Gestion des ressources humaines d'un service d'urgence en période épidémique

Omar El Rifai Sierra

► To cite this version:

Omar El Rifai Sierra. Gestion des ressources humaines d'un service d'urgence en période épidémique. Autre. Ecole Nationale Supérieure des Mines de Saint-Etienne, 2015. Français. <NNT : 2015EMSE0802>. <tel-01301000>

HAL Id: tel-01301000

https://tel.archives-ouvertes.fr/tel-01301000

Submitted on 11 Apr 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

NNT : 2015 EMSE 0802

THÈSE

présentée par

Omar El-Rifai

pour obtenir le grade de
Docteur de l'École Nationale Supérieure des Mines de Saint-Étienne

Spécialité : Génie Industriel

GESTION DES RESSOURCES HUMAINES D'UN SERVICE D'URGENCE EN PERIODE EPIDEMIQUE

soutenue à Saint-Etienne, le 24 Novembre 2015

Membres du jury

Président :	Alain Quilliot	Professeur, Institut Sup. d'Informatique, Aubiere
Rapporteurs :	Slim HAMMADI	Professeur Ecole, Centrale de Lille, Lille
	Christian ARTIGUES	Directeur de Recherche, LAAS, Toulouse
Examinateur(s) :	Nabil ABSI	Maître de Recherche, EMSE, Gardanne
Directeur(s) de thèse :	Xiaolan XIE	Professeur, EMSE, Saint-Etienne
	Thierry GARAIX	Maitre assistant, EMSE, Saint-Etienne

Spécialités doctorales :	Responsables :
SCIENCES ET GENIE DES MATERIAUX	K. Wolski Directeur de recherche
MECANIQUE ET INGENIERIE	S. Drapier, professeur
GENIE DES PROCEDES	F. Gruy, Maître de recherche
SCIENCES DE LA TERRE	B. Guy, Directeur de recherche
SCIENCES ET GENIE DE L'ENVIRONNEMENT	D. Graillot, Directeur de recherche
MATHEMATIQUES APPLIQUEES	O. Roustant, Maître-assistant
INFORMATIQUE	O. Boissier, Professeur
IMAGE, VISION, SIGNAL	JC. Pinoli, Professeur
GENIE INDUSTRIEL	A. Dolgui, Professeur
MICROELECTRONIQUE	

EMSE : Enseignants-chercheurs et chercheurs autorisés à diriger des thèses de doctorat (titulaires d'un doctorat d'État ou d'une HDR)

AVRIL	Stéphane	PR2	Mécanique et ingénierie	CIS
BATTON-HUBERT	Mireille	PR2	Sciences et génie de l'environnement	FAYOL
BENABEN	Patrick	PR1	Sciences et génie des matériaux	CMP
BERNACHE-ASSOLLANT	Didier	PR0	Génie des Procédés	CIS
BIGOT	Jean Pierre	MR(DR2)	Génie des Procédés	SPIN
BILAL	Essaid	DR	Sciences de la Terre	SPIN
BOISSIER	Olivier	PR1	Informatique	FAYOL
BORBELY	Andras	MR(DR2)	Sciences et génie de l'environnement	SMS
BOUCHER	Xavier	PR2	Génie Industriel	FAYOL
BRODHAG	Christian	DR	Sciences et génie de l'environnement	FAYOL
BURLAT	Patrick	PR2	Génie Industriel	FAYOL
COURNIL	Michel	PR0	Génie des Procédés	DIR
DARRIEULAT	Michel	IGM	Sciences et génie des matériaux	SMS
DAUZERE-PERES	Stéphane	PR1	Génie Industriel	CMP
DEBAYLE	Johan	CR	Image Vision Signal	CIS
DELAFOSSE	David	PR1	Sciences et génie des matériaux	SMS
DESRAYAUD	Christophe	PR2	Mécanique et ingénierie	SMS
DOLGUI	Alexandre	PR0	Génie Industriel	FAYOL
DRAPIER	Sylvain	PR1	Mécanique et ingénierie	SMS
FEILLET	Dominique	PR2	Génie Industriel	CMP
FOREST	Bernard	PR1	Sciences et génie des matériaux	CIS
FORMISYN	Pascal	PR0	Sciences et génie de l'environnement	DIR
FRACZKIEWICZ	Anna	DR	Sciences et génie des matériaux	SMS
GARCIA	Daniel	MR(DR2)	Génie des Procédés	SPIN
GERINGER	Jean	MA(MDC)	Sciences et génie des matériaux	CIS
GIRARDOT	Jean-jacques	MR(DR2)	Informatique	FAYOL
GOEURIOT	Dominique	DR	Sciences et génie des matériaux	SMS
GRAILLOT	Didier	DR	Sciences et génie de l'environnement	SPIN
GROSSEAU	Philippe	DR	Génie des Procédés	SPIN
GRUY	Frédéric	PR1	Génie des Procédés	SPIN
GUY	Bernard	DR	Sciences de la Terre	SPIN
GUYONNET	René	DR	Génie des Procédés	SPIN
HAN	Woo-Suck	CR	Mécanique et ingénierie	SMS
HERRI	Jean Michel	PR1	Génie des Procédés	SPIN
INAL	Karim	PR2	Microélectronique	CMP
KERMOUCHE	Guillaume	PR2	Mécanique et Ingénierie	SMS
KLOCKER	Helmut	DR	Sciences et génie des matériaux	SMS
LAFOREST	Valérie	MR(DR2)	Sciences et génie de l'environnement	FAYOL
LERICHE	Rodolphe	CR	Mécanique et ingénierie	FAYOL
LI	Jean Michel		Microélectronique	CMP
MALLIARAS	Georges	PR1	Microélectronique	CMP
MOLIMARD	Jérôme	PR2	Mécanique et ingénierie	CIS
MONTHEILLET	Franck	DR	Sciences et génie des matériaux	SMS
PERIER-CAMBY	Laurent	PR2	Génie des Procédés	DFG
PIJOLAT	Christophe	PR0	Génie des Procédés	SPIN
PIJOLAT	Michèle	PR1	Génie des Procédés	SPIN
PINOLI	Jean Charles	PR0	Image Vision Signal	CIS
POURCHEZ	Jérémy	CR	Génie des Procédés	CIS
ROUSTANT	Olivier	MA(MDC)		FAYOL
STOLARZ	Jacques	CR	Sciences et génie des matériaux	SMS
SZAFNICKI	Konrad	MR(DR2)	Sciences et génie de l'environnement	CMP
TRIA	Assia		Microélectronique	CMP
VALDIVIESO	François	MA(MDC)	Sciences et génie des matériaux	SMS
VIRICELLE	Jean Paul	MR(DR2)	Génie des Procédés	SPIN
WOLSKI	Krzystof	DR	Sciences et génie des matériaux	SMS
XIE	Xiaolan	PR0	Génie industriel	CIS

ENISE : Enseignants-chercheurs et chercheurs autorisés à diriger des thèses de doctorat (titulaires d'un doctorat d'État ou d'une HDR)

BERGHEAU	Jean-Michel	PU	Mécanique et Ingénierie	ENISE
BERTRAND	Philippe	MCF	Génie des procédés	ENISE
DUBUJET	Philippe	PU	Mécanique et Ingénierie	ENISE
FORTUNIER	Roland	PR	Sciences et Génie des matériaux	ENISE
GUSSAROV	Andrey	Enseignant contractuel	Génie des procédés	ENISE
HAMDI	Hédi	MCF	Mécanique et Ingénierie	ENISE
LYONNET	Patrick	PU	Mécanique et Ingénierie	ENISE
RECH	Joël	MCF	Mécanique et Ingénierie	ENISE
SMUROV	Igor	PU	Mécanique et Ingénierie	ENISE
TOSCANO	Rosario	MCF	Mécanique et Ingénierie	ENISE
ZAHOUANI	Hassan	PU	Mécanique et Ingénierie	ENISE

PR 0	Professeur classe exceptionnelle
PR 1	Professeur 1ère classe
PR 2	Professeur 2ème classe
PU	Professeur des Universités
MA (MDC)	Maître assistant
DR	Directeur de recherche

Ing.	Ingénieur
MCF	Maître de conférences
MR (DR2)	Maître de recherche
CR	Chargé de recherche
EC	Enseignant-chercheur
IGM	Ingénieur général des mines

SMS	Sciences des Matériaux et des Structures
SPIN	Sciences des Processus Industriels et Naturels
FAYOL	Institut Henri Fayol
CMP	Centre de Microélectronique de Provence
CIS	Centre Ingénierie et Santé

Mise à jour : 07/04/2013

Document édité le : 11 février 2016

Remerciements

Un long et beau chapitre de ma vie se clôt avec cette thèse. Ce chapitre n'aurait pas pu être écrit sans l'aide précieuse de plusieurs personnes qui ont, chacune à sa façon, apporté les éléments nécessaires pour m'aider à mener à bien ce projet.

Je voudrais d'abord remercier tous les organismes impliqués dans ce projet et qui ont rendu cette thèse possible (ANR, Ecole Nationale Supérieure des Mines de Saint-Etienne et LIMOS). Je voudrais en même temps remercier mes encadrants pour leur investissement, leur confiance et leur support constant. Je remercie Xiaolan pour sa lucidité, sa rigueur et son intelligence qui ont été des exemples à suivre. Je remercie aussi tout particulièrement Thierry pour tous les échanges que nous avons eu, sa générosité d'esprit et son aide toujours renouvelée face aux difficultés que j'ai rencontrées.

Mes remerciements vont aussi à Vincent Augusto, Marianne Sarazin et Rémy Brunon qui ont également participé à ce travail, soit directement soit en m'ouvrant des portes et me donnant accès à des informations précieuses. Ils ont également su remettre en question mon point de vue en partageant leur savoir sur le système de santé.

Je remercie par ailleurs mes rapporteurs Christian Artigues et Slim Hammadi qui ont montré avec beaucoup de pertinence les points forts et les limites de mon travail. Je remercie aussi les membres du jury Alain Quilliot et Nabil Absi pour avoir participé à ce travail et m'avoir donné des pistes pour la suite. D'un point de vue plus personnel, je remercie particulièrement mes parents pour leur amour infini et inconditionnel. Je remercie aussi mes amis, à la fois ceux de longue date et ceux que j'ai rencontrés pendant ces trois années de thèse et qui ont à leur façon rendu ce travail possible.

David toujours patient et compréhensif et compagnon de musique. Fanette amoureuse de culture et toujours au rendez-vous pour des discussions ou des sorties. Baptiste avec qui je me suis bien entendu depuis le début et qui a vite rejoint le rang des musiciens. Fanny rockeuse dans l'âme et toujours à l'écoute d autres points de vue. Boris, éternel bon vivant et graine de poète. Klervi grande sportive et toujours de bonne humeur. Pierre, Armelle et Salim pour toutes les conversations autant sur des équations mathématiques que sur le goût du café. Sans oublier toutes les autres personnes de l'école des Mines en général et membres du CIS en particulier avec qui j'ai partagé plusieurs repas, sorties et pauses café et qui ont fait que mon passage soit rempli de bonne humeur.

Je n'oublie pas non plus Ali, Hadi, Marisol, Victor, Olfa, Lucas, Aya, Jorge, Patricio, Yves, Julia, Canan, Claire, Karine, Christelle, et Fab pour leurs amitiés aussi diverses qu épanouissantes.

J'espère que ce travail sera utile et servira à démêler une problématique à la fois complexe et critique.

Table des matières

Table des matières v

Table des figures viii

Liste des tableaux x

1 Introduction 1

1.1 Contexte de l'Etude 2

1.2 Organisation des Services d'Urgences 5

1.2.1 Systématisation des Activités 6

1.2.2 Les Acteurs 7

1.2.3 Les Flux aux Urgences 7

1.2.4 Les Données 9

1.3 Objectifs et Méthodes 9

1.3.1 La Modélisation 9

1.3.2 Prise en Compte des Aléas 10

1.4 Organisation de la Thèse 11

2 Overcrowding in Emergency Departments : Literature Review 13

2.1 Patient Flow Management 15

2.2 Resources Management 17

2.2.1 Equipment Management 17

2.2.2 Workforce Planning 18

2.3 Modeling the Demand 21

3 Capacity Planning in Hospitals during Epidemics 25

3.1 Introduction 27

3.2 Literature Review 27

3.3 Problem Description 28

3.3.1 Optimal Characterization 29

3.3.2 How to compute an optimal solution 31
3.3.3 Normal Approximation 31
3.3.4 Upper Bound Constraints 32
3.3.5 Minimization of the largest expected capacity shortage 35
3.4 Numerical Examples 36
3.4.1 Epidemic Simulations 37
3.4.2 Demand Distribution 39
3.4.3 Capacity Upper Bounds Influence 39
3.4.4 Min Max Criterion 39
3.5 Conclusion 41

4 A Stochastic Optimization Model for Shift Scheduling in Emergency Departments 43
4.1 Introduction 45
4.2 Workforce Scheduling in Emergency Departments 46
4.3 Emergency Department Queuing Network 48
4.4 Mathematical Formulation 50
4.5 Simulation Model 55
4.6 Numerical Experiments 56
4.6.1 Parameters Settings and Real-World Data 56
4.6.2 Analysis of optimized schedules 59
4.6.3 Validation of the Robustness of the Schedules 63
4.6.4 Week to Weekend Transition 67
4.7 Conclusion 69

5 Proactive On-call Scheduling during a Seasonal Epidemic 71
5.1 Introduction 73
5.2 Literature Review 74
5.3 Workload Patterns 75
5.4 Staff Allocation Model 76
5.4.1 On-Call Duties 78
5.4.2 Mathematical Formulation 78
5.5 Numerical Experiments 82
5.5.1 Design of Experiments 83
5.5.2 Demand Generation 84
5.5.3 Cyclic Schedules 84
5.5.4 Non-cyclic Schedules 88
5.5.5 Relaxed Constraints on Workload Regulations 89
5.6 Conclusion 89

6 Dynamic Management of Overtime Hours 91
6.1 Introduction 93
6.2 Literature Review 93
6.3 Problem Description 94
6.3.1 Demand Representation 95
6.4 Markov Decision Process 96
6.4.1 Decision Variables 96
6.4.2 Marginal Cost 97
6.4.3 Value Function 97
6.5 Policy Approximations 98
6.5.1 Upper Bound 98
6.5.2 Lower Bound 99
6.5.3 Dynamic Approximation Scheme 100
6.6 Experimental Results 100
6.6.1 Model Parameters 101
6.6.2 Policy Approximation 102
6.6.3 Long Planning Horizon 106
6.7 Conclusion 107

7 Synthèse et Conclusions 109
7.1 Des modèles non-exhaustifs qui doivent capter l'essentiel du système à optimiser 109
7.2 Synthèse de la Gestion des Ressources Humaines 110
7.3 Évaluation par Scénarios 111
7.4 Perspectives : Vers des Résultats Pratiques 111

Bibliographie 113

A Compte Rendu de la visite du CHRU de Lille 121
A.0.1 Principales Maladies 121
A.0.2 Le services des urgences 121
A.0.3 Les ressources humaines 122
A.0.4 Les Outils informatiques 123
A.0.5 Les goulots d'étranglement principaux et les pics d'affluence . . . 123
A.0.6 Dispositifs mis en place pour désengorger les urgences 123
A.0.7 Autres stratégies envisagées 124
A.0.8 Services Auxiliaires 126

Table des figures

1.1 Les différents partenaires du projet HOST (Présentation ANR) 4
1.2 Les différents "Work Package" du projet HOST (rapport technique HOST) 5
1.3 La prise en charge d'un patient au service d'urgences pédiatriques (rapport technique HOST) 8

3.1 The cumulative distribution functions of three periods along with the optimal solution corresponding to α^*. 30
3.2 One iteration of the algorithm to find the optimal bounded capacity distribution 35
3.3 Example of optimal solution for the min max allocation problem with three periods 36
3.4 Expected value of demand in three types of epidemic scenario 37
3.5 Mean and standard deviation of 10 instances of the solution with C=250 38
3.6 Expected value as a function of C with epidemic data and normal approximations. 39
3.7 Capacity distribution with bound for $C = 190$ and $C = 200$. 40
3.8 Minimizing the expected sum over the planning horizon using solutions from **P1** and **P2**. 40
3.9 Minimizing the maximum shortage using solutions from **P1** and **P2**. . . . 41

4.1 Preliminary analysis of the data from the pediatrics ED of Lille Hospital 46
4.2 A basic queue model that represents a process in an ED 49
4.3 Process chart of patients in the ED 50
4.4 The evolution of waiting time with the number of scenarios used in the optimization model 58
4.5 Resulting staffing levels with shifts fixed to 8 hours 63
4.6 Resulting staffing levels for shifts varying from 4-12 hours 63
4.7 Resulting staffing levels for shifts varying from 1-12 hours 64
4.8 Expected sojourn time for the different acuity classes in normal conditions 69
4.9 Expected sojourn time for the different acuity classes in epidemic conditions 69

5.1 Number of patients arriving per week for the years 2011 and 2012 75
5.2 Number of patients per week hospitalized after the ED 75
5.3 Number of patients per week with sojourn time greater than 2.5 hours . . 76
5.4 Difference of arrivals between night and day 76
5.5 Day and night demand for three types of epidemic scenarios 85
5.6 Difference in objective value with and without on-calls for different values of α in the three scenarios 85
5.7 Costs of cyclic schedules in the mild, moderate and severe scenarios . . . 87
5.8 Effective work reduction with on-call vs without on-call for cyclic schedules 87
5.9 Difference in objective value with and without on-calls for non-cyclic schedules 88
5.10 Costs of non-cyclic schedules in the mild, moderate and severe scenarios . 88
5.11 Effective work reduction with on-call vs without on-call for non-cyclic schedules 89

6.1 M/M/N queue that allows us to calculate the excess workload distribution at the end of each shift 96
6.2 A framework for using the lower bound approximation in a dynamic setting 100

A.1 Évolution des syndromes grippaux et des diarrhées aigues pour l'année 2010(fin)/2011.Source : Réseau Sentinelles 121

Liste des tableaux

2.1 Parameter Values for the Epidemic Model . . . 23

3.1 Influence of the number of samples used on the stability of the results . . 38

4.1 The parameters of the scheduling model. . . . 52
4.2 The variables of the scheduling model. . . . 53
4.3 Parameters of the system and of the optimization algorithm . . . 58
4.4 Parameters for the scheduling strategies . . . 59
4.5 Optimization results for the weekend scenarios . . . 60
4.6 Simulation results for the weekend schedules . . . 62
4.7 Simulation results for weekend schedules during an epidemic period . . . 65
4.8 Simulation results for week schedules during an epidemic period . . . 66
4.9 Simulation results for schedules spanning an entire week . . . 68
4.10 Simulation results for schedules spanning an entire week during an epidemic period ϵ . . . 68

5.1 Parameters of the model. . . . 79
5.2 Variables of the model . . . 80
5.3 Model Parameters . . . 83
5.4 Numerical stability through different number of samples . . . 83
5.5 Results for cyclic schedules with on-call duties . . . 86
5.6 Results for cyclic schedules without on-call duties . . . 86
5.7 Relaxed weekly rest constraint for moderate scenario . . . 89

6.1 Problem variables . . . 97
6.2 The number of MDP states as a function of the overtime hours allowed by physician with 6 physicians . . . 101
6.3 Morning shift (8-16h) arrival averages . . . 101
6.4 Afternoon shift (16-24h) arrival averages . . . 102
6.5 Night shift (24h-8h) arrival averages . . . 102

6.6 Physicians schedule with preferences considerations 102
6.7 Comparison between the optimal policy and the LP approximations for the "adjusted" schedule 104
6.8 Comparison between the optimal policy and the LP approximations for the "fixed" schedule 104
6.9 Comparison between the optimal policy and the LP approximations for the "manual" schedule 105
6.10 Results obtained with the approximation scheme for a 4 weeks planning horizon 106

Chapitre 1

Introduction

Sommaire

1.1 **Contexte de l'Etude** **2**
1.2 **Organisation des Services d'Urgences** **5**
1.2.1 Systématisation des Activités 6
1.2.2 Les Acteurs 7
1.2.3 Les Flux aux Urgences 7
1.2.4 Les Données 9
1.3 **Objectifs et Méthodes** **9**
1.3.1 La Modélisation 9
1.3.2 Prise en Compte des Aléas 10
1.4 **Organisation de la Thèse** **11**

La majeure partie de cette thèse est rédigée en anglais car elle se base sur deux articles publiés dans des revues avec comité de lecture et deux chapitres rédigés comme des articles non encore soumis. Pour suivre les règles de l'école doctorale, nous avons rédigé également une introduction générale, une conclusion générale et un résumé de chaque chapitre en français. Une revue de la littérature est également presentée après l'introduction.

L'industrialisation des hôpitaux en France s'est développée à un rythme plus lent que celui des industries privées. Alors que, dès le début du XIX[e] siècle, dans le secteur privé le travail commençait à s'organiser selon la “Scientific Work Organisation” de F.W Taylor, les hôpitaux, quant à eux, étaient structurés en petites unités avec un personnel très versatile. Ce n'est que plus tard, avec l'augmentation des contraintes budgétaires, que le processus d'industrialisation s'est également étendu aux hôpitaux. En 1992, il existait déjà environ 1000 hôpitaux publics en France employant environ 60000 personnes. En termes de dépenses, les ressources humaines totalisent aujourd'hui près de 70% du budget total des hôpitaux [Gonnet and Lucas, 2003].

Le service d'urgence (SU) tel qu'il est décrit dans ce document, est une structure médicale qui fait partie intégrante de l'hôpital et dont le rôle est de soigner les patients qui s'y présentent, sans discrimination aucune. Comme dans toute industrie, pour que le processus des urgences soit efficace, il faut que l'organisation interne soit fondée sur une compréhension de tous les échanges nécessaires. En effet, la combinaison de personnel hautement qualifié et de machines sophistiquées crée une interdépendance entre les différents acteurs.

Un des principaux problèmes auxquels les SU doivent faire face est celui de leur engorgement qui mène à des situations de tension. Nous définissons la tension comme “un régime transitoire résultant d'un déséquilibre entre le flux de charges en soins (patients) et la capacité de prise en charge (personnels et ressources physiques)”. Un manque de ressources ou une gestion sous-optimale des ressources peut en être une cause. De même, une augmentation des entrées aux urgences peut conduire à un déséquilibre entre la capacité et la demande. L'augmentation de la demande peut être due à des facteurs économiques, démographiques ou sociaux. Afin de mieux comprendre les situations de tension, il est nécessaire de caractériser l'organisation et le fonctionnement des SU ainsi que de définir les différents acteurs et les échanges qui ont lieu. Cette thèse s'inscrit dans le cadre d'un projet national intitulé “Hôpital : Optimisation, Simulation et évitement des Tensions” (HOST) qui vise à faire face à ce problème d'engorgement. Les difficultés rencontrées sont nombreuses à cause du décalage inévitable entre théorie et pratique et de la complexité inhérente à l'organisation des SU.

1.1 Contexte de l'Etude

Le problème de tensions aux urgences est un problème à l'échelle mondiale comme en témoignent de nombreuses publications scientifiques à ce sujet (c.f Chapitre 2). Au niveau européen, une conférence sur l'aide médicale urgente s'est tenue les 14 et 15 mars 2005 réunissant plus de 350 participants dans le but de favoriser les échanges sur des méthodes d'amélioration du système des urgences. Dans cette partie nous définissons les causes principales de l'engorgement des SU en France ainsi que le but précis de cette étude.

Les urgences sont souvent perçues comme une alternative gratuite aux médecins de ville qui sont ouvertes 24 heures sur 24 et 7 jours sur 7. D'un point de vue économique, la croissance du PIB en France est de plus en plus faible avec une croissance moyenne presque nulle entre 2007 et 2012 [L'Express, 2014]. De même, le pouvoir d'achat des ménages en moyenne stagne. Ce ralentissement de l'économie n'explique cependant pas à lui seul l'augmentation des entrées aux urgences. En effet, la proportion de patients aux urgences faisant appel au dispositif CMU-C d'urgence (dispositif visant à faciliter l'obtention des bénéfices de la protection complémentaire de santé aux personnes en situation précaire) reflète la proportion de la population nationale en situation précaire. Cependant, on constate depuis 1980 une constante augmentation de la population. Cette augmentation s'est accompagnée d'une croissance de l'espérance de vie moyenne au cours des cinq dernières années. Aujourd'hui des études recensent environ 65 millions d'habitants en France métropolitaine [Insee, 2014]. Parallèlement, la fréquentation des services d'urgences a augmenté constamment de 1990 à 2012 passant de 7 à 18,4 millions de passages avec une augmentation d'environ 30% au cours des dix dernières années [des Comptes, 2014].

Cependant, outre l'augmentation de la population, les raisons des engorgements sont diverses. Le même rapport [des Comptes, 2014] montre que deux personnes sur dix déclarent être venues aux urgences à cause d'un manque d'alternatives (médecin traitant non disponible et difficultés pour obtenir un rendez-vous pour des examens complémentaires). Une majorité déclare également être venue pour les avantages qu'offrent les urgences sur les alternatives disponibles (la gratuité des soins et les horaires de nuit). Pour faire face à l'augmentation de la fréquentation des urgences, une aide financière d'une hauteur de 500 millions d'euros a été allouée entre 2004 et 2008 aux SU par l'état. Cette aide financière était destinée à améliorer les conditions de travail en augmentant les ressources matérielles et l'effectif médical. Néanmoins, des progrès restent à faire et notamment au niveau de l'organisation interne [des Comptes, 2014].

La littérature scientifique définit plusieurs causes d'engorgement. L'une d'elles, récurrente, est le manque de lits d'aval. En effet, les médecins urgentistes ont souvent du mal à trouver un lit d'hospitalisation pour les patients sortant des urgences. Des dispositifs ont toutefois été mis en place pour résoudre ce problème. Ainsi, il existe dorénavant aux urgences des structures d'hospitalisation à courte durée pour garder les patients dont la durée de séjour ne dépasse pas les 24 heures. De plus, l'ouverture d'un poste de gestionnaire de lits a été pensé dans la même optique. Un autre problème majeur est celui du manque d'effectif médical. Malgré l'augmentation de ce dernier entre 2001 et 2011, (augmentation de 72,7% pour les médecins et de 56,3% pour les infirmiers) le problème des ressources humaines reste un problème fréquemment cité.

Le projet ANR HOST au sein duquel s'inscrit cette thèse vise à étudier et résoudre le problème des tensions aux urgences hospitalières. Le projet cherche en premier lieu à définir rigoureusement la notion de tension, puis d'utiliser des outils mathématiques pour

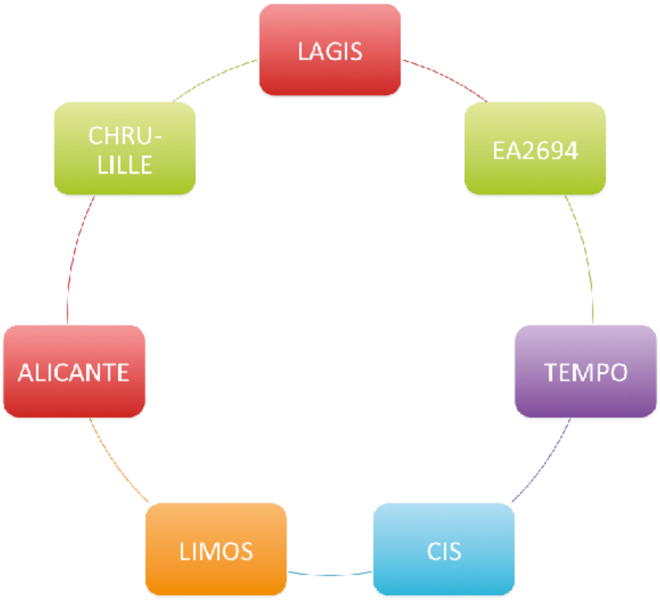


Figure 1.1 – Les différents partenaires du projet HOST (Présentation ANR)

proposer des solutions qui prennent en compte toute la complexité du système de santé. Le contexte médical du projet se base sur les périodes épidémiques dans les services d'urgence pédiatrique. En effet, malgré la récurrence des phénomènes d'engorgement, ils sont limités dans le temps (périodes hivernales) et requièrent en conséquent des solutions à courtes durées. La difficulté de trouver des solutions efficaces est principalement due à l'incertitude et à la dynamique des flux caractéristiques des SU. D'une année à l'autre, l'intensité des épidémies, et donc du nombre de patients se présentant aux urgences, peut varier radicalement. De plus, le volume journalier de patients ne suffit pas à prédire la charge de travail associée. Les caractéristiques des patients et leur ordre d'arrivée impactent fortement la prise en charge.

