique et la mise en place d'indicateurs de suivi du goulot, n'exigent aucun investissement additionnel : il s'agit de changer la règle d'ordonnancement du flux, non d'augmenter les ressources.

L'illustration empirique conduite à l'Hôpital Mohammed Ben Abdallah de Mohammedia (85 486 admissions, 323 jours d'observation continue, un ratio attente/traitement de 6,4 à l'étape des soins) a permis d'étayer empiriquement la pertinence analytique du cadre proposé et d'en démontrer la pertinence opérationnelle pour le contexte hospitalier public marocain. Ces données constituent, à notre connaissance, l'une des bases empiriques les plus complètes

publiées sur un service d'urgences marocain, et leur valeur réside autant dans ce qu'elles révèlent de la situation actuelle que dans ce qu'elles rendent possible : une démarche de transformation organisationnelle fondée sur des faits, et non sur des impressions.

Ce travail comporte naturellement des limites qu'il convient de reconnaître sans les minimiser. La démarche présentée ici est, pour partie, théorique et intégrative : la validité du modèle proposé reste à tester dans d'autres services d'urgences marocains présentant des configurations différentes, notamment en termes de volume, de localisation et de ressources disponibles. Par ailleurs, si l'illustration empirique est fondée sur des données robustes, l'article ne prétend pas à la généralisation statistique : il offre une démonstration de principe, pas une loi universelle.

Ces limites ouvrent trois perspectives de recherche hiérarchisées. En priorité, la mise en oeuvre et l'évaluation longitudinale du protocole de triage co-construit avec les équipes soignantes permettrait de confirmer empiriquement les gains attendus et d'en affiner les conditions de succès. À moyen terme, l'application du modèle intégratif à d'autres établissements du réseau hospitalier public marocain, notamment en milieu rural ou dans des configurations de flux différentes, permettrait d'en tester la transférabilité. À plus long terme, la construction d'indicateurs de performance standardisés pour les urgences marocaines et maghrébines ouvrirait la voie à des études comparatives inter-établissements et à une contribution plus explicite au débat international sur le patient flow dans les pays à revenu intermédiaire.

Références

- (1). Asplin, B. R., Magid, D. J., Rhodes, K. V., Solberg, L. I., Lurie, N., & Camargo, C. A. (2003). A conceptual model of emergency department crowding. *Annals of Emergency Medicine*, 42(2), 173–180.
- (2). Bhattacharya, A., Varshney, U., & Singh, N. (2010). Applying the theory of constraints to an emergency department. *International Journal of Health Care Quality Assurance*, 23(5), 483–497.
- (3). Broyles, J. R., & Cochran, J. K. (2011). Estimating business loss to a hospital emergency department from patient diversion using queuing models. *Engineering Management Journal*, 23(1), 35–47.
- (4). Cox, D. R., & Smith, W. L. (1961). *Queues*. Methuen.
- (5). de Vries, J., & Huijsman, R. (2011). Supply chain management in health services: An overview. *Supply Chain Management: An International Journal*, 16(3), 159–165.
- (6). Dumas, M., La Rosa, M., Mendling, J., & Reijers, H. A. (2013). *Fundamentals of business process management*. Springer.
- (7). El Fakir, S., Tachfouti, N., Lahmini, W., Zerrouq, B., Slaoui, M., & Nejari, C. (2016). Caractéristiques des patients admis aux urgences d'un hôpital universitaire au Maroc. *The Pan African Medical Journal*, 24(1). <https://doi.org/10.11604/pamj.2016.24.52.8631>
- (8). Goldratt, E. M. (1990). *Theory of constraints*. North River Press.
- (9). Goldratt, E. M., & Cox, J. (2004). *The goal: A process of ongoing improvement* (3rd ed.). North River Press.
- (10). Green, L. V. (2006). Queueing analysis in healthcare. In R. W. Hall (Ed.), *Patient flow: Reducing delay in healthcare delivery* (pp. 281–307). Springer.
- (11). HCP. (2022). *Situation démographique du Maroc 2021*. Rabat : Haut-Commissariat au Plan.
- (12). Haboubi, H., Hammadi, L., Jidane, S., Ballouti, A. E., & Belyamani, L. (2024). Hospital supply chain 4.0 performance: A systematic literature review. In M. Ezziyyani et al. (Eds.), *AI2SD 2023. Lecture Notes in Networks and Systems* (vol. 905). Springer. https://doi.org/10.1007/978-3-031-52385-4_27