Le projet est une collaboration entre sept partenaires que nous pouvons voir à la Figure 1.1 à compétences complémentaires.

L'équipe d'accueil santé publique de l'université de Lille est principalement responsable de la coordination du projet, de la définition d'indicateurs de tension à partir des données médicales et de l'évaluation de l'ergonomie de l'interface des logiciels proposés. L'équipe LAGIS de l'école centrale de Lille, quant à elle est responsable de l'élaboration d'un système pour l'anticipation des tensions et la modélisation d'actions correctives possibles. L'équipe PSI du laboratoire TEMPO participe à la modélisation d'indicateurs de tension et au développement de stratégies d'évitement des situations de tension via la planification des ressources et la synchronisation des flux. L'équipe ALICANTE est impliquée en assistance à la maîtrise d'ouvrage et d'oeuvre. Enfin le CHRU de Lille s'occupe de l'organisation du projet et de l'extraction des données. Une schéma récapitulatif des différentes tâches du projet est représenté à la Figure 1.2.

L'objectif de cette thèse est de proposer des solutions stratégiques, tactiques et opéra-

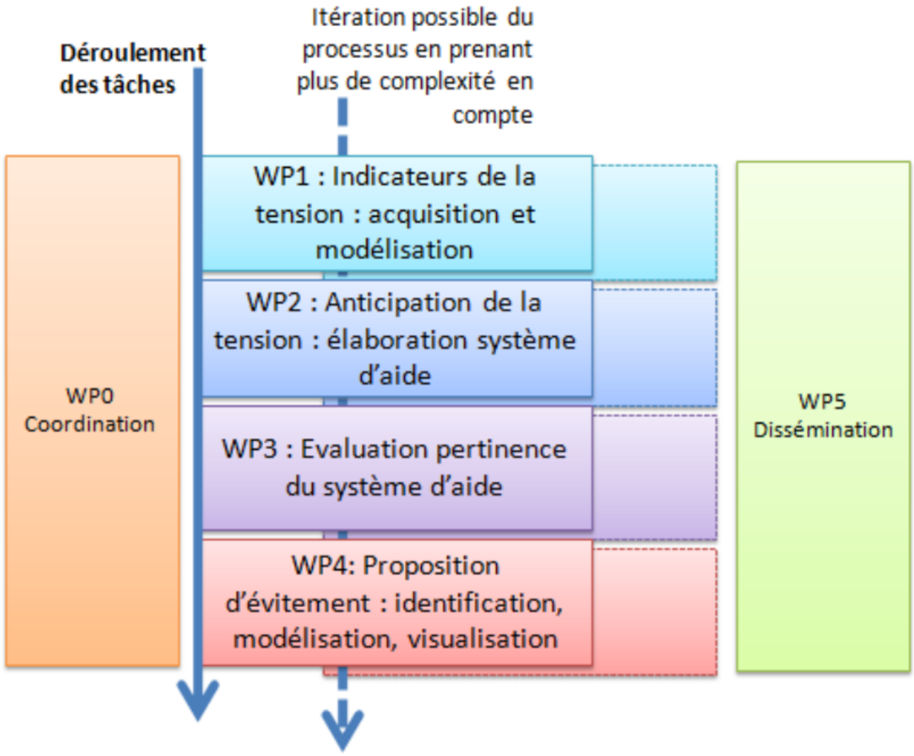


FIGURE 1.2 – Les différents "Work Package" du projet HOST (rapport technique HOST)

tionnelles de gestion des ressources en prenant en compte les aléas caractéristiques des urgences. Nous situons ainsi notre travail principalement dans les "Work Package" (WP) 2 et 4 du projet. Plus particulièrement, nous avons proposé une méthode stochastique de modélisation de la demande pendant les périodes épidémiques (c.f Chapitre 2) contribuant ainsi au WP2. Puis, par rapport au WP4, nous avons proposé plusieurs systèmes d'aide à la décision dans les Chapitres 3 à 6.

1.2 Organisation des Services d'Urgences

D'un point de vue organisationnel, les hôpitaux sont considérés comme des industries de services. Les industries de services diffèrent radicalement des industries de productions [Miles, 2008]. En effet, les industries de services se caractérisent par une demande très variable dont le traitement ne peut pas être remis à plus tard ainsi qu'un rendu qui ne peut pas être stocké. De plus, il existe des particularités qui différencient les hôpitaux du reste des industries de services. Par exemple, les résultats obtenus dans les hôpitaux sont difficilement quantifiables parce que souvent compliqués à analyser. En effet, l'état de santé d'un patient ne dépend pas uniquement de la qualité du service mais également de sa condition physique et de plusieurs facteurs extérieurs non ou peu prévisibles. Une caractéristique des hôpitaux est également que le client (dans ce cas le patient) est en situation de dépendance radicale envers le personnel. Le patient a généralement très peu de choix en ce qui concerne les démarches médicales nécessaires à son traitement. Cette dépendance peut créer des réactions émotionnelles. Une situation difficile à la fois pour le patient et pour le personnel.

Pour arriver à bien gérer un SU, il est nécessaire d'avoir l'avis des experts et des professionnels qui travaillent au jour le jour et se heurtent aux problèmes quotidiens. Cependant, et paradoxalement, le personnel médical a souvent peu de recul par rapport aux pratiques déjà mises en place même si elles sont loin d'être optimales. De même, l'administration peut ralentir les changements dans la mesure où elle ne voit pas forcément le besoin d'une réorganisation du service vu que les problèmes du quotidien sont souvent absorbés par la résilience naturelle des ressources humaines.

1.2.1 Systématisation des Activités

Plusieurs aspects de l'organisation aux urgences sont agencés de façon méthodique. Ainsi l'organisation méthodique des processus de soins s'insère dans ce qu'on appelle "industrialisation" au sens large. Celle-ci concerne une grande partie de l'activité des urgences, depuis la trajectoire des patients jusqu'au processus intellectuel de diagnostic des patients à travers des protocoles stricts. Dans cette partie nous relevons deux points d'organisation qui nous serviront par la suite dans notre démarche de recherche de solutions au problème d'engorgement.

La gestion des ressources humaines, le principal sujet de cette étude, se fait de manière systématique. Des plannings mensuels ou hebdomadaires sont créés par le responsable des ressources humaines et distribués aux employés avec un temps d'avance. La gestion des plannings est un problème complexe du fait que les SU fonctionnent en continu pour une activité incertaine. Ainsi, les postes des ressources humaines doivent êtres échelonnés afin d'assurer une qualité de service acceptable tout au long de la journée. De plus, le code du travail et les politiques de chaque établissement imposent des contraintes de distribution des postes.

Du côté des patients, une classification standardisée de leur état de gravité permet d'établir une relation entre les besoins des patients et les délais raisonnables pour obtenir des soins. A travers cette classification, une hiérarchisation des patients est ainsi établie. Dans les cas les plus graves où le diagnostic vital est engagé, le patient est directement dirigé vers une unité de réanimation et ne suit pas le parcours typique. Cette classification est normalement définie par l'infirmier d'accueil et d'orientation dès l'arrivée du patient. De plus, la classification des patients rentre dans un autre processus de systématisation, qui est la tarification à l'activité (T2A). Depuis 2003, les hôpitaux sont rémunérés par rapport aux nombres et types d'interventions médicales effectuées au cours de l'année. Des frais sont ainsi définis pour chaque type de séjour à l'hôpital.

1.2.2 Les Acteurs

Les patients sont les acteurs clés des SU. En entrant aux urgences, ils peuvent être en situation critique et nécessiter des soins immédiats. Le rôle principal des urgences est ainsi de s'occuper des besoins médicaux des patients et de s'assurer de leur bien-être durant leur séjour. De ce fait, les critères de qualité des SU doivent être liés à la qualité de service des patients qui est assurée principalement par les médecins et les infirmiers. Il est indispensable de maintenir une bonne cohérence entre les flux des patients et les flux des ressources humaines.

Les médecins ont pour rôle d'assurer le diagnostic, de demander des examens complémentaires et de soigner les patients. Les médecins ont aussi la tâche d'établir les dossiers des patients et d'analyser leur historique médical (allergies, historique des opérations effectuées, etc). Enfin, en tâches annexes, l'hôpital est également tenu de communiquer l'état de santé des patients à leurs proches, de trouver un lit d'aval pour les patients qui vont être hospitalisés et d'aider les infirmiers à mieux trier les patients à l'arrivée. Les médecins se divisent en internes, médecins juniors et seniors. Il y a également des médecins spécialistes qui peuvent prendre des astreintes pour les urgences. En ce qui concerne les infirmiers, leur rôle principal est de soigner les patients en suivant les recommandations des médecins. Ils ont aussi la tâche de surveiller les signes vitaux des patients et de participer à la recherche des lits d'hospitalisation.

1.2.3 Les Flux aux Urgences

On peut relever trois types de flux aux SU : le flux de patients, celui des ressources humaines, et le flux d'informations. Le premier flux définit les différentes étapes qu'empruntent les patients au cours de leur visite aux urgences. Il est important de bien le caractériser parce qu'il nous permet d'analyser l'évolution de la charge de travail des différentes ressources à travers le temps. En France, les patients arrivent par leurs propres moyens ou sont transportés par un organisme d'urgence, qui peut être le SAMU et le SDIS. Une fois aux urgences, les patients suivent une trajectoire qui dépend de leur état de santé et de leurs pathologies. De façon schématique, la trajectoire des patients est la suivante : un patient est d'abord examiné par un infirmier d'accueil pour un diagnostic rapide et surtout pour déterminer s'il a besoin de soins vitaux. S'il s'avère que le patient est en état critique, il est emmené vers une unité de réanimation. Sinon, ce qui constitue la grande majorité des cas, le patient patiente dans la salle d'attente jusqu'à ce qu'un médecin se libère. Après le premier diagnostic établi, les patients peuvent être amenés à réaliser des examens complémentaires. Les tests sanguins sont effectués par des infirmiers et les examens d'imagerie par des techniciens de laboratoire. Une fois le diagnostic bien posé, les patients sont soignés par les infirmiers. À la sortie des urgences, si leur état de santé l'exige, les patients peuvent être admis dans d'autres unités de l'hôpital. Les urgences en France sont constituées de deux unités qu'on peut distinguer. L'unité externe

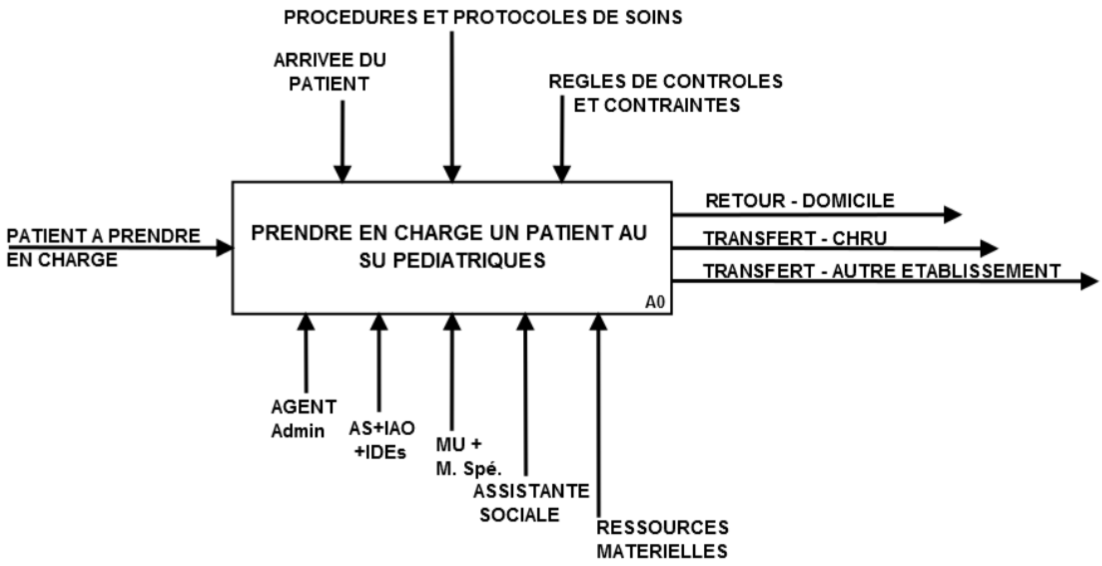


FIGURE 1.3 – La prise en charge d'un patient au service d'urgences pédiatriques (rapport technique HOST)

dont le fonctionnement vient d'être décrit, accueille les patients qui ne sont pas internés à la sortie des urgences. D'un autre côté, l'unité d'hospitalisation de courte durée (UHCD) est une unité transitoire entre les urgences externes et les services d'aval. Cependant les ressources humaines sont souvent mise en commun. D'autant plus pendant les périodes de tension comme les périodes épidémiques.

Le flux de ressources humaines est le flux d'arrivées et de départs du personnel hospitalier. Une synchronisation du flux des différentes ressources est nécessaire pour faciliter l'échange et la communication du travail. Par exemple, il est nécessaire qu'aux heures de pointe, il y ait à la fois des médecins, des techniciens de laboratoire et des infirmiers pour faire face efficacement à la charge de travail. De même, les heures de sortie des ressources doivent être plus ou moins synchronisées avec les heures d'arrivée aux postes suivants pour permettre un transfert d'informations sur les patients qui restent d'un poste à l'autre. Un diagramme représentant la prise en charge des patients dans un service d'urgences pédiatrique est montré à la Figure 1.3.

Ainsi, l'échange entre les différents acteurs est souvent nécessaire à la prise en charge des patients. Celui-ci se fait de plus en plus sous forme numérique. Des informations sont saisies à l'arrivée des patients par les infirmiers d'accueil et d'orientation, puis par les médecins au cours de la visite. De même les résultats des examens complémentaires font partie du dossier des patients. Il est alors fondamental d'avoir des dossiers sur l'historique médical des patients et des informations sur leurs coordonnées afin d'effectuer un diagnostic précis et à des fins administratives. Il existe actuellement des systèmes qui permettent de partager facilement les informations entre les différents services et ainsi gagner du temps et avoir une vision globale de l'historique des patients.

1.2.4 Les Données

En France, les hôpitaux conservent l'historique des visites dans un système d'information (c.f Annexe A). En particulier dans les SU pour lesquels les bases de données contiennent, entre autres, la date d'arrivée et de sortie des patients, le diagnostic établi et les examens complémentaires de chaque visite. Ces données permettent d'analyser les tendances générales des arrivées aux SU. Dans cette partie, nous présentons quelques courbes caractéristiques des données à disposition.
Tous les SU visités s'accordent sur le fait que la charge de travail est variante au cours de l'année : par exemple en hiver, il y a une augmentation de la charge de travail due à l'arrivée des patients infectieux. Afin d'utiliser cette information dans nos modèles, il est important de quantifier l'augmentation de la charge de travail et de la lier à un changement dans les données. Cependant, le nombre de patients aux urgences n'est pas le seul indicateur de l'augmentation de la charge de travail. En effet, dans les données disponibles au service d'urgences pédiatriques de Lille, on ne remarque pas une augmentation du nombre d'arrivées pendant la période hivernale (c.f Chapitre 5). Toutefois, d'autres indicateurs, comme la durée des visites aux SU et la destination des patients à la sortie des urgences, nous montrent bien que le problème est particulièrement accentué pendant les périodes hivernales. De plus, au cours d'une même journée, le flux des patients est différent. On observe ainsi des pics d'activités en fin de journée ($\approx$ 18h-23h) et une très faible activité sur la période de nuit.

1.3 Objectifs et Méthodes

Un travail de terrain, en complémentarité avec une recherche bibliographique, a permis de choisir les problématiques traitées au début de cette thèse. Dans les SU de Lille, Firminy et de Saint-Etienne que j'ai pu visiter et y interroger le personnel, le problème de manque de ressources humaines est ressenti pendant les périodes hivernales. Ainsi la problématique de gestion des ressources humaines a été retenue et investiguée au niveau stratégique (Chapitre 3), tactique (Chapitre 4 et 5), et au niveau opérationnel (Chapitre 6). Ce travail de terrain a également permis de vérifier les hypothèses utilisées tout au long de cette thèse. Particulièrement, les données utilisées pour modéliser un système d'urgence ont été prélevées des bases de données du CHRU (Centre Hospitalier Régional Universitaire) de Lille et les paramètres discutés avec le personnel médical. Nous définissons dans ce qui suit les méthodes utilisées ainsi que les limites qui leurs sont inhérentes.

1.3.1 La Modélisation

Dans les chapitres suivants, nous utilisons des outils mathématiques, tels que les statistiques et l'optimisation dans le but d'apporter une réponse objective au problème de l'engorgement. Tout travail de développement d'outils d'aide à la décision implique un processus de modélisation. La modélisation est relativement simplifiée quand il s'agit d'objets quantifiables tels que des lits d'hospitalisation ou des instruments d'imagerie. Elle est plus délicate quand il s'agit d'objets non quantifiables tels que la satisfaction des patients ou le bien-être du personnel médical. Cette difficulté est particulièrement apparente dans le domaine médical où la santé des patients peut être mise en jeu. Il s'agit souvent de quantifier des éléments qui peuvent avoir des conséquences parfois mésurable à long terme et sans élément de comparaison si on avait agi différemment. Ainsi la qualité d'une solution doit nécessairement être sujette à une interprétation humaine et experte qui va au-delà des simples soucis d'efficacité. Nous assumons, en proposant des solutions dans les chapitres suivants, qu'une discussion s'engage face à une problématique qui ne peut pas prétendre à une solution purement technique.

À notre avis, le meilleur moyen d'échapper à une mauvaise interprétation des résultats est de bien expliciter les hypothèses choisies et la part de subjectif dans les choix de modélisation. Pour bien modéliser nos problèmes nous avons également suivi les conseils et le vécu du personnel médical. Par exemple, nous savons que les médecins passent une part de leurs temps à s'occuper des familles des patients. Cette tâche ne fait pas partie du processus de soin mais elle est essentielle d'un point de vue humain. À ce sujet, nous avons donc choisi de modéliser un temps de service lié aux tâches annexes dans notre modèle de simulation du Chapitre 3. De plus, le fonctionnement des SU est dynamique. Ainsi, les arrivées et les durées des visites doivent être estimées. Nos approximations se fondent sur les bases de données et l'expérience du personnel médical.

1.3.2 Prise en Compte des Aléas

Il est rare que l'aléatoire soit entièrement absent des problèmes pratiques de recherche opérationnelle. En particulier, aux SU, l'arrivée des patients, l'évolution de leur état de santé et la performance des médecins peuvent difficilement être représentés par des processus déterministes. Tout au long de cette thèse, nous traitons ces phénomènes par l'analyse des données historiques qui nous permettent de dégager des modèles et de mettre en avant ainsi une représentation fonctionnelle de l'aléa. Par exemple, nous observons que les arrivées aux SU sont en moyenne plus élevées à midi et en fin de journée que le reste du jour. De même, il y a une augmentation du nombre d'arrivées les jours fériés. À partir de ces observations, nous utilisons des lois d'arrivées probabilistes pour représenter les arrivées. De même, la performance des médecins et les ressources disponibles influent sur le temps de service accordé aux patients. Ces informations sont agrégées dans nos modèles mais traitées avec des variables aléatoires. Le concept d'aléa est difficilement définissable rigoureusement. Nous utilisons cependant le concept d'aléa pour tout événement dont la chaîne de causalité n'est pas connue. Cependant, même si un phénomène n'est pas entièrement défini, nous pouvons définir la plupart du temps une palette de résultats possibles. Cette palette constitue mathématiquement l'univers d'une variable aléatoire.

Le nombre de patients atteints par un épidémie peut être quantifiée par des séries temporelles. Cette représentation permet d'analyser les propriétés statistiques de la série telle que l'espérance et la variance. Mis à part les éléments bien définis, il existe toujours un bruit blanc. Le bruit blanc définit les éléments purement aléatoires de la série. Les études statistiques cherchent à séparer ce bruit blanc des éléments réguliers. Une fois les paramètres définis, des méthodes de régression peuvent être utilisées pour estimer le comportement de la série à des horizons de temps différents.

Cependant, dans notre approche, nous avons besoin uniquement des courbes de tendances des périodes épidémiques et non de prévisions précises. En effet, les méthodes présentées dans les chapitres suivants sont conçues pour ne pas dépendre d'une période hivernale en particulier mais de fonctionner pour toute période présentant les même tendances. Nous utilisons donc des modèles épidémiologiques. Les modèles épidémiologiques sont utilisés pour étudier la propagation d'une épidémie dans une population. Dans ces modèles à compartiments, une population est divisée en différents compartiments selon l'état de santé des individus. Les données historiques nous servent à paramétrer nos modèles épidémiologiques et avoir des courbes de tendances représentatives de la région étudiée. A partir de ces courbes de tendances, nous générons des scénarios types de demande aux urgences et les utilisons pour évaluer nos solutions.

1.4 Organisation de la Thèse

Le Chapitre 2 présente une revue de la littérature sur le thème de l'engorgement aux SU. Les causes les plus fréquemment citées ainsi que les solutions envisagées y sont détaillées. Ce chapitre introduit scientifiquement les chapitres suivants et donne un aperçu sur les problématiques que nous avons traitées. Le reste de la thèse est présenté sous forme d'articles. En particulier les Chapitres 4 et 5 ont été publiés au cours de la thèse et les Chapitres 3 et 6 sont écrits dans la perspective de les soumettre à des revues scientifiques.

Le Chapitre 3 traite du problème du dimensionnement des ressources en période épidémique. Le problème de gestion des ressources humaines est ainsi introduit d'un point de vue stratégique. Les hypothèses posées sont aussi génériques que possible et la solution proposée est d'ordre analytique. Puis, nous traitons du problème d'optimisation des postes de travail du personnel hospitalier au Chapitre 4. Les postes de travail sont optimisés de façon à prendre en compte les interactions entre les différentes ressources et les arrivées aléatoires des patients. Dans le Chapitre 5, nous analysons l'option des heures d'astreintes aux urgences comme solution temporaire pendant les périodes hivernales. Les avantages et les inconvénients sont présentés et quantifiés. Finalement, au Chapitre 6 nous étudions le problème d'affectation des heures supplémentaires aux médecins pour traiter du problème des engorgements au jour le jour. Plusieurs politiques de répartition sont présentées et comparées numériquement. En annexe (Annexe A), nous trouvons un compte rendu de la première visite au CHRU de Lille.

Chapitre 2

Overcrowding in Emergency Departments : Literature Review

Résumé en français du chapitre : Revue de la Littérature sur l'Engorgement aux Urgences

Nous présentons dans ce chapitre un état de la littérature sur le thème d'engorgement aux services d'urgences hospitaliers. Le travail effectué au cours de cette thèse est ainsi positionné par rapport aux études existantes dans le but de souligner les différences et les particularités de notre étude. Nous définissons les principales méthodes utilisées et les regroupons selon deux axes : les études liées aux flux des patients et celles qui traitent de la gestion des ressources. En ce qui concerne le flux des ressources, nous faisons la différence entre les flux rentrants et les flux sortants des urgences. Les études liées aux flux rentrants abordent à la fois le thème de gestion des ambulances et celui de l'acheminement des patients admis aux urgences. Deux solutions récurrentes dans cet axe sont le déroutement des ambulances et l'ouverture d'une filière rapide pour les patients sans pathologie grave. En ce qui concerne l'optimisation des flux sortants, une des solutions typiquement proposée pour améliorer le flux de sortie des patients est l'optimisation des lits d'aval. Il s'agit de gérer la capacité en lits dans les services hospitaliers qui peuvent accueillir des patients sortant des urgences afin de prendre en compte l'incertitude liée au volume de patient. Pour ce qui est de la gestion des ressources, nous définissons deux sous catégories de problèmes. Les problèmes de gestion des ressources humaines et ceux liés aux examens complémentaires. Nous détaillons uniquement les études sur la gestion des ressources humaines afin de comparer les différentes méthodes et introduire les travaux présentés dans les chapitres suivants. Dans un premier temps, nous faisons abstraction du contexte des services d'urgences et nous étudions la gestion des ressources humaines d'un point de vue méthodologique. Les particularités des services d'urgences sont ensuite présentées avec les difficultés qu'elles engendrent. Nous examinons ensuite les études qui se sont faites sur le thème et relevons les lacunes apparentes pour mener à bien notre pro-

jet. Finalement, nous nous intéressons au problème de modélisation de la demande aux services d'urgences hospitaliers. La modélisation de la demande est fondamentale pour caractériser les périodes de tension et proposer des solutions pertinentes et réalistes. Les principaux outils que nous utilisons dans cette thèse pour modéliser la demande sont les modèles d'épidémiologie. Nous définissons ainsi dans une dernière partie les différents modèles d'épidémiologie présents dans la littérature et les applications faites à partir de ces modèles.

Sommaire

2.1 Patient Flow Management 15
2.2 Resources Management 17
2.2.1 Equipment Management 17
2.2.2 Workforce Planning 18
2.3 Modeling the Demand 21

Overcrowding is defined as a situation where the demand is greater than the service capacity. In our context, the demand is the workload incumbent on the ED resources. There are many indicators of overcrowding in EDs but no consensus on one across hospitals. For example, when the waiting time of patients becomes prohibitively large, it is an indication that the service capacity is not sufficient. Another indicator can simply be a ratio between the patients and the resources present. Depending on the cause, ED overcrowding can require long term strategical solutions or require more operational solutions. For example, demographic changes in a population are slow to come by. As such, dimensioning a unit or hiring new resources is not needed frequently. Conversely, epidemic periods can be unexpectedly difficult. In such cases, operational solutions are needed to cover the surge demand.

Patients misinformation and inappropriate use of EDs have been cited frequently as major causes to ED overcrowding [Afilalo et al., 2004, Howard et al., 2005, Andersen, 1995]. Indeed, non-urgent visits from patients who could have sought health care elsewhere without a risk to their health condition occupy resources in vain. What is more, non-urgent visits are often from frequent flyers that show up to EDs regularly for no reason.

Seasonal epidemics is another frequently cited cause of overcrowding [See et al., 2009, Rico et al., 2007, Rico, 2009, Bienstock and Zenteno, 2012]. During seasonal epidemics, a large part of the population is afflicted by digestive or respiratory illnesses. These pathologies are particularly problematic in infants and elderly people because they are more fragile. As the context of this study is that of a pediatric emergency department, we are interested in studying overcrowding caused by seasonal epidemics.

ED overcrowding has be addressed in many fields of study (Operations Research, Mathematics, Management Science, Medicine, Epidemiology, Statistics etc). Consequently, the literature on the topic is immense [Guo and Harstall, 2006,Hoot and Aronsky, 2008,Wargon et al., 2009]. We focus in this chapter on solutions which tackle the organization inside EDs. In Section 2.1 we review the works that have addressed patient flow management problems. Then in Section 2.2, we review works on resource management. A particular attention is given to articles that address human resource management problem as it is the main focus of this thesis. Finally, in Section 2.3, we give an overview on the different demand modeling methods used in the literature and how we used them in the rest of this thesis.

2.1 Patient Flow Management

Patient flow management is the problem of distributing patients' workload equitably across space and time. From a strategic point of view, facility location and sizing problems have been addressed to better equate the demand in a region to the service levels [Daskin and Dean, 2004,Hajipour et al., 2014,Jia et al., 2007,Mestre et al., 2015]. For example, in a paper by Hajipour et al. [Hajipour et al., 2014] a location-allocation problem is presented to minimize the aggregate waiting time of customers and minimize the maximum idle time of facilities.

The case of health care facilities is considered in several studies as well. For example, a review of the main models used in health care is analyzed in a paper by Daskin and Dean [Daskin and Dean, 2004]. Similarly, Mester et al. analyze the specific case of hospital network planning [Mestre et al., 2015]. In particular, they take into account the uncertainty typical of the health-care industry and possible technological advances. A paper by Jia et al. [Jia et al., 2007] addresses the specific case of large-scale emergencies in a location problem for medical supplies. Large scale emergencies such as natural disasters and pandemic crises require special considerations because they are punctual phenomenons that have a big impact on the coverage rate of the proposed solution. In Chapter 3, we propose a framework for right-sizing a facility in the hospital by taking into account fluctuating demand patterns. The framework is generic enough to account for epidemic periods without difficulty.

On the operational side, patient flow management methods such as ambulance diversion and patients' pathway management are common in EDs. Ambulance diversion occurs when the closest ED facility to the patient is overcrowded and asks that patients be diverted to neighboring facilities. There is usually no systematic rule that defines the proper use of diversion. Diversion is based on the perception of decision makers and liable to misuse [Vilke et al., 2004, Patel and Derlet, 2006]. There are many disadvantages to ambulance diversion. For starters, patients take more time to arrive to an ED and hence risk a deterioration of their health condition. Moreover, neighboring facilities have the risk of being overcrowded in turn. Vilke et al. [Vilke et al., 2004] study the effect of avoiding diversion in one facility on neighboring facilities. Their analyzes shows that ambulance diversion can lead to an oscillatory phenomenon where diversion in one facility induces diversion in another facility. Patel and Derlet [Patel and Derlet, 2006] describe a region-wide program to reduce ambulance diversion. The program is based on a more structured use of ambulance diversion and a synchronization between different neighboring facilities.

After patients arrive at the ED, they go through a series of procedures that vary slightly from patient to patient. One idea for improving responsiveness in EDs is to study the patients' pathway and possibly improve its design. A solution that has been studied extensively in EDs is opening a fast track for patients with less serious symptoms [Sven et al., 2011]. Fast tracks have been shown for the most part to improve waiting time and length of stay. For instance, in a study by Sanchez et al. [Sanchez et al., 2006] a cohort study showed that opening a fast track can lead to a 50% reduction in the waiting time of patients and a 10% reduction in the length of stay.

Other, less successful streaming mechanisms exist. For example, Saghafian et al. [Saghafian et al., 2012] propose a patient-streaming mechanism that separates patients on the basis of whether they are expected to be hospitalized or not. The authors find that this streaming mechanism is most useful when the workload in EDs is important and there is a good share of variability in the percentage of admitted patients from day to day. Their findings also suggest that this mechanism works best when the resources are shared among the different streams. A simulation study has been proposed by Marmor et al. to compare different streaming mechanism and help decision makers fit a streaming mechanism to their particular ED [Marmor et al., 2012]. Their work also finds that a fast-track mechanism is best in the case of an aging population.

Even after patients have been treated in the ED, organizational problems can still arise. For instance, when after an ED visit, patients need to be transferred to other departments in the hospital, physicians have to find available beds for the patients in downstream units. This process is time consuming and studies have dealt with the problem of optimal appointment scheduling in downstream units to account for these emergency arrivals [Howell et al., 2008, Mazier, 2010]. Howell et al. suggest in a comparative study [Howell et al., 2008] that a hospital-wide bed management strategy could improve ED throughput. The authors find that during an intervention period where hospital's beds management was centralized, the number of hours when the ED was on alert due to lack of intensive care unit beds decreased by 27%. In the same line, the thesis of Mazier [Mazier, 2010] proposes a solution to the problem of patient flow management in EDs' downstream units by solving a succession of decision problems.

2.2 Resources Management

Besides patient flow management strategies, resource management strategies organize resources to better handle the workload. In a sense, the two solution categories are intertwined because modifying the patient flow necessarily changes the resources' work patterns and vice versa. However, when examining resource management strategies we consider constraints on the hospital's resources that are not necessarily taken into account elsewhere. We divide the solutions the resource management solutions into human resource management strategies and equipment management strategies.

2.2.1 Equipment Management

EDs are not isolated entities inside the hospital. As such, ED efficiency has been shown to depend on the efficient operation of units connected to it [Mazier, 2010]. As we have seen in the previous section, bed management in downstream units can improve the output flow of EDs. Similarly, appointments scheduling in operating rooms or imaging units has a direct impact on the quality of care in EDs. In this Section we give a quick overview of some of the works that have been done on the topic and we refer to the thesis of Mazier et al. [Mazier, 2010] for a more detailed examination of the related works. The two main topics in that domain are resources management, and bed management. Each of these topics has been examined under different perspectives and solved for different time horizons.

About one third of patients that visit EDs go through auxiliary exams. The purpose of these exams is to confirm or repudiate the diagnosis established by physicians. The exams can be simple blood tests or imaging exams. However, the imaging equipment used by EDs is often shared with other departments in the hospitals and are very solicited. Unlike EDs, other departments typically schedule their patients ahead of time. Since the imaging equipment is expensive, it is preferable that there is no idle time on these resources. Consequently, hospitals tend to want to book a large time slot for non-urgent patients. This behavior is problematic for urgent patients that have to wait accordingly. As a result, some resources management studies propose solutions taking into account emergency arrivals. A study by Patrick and Puterman [Patrick and Puterman, 2007] present an model to find reservation policies for diagnostic services taking into account uncertain demand. Another work by Patrick et al. proposes a dynamic scheduling method that considers patients with different priorities [Patrick et al., 2008].

Similarly, bed management strategies are complicated by the fact that emergency arrivals are unpredictable. On this topic, a paper by Huang [Huang, 1995] analyzes in-patient bed requirement considering both elective and emergency beds. A queuing model is proposed to include the variations typical of emergency arrivals. Harper and Shahani [Harper, 2002] present simulation models with different patient flows to examine the effects of capacity planning decisions on bed occupancy and refused admissions rates. Bekker and Bruin for their part present an analysis of the impact of arrival patterns on bed occupancy using queuing models [Bekker and De Bruin, 2010]. The arrival patterns in their case are the hourly and weekly arrivals patterns of patients. A similar study can be considered for epidemic periods which also displays specific arrival patterns. In the same way, treatment rooms can also be over solicited. For instance, in Dutch hospitals, treatment rooms in EDs are shared by two doctors [van de Vrugt and Boucherie, 2015] and thus assignment strategies can be used to reduce the waiting time of patients.

2.2.2 Workforce Planning

In general, human resources are scarce and expensive resources. Additionally, in hospitals, physicians and nurses are responsible for the health condition of patients. For these reasons, the problem of human resources planning has been the topic of a host of papers [Rico et al., 2007,Rico, 2009,Bienstock and Zenteno, 2012,Garg et al., 2012,Beaulieu et al., 2000, Brunner and Edenharter, 2011, Brunner et al., 2009, Green and Soares, 2006,Izady and Worthington, 2012,Sinreich et al., 2012,Mobasher et al., 2011,Xiao et al., 2010,Yankovic and Green, 2008,Yankovic and Green, 2011]. An early paper by Tien and Kamiyama [Tien and Kamiyama, 1982] reviews some of the seminal works on employee scheduling. The authors define 5 stages to employee scheduling problems and explain the different solution procedures used in the literature. A more recent paper by Van den Bergh et al. [Van den Bergh et al., 2013] classifies previous works on employee scheduling according to the methods used and the contexts of each paper. The authors also compare different review papers done in the past. For our purpose, we break down the problem of human resources scheduling into 4 sub-problems that overlap more or less depending on the approach.

The first problem is that of demand modeling. Finding an accurate way to model the demand is important to adequately represent the workload per period of time. For example, in our case, it is important to take into account the bell shaped pattern of the demand across the epidemic horizon. We explain in more details in Section 2.3 how we modeled the demand in the rest of this thesis. Once the demand is modeled, a staffing problem is solved to find staffing levels that match the demand given the overall workforce. However, finding optimal staffing levels is not enough if it is not linked to work regulations. Indeed it is not always easy to adjust the work days and hours of employees with those of optimal staffing levels required per period. For that reason, the third problem is that of shift scheduling. Shift scheduling is the process of assigning different types of predetermined shifts to match more or less the staffing levels. The shifts are designed in a way to ensure work regulations are included.

In the health-care domain, the literature on nurse scheduling has attracted researchers from early on. For example, in 1976, Miller et al. [Miller et al., 1976] describe a nurse scheduling problem and study the trades-off between nurses' preferences and staffing coverage. Similarly, another paper [Warner, 1976] formulates a nurse scheduling system including nurses' preferences in the objective function. More recent papers have addressed increasingly complicated constraints and larger problem instances by applying state of the art optimization tools. For instance, Guthjhar and Rauner [Gutjahr and Rauner, 2007] use an ant colony optimization (ACO) approach to solve the nurse scheduling problem dynamically for the next day. Brunner and Edenharter [Brunner and Edenharter, 2011] for their part formulate a mixed-integer program (MIP) to determine optimum staffing levels and shift assignments for physicians with different skill sets. However in that paper, there is no clear description on the method to obtain the demand per period. In Chapter 4 in order to determine the shift schedules of employees, we describe patients trajectories so as to model the demand explicitly.

Some studies have addressed the problem of employee scheduling specifically in EDs [Green and Soares, 2006, Ingolfsson et al., 2002, Yankovic and Green, 2011, Beaulieu et al., 2000, Carter and Lapierre, 2001, Tan et al., 2013]. Among these papers, most use queuing approaches to model the ED efficiently. In a paper by Green et al. [Green and Soares, 2006], the authors use queuing analysis to study the impact of staffing levels on the proportion of patients that leave without waiting to be diagnosed by a physician. The primary difficulty they encounter in the ED context is that the demand is time dependent. In other words patients' arrival patterns change according to the hour of the day and the day of the week. Similarly, these patterns change during seasonal epidemics. The first thing that comes to mind to resolve this issue is to divide the problem into a series of sub-problems that reflect the different arrival patterns. However, the authors show that in the context of the ED assuming each period is independent with a stationary demand does not work because the peak congestion hours do not exactly coincide with the arrival rates. Indeed, patients' length of stay in the ED is longer than the change in arrival rates. This phenomenon creates a lag between the peak arrival periods and the peak congestion periods. The authors then use a modified approach that integrate this lag in order to gain insights into the impact of staffing level on the service quality. Similarly, a paper by Ingolfsson et al [Ingolfsson et al., 2002] tries to find both optimal staffing levels and shift schedules that meet the work constraints by taking into account the time-varying queuing effects. A later paper by Yankovic and Green [Yankovic and Green, 2011] develop a queuing model to evaluate the impact of different work parameters in the ED on the staffing levels required to attain a level of performance. One limitation of these studies is that the workforce is considered homogeneous. It is clearly not the case in EDs where patients are served by different resources (doctors, nurses, lab technicians, etc..).

A few articles did consider the case different servers interacting at different stages during the patients' visit [Izady and Worthington, 2012,Sinreich and Jabali, 2007,Sinreich et al., 2012]. For example, a article by Izady and Worthington [Izady and Worthington, 2012] studies the EDs scheduling problem by modeling the system as a multi-class queuing network where classes of patients represent different patients with different estimated trajectories through the system and the different servers are the different type of human resources to be scheduled (physicians, nurses, technicians). They use a network extension of the staffing law to determine optimal staffing levels throughout the day and then apply a MIP to calculate shift schedules that meet the staffing required. Sinreich et al [Sinreich et al., 2012], for their part, develop an iterative scheme for finding shift schedules to the critical resources in the ED. At each iteration, a different resource is handled and an optimization model is solved to adjust the capacity accordingly. In Chapter 4, we use a similar queuing network representation of EDs but solve our staffing and shift scheduling problem simultaneously using a stochastic MIP formulation of the problem.

The last stage in the employee scheduling literature is to assign employees' to shift schedules. Of course, in many papers this problem has been solved along with the shift-scheduling problem. However, the assignment problem requires a very clear view on the potential preferences and constraints that each employee has. Furthermore, most papers dealing with the assignment phase in details have simplified demand models. For our purpose, we allow certain employee specific criteria in our models but do not deal with the assignment phase explicitly. For example, in Chapter 5 when finding staff allocations schemes that include on-call duties, we consider employee specific parameters for on-call hours. Similarly, in Chapter 6 when defining the assignment of overtime hours, the number of overtime hours to be assigned is employee specific which allows for decision makers to take into account employee specific constraints and preferences.

2.3 Modeling the Demand

Representing the demand is a key for any improvement to be made in the EDs. Conversely, the stakes from misrepresenting the demand is high as it can lead to erroneous readings of the overcrowding causes. There is not straightforward definition of "demand" in EDs because each entity (physician, nurses, administrative staff) perceives demand differently. Nonetheless, a common starting point to quantify the demand is the number of patients present in the ED at any given moment. In some cases, the number of patients might not be enough because it overshadows important points such as the length of stay and the resources needed. A good characterization of the demand is thus one that takes into account the key elements that we are trying to influence. For instance, if we want to determine the number of receptionists needed to have patients wait less than 5 minutes on average at the reception desk, an estimate on the number of arrivals might be enough. This is because the registration process is relatively straightforward and does not involve many resources. However, for determining optimal physicians' staffing levels, the service times have to be taken into account as well.

At a strategic level, demand estimation helps right-sizing the department. More operationally, demand estimation can shed a light on the problematic steps in a patient trajectory. For example, if we follow patients' trajectory throughout the ED, we might find that the peak demand hours for lab technicians is different than the peak demand hours for physicians or nurses because of the non-negligible service times. Generic forecasting models are often used when studying EDs and are believed to help decision makers take more informed decisions regarding the management of resources [Jones and Thomas, 2008, McCarthy et al., 2008]. For example, a study by Jones et al. [Jones and Thomas, 2008] compares the accuracy of several forecasting methods such as linear regression and artificial neural networks in the context of EDs. McCarthy et al. [McCarthy et al., 2008] on the other hand, propose a forecasting methodology that incorporates "temporal, climatic and patient factors". The problem with these forecasting methods is that it is difficult to understand the significance of the parameters that are used in the model. Indeed in the context of epidemics, it is hard to establish a link between the type of epidemic and the estimations made using forecasting models.

An alternative way to represent arrival patterns during epidemic season are based on epidemic models. These models have been often used in the operations research literature to represent the demand in health-care settings and link that demand to epidemic parameters.

Epidemic models help us study the dynamic of disease propagation inside a population. Given certain parameters of the epidemic of interest, these models can be used to simulate the epidemic propagation and follow the evolution of the number of infected individuals across days of the epidemic horizon. From the number of infected individuals we estimate the number of individuals that arrive to the ED on any given day.

The diseases of interest in this study are those that posses a seasonal pattern and that cause influenza like symptoms. These type of disease can be grouped under the general term of Influenza-like-Illnesses (ILI). In the same lines as [Allen and Burgin, 2000] we use throughout this thesis a discrete-time stochastic SEIR model implemented as a Markov chain with finite state space. The SEIR model is an epidemic model that considers four categories of individuals during disease propagation (Susceptible S, Exposed E, Infected I, and Recovered R). Initially, the model starts with a few infected individuals. Then individuals evolve through the four stages following a dynamic translated in differential equations. Throughout the thesis, we use a SEIR models to generate our test instances. The following assumptions are made for all the models we used.

1. We consider a closed population of size N where demographic changes such as births, deaths (natural) and emigrations are ignored. We think this assumption is reasonable during seasonal epidemic crises when demographic changes are really negligible.
2. Only one epidemic wave occurs at a time. It is possible and likely that ED experience two or more epidemic waves in one year, however we assume those phenomena independent or are absorbed by the uncertainty in epidemic crises.
3. The population is assumed homogeneous in the sense that individuals are equally likely to get infected during an epidemic wave. This assumption is partly justified by noting that hospitals often have a separate pediatric ED. Therefore, the population groups that are most different with regards to contact rates (adults and children) are necessarily studied independently. The data we have from our partner hospital is specific to a pediatric ED. As such, having only one population group is not really problematic.
4. There is an incubation period during which individuals are exposed but not yet infectious. Individuals that recover from the virus are conferred immunity to the disease and are not susceptible again. This is typical of disease such as influenza and gastroenteritis even if the incubation period varies in length.

The power of epidemic models is then that we can parametrize the models to fit specifically the disease we want to consider. In Table 2.1 we note down the parameters that can be modified in an SEIR model to give an overview on what type of test instances can be created.

There are many example uses of epidemic models in the literature. Parvin et al [Parvin et al., 2012] study the problem of finding health policies to treat two stage-contagious diseases. An epidemic model is used to represent the different states a population can be in and a framework is developed to help decision makers allocate their budget optimally.

Parameter	Description
p	Contagion rate of the epidemic
μ_E^{-1}	Latent period
μ_R^{-1}	Infectious period
N	Total Population Number
ϵ	Initial Number of Infected Individuals

TABLE 2.1 – Parameter Values for the Epidemic Model

Yarmand et al [Yarmand et al., 2014] deal specifically with the problem of vaccine allocation to different regions. Similarly, they simulate the disease spread with an epidemic model and use it as a base for decision making. Finally, Bienstock and Zenteno [Bienstock and Zenteno, 2012] study the problem of surge resources during influenza epidemics by relying on an SEIR epidemic model.

Unlike other studies, we did not "hard-wire" the epidemic model to the decision making process. The results in later chapters are agnostic of the mechanism of the epidemic model. This renders the results more general and usable in different settings. For example the models in chapter 3 and 5 use simple random variables to represent the demand. We chose to generate test instances using the epidemic model because they reflect the process of epidemic that is pertinent in our case.

Chapitre 3

Capacity Planning in Hospitals during Epidemics

Résumé en français du chapitre : Dimensionnement des ressources aux urgences pendant une période épidémique

Dans ce chapitre nous étudions le problème de dimensionnement des ressources dans les hôpitaux dans un contexte d'épidémies saisonnières. Pendant les saisons épidémiques, les hôpitaux sont souvent surchargés mais il est difficile de prévoir en avance et avec certitude la répartition de la demande sur l'horizon épidémique. De ce fait, il devient difficile de définir des niveaux adéquats de ressources sur la période épidémique. Par exemple, une estimation sur le nombre de lits d'urgence nécessaires permettrait de réserver des lits dans d'autres services si le besoin se présente. De même, une estimation du nombre de médecins requis sur l'horizon épidémique permettrait de gérer le planning avec anticipation.

Afin de répondre à ce problème, nous proposons une formulation mathématique simplifiée modélisant la demande en variable aléatoire arbitraire. Nous ne faisons pas d'hypothèses a priori sur la variable aléatoire et laissons ainsi le choix de représenter la variable par différentes méthodes. Nous proposons ensuite une relaxation Lagrangienne du problème et définissons une politique optimale de répartition des ressources sur l'horizon épidémique. Afin de rendre la formulation plus réaliste, nous ajoutons ensuite des contraintes de bornes au modèle. Ainsi, nous imposons que la solution optimale respecte une limite de ressources à distribuer par période. Le nouveau problème est résolu de façon similaire en utilisant une relaxation Lagrangienne et une nouvelle politique optimale est trouvée. Finalement, nous proposons quelques exemples numériques pour concrétiser nos propos et illustrer les solutions obtenues. Nous comparons les solutions avec des demandes générées par un modèle épidémiologique et des demandes qui suivent une loi aléatoire.

Sommaire

3.1 Introduction 27
3.2 Literature Review 27
3.3 Problem Description 28
3.3.1 Optimal Characterization 29
3.3.2 How to compute an optimal solution 31
3.3.3 Normal Approximation 31
3.3.4 Upper Bound Constraints 32
3.3.5 Minimization of the largest expected capacity shortage 35
3.4 Numerical Examples 36
3.4.1 Epidemic Simulations 37
3.4.2 Demand Distribution 39
3.4.3 Capacity Upper Bounds Influence 39
3.4.4 Min Max Criterion 39
3.5 Conclusion 41

3.1 Introduction

In France expansions of hospitals are subsidized by the government. With the reduction of government subsidies and the strict "fee for service" rates, hospitals are struggling to stay operational. Thus, there is an urgent need to right-size the facilities by determining appropriate capacity levels. Capacity can be thought of in terms of human resources or equipment capacity. In this chapter we study a capacity planning problem in the context of seasonal epidemics. During seasonal epidemics the problem of right-sizing the facilities is accentuated because the demand is very fluctuating. Indeed, the demand dependends, among other things, on the type of the epidemic and the size of the hospital. It is then important for decision makers to know ahead of time whether the capacity available for the epidemic period is enough to ensure a good coverage of the demand. In order to right-size the facilities, we formulate a capacity allocation problem and find optimal allocation policies. Using these allocation policies we can estimated the coverage rate given a initial global capacity C. Because we are working in the epidemic context, we make no assumptions on the demand other than that it is estimated by a random variable with a monotonically increasing cumulative distribution. Optimal allocation policies are analytically characterized and numerical examples are used to illustrate the coverage rates in concrete examples. In Section 3.2, we go through the related literature on capacity management in crisis situations. Then, in Section 3.3 we describe the problem formally and propose a mathematical model. Finally, in Section 3.4 we test our solutions on different epidemic scenarios.

3.2 Literature Review

In terms of capacity, hospitals have three types of resources to manage : beds, human resources and medical equipments [Brandeau et al., 2004]. Adequately managing each of these three resources is fundamental because of their impact on the service quality. Many studies have addressed the issue of capacity management from various angles and at different planning horizons. In particular at the strategic level, the issue of right-sizing facilities has been addressed in several studies [Rais and Viana, 2011].

From early on, mathematical models have been used in hospitals to better calibrate the capacity and organize its operations. It is generally agreed upon that average values and ratios are insufficient to appropriately determine the needs in capacity. Indeed, average values are misleading because of uncertainty in the demand [Brandeau et al., 2004, Harper, 2002]. Hence, different approaches ranging from queuing theory [Jennings, 1996, Green et al., 2001, Green and Soares, 2006, Bruin et al., 2009, Cochran and Roche, 2009, Yankovic and Green, 2011, Izady and Worthington, 2012] to linear programming and simulation [Harper, 2002, Omar et al., 2014, Brunner and Edenharter, 2011] have been used for the calculation of capacity in hospital settings. For example, Harper [Harper, 2002] proposes an integration of a patient classification system and simulation to accurately determine the needed capacity in different hospital resources. Similarly, Yankovic et al. [Yankovic and Green, 2011] represent a clinical unit using queues with two type of servers : nurses and beds. They show that the dynamics of the system and interplay between resources is important in calculating good staffing levels.

Some studies have looked into the impact of surge demand and epidemics on capacity planning [Valdmanis et al., 2010, Arora et al., 2010, Rico, 2009, See et al., 2009]. For example, Valdmanis et al. measure capacity utilization in hospitals following a major disaster using data envelopment analysis [Valdmanis et al., 2010]. Their findings suggest that some services in the hospitals studied are not capacited to absorb surge demand following a disaster. Similarly Arora and al. deal with the resource allocation problem following a health emergency [Arora et al., 2010]. They perform a cost benefit based optimal to best provide regional aids by distributing vaccines. However all of these studies have approached the problem from the point of view of specific settings and did not look at the problem in its generic form. As such, in papers using queuing theory, assumptions on the demand distributions and resources service times are made. Similarly, in surge demand studies, problem specific constraints are introduced.

In this work we propose a formulation of the capacity allocation problem with little assumptions with the purpose of having a generic framework to right size facilities. Based on our formulation, we characterize the solution space of the problem. We make no assumptions on the demand other than that it is quantifiable using random variables with a monotonically increasing cumulative distribution. Having little assumption on the distribution of the demand ensures that the characterization is relevant in many settings where the workload can be roughly estimated. We then add an upper bound constraints at each period to make the solutions more realistic. The contributions of this chapter are twofold :

1. Define an optimal capacity allocation policy using little assumptions
2. Propose a framework to right-size facilities using the policy

3.3 Problem Description

In hospitals there are certain characteristic demand patterns. These patterns are related to the arrival tendencies of patients. For example the weekday arrivals are different than week-ends' arrivals. Similarly, night arrival patterns differ from day arrival patterns (c.f Chapter 5). During seasonal epidemics, the demand is known to increase as the epidemic expands then decrease with the extinction of the epidemic. Heedless to this fact, hospitals usually follow a uniform allocation of resources across the year. In this section we propose a mathematical model to find capacity allocation policies more in line with the observed epidemic patterns.

Let d_t be a random variable representing the work demand required on periods $t = 1..T$ of the planning horizon. Let c_t be the capacity allocated to period t to meet this demand. When dealing with human resources, capacity usually refers to man-hours available. Man-hours can be decomposed into regular man-hours, overtime and surge staff man-hours. For the sake of simplicity, we ignore those distinctions and only assume one type of capacity in this paper. Note that such a model remains relevant when different types of resources are independently scheduled.

In hospital settings, a common scenario is to have a fixed number of resources C. This is commonly the case during epidemic seasons which are limited in time and during which it is difficult to change the number of resources. Consequently, the total capacity is an parameter in our problem and the objective is to minimize the capacity shortage across the horizon. The shortage is computed as the expected positive difference between demand and capacity over all periods. In the rest of this chapter we study the problem and extensions formulated as follows :

$$\min_{c_t} \sum_t \mathbb{E}[(d_t - c_t)^+] \tag{3.1}$$

$$\sum_t c_t = C \tag{3.2}$$

Equivalently, the objective function can be written as $\mathbb{E}\sum_t[(d_t - c_t)^+])$. The notation $(.)^+$ in objective (3.1) represents the positive part of the function. Constraint (3.2) sets the global capacity. In other words, C is the total of resources available for the planning horizon. As all terms of the objective function are decreasing functions, an inequality symbol $\leq$ can replace equality in constraint (3.2) without changing optimal value. In order to solve this problem, we study in the next section the Karush-Kuhn-Tucker optimality conditions given an arbitrary distribution of the demand and derive a generic property for our solution at optimality.