- (13). Hoot, N. R., & Aronsky, D. (2008). Systematic review of emergency department crowding: Causes, effects, and solutions. *Annals of Emergency Medicine*, 52(2), 126–136.
- (14). Kendall, D. G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *Annals of Mathematical Statistics*, 24(3), 338–354.
- (15). Kershaw, R. (2000). Using TOC to 'cure' healthcare problems. *Management Accounting Quarterly*, 1(3), 22–28.
- (16). Kemmis, S., & McTaggart, R. (1988). *The action research planner* (3rd ed.). Deakin University Press.
- (17). Le Moigne, J.-L. (1977). *La théorie du système général : Théorie de la modélisation*. Presses Universitaires de France.
- (18). Litvak, E., & Bisognano, M. (2011). More patients, less payment: Increasing hospital efficiency in the aftermath of health reform. *Health Affairs*, 30(1), 76–80.
- (19). Little, J. D. C. (1961). A proof for the queuing formula: $L = \lambda W$. *Operations Research*, 9(3), 383–387.
- (20). Lynn, S. G., & Kellermann, A. L. (1991). Critical decision making: Managing the emergency department in an overcrowded hospital. *Annals of Emergency Medicine*, 20(3), 287–292.
- (21). Marty, A. (2003). *Approche systémique et organisation hospitalière*. Presses de l'École des Hautes Études en Santé Publique.
- (22). Morley, C., Unwin, M., Peterson, G. M., Stankovich, J., & Kinsman, L. (2018). Emergency department crowding: A systematic review of causes, consequences and solutions. *PLOS ONE*, 13(8), e0203316. <https://doi.org/10.1371/journal.pone.0203316>
- (23). Mazzocato, P., Savage, C., Brommels, M., Aronsson, H., & Thor, J. (2010). Lean thinking in healthcare: A realist review of the literature. *Quality and Safety in Health Care*, 19(5), 376–382.
- (24). MSPS. (2022). *Bulletin épidémiologique annuel du Maroc*. Rabat : Ministère de la Santé et de la Protection Sociale.
- (25). CESE. (2023). *Améliorer la prise en charge des urgences médicales au Maroc*. Rabat : Conseil Économique, Social et Environnemental. Repéré à <https://www.cese.ma/media/2023/06/ebook-Avis-urgences-medicales-au-Maroc-2-vf.pdf>
- (26). Deming, W. E. (1986). *Out of the crisis*. Massachusetts Institute of Technology, Center for Advanced Engineering Study.
- (27). SMMU. (2020). *Livre Blanc sur les services d'urgences au Maroc*. Casablanca : Société Marocaine de Médecine d'Urgence.
- (28). OMS. (2023). *Profil sanitaire du Maroc. Statistiques mondiales de la santé 2023*. Genève : Organisation Mondiale de la Santé.
- (29). Plsek, P. E., & Greenhalgh, T. (2001). Complexity science: The challenge of complexity in health care. *BMJ*, 323(7313), 625–628.
- (30). Radnor, Z. J., Holweg, M., & Waring, J. (2012). Lean in healthcare: The unfilled promise? *Social Science & Medicine*, 74(3), 364–371.
- (31). Ronen, B., Pliskin, J. S., & Pass, S. (2018). *The hospital and clinic improvement handbook: Using Lean and the Theory of Constraints for better healthcare delivery*. Oxford University Press.
- (32). Rivard-Royer, H., Landry, S., & Beaulieu, M. (2002). Hybrid stockless: A case study. *International Journal of Operations & Production Management*, 22(4), 412–424.

- (33). Rohleder, T. R., Cooke, D. L., Rogers, P., & Ledger, J. (2011). Coordinating health services: An operations management perspective. *Socio-Economic Planning Sciences*, 45(4), 147–155.
- (34). Vissers, J., & Beech, R. (Eds.). (2005). *Health operations management: Patient flow logistics in health care*. Routledge.
- (35). Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). Sage Publications.
- (36). Ziat, A., Sefiani, N., Reklouï, K., & Azzouzi, H. (2019). A generic framework for hospital supply chain. *International Journal of Healthcare Management*, 13(sup1), 488–495. <https://doi.org/10.1080/20479700.2019.1603415>
- (37). Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). Sage Publications.
- (38). von Bertalanffy, L. (1968). *General system theory: Foundations, development, applications*. George Braziller.
- (39). Womack, J. P., & Jones, D. T. (1996). *Lean thinking: Banish waste and create wealth in your corporation*. Simon & Schuster.
- (40). World Bank. (2022). *World Development Indicators: Morocco*. Washington DC: World Bank Group. Repéré à <https://data.worldbank.org/country/morocco>