3.3.1 Optimal Characterization

In order to solve Problem (3.1)-(3.2), we let $L(\mathbf{c}, \lambda)$ be the Lagrangian function of our problem with λ the Lagrange multiplier for the equality constraint (3.2) and $\mathbf{c}$ the vector of c_t variables. Let $\mathbf{c}^*$ and λ^* be optimal solutions, then the gradient of the Lagrangian function $L(\mathbf{c}, \lambda)$ must vanish at these points. At optimum we must have that :

$$\nabla\ L(\mathbf{c}, \lambda) = \nabla(\sum_t \mathbb{E}[(d_t - c_t)^+] + \lambda(\sum_t c_t - C)) = 0 \tag{3.3}$$

Expanding on the necessary condition (3.3) we obtain :

$$\frac{\partial L(\mathbf{c}, \lambda)}{\partial c_t} = \frac{\partial \mathbb{E}[(d_t - c_t)^+]}{\partial c_t} + \lambda = 0,\ \forall t \tag{3.4}$$

$$\frac{\partial L(\mathbf{c}, \lambda)}{\partial \lambda} = \sum_t c_t - C = 0 \tag{3.5}$$

Let $\mathbf{I}$ be the indicator function. The partial derivatives with respect to c_t of equations (3.4) can be rewritten as :

$$\mathbb{E}[-\mathbf{I}(d_t - c_t \geq 0)] + \lambda = 0 \tag{3.6a}$$
$$-P(d_t \geq c_t) + \lambda = 0 \tag{3.6b}$$
$$-(1 - F_{d_t}(c_t)) + \lambda = 0 \tag{3.6c}$$
$$F_{d_t}(c_t^*) = 1 - \lambda^* = \alpha^* \tag{3.6d}$$

With $F_{d_t}(c_t) = P(d_t \leq c_t)$ being the cumulative distribution function of our variables. The result in equation(3.6d) say that in order to reach an optimum capacity allocation, the coverage probability of the demand α must be the same across all periods. In other words, the capacity allocated to each period must be such that the probability of being short in resources is equal across all days. An example solution with three periods is illustrated in Figure 3.1. In that figure, the curves represent the cumulative distribution functions of the demand for periods $t = 1..3$. For a given value of α, we can identify the corresponding levels c_t by projecting a line on the x-axis. We refer to this optimal allocation policy as **P1**.

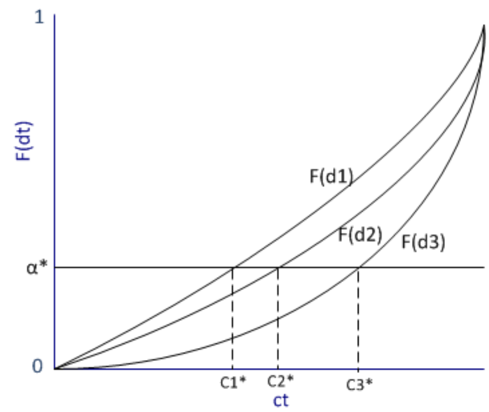


FIGURE 3.1 – The cumulative distribution functions of three periods along with the optimal solution corresponding to α^*.

Equation (3.6d) and the total capacity constraint (3.2) constitute necessary but not sufficient conditions for optimality. In order to show that the above characterization is sufficient for optimality, we show that in our context a solution $\mathbf{c}$ that satisfies these conditions is unique. We first justify why we can restrict our study to a demand quantifiable using a random variables with a monotonically increasing cumulative distribution.

Assumption 1. *The cumulative distribution function of the demand for any period is monotonically increasing.*

Explanation :
By definition, in order for the cumulative distribution function not to be monotonically increasing, there must exist some interval $[a, b]$ where the probability density function of d_t is nil. However, in the context studied this means that the probability of having demand fall in a sub interval $[a_1, b_1] \subset [a, b]$ is nil. For emergency department arrivals, it seems reasonable to assume that such cases are not realistic.

Theorem 1. *A solution vector that satisfies equations* (3.6d) *for each period t and constraint* (3.2) *is globally optimal.*

Proof From Assumption 1, for a given α there is only one solution vector $\mathbf{c}$ that satisfies equation (3.6d), as the sum of strictly increasing functions is also strictly increasing. Then $\mathbf{c}^*$ satisfying **P1** is a global optimum. □

3.3.2 How to compute an optimal solution

In practice, the solution can easily be obtained using a bisection method on the coverage probability using the cumulative distribution functions of the demand F_{d_t}. We first set $\alpha = 0.5$ and evaluate the corresponding values of c_t for each period. We sum the values of c_t and compare the sum to the global capacity C. If the sum is greater than the global capacity, then we know we have to decrease the coverage ratio. Inversely, if the sum of c_t is less than the global capacity, then we know we can increase the value of α. The stopping criteria of the bisection method is the satisfaction of constraint (3.2) when the solution vector $\sum_t c_t^* = C$.

3.3.3 Normal Approximation

If the demand in the context studied can be approximated by a normal distribution, then a similar reasoning can be followed to analytically find the optimal capacity allocation. Let :

$$d_t \sim \mathcal{N}(\mu_t, \sigma_t) \tag{3.7}$$

We can write the demand in standardized form for each period t as follows :

$$\forall t, \frac{d_t - \mu_t}{\sigma_t} = \mathcal{N}(0, 1) = z \tag{3.8}$$

From Equation (3.6d) we know that :

$$P(d_t \leq c_t^*) = \alpha^*, \forall t \tag{3.9}$$

Rewriting this with the normal demand we obtain :

$$\alpha^* = P(\frac{d_t - \mu_t}{\sigma_t} \leq \frac{c_t^* - \mu_t}{\sigma_t}), \forall t \tag{3.10a}$$

$$\alpha^* = P(z \leq \frac{c_t^* - \mu_t}{\sigma_t}),\ \forall t \tag{3.10b}$$

$$\alpha^* = \phi(\frac{c_t^* - \mu_t}{\sigma_t}),\ \forall t \tag{3.10c}$$

As the cumulative distribution function of $\mathcal{N}(0,1)$, $\phi(.)$ is monotonically increasing, we can consider $g(\alpha)$ the inverse of $\phi(.)$, then :

$$g(\alpha^*) = \frac{c_t^* - \mu_t}{\sigma_t},\ \forall t \tag{3.11a}$$

$$c_t^* = \mu_t + g(\alpha^*)\sigma_t,\ \forall t \tag{3.11b}$$

$$\sum_t c_t^* = \sum_t \mu_t + g(\alpha^*) \sum_t \sigma_t \tag{3.11c}$$

$$g(\alpha^*) = \frac{C - \sum_t \mu_t}{\sum_t \sigma_t} \tag{3.11d}$$

Finally, $g(\alpha^*)$ is a obtained as a simple function of the total capacity, the mean values and the standard deviations of random variables that define the demand. Replacing $g(\alpha^*)$ by its value in equations 3.11b, gives the optimal solution.

3.3.4 Upper Bound Constraints

So far the model proposed does not restrict the distribution of the capacity in any way. In practice, it is normal for the capacity allocation to be constrained because of labor law and practical constraints. In particular, it seems reasonable to have an upper bound on the number of resources allocated per period. Having this type of constraint also prevents strategies that concentrate the capacity in certain periods and neglect the rest. Let m be an upper bound on the number of resources that can be allocated at period t. Constraints (3.14) are added to the mathematical model of the problem defined by equations (3.1)-(3.2) to give the problem defined by equation (3.12)-(3.14) :

$$\min_{c_t} \sum_t \mathbb{E}[(d_t - c_t)^+] \tag{3.12}$$

$$\sum_t c_t = C \tag{3.13}$$

$$c_t \leq m_t,\ \forall\ t \tag{3.14}$$

We assume here that the total capacity available exceeds the sum of upper bounds ($\sum_t m_t \geq C$) ; otherwise the problem defined by equations (3.12) (3.14) is infeasible. Note that this assumption allows to replace equality by inequality ($\leq$) in constraint (3.14).

The optimality conditions on the Lagrangian function related to problem defined by equations (3.12)-(3.14) are :

$$\nabla\ L(\mathbf{c},\lambda,\gamma) = \nabla(\sum_t \mathbb{E}[(d_t - c_t)^+] + \lambda(\sum_t c_t - C) + \sum_t \gamma_t(c_t - m_t)) = 0 \tag{3.15a}$$

Following the same lines of reasoning than previously, at optimum the following equations resulting from the partial derivative of the Lagrangian with respect to c_t have to be respected :

$$\begin{aligned} \frac{\partial L}{\partial c_t} &= \frac{\partial \mathbb{E}[(d_t - c_t)^+]}{\partial c_t} + \lambda + \gamma_t = 0,\ \forall\ t \\ \frac{\partial L}{\partial \lambda} &= \sum_t c_t - C = 0 \end{aligned} \tag{3.16}$$

Additionally, equations (3.17) result from the inequality constraints 3.14.

$$\gamma_t\ (c_t - m_t) = 0,\ \forall\ t \tag{3.17}$$

The Karush-Kuhn-Tucker optimal conditions of the system are then :

$$F_{d_t}(c_t) = 1 - \lambda + \gamma_t = 0,\ \forall\ t \tag{3.18a}$$

$$\sum_t c_t = C \tag{3.18b}$$

$$\gamma_t(c_t - m_t) = 0,\ \forall\ t \tag{3.18c}$$

$$\gamma_t \geq 0,\ \forall\ t \tag{3.18d}$$

In order to find the optimum, we need to determine the periods which belong to T_0 and those that belong to T_1 such that :

$$\begin{cases} T_0 = \{t : c_t = m_t\} \\ T_1 = \{t : c_t < m_t\} \end{cases}$$

We note that optimal conditions for periods of T_1 are similar that ones of the problem defined by equations (3.1)-(3.2). Proposition 2 will allow to exploit results obtained by **P1** to build a solution algorithm for the problem defined by equations (3.12)-(3.14).

Theorem 2. *If in the solution of the non-constrained problem $\mathbf{c}'$ we have that $c'_t > m_t$ for period t, then in the constrained problem's optimal solution $t \in T_0$.*

Proof Let $\mathbf{c}^*$ be an optimal allocation in the constrained problem with $c^*_a > c^*_b$. If we increase the value of c^*_a and decrease the value of c^*_b by a small Δ to derive solution $\mathbf{c}$, we can write out the objective function as follows :

$$\begin{aligned} f(\mathbf{c}) &= \sum_t \mathbb{E}[(d_t - c_t)^+] \\ &= A + \mathbb{E}[(d_a - (c_a^* + \Delta))^+] + \mathbb{E}[(d_b - (c_b^* - \Delta))^+] \end{aligned}$$

where A does not depend on Δ.
Taking the derivative of f with respect to Δ we obtain :

$$\begin{aligned} \frac{\partial f(\mathbf{c})}{\partial \Delta} &= \frac{\mathbb{E}[\partial(d_a - (c_a^* + \Delta))^+]}{\partial \Delta} + \frac{\mathbb{E}[\partial(d_b - (c_b^* - \Delta))^+]}{\partial \Delta} \\ &= \mathbb{E}[-\mathbf{I}(d_a \geq c_a^* + \Delta] + \mathbb{E}[\mathbf{I}(d_b \geq c_b^* - \Delta)] \\ &= -P(d_a \geq c_a^* + \Delta) + P(d_b \geq c_b^* - \Delta) \\ &= -1 + F_{d_a}(c_a^* + \Delta) + 1 - F_{d_b}(c_b^* - \Delta) \\ &= F_{d_a}(c_a^* + \Delta) - F_{d_b}(c_b^* - \Delta) \end{aligned}$$

A positive derivative means that the perturbation of Δ is not improving the solution $\mathbf{c}^*$, since we have a minimization problem. The derivative is positive if $F(c_a^*+\Delta) > F(c_b^*-\Delta)$. We said that $c_a^* > c_b^*$ and $F(c_a^*) = F(c_b^*)$. This implies $F(c_a^* + \Delta) > F(c_b^* - \Delta)$. It is then undesirable to decrease the capacity of any period b in the opposite direction than that of the optimum. In the constrained problem, this means that when all periods t such that $c_t^{'} > m_t$ (for problem (3.1)-(3.2)) are on the boundary ($c_t = m_t$), the capacity in the remaining periods can only increase. $c_t^* = m \ \forall \{\ t : c_t^* > m_t\}$

□

Algorithm 1 finds the periods that belong to T_0 and then apply on the remaining periods T_1 the optimal allocation policy found in Section 3.3.1 defined earlier to obtain an optimal capacity allocation. We argue that the solution obtained using this algorithm are globally optimal.

Algorithm 1 Bounded allocation problem solution

Require: $\mathbf{c}^{'}$ the optimal allocation for the unconstrained problem
$\mathbf{c} \leftarrow \mathbf{c}^{'}$
while $\exists t$ such that $c_t > m_t$ **do**
    $c_t = m_t$
    $T_0 = T_0 \cup t$
    $T_1 = T \setminus T_0$
    $\mathbf{c}$ = Policy **P1** on restricted problem to T_1 with capacity $C - \sum_{t \in T_0} m_t$
end while
$\mathbf{c}^* \leftarrow \mathbf{c}$

We illustrate an example solution in Figure 3.2 where the dotted vertical line represents the upper bound m_t. Before using the upper bound constraints, the optimal solution had $c_3 > m$. The new solution sets $c_3 = m$ and then optimizes the rest of the points with the global capacity updated using policy $P1$. If any of these remaining point crosses the upper bound once optimized again, the process is repeated until a feasible solution is found. Following Proposition 2, this feasible solution is the optimal solution.

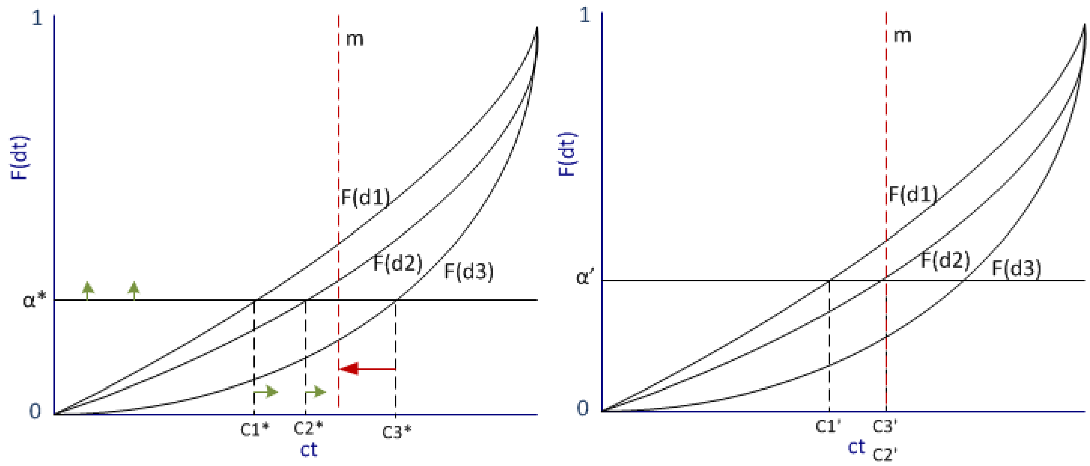


FIGURE 3.2 – One iteration of the algorithm to find the optimal bounded capacity distribution

3.3.5 Minimization of the largest expected capacity shortage

When working with the sum of expected values as we did in Section 3.3, the solutions obtained respond well on average to fluctuating demand. However minimizing the sum can result in unsatisfactory coverage levels in some periods. For this reason, we propose in this section a formulation with an alternative objective function. The objective function hedges against exceptional circumstances by minimizing the worst coverage over all the periods. Below is the formulation of a min max optimization problem.

$$\underset{c_t}{min}\ \underset{t}{max}\ \mathbb{E}[(d_t - c_t)^+] \tag{3.19}$$

$$\sum_t c_t \leq C \tag{3.20}$$

$$c_t \leq m_t,\ \forall\ t \tag{3.21}$$

We show next that with criterion (3.19) planning the optimal capacity allocation consists in having the expected value equal for all periods where the capacity is not nil. This is illustrated in Figure 3.3.

Theorem 3. *The optimal solution value E of the problem in equations* (3.19)-(3.21) *can be obtained with periods with equal expected capacity shortages or nil capacity :*

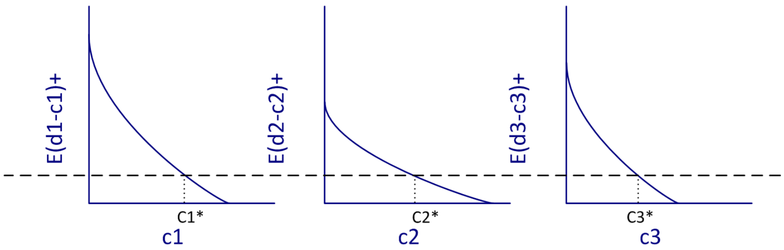


FIGURE 3.3 – Example of optimal solution for the min max allocation problem with three periods

$$\forall t \in T \begin{cases} c_t = 0, \ \ \textit{if } \mathbb{E}[d_t] \leq E \\ \mathbb{E}[(d_t - c_t)^+] = E, \qquad \textit{otherwise} \end{cases}.$$

Proof Let c^* be an optimal solution, where the maximal expected value is met on period a, *i.e.*, $E = \mathbb{E}[(d_a - c_a^*)^+]$ and there exists $b \in T$ such that $\mathbb{E}[(d_a - c_a^*)^+] > \mathbb{E}[(d_b - c_b^*)^+] > 0$ (if not, the property is satisfied). Then, there exists $\Delta > 0$ such that $\mathbb{E}[(d_b - (c_b^* - \Delta))^+] = E$ or $\mathbb{E}[d_b] \leq E$. This procedure can not increase the objective value and can be repeated until a solution that satisfies the properties stated in the proposition. □

Numerically, we can obtain c^* by approximating the functions $\mathbb{E}[(d_t - c_t)]$ for discrete values of c_t and using a bisection method until Theorem 3 is satisfied. An upper bound can be obtained from any feasible solution. As it is shown in Proposition 4 setting each capacity c_t to its limit m_t gives a lower bound.

Theorem 4. *The objective value of the min max problem is lower bounded by the greatest expected capacity shortage at the capacity limitation over all periods.*

$$LB = \max_t \{\mathbb{E}[(d_t - m_t)^+]\} \tag{3.22}$$

Proof Because of decreasing of all terms of the objective function related to each period, the optimal solution of Problem (3.19) ;(3.21) (where constraint (3.20) is removed) gives LB by saturation of constraints (3.21), and is a lower bound for Problem (3.19)-(3.21). □

3.4 Numerical Examples

The analytical results obtained in this work depend on the probability density function of the demand for each planning period. Consequently, a strong hypothesis of this work is that there is some mean of estimating the demand in hospitals during epidemic seasons. These estimations are usually based on historical data of hospitals and can be adjusted according to the expertise of the physicians who evaluate the criticality of the epidemic.

The purpose of the following experiments is to illustrate the analytical solutions with concrete examples. We use demand generated using an epidemic simulation model. Three different epidemic scenarios corresponding to different acuteness of the outbreak are considered. The epidemiological models used have three different contact rates (mild = 0.4, moderate = 0.6 and severe = 0.8) [Britton, 2010].

The hospital's demand is estimated from the epidemic simulation by multiplying the number of infected individuals by a coefficient. This coefficient represents the resources needed to service the patients visiting the ED of interest. The resources considered can be of different type, we can imagine dimensioning the beds or determining the number of additional shifts required to service influenza patients. The coefficient itself is a random variable.

The scenarios are illustrated in Figure 3.4. In the mild scenario, the overall average demand is of 8.4 ± 1.5 units in average. In the moderate and severe scenario we have a mean and average standard deviation of 9.9 ± 2.1 and 10.9 ± 2.5 respectively. The standard deviation is a bit larger during the periods with more demand. This is to illustrate the fact that there is more uncertainty during days with peak demand.

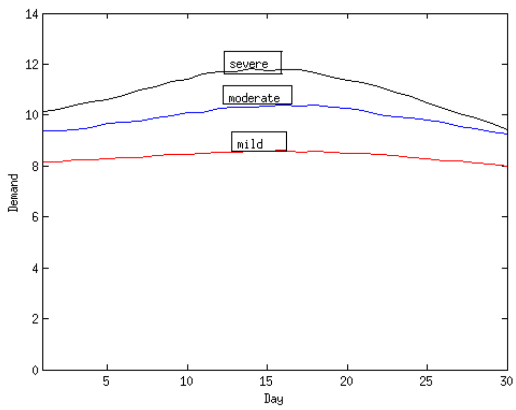


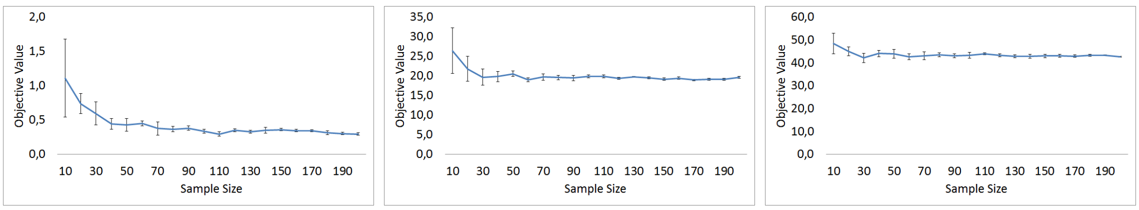
FIGURE 3.4 – Expected value of demand in three types of epidemic scenario

3.4.1 Epidemic Simulations

In the scenarios above, we do not assume a known distribution for our demand but instead use simulations to generate sample demand patterns. Then we can only approximate $F_t(d_t)$ and the number of epidemic realizations used for this approximation impact the quality of the solution. Table 3.1 shows the value of the objective function for different number of samples $|\xi|$ and a global capacity C of 250 resource units. For each number of sample tested, we ran the experiment 10 times and noted the mean capacity as well as the standard deviation of the results. The results start to stabilize for 100 samples. We use in the rest of the tests 5000 samples to generate empirical cumulative distribution functions.

$\|\xi\|$	0.4	0.6	0.8
10	1,1 ± 0,57	26,5 ± 5,82	48,5 ± 4,43
20	0,7 ± 0,15	21,8 ± 3,17	45,1 ± 1,91
30	0,6 ± 0,17	19,7 ± 2,04	42,2 ± 2,05
40	0,4 ± 0,08	19,9 ± 1,29	44,2 ± 1,46
50	0,4 ± 0,09	20,6 ± 0,65	43,9 ± 1,9
60	0,5 ± 0,04	19,1 ± 0,51	42,8 ± 1,27
70	0,4 ± 0,1	19,8 ± 0,81	43,1 ± 1,68
80	0,4 ± 0,04	19,7 ± 0,56	43,6 ± 0,8
90	0,4 ± 0,03	19,5 ± 0,64	43,2 ± 0,79
100	0,3 ± 0,03	19,9 ± 0,38	43,4 ± 1,25
110	0,3 ± 0,03	19,9 ± 0,32	44 ± 0,47
120	0,4 ± 0,02	19,4 ± 0,22	43,3 ± 0,65
130	0,3 ± 0,02	19,8 ± 0,14	42,9 ± 0,72
140	0,4 ± 0,04	19,5 ± 0,23	43 ± 0,81
150	0,4 ± 0,02	19,2 ± 0,3	43,1 ± 0,74
160	0,3 ± 0,02	19,5 ± 0,32	43,2 ± 0,72
170	0,3 ± 0,02	19 ± 0,16	43 ± 0,48
180	0,3 ± 0,03	19,2 ± 0,22	43,4 ± 0,41
190	0,3 ± 0,02	19,2 ± 0,25	43,3 ± 0,26
200	0,3 ± 0,02	19,7 ± 0,23	42,7 ± 0,22

TABLE 3.1 – Influence of the number of samples used on the stability of the results



(a) $\alpha = 0.4$ (b) $\alpha = 0.6$ (c) $\alpha = 0.8$

FIGURE 3.5 – Mean and standard deviation of 10 instances of the solution with C=250

3.4.2 Demand Distribution

As we said earlier, the quality of our solutions depend on the quality of our characterization of the demand. In this section, we show how the objective function in equation (3.1) evolves for different values of C using the epidemic scenarios and a normal approximation of the demand. Figure 3.6 illustrates the total shortage as a function of the global capacity C. This informs us on the influence of the capacity to ensure a certain quality of service. To recap, the objective function expresses the sum of the uncovered demand over all the periods. We see that using normal demand approximations instead of epidemic simulation we have results with mean relative error of 0.8 units.

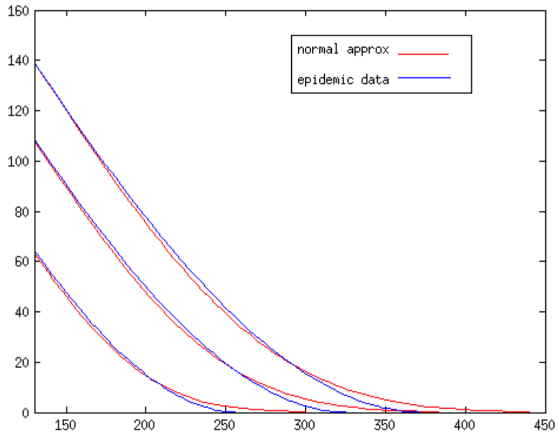


FIGURE 3.6 – Expected value as a function of C with epidemic data and normal approximations.

3.4.3 Capacity Upper Bounds Influence

As was explained in Section 3.3.4, the optimal allocation policy with upper bounds on the capacity to distribute per period can be calculated easily from policy **P1**. We show in Figure 3.7 an example distribution of capacity with a limit of 7 units of resources for each period ($m_t = 7, \forall t \in T$). When the global capacity is 190 units, the resulting distribution of capacities follows policy **P1** as no period requires a demand greater than 7 units. However, if the global capacity is 200 units, periods 9 to 21 get assigned 7 units of capacity and the remaining periods are optimised again using policy **P1**.

3.4.4 Min Max Criterion

To test the effectiveness of the alternative min max criterion, we evaluate solutions obtained with it using the first objective and vice versa. In other words, we analyze the efficiency of the solutions under different conditions than they were optimized for. In Figure 3.8, the expected value of the sum of the shortages is plotted against the global

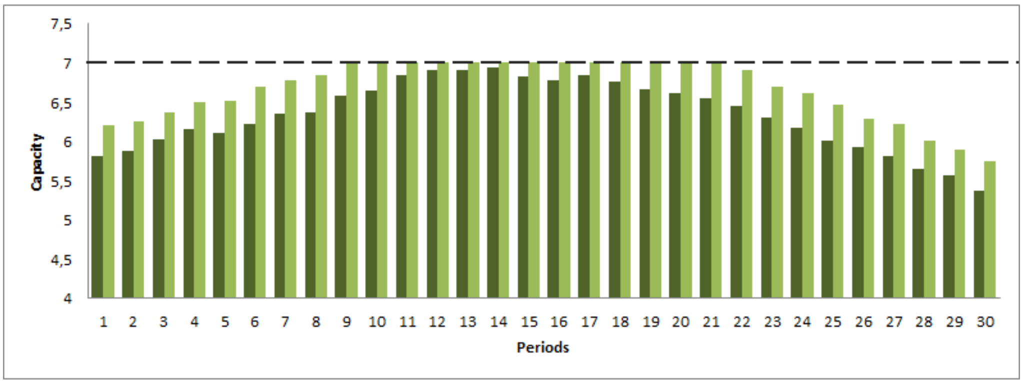


FIGURE 3.7 – Capacity distribution with bound for $C = 190$ and $C = 200$.

capacity for the solutions obtained with policy **P1**. Graphically, the two curves almost overlap with a slight dominance from the actual **P1** solutions. This indicates that the solutions with **P2** – the policy for the min-max criterion – can lead to good solutions evaluated with the sum. We note however that we allowed for the min max policy to distribute the entire capacity across the periods even if it wasn't necessary to minimize the min max. In other words, we replaced equation 3.20 with an equality equation.

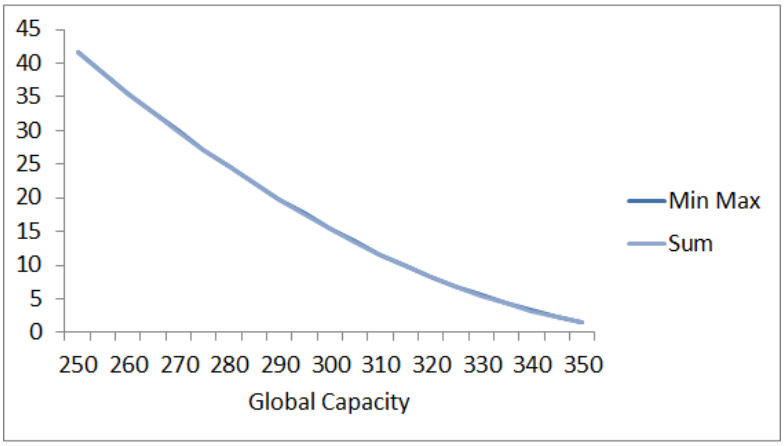


FIGURE 3.8 – Minimizing the expected sum over the planning horizon using solutions from **P1** and **P2**.

Similarly, in Figure 3.9, we plot the value of the maximum shortage against the global capacity C for solutions obtained with policy **P1** and policy **P2**. In the same way, we notice that solutions optimized with policy **P1** are efficient under the min max objective. As the objective of policy **P2** is to minimize the maximum value and not the sum, the y-axis on Figure 3.9 is much smaller. In reality the results in Figure 3.8 and 3.9 are very similar.

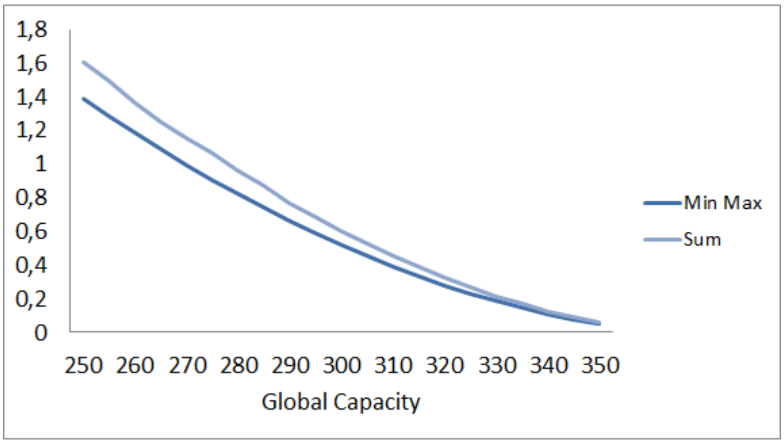


FIGURE 3.9 – Minimizing the maximum shortage using solutions from **P1** and **P2**.

3.5 Conclusion

Decision makers need a tool to evaluate shortage risks in their unit. One way to evaluate shortage risks is to compare the total capacity of the unit to the expected demand. However, during epidemic season, the demand is not stationary and a simple comparison between the total capacity and total expected demand is not sufficient. For this reason, we presented in this chapter a strategic capacity planning formulation to find good allocation policies across the epidemic horizon and then properly compare the unit's capacity with the expected demand.

We first analyzed the simple case were the objective is to minimize the sum of expected shortage across the horizon. An optimal solution for this problem was characterized analytically and an algorithm for applying the solution on real world data was discussed. Then an upper bound constraint by period was added. These constraints reflect the fact that resources (like man-hours) are necessarily limited on a given period (day). A policy was then defined for this new problem and its optimality proven. Finally, a min max formulation was studied to propose an alternative objective function to avoid extreme cases.

Once the policies are defined, a proper evaluation of the shortage risks for a given capacity level can be made. Our tests show that for the purpose of right-sizing a unit, both policies can give us good estimates on the shortage risks given that demand is accurately forecast. We believe that the analytical solutions proposed in this chapter can serve as a quick way to evaluate certain strategic risks and benefits of right-sizing a facility.

Chapitre 4

A Stochastic Optimization Model for Shift Scheduling in Emergency Departments [1]

Résumé en français du chapitre : Optimisation des postes de travail journalier dans un service d'urgence

Suite à une analyse des bases de données du service d'urgences de Lille, nous remarquons un décalage entre le nombre de patients présents aux urgences et l'effectif médical aux différentes heures de la journée. Particulièrement, le pic d'activité observé le soir n'est pas pris en compte dans l'élaboration des emplois de temps du personnel médical. Nous partons de cette observation pour étudier le problème de gestion des postes de travail des médecins et des infirmiers dans les services d'urgences. Une revue de la littérature permet de dégager plusieurs difficultés dans le problème de gestion des postes qui sont spécifiques au contexte des urgences. Dans les SU, l'interaction entre les différentes ressources humaines est impérative pour bien prendre en charge les flux des patients. Ainsi, les emplois du temps des médecins et des infirmiers doivent s'accorder pour que l'effectif médical soit suffisant à tout moment de la journée. De plus, les arrivées des patients aux urgences sont aléatoires mais la part d'aléa évolue tout au long de la journée. Il est donc nécessaire d'utiliser d'autres lois d'arrivées pour le soir que pour le matin par exemple. À cause de ces difficultés, les études consacrées à la gestion de l'effectif médical en cours de journée ont pour la plupart divisé le problème en deux. D'une part, on détermine les niveaux de ressources nécessaires pour chaque heure de la journée et ensuite des postes de travail sont distribués aux médecins et infirmiers afin de coller au mieux au niveau de ressources pré-calculées. Les méthodes numériques souvent utilisées peinent cependant à intégrer l'évolution des arrivées des patients sur les heures de la journée. Nous présentons dans ce chapitre un modèle d'optimisation stochastique qui intègre les deux volets du problème et permet de gérer l'évolution des arrivées de façon rigoureuse.

Pour développer le modèle d'optimisation, nous avons tout d'abord utilisé le formalisme

1. Article publié dans Health Care Managment Science

des files d'attente afin de modéliser le parcours de soins des patients aux urgences et définir les étapes clés. Puis, nous avons traduit la logique de ce modèle de files d'attente sous forme de programme linéaire stochastique en nombres entiers. Le parcours des patients dans le système est représenté par des contraintes linéaires. Un des avantages de cette modélisation est qu'elle permet d'optimiser le niveau de ressources et la distribution des postes simultanément pour les médecins et les infirmiers. Une fois les solutions obtenues, nous évaluons les plannings grâce à un modèle de simulation à événements discrets. Ce modèle colle plus finement à la réalité que le modèle d'optimisation. Par exemple, contrairement au modèle d'optimisation, le modèle de simulation catégorise les patients selon leur degré d'urgence médicale. Un plan d'expérience nous permet ainsi de tester les plannings obtenus et de conclure sur une amélioration nette du temps d'attente des patients avec des plannings optimisés.

Sommaire

4.1 Introduction **45**
4.2 Workforce Scheduling in Emergency Departments **46**
4.3 Emergency Department Queuing Network **48**
4.4 Mathematical Formulation **50**
4.5 Simulation Model **55**
4.6 Numerical Experiments **56**
4.6.1 Parameters Settings and Real-World Data 56
4.6.2 Analysis of optimized schedules 59
4.6.3 Validation of the Robustness of the Schedules 63
4.6.4 Week to Weekend Transition 67
4.7 Conclusion **69**

4.1 Introduction

Emergency departments are the crux of hospitals as they serve as a center for medical treatment and as a continuously operating portal for inpatient admissions [Schuur and Venkatesh, 2012]. It is important when considering the ED, to provide patients with timely health care while maintaining a healthy work environment for the personnel. Further ED are facing the challenge of providing a good quality of service while being cost-effective.

Increasing demand for health services in recent years and a reduction of the number of resources in the health industry due to budget cuts have led to chronic overcrowding in the ED [Jones et al., 2009]. From the patients' point of view, an overcrowded emergency department leads to excessive waiting times. While from a health specialist point of view, it can be the cause of an unfriendly work environment [Rogers et al., 2004]. One of the major consequences of overcrowding with regards to the Operations Research community is a growing interest in developing strategies that make the most efficient use of the available resources [Ernst et al., 2004]. In France, a recent report of the first national ED conference which gathered health practitioners and government officials decreed the urgency of the problem of long waiting times [Ricard Hibon and Petit, 2012]. The report describes several strategies to reduce the potential bottlenecks that cause overcrowding which can also be found in the literature [Sven et al., 2011, Akkerman and Knip, 2004, Bruin et al., 2009]. Among the recommendations, the authors mention the need to improve patient flow, the need to organize inpatients bed resources and the need to make a better use of available resources.

As ED work around the clock, a turnover of health care professionals is necessary in order to ensure an appropriate service level at all time. In the context of emergency departments, the work demand across the day is subject to the rate of arrivals of patients. In turn, the arrival rate of patients is subject to stochastic variations. These variations make the problem of organizing the workforce difficult.

A look at the number of patients and the staff level in Figure 4.1 observed at our collaborating hospital clearly shows mismatch of patients flow and personnel schedules. At every time period, the activity is estimated with the average number of patients present in the ED. Figure 4.1 clearly shows a peak of activity in the evenings. The distribution of staffs on duty does not appropriately match the peaks of activities observed in the ED. These observations prompted us to design a personnel scheduling model to meet the work demand accurately while satisfying the work constraints imposed.

In this paper, we present a human resources planning model for ED in order to explore the various human resources planning options and to take into account their robustness to face uncertainties and time-dependent patient arrival rate. The main efficiency criterion considered is the total patient waiting time. While other criteria are possible, total waiting time is a good indicator of patient service quality and ED crowding.

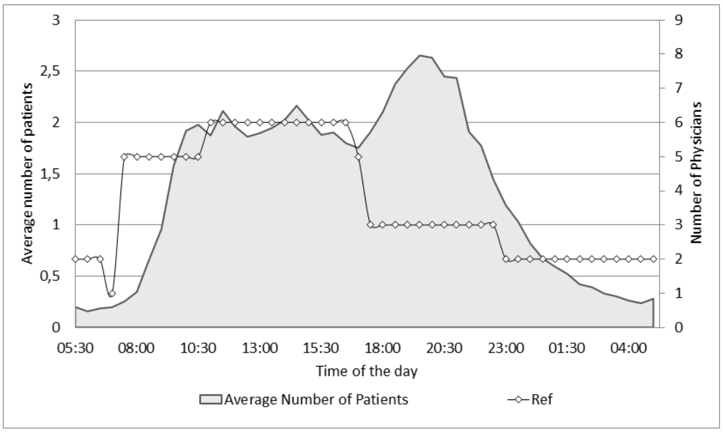


FIGURE 4.1 – Preliminary analysis of the data from the pediatrics ED of Lille Hospital

Personnel schedule is determined by a high-level stochastic programming model in which patient waiting times are approximated by a discrete time model and legal constraints for personnel scheduling are taken into account. The stochastic programming model is solved by a deterministic sample average approximation (SAA) [Kleywegt et al., 2002] and mixed-integer programming. The resulting personnel schedules are evaluated with a more realistic ED discrete-event simulation model.

4.2 Workforce Scheduling in Emergency Departments

With the growing work constraints and the expansion of work facilities, the employee scheduling problem has become, for most industries, manually intractable. As early as 1954, mathematical models have been used in order to computationally assist industries to schedule their employees to work [Edie, 1954]. Employee scheduling is often decomposed into several sub-problems [Ernst et al., 2004]. In a dynamic environment where the workload is variable, the first task is to be able to forecast the work demand for any scheduling horizon [McCarthy et al., 2008, Jones and Thomas, 2008, Jones et al., 2009]. Once a work demand estimate is known, the decision maker has to determine an optimal staffing level for each scheduling period. After that, the staffing level is used as a basis for creating schedules that meet the staffing level appropriately. Finally, an assignment phase matches employees to the schedules created.

One of the peculiarities in scheduling studies conducted for health care systems is the focus on multi-skill scheduling. The work demand in health care systems needs to be addressed by appropriately qualified personnel and cannot be distributed arbitrarily between the resources [Warner, 1976, Jaumard et al., 1998]. This multi-skill characteristic adds yet another layer of complexity to the workforce scheduling in ED that is often overlooked. We identify two main types of human resources in this study (physicians and nurses) and schedule them simultaneously. Simultaneous scheduling of different resources means having the work of one type of resource impact the schedules of the other.

With regards to the staffing problem, the work demand in the ED cannot be determined easily and depends on the arrivals of patients in the services and on the interactions between different resources. Literature on staff scheduling in the ED often considers stochastic demand patterns in order to address the uncertainty in the arrival of patients [Izady and Worthington, 2012, Sinreich et al., 2012]. Sinreich and Jabali [Sinreich and Jabali, 2007] propose a simulation model coupled with an linear programming model to downsize emergency departments while keeping the patients' length of stay (LOS) constant. The simulation model dictates the required staffing profile while the linear programming model chooses the schedules that matches this staffing profile best. Zeltyn et al. [Zeltyn et al., 2011] propose a unique simulation model to address the human resources allocation problem on multiple decision levels (operation, tactical and strategic). On the tactical level, they use two strategies : the Rough Cut Capacity Planing (RCCP) technique and the Offered Load approach. The later being a refinement of the former that allows to split the workload. The staffing level is computed hourly through the analysis of a simple queuing system. Queuing theory has also been extensively used for determining staffing levels because it allows the evaluation of quality of service criteria such as the waiting time of patients [Green and Soares, 2006, Jennings, 1996]. However, analytic formulas for determining optimal staffing levels are often limited with regards to the complexity of the system they can analyze. As such, no analytical formula exists for optimally determining staffing level in systems with time-dependent stochastic arrival rates but instead numerical approximations have been used to alleviate this problem [Vile, 2013]. For example Ingolfsson et al. [Ingolfsson et al., 2002] use an integrated queuing and optimization model to investigate the effect of time-varying arrival rates in the context of workforce scheduling. The authors present a heuristic to construct potential schedule then use queuing theory to evaluate the resulting service levels. Izady and Worthington [Izady and Worthington, 2012] use queuing theory and simulation in an iterative staffing scheme to determine schedules for an ED modeled as an infinite server network with different tasks and processes. After an appropriate staffing level is determined, an ILP is employed to find shifts that minimize the over-staffing and under-staffing. Cochran and Roche [Cochran and Roche, 2009] suggest a queuing network approach for capacity planning of beds and staff in an ED. Their methodology is based on a multiple-customer class queuing analysis to increase ED access. Seasonal variation is of patient arrival patterns is considered.

In this work, the ED is represented as a multi-server multi-stage time-dependent priority queue system with re-entrance, Poisson distributed arrivals and exponential service times. Physicians and nurses are the servers whose cyclic schedules are to be determined. The planning horizon is discretized into time periods and a different mean arrival rate is given for every period of the day. We choose the expected total patient waiting time as the main criteria to optimize as it is often associated with patients' and staff well-being and is a good indicator of ED crowding. In order to address the complex nature of the problems, many studies addressed each sub-problem in the employee workforce scheduling independently by overlooking the inherent interaction between them [Vile, 2013]. As such, the optimal staffing level is often considered separately from the scheduling problem. This approach neglects complex waiting queues generated by uncertainties and changing patient arrival rate, which is crucial for ED crowding. In this work, we consider the staffing level as part of the scheduling problem and solve them both simultaneously. Integrating the two sub-problems allows us to schedule employees while reproducing the work dynamics of the system and simulating the patient flow.

The structure of the work is divided into two parts. First, we build an optimization model that finds staff schedules with operational constraints. Then, we evaluate more accurately our solution in a simulation model to measure the value of some of the assumptions in the optimization model. The optimization problem is solved by a Sample Average Approximation (SAA) solution methodology by considering different possible scenarios or realization of patient flows in the ED and optimizing the workforce scheduling problem considering all the different realizations. The remaining of the paper is organized as follows. We first describe the ED model in details in Section 4.3, then explain the optimization and the simulation models along with the assumptions that we consider in each in Sections 4.4 and 4.5 respectively. Then, we present numerical experiments in Section 4.6 and conclude in Section 4.7.

4.3 Emergency Department Queuing Network

In order to dynamically match the staff schedules and patient flow, we include in our system the dynamics of the patient flow through the ED. We present a queuing system with the key treatment steps and resources of an ED. Patients can follow different trajectories through the ED depending on their pathology and their severity level. A simplified scheme for their trajectory is as follows. Patients are first assessed by a nurse to determine whether they need vital medical care, a step often called triage. If they need vital care, they are taken in priority to a resuscitation room. Otherwise, the patients wait in a waiting room until a physician is available to examine them. After the first assessment, patients are either directly treated by a nurse or are sent for auxiliary exams to confirm the diagnostic. The auxiliary exams range from a simple blood test to time-consuming imaging tests. After the auxiliary exams are performed, the patients are examined again by a physician and finally treated by a nurse before being discharged or admitted into another department in the hospital [Izady and Worthington, 2012]. The second assessment leads to re-entrance of patients to the same resources. Priority is given to patients that are being assessed for the second time by physicians over patients arriving for the first time.

Each process in the ED can be modeled as a queue where jobs to be served are the patients and the hospital staff are the servers that handle these jobs. Figure 4.2 illustrates a basic queuing system. On the left side, we have patient arrivals into the system which follow an inhomogeneous Poisson distribution of parameter λ_t. Once in the system, they wait in a queue before being served by one of the servers of the system. The service time is assumed to be exponentially distributed in this paper. This corresponds to an $M_t/M/c$ queue model where M stands for the Markovian assumptions on the arrival and the service rates and c is the number of servers available.

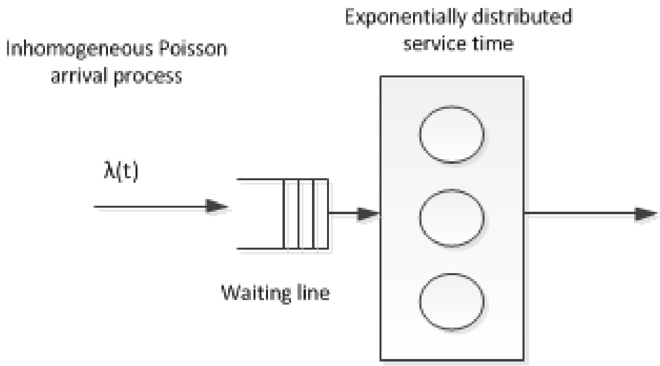


FIGURE 4.2 – A basic queue model that represents a process in an ED

From the above, each ED has three main activities : assessment, treatment and auxiliary exams. The two health care providers that we model in the system are physicians and nurses. Physicians perform health assessments and examine the patients before and after having auxiliary exams. While nurses handle the final treatment administered to patients. As such, physicians are the servers for the two assessment queues of our system and nurses are the servers for the last queue. The overall system is illustrated in Figure 4.3. We do not schedule the lab technicians as they are often resources not entirely dedicated to the ED but instead delay patients that need to have auxiliary exams done.

Apart from the patients care pathway modeling, the queuing model also accounts for the stochastic nature of the work in the ED. The arrival rate of patients greatly varies during a day with peak arrival during the evenings and low arrival during the night. Service time needed to treat each patient can-not be known in advance. The stochastic service duration used in this paper models the different needs of patients. The number of patients arriving for each period of our planning horizon is parametrized by mean arrival rates that can be estimated from past data. Service distributions are estimated based on mean service times observed and found in the literature.

In Section 4.4, we present the mathematical formulation that we derived from the system specifications explained in this section. Some simplifying assumptions are made and detailed to solve the problem in reasonable time. Nonetheless, the solutions are evaluated using a more realistic simulation model whose specifications are explained in Section 4.5.

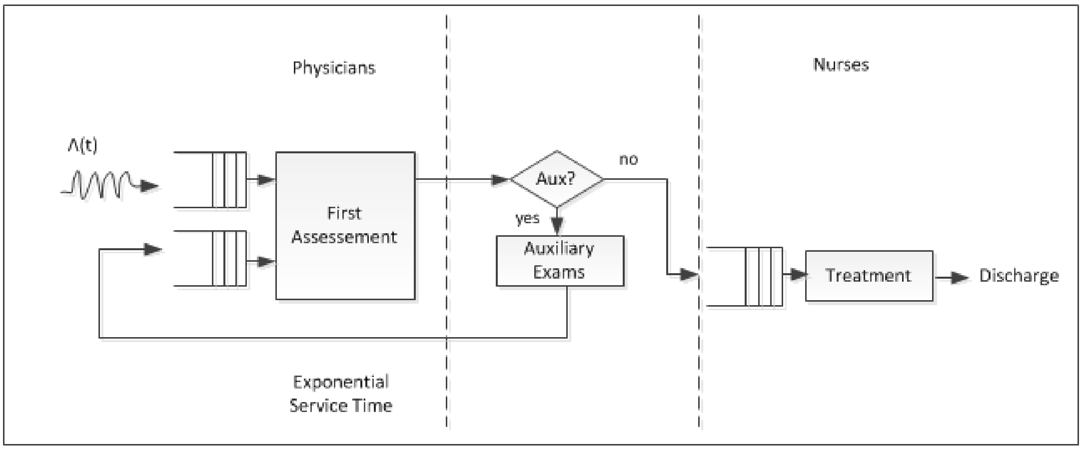


FIGURE 4.3 – Process chart of patients in the ED

4.4 Mathematical Formulation

In this section, we present a mathematical formulation that is based on the details discussed in section 4.3. The objective is to optimize a waiting time performance indicator that we approximate with discrete time variables. As patients' waiting time is heavily influenced by the number of personnel present in the ED at any given time, the model finds optimal schedules over the planning horizon. For the time being, we are not interested in assigning the schedules to actual health personnel but only in finding an optimized schedule that meets the objective of minimizing the average total waiting time. In order to keep the model realistic, we allow for a limited number of different hours in which the staff can start working.

The number of patients arriving at each period t for each queue q is denoted $A_{t,q}(\omega)$. This symbol is overloaded in our model as it both designates the Poisson distributed arrival to the first two queues and the variable designating the arrival of patients to the nurses' queue. After the patients enter a queue, they are counted as waiting in the variable $W_{t,q}(\omega)$. Patients that are served in a queue, are counted in the variables $O_{t,q}(\omega)$. Then, parameter $L_{i,n,t}(\omega)$ determines the service capacity of the staff at each period. More specifically, the parameter determines the number of patients that n resources of type i can serve during period t. The ω index designates one scenario of patients' arrival and visit to the ED. The decision variables of the model $x_{i,n,t}$ define the number of health care providers working during period t.

To model the operational constraints, we use several parameters. A lower bound LBD and an upper bound UBD specify the possible lengths of the shifts that can be assigned to the personnel. The number of different hours of the day in which personnel of type i can start working is denoted by d_i. An upper bound TT_i is set on the number of total working hours that can be scheduled for resources of type i. The delay at the auxiliary exams is designated by δ. Finally some variables, $p_{i,t}$, $s_{i,t}$, $e_{i,t}$ and $z_{i,t}$ are used to facilitate the expression of shift constraints.

Some modeling assumptions are made in order to be able to solve the problem efficiently. A discrete-time approximation of the waiting time is calculated as $W_{t,q}$ minus $O_{t,q}$. What follows is that patients arriving during one period can only be treated in the next. Also, the waiting time of patients during the period they arrive is not counted. The patients are divided into two classes, those that do not require auxiliary exams and those that do. Each patients' class has a predefined trajectory from the start. This allows us to avoid having to route the patients to their appropriate queue after they have been assessed. This assumption is justifiable inasmuch as we do not care about each patients' trajectory individually in the model. Instead, we keep track of the total number of patients in each queue of the system at each period t. We also assume the system is initially empty. We make this assumption by starting the scheduling horizon at an hour when it is reasonable to say that the system is clear. As of the data available to us, this is most true around 5 :30 in the morning. Furthermore, we do not distinguish between the different health care providers with the same skills and consider all physicians to be homogeneous and all nurses to be homogeneous. For the sake of simplicity, we omitted the possible visit of patients to the vital care unit. As of the data available from the UHC of Lille for the year 2011 and 2012, patients that require vital prognosis only constitute 1.79% of the total patients and those in need of resuscitation merely 0.45%. Finally, auxiliary exams are always assumed to last for a fixed duration. Of course, these assumptions are relaxed in the simulation model and evaluated for consistency in the numerical experiments.

Objective :

$$min \underset{\omega \in \Omega}{\tilde{\mathbb{E}}} [\sum_t \sum_q (W_{t,q}(\omega) - O_{t,q}(\omega))] \tag{4.1}$$

Subject to :

$$W_{1,q}(\omega) = A_{1,q}(\omega) \qquad \forall q \in \{1,2\} \tag{4.2}$$

$$W_{1,q}(\omega) = 0 \;\; \forall q \notin \{1,2\} \tag{4.3}$$

$$O_{1,q}(\omega) = 0 \;\; \forall q \tag{4.4}$$

$$W_{t,1}(\omega) = W_{t-1,1}(\omega) + A_{t,1}(\omega) + O_{max(t-\delta,1),2} - O_{t-1,1}(\omega) \quad \forall t \;\; \in [2..T] \tag{4.5}$$

$$W_{t,2}(\omega) = W_{t-1,2}(\omega) + A_{t,2}(\omega) - O_{t-1,2}(\omega) \quad \forall t \;\; \in [2..T] \tag{4.6}$$

Parameter	Meaning	Domain
t	The different periods of the day	$[1..T]$
n	Number of resources of each type	$[1..N_i]$
i	The different type of resources available	{1,2}
q	The queue number	[1..Q]
d_i	Maximum number of different starts for each resource	$[1..S_i]$
TT_i	Number of working hours available to schedule for resource of type i	
LBD	Lower bound on the shift length allowable	[1..47]
UBD	Upper bound on the shift length allowable	[2..48]
δ	The delay at the auxiliary exams	{1,2}
$A_{t,q}(\omega)$	Poisson distributed arrivals in period t for queue q	$\mathbb{N}+$
$L_{i,n,t}(\omega)$	Sum of exponentially distributed consultation times for n resources of type i at period t	$\mathbb{N}+$

TABLE 4.1 – The parameters of the scheduling model.

Variable	Meaning	Domain
$x_{i,n,t}$	$=1$ if there are exactly n resources of type i working at period t	{0,1}
$p_{i,t}$	Number of resources of type i present at period t	$[1..N_i]$
$s_{i,t}$	Number of resources of type i starting at period t	$[1..N_i]$
$e_{i,t}$	Number of resources of type i ending at period t	$[1..N_i]$
$z_{i,t}$	Indicator variable that is equal to 1 if at least one resource of type i starts working at t	{0,1}
$W_{t,q}(\omega)$	Number of patients that are waiting in queue q a the start of period t	$\mathbb{R}+$
$O_{t,q}(\omega)$	Number of patients in queue q that have been served during period t	$\mathbb{R}+$

TABLE 4.2 – The variables of the scheduling model.

$$W_{t,3}(\omega) = W_{t-1,3}(\omega) + O_{t-1,1}(\omega) - O_{t-1,3}(\omega) \quad \forall t \in [2..T] \tag{4.7}$$

$$O_{t,q}(\omega) \leq W_{t,q}(\omega) \quad \forall t \ \forall q \tag{4.8}$$

$$\sum_{q \in \{1,2\}} O_{t,q}(\omega) \leq \sum_{n=1}^{N_1} L_{1,n,t}(\omega) * x_{1,n,t}(\omega) \quad \forall t \tag{4.9}$$

$$O_{t,3}(\omega) \leq \sum_{n=1}^{N_2} L_{2,n,t}(\omega) * x_{2,n,t}(\omega) \quad \forall t \tag{4.10}$$

$$\sum_{t} p_{i,t} \leq TT_i \quad \forall i \tag{4.11}$$

$$\sum_{t} (s_{i,t} - e_{i,t}) = 0 \ \forall i \tag{4.12}$$

$$p_{i,t} = \sum_{t'=2}^{t} (s_{i,t'} - e_{i,t'}) + p_{i,1} \quad \forall t \in [1..T] \ \forall i \tag{4.13}$$

$$p_{i,1} = p_{i,T} + s_{i,1} - e_{i,1} \quad \forall i \tag{4.14}$$

$$\sum_{t'=t+1}^{mod(t+LBD,T+1)} e_{i,t'} \leq p_{i,t} - s_{i,t} \quad \forall t \in [1..T] \ \forall i \tag{4.15}$$

$$p_{i,t} \leq \sum_{t'=t+1}^{mod(t+UBD+1,T+1)} e_{i,t'} \quad \forall t \in [1..T] \ \forall i \tag{4.16}$$

$$\sum_{n=1}^{N_i} n * x_{i,n,t} = p_{i,t} \quad \forall i \ \forall t \tag{4.17}$$

$$\sum_{n=1}^{N_i} x_{i,n,t} = 1 \quad \forall i \ \forall t \tag{4.18}$$

$$z_{i,t} * N_i >= s_{i,t} \quad \forall i \ \forall t \tag{4.19}$$

$$z_{i,t} \leq d_i \quad \forall i \ \forall t \tag{4.20}$$

$$p_{i,t} \geq 1 \quad \forall i \ \forall t \tag{4.21}$$

$$p_{i,t} \geq s_{i,t} + 1 \quad \forall i \ \forall t \tag{4.22}$$

The objective function of the model (4.1) is a discrete time approximation of the total patients waiting time accumulated over all the scenarios. By summing over all the sample scenarios $\omega \in \Omega$ considered, we seek to obtain a good approximation of the estimated value of the actual waiting time of patients in the system. Constraints (4.2), (4.3) and (4.4) initialize the number of patients in the queues in each scenario. This assumption of empty patient queue at the beginning of a day corresponds to our field observation of several ED in which often no patient remains late in the night such as at 5.30am. This is true whether the ED is busier during the day or not. The next set of equations of the problem are balance equations to track the movement of patients through the ED. As such, (4.5), (4.6) and (4.7) update the number of patients in each queue at every time period.

Constraints (4.8) specify a logical bound for the number of patients served during a period. This number can-not be larger than the number of patients already in the queue. Then, constraints (4.11) limit the total number of staffing hours available for each type of resource. Constraints (4.12) through (4.14) are logical constraints that define the range of values the variables p, s and e can take. Then, constraints (4.15) and (4.16) define the type of shifts that are allowed by specifying a lower bound and an upper bound on the shift lengths.

Constraints (4.17) and (4.18) assign a value to the variables x which is used to limit the number of patients served at every time period in constraints (4.9) and (4.10). Finally constraints (4.19) through (4.22) are operational constraints to make sure that there are at most d_i different moments of the day where resources of type i start working and that there is at least one resource of each type in every time period. We need to have an upper bound on the number of periods resources can start working in because physicians and nurses need to transmit patients' information to their colleagues before leaving the service. As such, constraints (4.22) make sure that the shifts of the different resources overlap.

4.5 Simulation Model

In order to evaluate the validity of the assumptions and simplifications made in the mathematical formulation, we test the solutions obtained using a simulation model. The simulation model includes several features not considered in the optimization model for a more accurate representation of the ED. The simulation improves over the mathematical formulation on several fronts : system specification, patients specification and resources activities.

As in the optimization model, exponential service time and Poisson arrival rates are used in the simulation model. However, the simulation uses a discrete event paradigm instead of the discrete time model paradigm in the optimization. This is important as one of the key assumptions in the optimization problem is that a discrete time model can serve to calculate patients' waiting time accurately. Second, the vital care pathway is no longer ignored. Patients requiring vital prognosis are attended by a physician and a nurse immediately. Auxiliary exams are divided into two categories : biological exams and diagnostic imaging exams. The biological exams make for 55% of the examinations.

Patients on the other hand are no longer homogeneous but are classified according to their acuity level (from 1 to 5 in increasing order of severity). The classification scheme used is that of the french hospitals known as CCMU (*Classification Clinique des Malades des Urgences*). This allows for a sickest-first prioritization in the different queues of the ED. Additionally, the patients that are classified as CCMU 4 or CCMU 5 skip the regular route and are treated in priority in the vital care pathway. These patients are served by both a nurse and a physician for a random duration.

Finally, with regards to resources, physicians are no longer solely occupied by patients. Instead, they also have to respond to administrative tasks such as searching patients' historical data or finding downstream beds for inpatients. Beds are not explicitly modeled in the system. Instead, patients waiting in queues are considered to be assigned either a room or a hallway bed. Administrative tasks occur at random intervals but follow the same distribution laws as patient arrivals. Having the same distribution makes sense if we think that administrative tasks are more likely to occur when patients enter the system. Nurses no longer solely handle patients' treatments but also take care of the biological exams.

This improved ED model is used to evaluate the solutions obtained with the optimization model. Important system statistics such as the estimated patients' waiting time and the time spent in the system can be extracted and a confidence interval for these data can be calculated.

4.6 Numerical Experiments

Numerical experiments have been conducted to show the potential reduction of patients' waiting time in an ED by adjusting the workforces' shift schedules. The parameters settings for the optimization model, the simulation model, and the real-world benchmark are described in Section 4.6.1. In Section 4.6.2, we check that the optimization model matches enough key features of the queuing system to provide suitable solutions. In order to do so, the schedules obtained by running the optimization model are evaluated and compared to simulation tests. The performances of different scheduling strategies are also compared. The strategies differ in the constraints on the employees' shifts. In Section 4.6.3 the robustness of schedules facing significant demand increase is evaluated and in Section 4.6.4 the transition from week schedules to week-end schedules is investigated. The optimization model is solved using a commercial solver (IBM Ilog Cplex 12.5). All tests were performed on an Intel(R) Xeon(R) 4 cores CPU E5520 @ 2.27GHz with 8 MB of cache memory and 8 GB of RAM. The evaluation is done using AnyLogic 6.9.

4.6.1 Parameters Settings and Real-World Data

The parameters used in the experimental section are based on data for the years 2011-2012 of patients' entries in the pediatrics' ED of the University Hospital Center (UHC) of Lille France. We extracted from these data an average number of patients' arrivals per half-hour of the day and used them as parameters for the Poisson distributions of each period. The length of the period also called time step $\Delta = 30$ minutes is chosen primarily to coincide with the shifts change marks used in the hospital. Also, the patients' arrival averages are calculated according to length of the time step. Average values would be less significant if they relied on less data. As the data available only contain arrival and departure times, we did not determine the service duration of patients in the same manner as the arrivals. Instead, we relied on average service durations found in the literature and on experts' opinion to calibrate these parameters. However, the patients' lengths of stays obtained with these service durations are of the same order of magnitude than those observed in the real data. Based on these parameters, we generate several possible scenarios for patients arrivals and service time.

In the UHC of Lille, the number of resources working during the week differs drastically from the number of resources working on the weekends. Therefore, two set of optimization tests are run. The first set of tests uses the total staff hours available during the week and the arrival distributions extracted for weekdays and the second set of tests relies on data for weekends. The total staff hours available serves as an upper bound on the number of hours scheduled. Each scenario in the optimization model is generated by obtaining a realization ω of the distributions of the arrival and service times. Once the different parameters of the models are set, the scenarios are the only input to the model.

After the schedules are optimized, we evaluate these schedules and the waiting time they produce in the simulation model. As the schedules we obtain are cyclical, the evaluation is done for several consecutive days. We introduce a warm-up period in order to reach stable results. Evaluating the performance of the schedules can be done on a large number of replications which allows the calculation a confidence interval. The parameters used for the experimental runs are summarized in Table 4.3.

To validate the pertinence of a scenario-based strategy, we ran three different instances of optimization tests. After obtaining the resulting schedules from each test, we evaluated the evolution of the expected waiting time in the simulation model using the resulting schedules. We see the results in Figure 4.4, where the average waiting time over all three tests are displayed. The average waiting time of the patients decreases with the increase of the number of scenarios used in the optimization. We notice that the expected waiting time stabilizes for about $|\Omega| = 100$ scenarios.

In terms of computational feasibility, our model strongly depends on the number of scenarios used for optimization and on the time step chosen. This is because the number of variables increases linearly with the number of scenarios and the number of periods. Fortunately, we showed that the simulated average waiting time converges with solutions obtained from 100 scenarios. From our tests, all solutions with 100 scenarios were obtained within 4 hours with a 0.11% relative gap in the worst case.

Other variables in the model also influence the running time but to a lesser extent. As such, the number of patients in the queues does not increase the number of variables but only the range the variables can take. On the other hand, the number of variables representing resources : $x_{i,n,t}$ increases as the number of resources increases but is independent of the number of scenarios. For example, in a large ED with about 200 patients daily and 10 to 15 resources of each type, the optimal solution is still obtained within the 4 hours of computation time.

Throughout the tests, the schedules obtained are compared against the reference schedule ref that is currently put in place in the ED of Lille. The first alternative we test are schedules obtained with shifts whose length is fixed to 8 hours for both physicians and nurses. The schedules obtained with this strategy give us an idea of how much we can improve by changing the starting times along with more restrictive shift constraints than what is currently practiced. Then, we keep the shifts lengths range from 4 to 12 hours as it is the case in the UHC of Lille but modify the starting times. The third and final alternative is to create shifts with no constraint on their lengths. In this last case, the

Parameter	Definition	Value
L	Length of scheduling horizon	24 hours
Δ	Length of period in the scheduling horizon	30 mins
Λ	Mean number of patients entering the system in one day	65(week), 66(we)
$\frac{1}{\mu}$	Mean service time per period	30 mins
σ	Number of scenarios used in the optimization	100
τ	The time limit set for the optimization model	4 hours

TABLE 4.3 – Parameters of the system and of the optimization algorithm

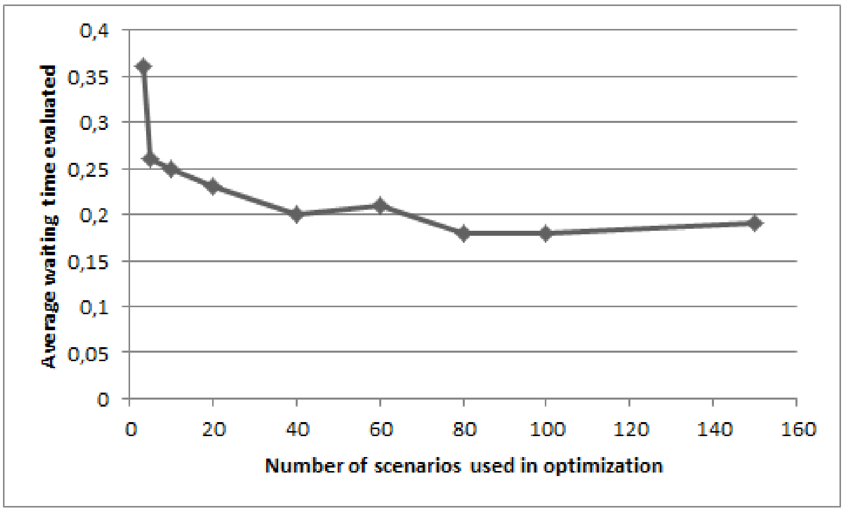


FIGURE 4.4 – The evolution of waiting time with the number of scenarios used in the optimization model

only constraint is the total number of hours worked. This last strategy is not realistic but merely allows us to compute a lower bound on the total waiting time of the patients in the optimization model. Constraints on every schedule strategies are given in Table 4.4. The parameters d_i and TT_i refer to the parameters presented in the optimization model.

Shifts	$LBD(h)$	$UBD(h)$	d_1	d_2	TT_1	TT_2
ref	4	12	7 (week), 5(weekend)	6	178 (week), 130 (weekend)	124
(4-12h)	4	12	7 (week), 5(weekend)	6	178 (week), 130 (weekend)	124
($8h$)	8	8	7 (week), 5(weekend)	6	178 (week), 130 (weekend)	124
$(1 - 24h)$	1	24	7 (week), 5(weekend)	6	178 (week), 130 (weekend)	124

TABLE 4.4 – Parameters for the scheduling strategies

4.6.2 Analysis of optimized schedules

We verify in a first part that the optimized schedules obtained can actually reduce the average waiting time of patients compared to the reference schedule when tested in simulation conditions. In order to do so, we allow for different assumptions on the shifts as explained previously. In all the alternatives tested, we allow the shifts to start on any period of the day but only permit a limited number of shift changes to occur as it is currently the case. Our analysis in this subsection focuses on weekends but results obtained for weekdays are similar and lead to the same conclusions.

Shifts	**W-O (h)**	$W_1 - O_1$ **(h)**	$W_2 - O_2$ **(h)**	$W_3 - O_3$ **(h)**	**Gap %**	**Time(mins)**
ref (1)	0.41	1730	2581	1131	0.00	0
(2)	0.39	1767	2175	1200	0.00	0
(3)	0.47	1919	2939	1289	0.00	0
8hrs (1)	0.27	1449	1415	683	0.02	240
(2)	0.26	1416	1311	738	0.01	240
(3)	0.31	1645	1690	735	0.11	240
4-12hrs (1)	0.21	1283	1101	431	0.00	20
(2)	0.20	1143	1133	415	0.00	34
(3)	0.24	1334	1356	522	0.00	42
1-24hrs (1)	0.21	1236	1109	469	0.00	180
(2)	0.20	1112	1130	400	0.00	53
(3)	0.24	1340	1342	523	0.00	51

TABLE 4.5 – Optimization results for the weekend scenarios

Table 4.5 summarizes the results obtained for schedules optimized to weekend conditions. The first column specifies the shifts type used. The number next to the shift type specifies different runs of the optimization problem, each with a different set of scenarios. $W - O$, is total objective value converted to average hour units per patient. This value gives an approximation on the expected waiting time as computed in the optimization model. The total waiting time associated with each queue, q in the optimization model is given in the columns $W_q - O_q$, for $q \in \{1, 2, 3\}$. Next, the Gap indicates the percentage distance between best lower and upper bounds obtained. An time limit of 240 minutes is used for each run of the optimization problem.

The simulation model of Section 4.5, is used to evaluate each solution schedule. The results of the evaluation are presented in Table 4.6. The simulation is run over 100 replications each of 10 successive days. Since we only have a unique reference schedule, only one row keeps track of the expected waiting time using the reference schedule. Table 4.6 first notes the patients' expected waiting time in $Wait$. The next columns indicated the expected and max waiting time in the physicians' queue (Q_{phys}) and nurses' final treatment queue (Q_{nurs}). The results consistently reduce the expected waiting time of patients in more flexible strategies. This is further confirmed by the similar behavior of the maximum expected waiting time over all replications, in columns $Qmax$. However, within this trend of reducing the average waiting time, the results are not very stable. Different set of scenarios in the optimization model result in different waiting time in the simulation model. For example with shifts of fixed length the waiting time varies between 79.94 minutes for the first set of scenarios and 87.28 for the second set. This can be explained by the approximations that were made and the fact that we only used a limited number of scenarios (σ= 100) to optimize.

Having schedules with very flexible shifts (in the range 1-24 hours) does not really improve the waiting time of patients compared to schedules with shifts in the range of 4-12 hours. Consequently, we note that modifying the starting times of the shifts is the factor that helps us most improve the expected waiting times, even with a constant total working hours capacity. As explained in Section 4.3, queues 1 and 2 are served by physicians and queue 3 is served by nurses. In the simulation model, the first two queues are grouped into a physician queue Q_{phys}.

Shifts	**Wait (mins)**	$\mathbf{Q}_{phys}$ **(mins)**	$\mathbf{Q}_{phys}$ **max (mins)**	$\mathbf{Q}_{nurs}$ **(mins)**	$\mathbf{Q}_{nurs}$ **max (mins)**
ref	125,23 ± 6,22	99,24 ± 5,6	977,07 ± 43,24	25,99 ± 1,5	422,26 ± 23,08
8hrs (1)	79,94 ± 4,85	54,66 ± 4,21	861,81 ± 46,89	25,28 ± 1,75	465,12 ± 38,88
(2)	87,28 ± 5,8	59,23 ± 5,13	910,35 ± 49,42	28,05 ± 1,87	499,15 ± 35,75
(3)	83,82 ± 7,61	58,73 ± 6,75	864,21 ± 52,76	25,09 ± 1,94	412,3 ± 39,43
4-12hrs (1)	62,43 ± 4,3	48,91 ± 4,08	832,22 ± 43,19	13,52 ± 1,01	315,84 ± 32,67
(2)	63,13 ± 4,27	48,54 ± 3,94	836,35 ± 47,46	14,59 ± 0,99	374,4 ± 34,51
(3)	64,46 ± 3,7	50,72 ± 3,38	842,91 ± 34,87	13,74 ± 0,85	330,66 ± 28,85
1-24hrs (1)	64,97 ± 4,58	51,06 ± 4,35	871,16 ± 45,72	13,91 ± 0,9	324 ± 30,8
(2)	64,55 ± 4,63	49,92 ± 4,16	855,5 ± 51,54	14,63 ± 1,02	369,36 ± 31,86
(3)	64,27 ± 4,16	48,18 ± 3,69	831,45 ± 42,44	16,09 ± 1,09	381,73 ± 28,08

TABLE 4.6 – Simulation results for the weekend schedules

Figures 4.5, 4.6 and 4.7 present example staffing levels obtained by optimizing the schedules following the different shift strategies. The reference schedule clearly misses the evening peak of arrivals while the other strategies closely match the arrival curve. We note that the solution with shifts varying from 4 to 12 hours and those varying from 1 to 24 hours are very close and match the ascending and descending slopes of the arrivals flow. These observations corroborate the results obtained in Tables 4.5 and 4.6.

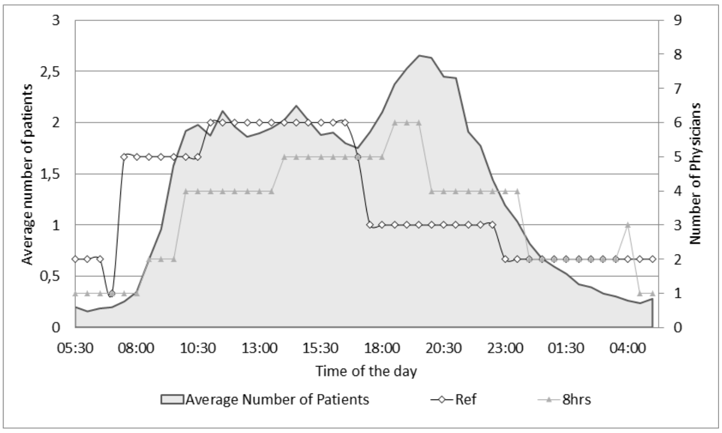


FIGURE 4.5 – Resulting staffing levels with shifts fixed to 8 hours

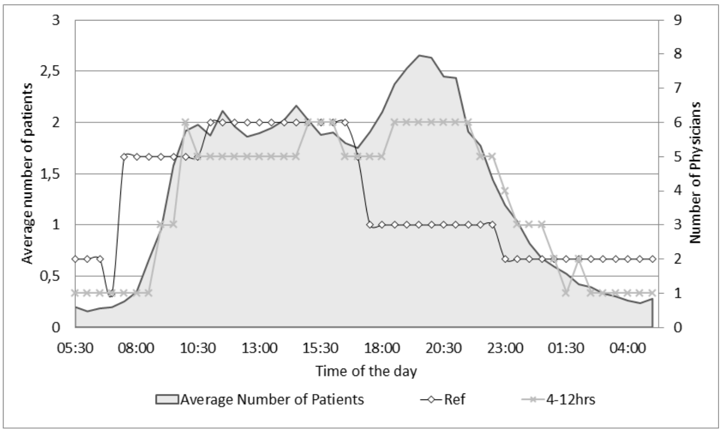


FIGURE 4.6 – Resulting staffing levels for shifts varying from 4-12 hours

4.6.3 Validation of the Robustness of the Schedules

In this set of simulation runs, we increment the number of patients arriving to the system and evaluate the schedules obtained through optimization again. In order to do so, we increment the average Poisson parameter for each period. This increase in the number of arrivals results in about 15% more patients per period. Since the schedules have not been optimized to perform under these conditions, the evaluation allows us to have an idea on the robustness of the schedules during epidemic periods.

The results of this set of tests are summarized in Tables 4.7 and 4.8. We notice from the results that the expected waiting time increases proportionally to the results of Section 4.6.2 across the different solution instances. Consequently, the optimized schedules

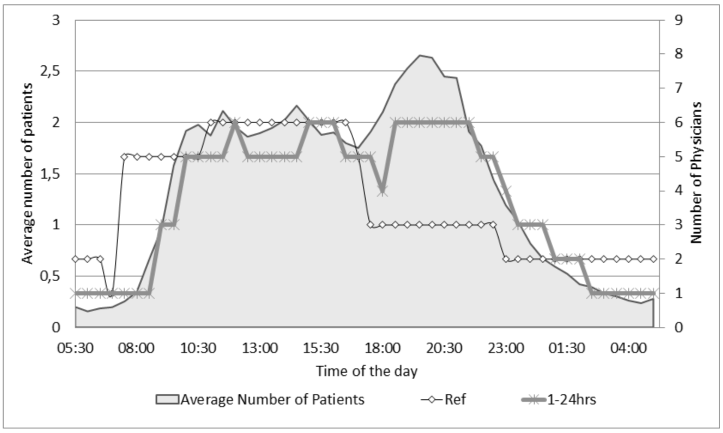


FIGURE 4.7 – Resulting staffing levels for shifts varying from 1-12 hours

still outperform the reference schedule in unfavorable conditions. This coherence in the waiting times is still true for each queue individually but we see a disproportionate increase in the average maximum waiting times for schedules constructed with (1-24h) constraints. This increase is understandable as the (1-24h) constraints specifically exploit the discrete-time approximation of the optimization model. Schedules with (1-24h) constraints can be expected to perform poorly under different conditions than those they were optimized for.

Shifts	**Wait**	$\mathbf{Q}_{phys}$	$\mathbf{Q}_{phys}$ **max**	$\mathbf{Q}_{nurs}$	$\mathbf{Q}_{nurs}$ **max**
ref	242,99 ± 13	199,2 ± 12,06	1323,22 ± 34,88	43,79 ± 2,61	559,87 ± 23,74
8hrs (1)	191,1 ± 13,32	148,81 ± 13,11	1268,17 ± 45,18	42,29 ± 3,38	651,89 ± 45,34
(2)	191,32 ± 14,34	147,3 ± 13,53	1263,06 ± 44,4	44,02 ± 2,76	659,27 ± 45,17
(3)	190,79 ± 13,6	149,63 ± 13	1264,16 ± 45,57	41,16 ± 2,75	633,62 ± 44,73
4-12hrs(1)	148,74 ± 10,86	125,47 ± 10,41	1218,3 ± 46,26	23,27 ± 1,85	457,79 ± 43,62
(2)	155,95 ± 12,04	133,56 ± 11,65	1258,86 ± 43,98	22,38 ± 1,43	494,64 ± 38,05
(3)	150,31 ± 10,85	128,77 ± 10,55	1242,86 ± 42,17	21,54 ± 1,59	477,88 ± 42,1
1-24hrs(1)	155,76 ± 13,59	132,37 ± 13,22	1211,44 ± 49,83	23,39 ± 1,69	482,04 ± 41,28
(2)	146,73 ± 11,67	122,78 ± 11,02	1212,04 ± 49,64	23,94 ± 1,56	501,33 ± 30,19
(3)	142,07 ± 11,26	116,78 ± 10,99	1180,36 ± 48,45	25,29 ± 1,65	525,05 ± 34,35

TABLE 4.7 – Simulation results for weekend schedules during an epidemic period

TABLE 4.8 – Simulation results for week schedules during an epidemic period

Shifts	**Wait**	$\mathbf{Q}_{phys}$	$\mathbf{Q}_{phys}$ **max**	$\mathbf{Q}_{nurs}$	$\mathbf{Q}_{nurs}$ **max**
ref	142,07 ± 11,26	116,78 ± 10,99	1180,36 ± 48,45	25,29 ± 1,65	525,05 ± 34,35
8hrs (1)	53,14 ± 3,17	22,53 ± 1,57	564,15 ± 42,14	30,61 ± 2,24	490,83 ± 37,81
(2)	51,62 ± 2,43	24,45 ± 1,45	580,12 ± 35,17	27,17 ± 1,65	502,22 ± 40,29
(3)	54,54 ± 3,44	21,44 ± 1,64	528,09 ± 42,14	33,1 ± 2,56	601,4 ± 36,75
4-12hrs(1)	37,5 ± 2,62	12,8 ± 0,92	498,01 ± 30,82	24,7 ± 2,27	541,19 ± 37,59
(2)	35,11 ± 2,19	11,41 ± 0,75	417,48 ± 30,38	23,7 ± 1,82	524,46 ± 35,12
(3)	40,63 ± 2,4	13,58 ± 0,86	510,86 ± 28,89	27,05 ± 2	598,06 ± 36,17
1-24hrs(1)	35,22 ± 1,95	11,94 ± 0,85	454,38 ± 30,59	23,27 ± 1,61	526,38 ± 33,41
(2)	35,39 ± 2,13	11,22 ± 0,73	414,83 ± 31,08	24,18 ± 1,77	549,97 ± 37,27
(3)	40,91 ± 2,22	13,32 ± 0,8	499,48 ± 26,06	27,59 ± 1,88	602,06 ± 32,71

4.6.4 Week to Weekend Transition

One limitation of the previous results is that the cyclical schedules are evaluated independently for weekdays and weekends. In order to have simulation conditions that are closer to reality, the transition from a weekday schedule to a weekend schedule needs to be evaluated. Consequently, weekday and weekend schedules are concatenated and tested under both the average conditions for arrivals and the epidemic conditions.

The results in Tables 4.9 and 4.10 show that the transition from the week schedules to the weekend schedules can be done without real complications. The average waiting times observed when the concatenation of schedules is performed follow the results obtained previously. Indeed, the optimize schedules outperform the reference schedules under normal and epidemic conditions.

Shifts	**Wait**	$\mathbf{Q}_{phys}$	$\mathbf{Q}_{phys}$ **max**	$\mathbf{Q}_{nurs}$	$\mathbf{Q}_{nurs}$ **max**
ref	88,59 ± 4,7	62,59 ± 3,63	871,71 ± 45,88	25,99 ± 1,92	430,6 ± 30,93
4-12hrs (1)	38,52 ± 2,96	24 ± 2,4	652,18 ± 45,73	14,53 ± 1,07	367,98 ± 32,31
(2)	37,53 ± 2,84	22,7 ± 2,36	635,05 ± 46,81	14,83 ± 1,21	379,98 ± 34,39
(3)	38,48 ± 2,52	24,53 ± 2,27	654,31 ± 38,59	13,96 ± 0,97	384,82 ± 35,06

TABLE 4.9 – Simulation results for schedules spanning an entire week

Shifts	**Wait**	$\mathbf{Q}_{phys}$	$\mathbf{Q}_{phys}$ **max**	$\mathbf{Q}_{nurs}$	$\mathbf{Q}_{nurs}$ **max**
ref	140,58 ± 6,63	102,25 ± 5,09	1087,38 ± 42,46	38,33 ± 2,52	561,25 ± 43,66
4-12hrs(1)	69,31 ± 4,59	46,13 ± 3,74	941,59 ± 56,66	23,18 ± 1,9	504,82 ± 36,2
(2)	70,46 ± 4,61	43,86 ± 3,72	941,56 ± 49,76	26,6 ± 2,05	573,99 ± 38,44
(3)	71,09 ± 4,06	46,73 ± 3,32	974,81 ± 51,18	24,36 ± 2,12	508,22 ± 43,41

TABLE 4.10 – Simulation results for schedules spanning an entire week during an epidemic period ϵ

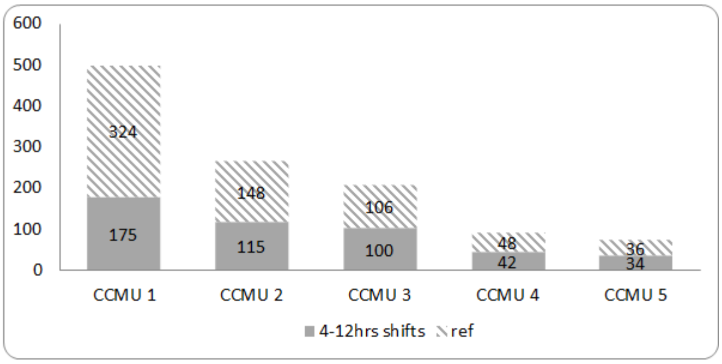


FIGURE 4.8 – Expected sojourn time for the different acuity classes in normal conditions

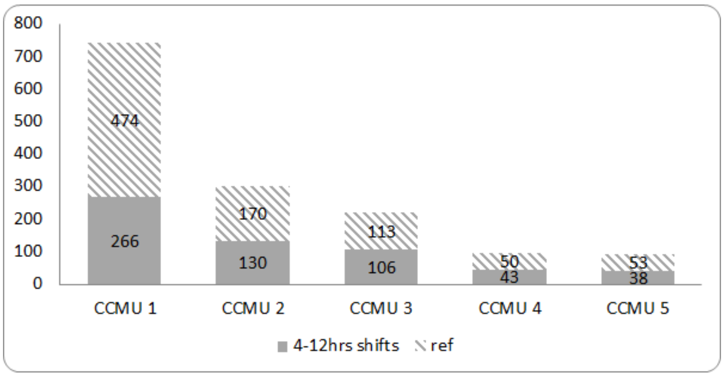


FIGURE 4.9 – Expected sojourn time for the different acuity classes in epidemic conditions

The results of the final set of tests are summarized in Figures 4.8 and 4.9. As we have mentioned, the schedules with shifts varying from 1 to 24 hours do not give a significant advantage over the schedules with shifts varying from 4 to 12 hours. Finally, Figures 4.8 and 4.9 illustrate the differences in sojourn time according to the CCMU of patients. Coherently, patients with higher priority spend less time in the system than those with lower priority.

4.7 Conclusion

We presented in this study a stochastic optimization formulation for an ED shift scheduling problem. Schedules of interest are computed once for some representative period such as winter or summer season or weekdays and weekends. During this period, the schedules are repeated daily. The formulation follows the flow of patients through the ED and takes into account the stochastic nature of arrivals and service times. The constraints used to follow the evolution of patients through the ED mirror those in stock management literature. The optimization model is solved using a deterministic Sample Average Approximation (SAA) method. This model assumes many simplifications in order to be tractable in a commercial linear programming solver. The solution schedules obtained by the optimization model are evaluated using a discrete event simulation model under various test conditions to verify the validity of the assumptions and the robustness of the solutions.

The simulation model used to evaluate the results has several improvements over the optimization model. Improvements such as patient prioritization and multiple tasks for resources (administrative tasks for physicians and biological exams for nurses) make the simulation model more realistic. The evaluation demonstrates that the assumptions made hold for optimizing the average waiting time. Nonetheless, the simulation model remains a simplification of the ED and hence lacks some elements that can be taken into account in future research. For example the service time of patients is approximated with an exponential distribution. Even if this approximation is very common in the literature, service times are usually not exponentially distributed. Hence, a possible extension to our work would be to test detailed patient service times to be collected from our partner hospital. Furthermore, excessive waiting time is only one aspect among many that define overcrowding in ED. Average waiting time is important and influences other criteria such as the maximum waiting time but other service quality criteria such as limited bed capacity contribute to overcrowding and should be considered in future research.

Real data from the UHC of Lille France are used to evaluate the quality and the performance of the schedules obtained. The proposed optimization approach allows to find promising cyclic schedules. The framework proposed is easily extensible and allows the integration of several strategical and operational constraints that are not possible using other methods.

Chapitre 5

Proactive On-call Scheduling during a Seasonal Epidemic [1]

Résumé en français du chapitre : Planification des postes d'astreinte aux urgences sur une période épidémique

Une fois la période hivernale commencée et les premières épidémies démarrées, les urgences ont très peu de recours disponibles au jour le jour pour faire face à une augmentation imprévue de la demande à part l'affectation d'heures supplémentaires au personnel. En particulier, les postes d'astreinte ne sont pas actuellement utilisés aux urgences. Cependant, les astreintes sont couramment mis en œuvre dans d'autres services de l'hôpital et rien dans les textes de lois n'interdit de les mettre en œuvre aux urgences. Une des raisons à cette exception aux urgences est l'appréhension du personnel médical qui assimile les astreintes à une augmentation des heures de travail effectives. Nous analysons dans ce chapitre les avantages et les inconvénients liés à ce type de postes dans le contexte particulier des épidémies hivernales et vérifions si cette appréhension est fondée. D'un point de vue fonctionnel, les postes d'astreinte requièrent deux prises de décisions. Dans un premier temps, le chef du service affecte des postes d'astreintes au personnel médical sur un horizon plus ou moins long. Puis, au jour le jour, le chef de service décide si les personnes en astreinte doivent être appelées ou non. Si une personne en astreinte est appelée, elle est rémunérée à un tarif d'heures supplémentaires. Cependant, si la personne en astreinte reste chez elle, elle est rémunérée à un tarif inférieur au tarif régulier et les heures passées en astreinte sont considérées comme des heures de repos. Pour refléter au mieux ces deux prises de décisions (le planning du personnel et l'appel ou non du personnel en astreinte), nous modélisons le problème sous forme de programme linéaire stochastique à deux étapes. La première étape modélise la prise de décision a priori en début de la période épidémique et la deuxième étape modélise les prises de décision au jour le jour. Cette modélisation nous permet de considérer plusieurs scénarios possibles de déroulement de l'épidémie et de garantir ainsi que la prise de décision initiale soit robuste face à des variations de la demande. Un plan d'expérience nous permet de conclure

1. Article publié dans Operations Research for Health Care.

par la suite que l'utilisation des heures supplémentaires permet de garantir une bonne couverture de la demande en diminuant à la fois le coût total et les heures effectives effectuées.

Sommaire

5.1 Introduction **73**
5.2 Literature Review **74**
5.3 Workload Patterns **75**
5.4 Staff Allocation Model **76**
5.4.1 On-Call Duties 78
5.4.2 Mathematical Formulation 78
5.5 Numerical Experiments **82**
5.5.1 Design of Experiments 83
5.5.2 Demand Generation 84
5.5.3 Cyclic Schedules 84
5.5.4 Non-cyclic Schedules 88
5.5.5 Relaxed Constraints on Workload Regulations 89
5.6 Conclusion **89**

5.1 Introduction

Demand surges are characteristic of Emergency Departments (EDs) during epidemic crises. The "white plan" is an emergency response plan defined in France to meet a sudden increase in activity of a hospital. This plan sets down corrective measures to be undertaken in crises situations. For example, health-care professionals on duty are retained, off-duty personnel are called in gradually and crisis management units are set up. However, the "white plan" is costly for hospitals and implies the application of several corrective measures simultaneously.

Because the "white plan" is difficult to put in place, a recurrent phenomenon is observed during seasonal epidemic periods in France. EDs are faced with surges in demand that are difficult to handle but that are not important enough to deploy the "white plan". These peaks are alleviated with empiric ad hoc strategies. For example, when faced with an unexpected increase of patients, the medical staff often stays overtime. These strategies often go unnoticed because no protocol formalizes them and the increase in demand is handled more or less. However, the adjustments can be costly for the hospitals and tiring for the personnel. Furthermore, resilience in the ED is limited and some situations can cause irremediable constraints. An increase in patients' waiting times or a decrease in the quality of service are possible consequences in such situations.

Hospitals spend a large part of their budget on human resources. It is thus imperative to have a strategy to manage the personnel adequately and even more so during epidemic crises. Some hospitals hire surplus personnel as a recourse to cover the rise in demand during peak seasons. Surplus personnel often need training, an adaptation time and are hired for a long period. For these reasons they are not a very flexible alternative. Additionally, EDs often schedule their permanent workforce relying on past experience. When no formal method is used, there is no guarantee that the resources are going to be deployed appropriately. Deploying resources at the wrong moment is problematic on several levels. When several resources are deployed too soon, the management of the epidemic peak will be difficult in later periods as no more surplus resources will be available. Similarly, if resources are called for too late, the early periods of the epidemic crises become difficult.

In this paper we study scheduling flexibility of the current work force as a way to better meet the seasonal epidemic demand. We analyze the impacts on the costs incurred by the hospital and their ability to better cover the demand. Staff scheduling is done robustly to account for uncertainties in the evolution of the epidemic. More specifically, we try an alternative to incrementing the workforce by proposing and evaluating an on-call staff allocation policy. The context of the epidemic season is first analyzed then a staff management strategy to efficiently cope with epidemics is proposed. Efficiency is evaluated with regards to patients' service level and labor cost. A two-stage stochastic programming model is developed to set the problem formally. In the first stage, only partial and incomplete data is available and decisions are made according to estimations of the demand. The second stage involves decisions that are made on a day-to-day basis.

5.2 Literature Review

Epidemic crises are often studied at the regional level where whole populations and many hospitals interact. Epidemic control and containment include issues ranging from awareness campaigns to the optimal distribution of medicines and vaccines [Flahault et al., 2006,Arora et al., 2010,Parvin et al., 2012,Yarmand et al., 2014]. For instance, in a study carried out by Arora et al. [Arora et al., 2010] a cost benefit based approach is used for developing a mutual-aid resource allocation strategy between different regions. Several hospitals share resources found in a central "warehouse", reminiscence of a warehouse-retailer supply chain. Similarly, in a study by Yarmand et al. [Yarmand et al., 2014], a central stockpile of vaccines is distributed among different regions using a two-stage stochastic programming approach. However, regional policies can not easily be used in individual hospitals. Particularly, it is difficult to adapt these strategies to a single ED department where the issue is not the distribution of resources across space but across time.

When considering ED departments independently, the problem of seasonal epidemics is different. Indeed, the main interest of EDs during seasonal epidemics is to effectively handle the increasing ED workload [Bienstock and Zenteno, 2012,Rico, 2009]. Few studies have addressed the problem of epidemics from a hospital perspective. It is important to note that during epidemic seasons, ED overcrowding is not strictly due to an increase in the volume of patients. Indeed other factors as the visit duration and acuity of the patients significantly increase ED workload [Sinclair, 2007,Stang et al., 2010].

Overcrowding affects many people, from the health-care staff to the patients and the people accompanying them. Resource management strategies to deal with overcrowding involve both human and material resources. The management of these resources relates to decisions on the number, the location and the time in which they should be deployed. As with all service industries, it is difficult for health care industries to make decisions during epidemics because of the uncertainty associated with the workload [Sasser, 1976].

In health-care industries, the continuous working environment and the different staff skills make the problem of resources allocation difficult. Consequently, there is an extensive number of articles written on the subject of resources scheduling and rostering [Ernst et al., 2004,Burke et al., 2004]. From as early as the 1970s, papers [Warner, 1976,Miller et al., 1976] dealt with the problem of nurse scheduling using mathematical programming. In the following years, more complex models accounted for constraints involving different shift types, skill mix, and fairness concerns [Jaumard et al., 1998, Beaulieu et al., 2000]. As the models complexity increased, the solution methodologies adapted. As such, articles using meta-heuristics and simulation techniques emerged [Gutjahr and Rauner, 2007,Bester et al., 2007]. For example, in [Gutjahr and Rauner, 2007], the authors propose an ant colony optimization approach for nurse scheduling so as to include constraints such as nurses' and hospitals' preferences and nurses qualifications. Finally, studies that have addressed the issue of on-call duties in EDs refer for the most part to specialists' on-call duties [Rudkin et al., 2004,Menchine and Baraff, 2008].

This study analyzes the situation of seasonal epidemics inside EDs. In particular, it addresses different managerial and operational issues involved during seasonal epidemics. The demand is estimated using a stochastic epidemiological model. We assume that an indicator for the start of the epidemic exists. As such, a threshold of workload inside the

ED can be defined beyond which, the system is considered to be overcrowded. After the problem context and assumptions are well defined, a robust ED staff allocation strategy is proposed. The model is a two-stage stochastic linear program solved using a Sample Average Approximation (SAA) approach. The model serves as a basis for the analysis of costs and impacts of different solution strategies. In particular we evaluate the impact of on-call duties on the coverage of epidemic demand. Finally we examine a real case study and compare our solutions with a staff scheduling policy in an ED in Lille France.

5.3 Workload Patterns

The workload inside an ED is fluctuating throughout the year. Usually, during epidemic seasons, hospital personnel experience an increase in the workload. It is generally admitted that this phenomenon occurs annually with varying amplitude. However, if we look at the weekly number of arrivals in Figure 5.1, we see that it is not obviously the case. Indeed, the number of arrivals seems more or less constant throughout the year except during the summer for about two months when there is a decrease in the number of arrivals.

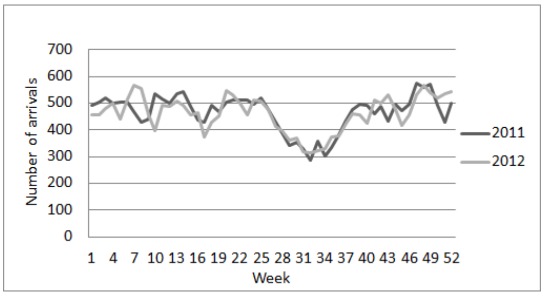


FIGURE 5.1 – Number of patients arriving per week for the years 2011 and 2012

This phenomenon led us to look into other factors that could influence the workload such as the sojourn time of patients and the destination of the patients after going through the ED. In Figure 5.2, we clearly see that the number of patients hospitalized after their ED visit increases during the last weeks of the year. This is also the case for the sojourn time which increases sharply during the same period as we see in Figure 5.3. Consequently, these data corroborate the need for additional capacity during epidemic season.

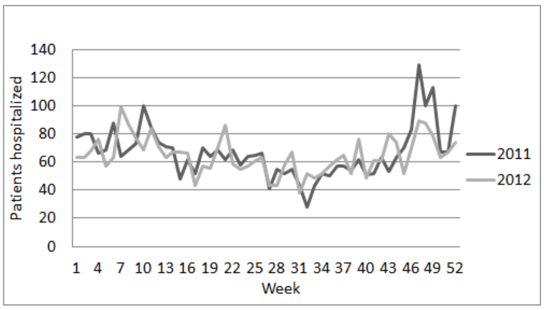


FIGURE 5.2 – Number of patients per week hospitalized after the ED

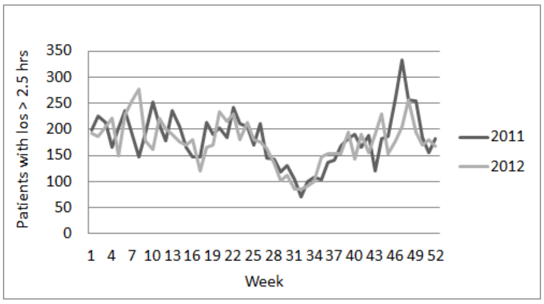


FIGURE 5.3 – Number of patients per week with sojourn time greater than 2.5 hours

Once the epidemic horizon is defined, we need to identify trends within it. According to the French law [2], the work done during the night has different rules than the work done during the day. For work to be considered as night work, it has to be performed during a period including the time from 9 PM to 6 AM. If we look at the number of arrivals to the ED during this night period Figure 5.4, we notice two things. The night workload is less fluctuating and lower than the day workload. Consequently, when defining possible epidemic scenarios, the night period needs to be distinguished from the day period.

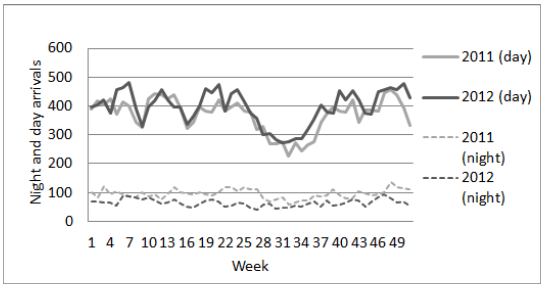


FIGURE 5.4 – Difference of arrivals between night and day

The number and type of patients in an ED does not directly map to a number of human resources. However, an increase in the workload calls for a parallel increase in resources if a good service quality is to be ensured. It is thus possible to estimate human resources demand given a good estimation of the epidemic development. In Section 5.4 a staff allocation model is presented to adapt the capacity according to the seasonal trends observed. Some assumptions are made in the allocation model and justified accordingly.

5.4 Staff Allocation Model

The allocation model we present in this section is proactive in the sense that decision makers take a decision prior to the realization of the epidemic. In reality, this can be achieved by efficient early epidemic declaration mechanisms. The first point of interest for an ED is to reduce overcrowding and retain a reasonable service level. Patients waiting excessively in EDs before being attended can experience severe health complications. Similarly, physicians can experience burnouts as a result of excessively stressful work patterns. From a hospital point of view, it is important to minimize the cost associated

2. http ://vosdroits.service-public.fr/particuliers/F573.xhtml

with recourse actions. The problem is not trivial because of financial and operational constraints and the stochastic nature of the workload.

As mentioned earlier, when there is excess workload, a natural tendency of care-practitioners is to modify their attitude and adapt to stressful situations at the expense of their own well being. The problem of optimal human resource allocation is then an actual concern in EDs during seasonal epidemics. However, extending the capacity of EDs is not easily implementable. A human resource management strategy is thus crucial for efficient service in EDs.

In order to develop a mathematical model suitable for the problem of staff allocation during epidemic crises, some assumptions are made. First we notice that during epidemics, the staff can also be exposed to infections. But as we are interested in seasonal epidemic that display mild infection patterns, we assume that the effects are negligible. In practice, even if a certain particular health-care practitioner is indisposed, managers find substitutes to replace him, albeit at a greater cost. Modeling the availability of staff makes sense in the case of large pandemics where the majority of the population is at risk.

Then, only one class of patients is considered. We assume that these patients are homogeneous in visit length, discharge and resources needed. A look at the data from the UHC of Lille shows that during epidemic seasons, the arrivals are primarily from epidemic patients. This peak in arrivals generates additional workload as these epidemic patients spend more time in the ED than the rest of the year. Some of this additional time is spent waiting in-between their diagnosis/treatment and the admission at the ward. Moreover, patients' destination following their visit is not important insofar as the resources deployed will not be responsible for the patients after their visit.

Another assumption is that based on the available information at the start of the epidemic, decision makers can characterize the resources needed across the planning horizon. Of course, these estimations are subject to errors and represented as random variables but the general trend can be forecast. In reality, at the time of the declaration of the epidemic, the strain of the virus is known and epidemiological models can be used to simulate the spread of the disease. Specifically, in this study we do not deal with special cases of epidemics where new viruses cause a pandemic. Rather, we examine cases of periodic overcrowding resulting from well studied pathologies such as known Influenza strains. These simulation models estimate the number of infected persons in a population per time period. It is then reasonable that ED decision-makers anticipate with a margin of error the number of resources required across the planning horizon.

We also make the assumption that resources can be scheduled independently. This implies that given a solution to the problem proposed, more operational decisions such a shift assignment with resource interactions and different patient flows are easier to take. The assignment of work periods to employees defines the day (or night) in which the employee is on duty but does not specify the type of shift he/she is on. We assume that given the 12 hours period in which an employee is working, it is always possible to come up with a shift assignment that satisfies individual constraints. This is possible if individual constraints are already ensured at the tactical level as we did in this study. Furthermore, since the workload remaining from a day to the other is very low because patients can not stay more than a day in EDs, the workload in the early morning is almost null and there is no accumulation of work.

5.4.1 On-Call Duties

Employees in France have one of three statuses : working, resting or on-call. This peculiar third status is defined as a period during which employees must be ready to intervene at work without being at the permanent and immediate disposal of the employer. When an employee is called in, the intervention periods including the transportation time are considered effective work hours. Otherwise, employees stay at home or in the vicinities and are paid a percentage of their hourly wage.

On-call duties are not widely applied in EDs yet. This reluctance to use on-calls comes from both practitioners and managers. For practitioners, on-call duties are associated with extra work hours as labor laws are sometimes circumvented. Managers on the other hand are put off by the strictness of the labor laws. The French laws for on-call duties are restrictive as they do not lead to an increase in the hospital's work capacity. Indeed, whether employees are on-call or not, the total work hours remains unchanged. In that sense, the only advantage of on-call duties is in the managerial flexibility they provide compared to regular duties. Specialized physicians are an exception as they often have on-call duties in the ED while working primarily in another department. In practice, some EDs use on-call duties as a surge mechanism in exceptional circumstances. Instead of counting work performed by employees who were on-call as effective work hours, they consider it to be paid "extra work". Of course, doing so means they bypass the law and effectively increment their capacity.

In reality, the on-call payment scheme can be beneficial for both employees and employers. Employers have staff on demand paying them a fraction of their hourly wage and employees are paid while staying at home. Additionally, if they called in to work, they are paid extra hours. Although flexible, this organizational scheme is not currently used in emergency departments in France for the reasons we stated above. However, this situation can change and according to a public service decree "The head of the hospital, after an agreement with the technical committee of the hospital, determines the list of activities, services and jobs that are concerned [with on-call duties]". This study analyzes the advantages of allowing for employees to be on-call during epidemic crises. For doing so, a comparison is made with regular ED schedules.

5.4.2 Mathematical Formulation

When an epidemic starts, ED managers know that demand will increase but have little or no information on the amplitude and duration of the epidemic. For that reason, it is logical to have some flexibility in the staff deployment plan. Decisions are made in two stages. The first stage decision is made at the start of the epidemic horizon and consists of contracting resources to be on-call or on a regular duty in period t. Second stage decisions are made successively at the beginning of each period and determine the number of resources to call to work among the ones that are on-call.

When on-call, physicians have a fixed hourly compensation w regardless of their actual wage. If these physicians are called in to work they are paid at overtime rates d. The distribution of the resources n among the T half day periods needs to respect the working time constraints imposed by the law. The rest of the employees is ensured by variables U_n that keep track of the worked periods in a week. Also parameters O_n impose an upper bound on the on-call duties possible and parameters H_n impose an upper bound on the

Parameter	Description
T	Length of the planning period, indexed by t
N	Total number of physicians, indexed by n
M	Minimum number of physicians on duty by period
b_t	Number of physicians needed in period t
c	Cost to put an employee on duty by period
w	Cost to put an employee on-call by period
d	Cost to call in an on-call personnel
W_n	Minimal number of duty periods of a physician n
O_n	Maximal number of on-call periods of n
H_n	Maximal number of night shifts allowed for resource n
α	Shortage cost coefficient

TABLE 5.1 – Parameters of the model.

night shifts allowed per resource n. The resources that can be allocated to on-call duties are those that are not on a regular shift during that day.

Consider the total number of patients arriving to the ED in period $t = 1, .., T$. Once this random process is realized, it is possible to estimate the number of physicians required to meet this workload. In our model, for each realization ξ, we denote $b_{nt}(\xi)$ the number of physicians to meet this demand. At the beginning of the planning horizon we take the decisions x_{nt} and y_{nt} to allocate physicians respectively on duty or on-call for periods $t \in 1..T$. The decisions are robust insofar as they are optimized with respect to several possible epidemic realizations. We denote the matrices containing the variables x_{nt} and y_{nt} by x and y respectively.

If a resource is scheduled to be on-call during period t, a decision $y'_{n,t}$ whether to call resource n is made for each period. This decision depends on a compromise between paying an extra fee or incurring a penalty α for not meeting the demand. Tables 5.2 and 5.1 summarize the different parameters and variables of the model.

Objective :

$$min \quad \sum_t \sum_n (c * x_{nt} + w * y_{nt}) + \mathbb{E}[Q(x, y, \xi)] \tag{5.1}$$

Subject to :

$$W_n \leq \sum_t x_{nt} \quad \forall n \tag{5.2}$$

Variable	Description
x_{nt}	$= 1$ if physician n is on duty in period t
y_{nt}	$= 1$ if physician n is on-call in period t
$y'_{nt}(\xi)$	$= 1$ if physician n is called in period t
U_{nt}	$= 1$ if physician n works at least one period during the three periods following t t
$l_t(\xi)$	Physician shortage in period t

TABLE 5.2 – Variables of the model

$$\sum_t y_{nt} \leq O_n \quad \forall n \tag{5.3}$$

$$\sum_{t \in night} (x_{nt} + y_{nt}) \leq H_n \quad \forall n \tag{5.4}$$

$$x_{nt+1} + y_{nt-1} + x_{n,t} + y_{n,t} \leq 1 \quad \forall n \ \ \forall t \tag{5.5}$$

$$\sum_{t'=t}^{min(t+2,T)} (x_{nt'} + y_{nt'}) \leq 3U_{nt} \quad \forall n \ \ \forall t \tag{5.6}$$

$$\sum_{t'=t}^{min(t+11,T)} U_{nt'} \leq 11 \quad \forall n \ \ \forall t \tag{5.7}$$

$$\sum_n x_{nt} \geq M \quad \forall t \tag{5.8}$$

$$U_{nt}, x_{nt}, y_{nt} \in \{0, 1\} \quad \forall n, \ \ \forall t \tag{5.9}$$

$Q(x, y, \xi)$ denotes the solution to the second stage problem where $Y'_t = \sum_n y'_{nt}$:

$$Q(x, y, \xi) = min \sum_t d * Y'_t(\xi) + \alpha l_t(\xi) \tag{5.10}$$

$$\sum_n y_{nt} \geq Y'_t(\xi) \quad \forall t \ \ \forall n \tag{5.11}$$

$$b_t(\xi) - Y'_t(\xi) - \sum_n x_{nt} \leq l_t(\xi) \quad \forall t \tag{5.12}$$

$$l_t(\xi) \geq 0 \quad \forall t \tag{5.13}$$

Planners are interested in obtaining a robust allocation scheme in the beginning of the planning horizon. The objective function (5.1) tries to find such a scheme by minimizing the total cost over the planning horizon while trying to meet as much demand as possible.

If EDs are overcrowded, i.e if the demand is above the service capacity, there are several consequences with regards to patients. Besides possible health complications, patients can both be delayed and leave without being seen. However, as we are dealing with decisions that improve patients throughput, it is much more likely in this context that patients remain in the ED even with increased waiting times. Patient's outcome is not explicitly stated, but an indicator of the amount of expected shortage is meaningful for decision makers who can relate to situations when resources were missing. We are interested in evaluating the impact of on-calls on schedules costs. For this reason, we assign a cost to regular duties as well as to on-call duties even if employees are not paid by duty but have a fixed salary.

The first set of constraints (5.2) ensure resources have to work on regular duties for at least W_n periods. This prevents the allocation model from assigning only on-call duties. Constraints (5.3) set an upper bound on the number of on-call duties that are assigned to the resources. Similarly, this prevents the allocation model from assigning more on-call duties than what is legally allowable. Next, constraints (5.4) set an upper bound on the number of night shifts that employees can have during the epidemic horizon. Since this limit depends on the work history, contract or preferences of each individual employee, they are indexed by n. Constraints (5.5) state that an employee can not be assigned to a regular duty and an on-call duty on the same period. Furthermore, an employee can not be assigned to duties on two consecutive periods. Admittedly the constraints on the on-call duties are over conservative as on-call duties are counted as effective work hours only if an employee is called. The actual constraints as imposed by law involve the variables y'_{nt} rather than y_{tn}. However, since the variable y'_{nt} refer to decisions that are scenario-dependent, having a constraint on these over optimizes the solution as if the future realizations were known while taking the decisions. Constraints (5.6) and (5.7) state that employees must rest for at least 3 consecutive periods in a week. This rest corresponds to the 36 hours of weekly rest imposed by the labor law. We first go through all three consecutive periods and verify if one of those periods is a work period. If we find three consecutive rest periods (i.e $x_{nt'}$ and $y_{nt'}$ are zero) then the variable U_{nt} can be assigned 0. Otherwise the value of U_{nt} is 1. Constraints (5.7) then checks that there is at least one t in the following week (14 periods) where U_{nt} is not 1. Finally, constraints (5.9) ensure that the variables x_{nt}, y_{nt} and U_{nt} are binary.

The second stage objective (5.10) is to find the optimal number of resources to call for each epidemic scenario given there is a penalty on the resource shortage. Constraints (5.11) are logical constraints that make sure that only employees that are on-call can be called. Constraints (5.12) are demand satisfaction constraints with the variable $l_t(\xi)$ keeping track of the shortage on period t. The variable $l_t(\xi)$ is calculated from a random number of infected patients. In the model, an equivalent formulation is to have the demand in number of patients denoted as $b_t(\xi)$ and a constraint specifying :

$$l_t(\xi) = b_t(\xi) * \beta \tag{5.14}$$

With β being a ratio of resources to patients that we calculate based on information on the epidemic and physician's experience. The ratio β in our tests is a random variable itself because it depends on individual patients and the service rate. We chose to hide this equation and generate data independently leaving the possibility of another mean of demand generation. We know that queuing effects are negligible from a day to another since the workload at night is almost always completely absorbed. Finally, constraints (5.13) are positivity constraints for variables $l_t(\xi)$. We can show that the second stage objective reduces to the following two cases :

$$\begin{cases} Q(x,y,\xi) = \alpha l_t(\xi) & \text{if } d > \alpha \\ Q(x,y,\xi) = \alpha l_t(\xi) + dY_t'(\xi) & \text{if } d <= \alpha \end{cases} \tag{5.15}$$

In reality, in order to mirror the day-to-day behavior where decisions are taken without an accurate vision of the future, non-anticipativity constraints need to be satisfied. Optimal second stage decisions $Y_t'(\xi)$ are null if $d > \alpha$ and $Y_t' = \min(b_t(\xi), \sum_n y_{nt})$ otherwise. Consequently, these decisions are independent from future realizations.

5.5 Numerical Experiments

In the following set of tests we show the advantages and shortcomings of using alternative scheduling strategies and analyze whether they help us handle uncertainties during an epidemic season. All tests were run on an Intel(R) Xeon(R) 4 cores CPU E5520 @ 2.27GHz with 8 MB of cache memory and 8 GB of RAM. We choose a Sample Average Approximation method (SAA) approach to solve the model [Kleywegt et al., 2002]. The Sample Average Approximation (SAA) method is a solution methodology for stochastic discrete optimization problems based on random sampling of data from the input space. Random samples are generated and the expected value function is approximated by the corresponding sample average function. With the increase of sample size, the solution converges to the real optimum. The sample size is chosen to guarantee an acceptable margin of error. We used samples estimated based on epidemic simulations. After we parametrize the epidemiological model based on the epidemic data of interest, we calculate a number of infected persons in a population and deduce the expected demand from those. In Table 5.4 the stability of the model is tested for different number of samples realization. The column Obj refers to the average value of the objective function obtained from three sets of realizations in each case. As the objective value seems to stabilize around 100 samples, this number is used throughout for the remaining tests.

Unless otherwise stated, the default values of the parameters used for the tests are shown in Table 5.3. The choice of the cost parameters are motivated by labor laws stating that the cost of an on-call duty is 1/4 the cost of a regular duty if the employee is not called in for work. Otherwise the employee is paid extra hours. Furthermore, the value of the shortage parameter α was chosen so as to be higher than the cost of any work duty. This is consistent with equation 5.15 which states that otherwise the solution to the problem is trivial. We picked a 30 days planning horizon which corresponds more or less to the length of an epidemic based on physicians' experience. The number of on-call duties and night shifts are motivated by labor laws and illustrate a realizable set up given the length of the planning horizon. Finally, the value of M states that there should be at least one physician present at all times in the ED. This parameter can be adjusted but serves to illustrate one practical example.

Parameter	Value
c	4
w	1
d	4
α	7
T	60
I	13
W	10
O_n	10
H	10
M	1

TABLE 5.3 – Model Parameters

Ξ	Obj
50	1054.5
100	1070.9
200	1075.0
500	1064.3
1000	1068.4

TABLE 5.4 – Numerical stability through different number of samples

5.5.1 Design of Experiments

In order to examine the impact of a flexible planning strategy in facing epidemic cases, the cost of different planning alternatives are studied. In Section 5.5.2 we examine three types of epidemics : mild, moderate and severe. These scenarios serve to estimate the arrivals in the ED and consequently the number of resources needed. Furthermore, the shortage cost is a key parameter in our model as it corresponds to a penalty cost for being short in capacity. As the value of this penalty cost can not easily be estimated we investigate in our tests schedules with the value of α in a range of $+20\%$ to $+500\%$ of the cost of an on-call duty.

In Section 5.5.3 the impact of on-call duties on ED schedules is studied. To do that, we impose a weekly cyclic schedule constraint on regular duties and allow on-call duties to be distributed optimally. The cyclic schedule where the number of resources per period is fixed mimics the current scheduling practice. The results are compared to a schedule with no on-call duties. Next, in Section 5.5.4 we relax the cyclic constraints and analyze whether on-calls are still beneficial in the presence of non-cyclic duties. Having non-cyclic schedules makes sense in the case accurate demand forecast is available. Finally, we examine in Section 5.5.5 how the optimal distribution of duties differs in the case laws are circumvented.

5.5.2 Demand Generation

Influenza-like-illnesses (ILI) or acute respirator infection (ARI) are umbrella terms that refer to medical disorders whose symptoms resemble those of Influenza cases. These concepts are important inasmuch as they allow us to group certain pathologies and study their collective incidence on EDs. During epidemic seasons, ILI are the main sources of ED activity. Patients with disease like influenza, bronchiolitis and gastroenteritis add up to the regular volume of patients and significantly increase the workload. As we have said previously, the increase in workload is not primarily due to an increase in patients' volume but rather in an increase in the time spent in the ED and acuity of the patients visiting.

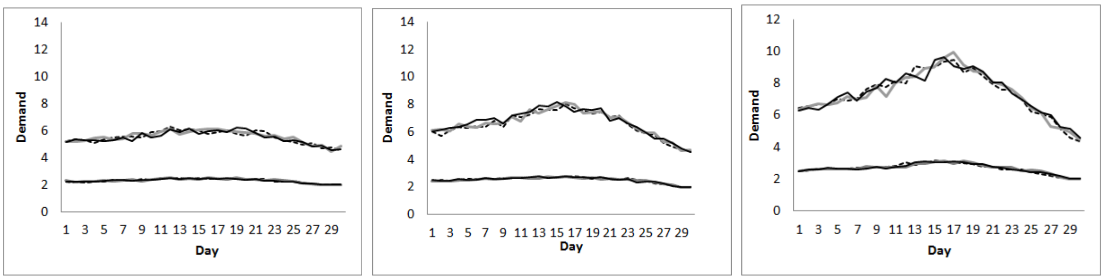
From year to year, epidemics differ in nature. This uncertainty leads to managerial complexities. EDs struggle to characterize the seasonal workload peak so as to organize their service and serve the demand better. From a modeling perspective, this uncertainty has to be taken into account if the model is to solve the difficulties encountered in the ED. There is uncertainty as to the type of epidemic and as to the intensity and fluctuation within one type of epidemic. A stochastic SEIR model is used to simulate realizations of the epidemic with the three different contact rates.

Running experiments with three different contact rates illustrate the solutions in different epidemic scenarios (mild = 0.4, moderate = 0.6 and severe = 0.8). In practice, the contact rate of an epidemic can be estimated by specialists if the type of the epidemic is known. The epidemic model for simulating the spread of disease is represented as a stochastic Markov chain with four compartments : Susceptible, Exposed, Infectious and Recovered which record the number of individuals in each category. The initial state has few infected individuals. At each time step (a couple of minutes), there is a probability that an individual goes to another state in the system. We chose to use these models to generate demand because they provide good approximations of the increase and reduction of infected population across time. As we are primarily interested in the workload associated with infected patients, we believe that estimating demand based on an epidemiological model provides good basis for planning. In this study, we are interested in the evolution of the number of individuals in the infectious category [Chowell et al., 2008]. Parameters of the SEIR model are fitted to the population of the area around the ED. Figure 5.5 shows the average number of required resources calculated for each type of epidemic for both day and night. The averages are calculated from 3 instances of 100 sample realizations each. Demand at night is lower than during the day and does not follow the epidemic as explained in Section 5.3.

Based on the number of infectious individuals in a population, ED arrivals and resources needed are calibrated for a particular hospital in Lille France. Consequently, the number of resources used for our tests allow us to cover in average all the demand of Figure 5.5a without any shortage.The three epidemic scenarios and the varying values of α are at the base of all the following experiments forming 15 test instances. We start in the next section by testing the benefits of on-calls compared to current scheduling practice.

5.5.3 Cyclic Schedules

So as to evaluate the impact of on-calls on the scheduling costs, we compare schedules with and without on-calls. In both cases, as in current practice, the schedules of regular



(a) mild (b) moderate (c) severe

FIGURE 5.5 – Day and night demand for three types of epidemic scenarios

duties are cyclic. This means that they are optimized for one week and repeated for the rest of the planning horizon. Table 5.5 and 5.6 show the results obtained for the moderate epidemic scenario.

The first column shows the percentage added to the cost α compared to the cost of an on-call duty. For example, for $\alpha = +20\%$ the penalty cost for the shortage is set to $w + d + ((w + d) * 20/100) = 6$. For each value of α, three instances are run each with 100 sample epidemic realizations. All tests that we ran finished in less than an hour for a relative optimality gap that was set to 0.01%. The next column shows the time in seconds to obtain the solution. After that the column “Fix” specifies the number of regular duties assigned in the optimal solution. Then “On-call” specifies the number of on-call duties assigned. Finally, the columns “Exp. Calls”' and “Exp. Short” show respectively the expected number of calls made to personnel that were on-call and the expected number of shortage across the horizon.

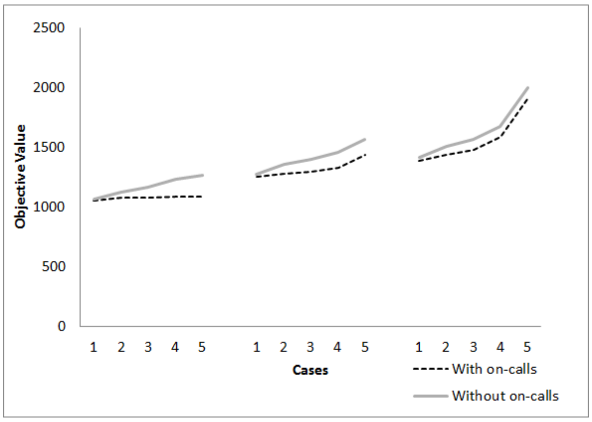


FIGURE 5.6 – Difference in objective value with and without on-calls for different values of α in the three scenarios

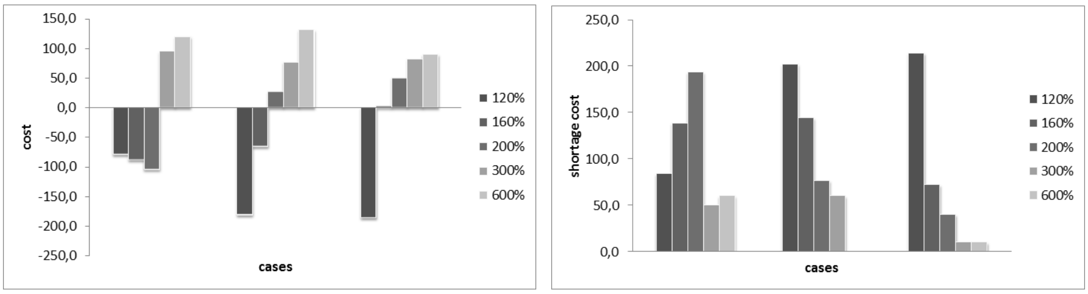
Figure 5.6 illustrates the difference in the objective value obtained for schedules with and without on-call duties for the three scenarios. We see that for all cases, the schedules with on-call duties dominate the schedules without. The values on the x-axis designate the different test cases (different values of α). So as to better understand how the plannings are configured we analyze the difference in costs for the objective functions. Figure 5.7a and Figure 5.7b show the reduction in the workforce and shortage costs with and without on-calls. A positive difference means that the schedules with on-calls are better. The

Scenario	α	Obj	Time (s)	Fix	On Calls	Exp. Calls	Exp Short
mild	1- +20 %	1055.8	57.8	191	43.3	28 ± 3.7	22.3 ± 4.6
	2- +60 %	1075.8	122.2	191	89.7	45.3 ± 6.1	4.3 ± 2.1
	3- +100 %	1081	204	191	104.7	49 ± 6.6	1 ± 1.3
	4- +200 %	1085.9	241.5	191	110.3	49.3 ± 6.9	0.3 ± 0.9
	5- +500 %	1090.1	280.1	191	116	50 ± 6.9	0 ± 0.5
moderate	1- +20 %	1254.3	137.6	209.7	77.7	51.3 ± 5.3	21 ± 4.8
	2- +60 %	1281.4	695.2	211	102	61 ± 6.6	10.7 ± 3.5
	3- +100 %	1296.1	1095.4	212.7	107.7	61.7 ± 6.8	8.7 ± 3.2
	4- +200 %	1330.2	1540.6	212.7	114.7	63.3 ± 7	6.7 ± 3
	5- +500 %	1433.8	2072.4	209.3	119.3	66.3 ± 7.2	6.3 ± 3
severe	1- +20 %	1389.4	469.2	218	81	57.3 ± 5	33.3 ± 6.4
	2- +60 %	1438.3	1105.8	222	98.7	64.7 ± 5.7	23.3 ± 5.6
	3- +100 %	1481.9	1375.5	222	104.3	66.3 ± 6	21.7 ± 5.6
	4- +200 %	1585.6	1727.2	224.7	104.3	65.3 ± 5.9	21 ± 5.5
	5- +500 %	1900.5	1673.1	226	104.7	64.3 ± 5.9	20.3 ± 5.5

TABLE 5.5 – Results for cyclic schedules with on-call duties

Scenario	α	Obj	Time (s)	Fix	On Calls	Exp. Calls	Exp Short
mild	1- +20 %	1062.3	40.3	210.0	0.0	0.0	36.3 ± 6.1
	2- +60 %	1124.9	43.6	236.7	0.0	0.0	21.7 ± 4.5
	3- +100 %	1167.1	44.1	240.0	0.0	0.0	20.3 ± 4.3
	4- +200 %	1232.3	45.5	291.7	0.0	0.0	3.7 ± 2
	5- +500 %	1269.7	42.8	300.0	0.0	0.0	2 ± 1.4
moderate	1- +20 %	1271.9	53.5	235.0	0.0	0.0	54.7 ± 7.5
	2- +60 %	1358.7	55.7	281.0	0.0	0.0	28.7 ± 5.7
	3- +100 %	1399.8	59.3	308.0	0.0	0.0	16.3 ± 4.6
	4- +200 %	1460.8	58.4	324.0	0.0	0.0	10.7 ± 3.7
	5- +500 %	1566.5	311.5	338.3	0.0	0.0	6.3 ± 3
severe	1- +20 %	1414.2	51.6	249.0	0.0	0.0	69 ± 8.8
	2- +60 %	1509.4	59.6	312.0	0.0	0.0	32.3 ± 6.7
	3- +100 %	1567.5	61.1	327.0	0.0	0.0	25.7 ± 5.9
	4- +200 %	1676.4	265.5	336.7	0.0	0.0	21.7 ± 5.4
	5- +500 %	1996.1	501.2	339.0	0.0	0.0	20.7 ± 5.4

TABLE 5.6 – Results for cyclic schedules without on-call duties



(a) Workforce cost reduction with on-call vs without on-call (b) Shortage cost reduction with on-call vs without on-call

FIGURE 5.7 – Costs of cyclic schedules in the mild, moderate and severe scenarios

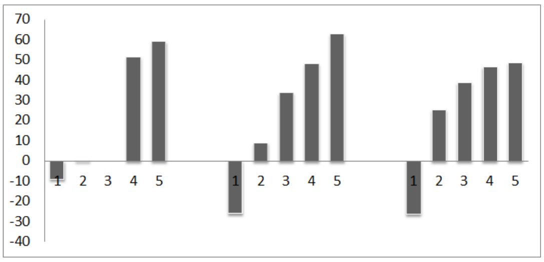


FIGURE 5.8 – Effective work reduction with on-call vs without on-call for cyclic schedules

workforce cost is calculated by multiplying the number of effective work hours by the unit cost of a work hour.

When the penalty on the shortage α is low, schedules with on-calls cover the demand better than schedules without on-calls. For example for $\alpha = +20\%$ There is an expected shortage of 54.7 resource-periods for the moderate epidemic scenario in the schedules without on-calls against 21 resource-periods in the schedules with on-calls. This reduction in the shortage costs compensates on the slightly higher workforce costs. As the value of α increases, the difference in shortage is reduced but the schedules with on-call gain on workforce costs. The improvements are due to the fact that on-call duties only cost a fraction of regular duties when they are not used. As such, only a fraction is paid for several epidemic realizations while covering the demand in the other cases.

To further understand the differences in the schedules obtained, we plot in Figure 5.8 the difference in actual worked hours between the schedules without and with on-calls. A positive difference means that the the schedules without on-calls have more effective worked hours. The difference is simply calculated as : $x[no\ on-call] - (x[on-call] + \mathbb{E}(Y'[on-call]))$. As the value of α and the number of on-call duties increase, employees work less effective hours. In other words, schedules with on-call duties are more likely to lead to more rest for the employees. For example, if we observe the last three cases in Table 5.5 and 5.6, we notice that in the cases with on-call duties, less man-hours are used with a better coverage. This improvement is due to the added coverage flexibility with on-calls.

5.5.4 Non-cyclic Schedules

In this section, we relax the cyclic constraints on the schedules. Figure 5.9 shows the resulting objective values obtained for schedules with and without on-call duties again. The values we observe are very similar to the costs obtained when the cyclic constraints are imposed.

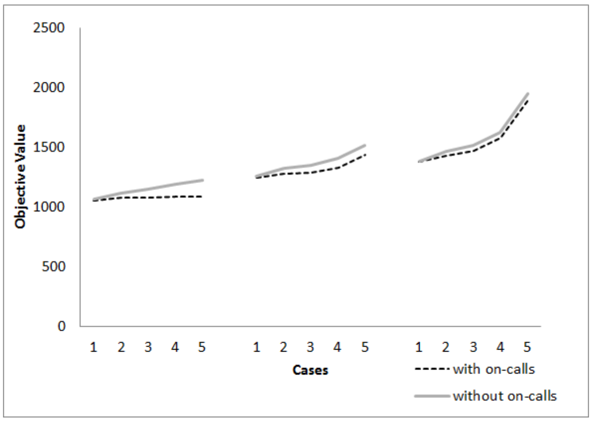
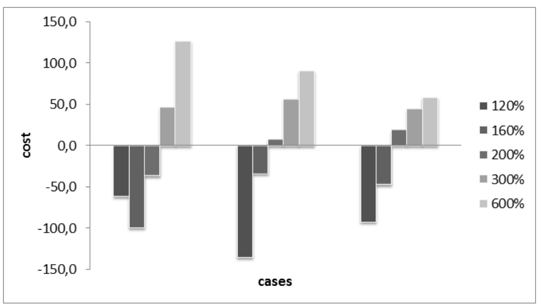


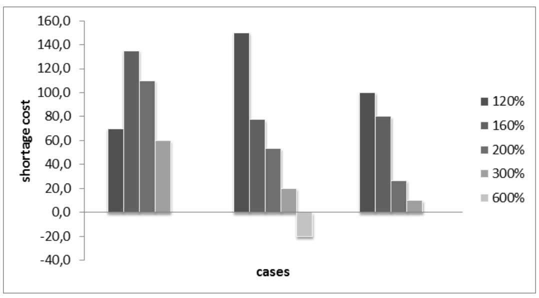
FIGURE 5.9 – Difference in objective value with and without on-calls for non-cyclic schedules

If we examine the details of the costs in the schedule in Figure 5.10a and 5.10b, we observe a similar behavior than the one in Section 5.5.3. When the value of α is low, the schedules without on-calls do not cover the demand as well as the schedules with on-calls. As α increases, the coverage becomes similar but the schedules with on-calls are cheaper. The results suggest that relaxing the cyclic constraint does not bring a lot of benefit over having optimized cyclic schedules.

Furthermore, in Figure 5.11 a similar pattern than the one observed in Figure 5.8 is seen. This similarity in the distribution of work duties suggests that the flexibility allowed by the regular duties are already exhausted in optimizing in a cyclic manner.



(a) Workforce cost reduction with on-call vs without on-call



(b) Shortage cost reduction with on-call vs without on-call

FIGURE 5.10 – Costs of non-cyclic schedules in the mild, moderate and severe scenarios

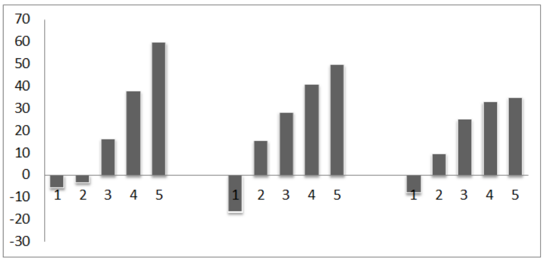


FIGURE 5.11 – Effective work reduction with on-call vs without on-call for non-cyclic schedules

5.5.5 Relaxed Constraints on Workload Regulations

In this section, we relax the weekly rest constraints relating to on-call duties and study the effect on the workforce cost and shortage costs of the resulting schedules.

Scenario	α	Obj	Time (s)	Fix	On Calls	Exp. Calls	Exp Short
moderate	1- +20 %	1248.7	69.3	224.0	66.3	41.3 ± 4.8	19.3 ± 4.5
	2- +60 %	1266.3	127.2	222.0	109.7	58 ± 7	4 ± 2
	3- +100 %	1271.2	170.3	220.3	123.3	61.7 ± 7.5	1 ± 1.5
	4- +200 %	1278.7	177.5	221.0	131.3	62.3 ± 7.7	0.3 ± 1
	5- +500 %	1279.6	202.6	220.3	139.7	63.7 ± 7.9	0 ± 0.3

TABLE 5.7 – Relaxed weekly rest constraint for moderate scenario

We observe in Table 5.7 that by relaxing the rest constraints on on-call duties, we allow for a better coverage of the demand (no shortage when α is large). More interestingly, the total effective work hours are reduced compared to schedules with tighter constraints. This corroborates our assertion that on-call duties can lead to a better coverage of the demand with less effective work hours being done.

5.6 Conclusion

There are very few choices ED managers have to reduce overcrowding during epidemic seasons. On-call duties are one of these possible actions that are readily available. On-calls are sometimes associated with extra work and unfriendly work conditions. We have analyzed in this paper the advantages and shortcoming of using on-call duties in an ED department for different epidemic scenarios.
To this end, we developed a stochastic model to compare different scheduling strategies. The schedules obtained are robust in the sense that they hedge against the uncertainty in the epidemic realization if the type of the epidemic is known. Given the objective of minimizing the total costs incurred and covering the demand as best as possible, using on-call duties resulted in schedules that dominated the regular schedules.
We recommend using on-calls as a recourse mechanism during seasonal epidemic periods as they increase the coverage while keeping the costs down and prevent idle hours of employees. For instance, when resource shortage is very expensive, using on-call duties

can lead to more than 30% reduction in effective work hours with schedules that are 10% less expensive. It is more desirable to use on-calls during periods with a lot of uncertainty on the demand. In the tests we ran this corresponded to day periods which had more variance. Nonetheless, in the scenarios considered, there is always over 60% of regular duties in the optimal schedules which indicates that on-call duties are not a substitute to regular duties but rather an efficient means to overcome increased variance in the workload.

Several extensions to this work are possible. The daily shift assignment problem is necessary for determining the optimal allocation of shifts to resources within the day. Also, we did not consider in this study the interaction between the resources but non negligible queuing effects are present when determining daily shift assignments. Furthermore, in this study we only focused on the workload resulting from epidemic patients, the workload patterns should be adjusted if different arrival trends are observed. For example, in some hospitals, the workload during the weekend is higher than during the week. This can easily be included when considering the demand scenarios to consider.

Chapitre 6

Dynamic Management of Overtime Hours

Résumé en français du chapitre : Optimisation dynamique d'une politique d'affectation d'heures supplémentaires aux urgences

Pendant les périodes de tension, il est courant que les médecins urgentistes restent en heures supplémentaires à la fin de leur poste pour finir de traiter les patients dont la prise en charge a déjà commencé. Cette pratique est d'autant plus courante pendant les périodes épidémiques où les pics d'activité sont fréquents. Cependant, le nombre d'heures supplémentaires que les médecins peuvent effectuer au cours d'une semaine est limité. Dans ce chapitre nous proposons une méthode pour trouver des politiques d'affectation des heures supplémentaires dans le contexte dynamique des urgences. Nous définissons une politique comme une fonction de décision définie sur l'ensemble des états possibles. Les états considérés indiquent la charge de travail encore à effectuer à la fin du poste ainsi que le nombre d'heures supplémentaires que chaque médecin peut encore faire dans la semaine. Nous modélisons dans un premier temps le problème sous forme de processus de décision Markovien. Cette modélisation a l'avantage de considérer tous les états possibles sans aucune simplification. Malheureusement, pour des instances réalistes du problème, il est impossible de résoudre le processus Markovien en un temps acceptable. Par conséquent, nous proposons par la suite des modèles approximatifs qui génèrent des politiques sous-optimales. Nous comparons finalement nos politiques de répartition d'heures supplémentaires sur des instances tirées des données du CHRU de Lille et notons les différences en termes de temps de calcul ainsi que par rapport à la fonction objectif.

Sommaire

6.1 Introduction **93**
6.2 Literature Review **93**
6.3 Problem Description **94**
6.3.1 Demand Representation 95
6.4 Markov Decision Process **96**

6.4.1 Decision Variables 96
6.4.2 Marginal Cost 97
6.4.3 Value Function 97
6.5 Policy Approximations 98
6.5.1 Upper Bound 98
6.5.2 Lower Bound 99
6.5.3 Dynamic Approximation Scheme 100
6.6 Experimental Results 100
6.6.1 Model Parameters 101
6.6.2 Policy Approximation 102
6.6.3 Long Planning Horizon 106
6.7 Conclusion 107

6.1 Introduction

A key to reduce overcrowding in EDs is an informed use of resources, both human and material. In this work we find dynamic policies for managing overtime hours in EDs. Overtime hours are used in exceptional circumstances to help react to unaccounted changes and perturbations in the schedules established. As overtime hours are limited, policies should be evaluated to use these resources as best as possible.

Other operational solutions to counter overcrowding in EDs are possible but are not currently used in the hospitals visited. Furthermore they have several drawbacks that make them less practical to use in the context of seasonal epidemics. For example, hiring additional workforce requires some planning time and is costly. Usually the hiring process is time consuming as the personnel needs training and adaptation time. Downstream beds management on the other hand requires data from other departments in the hospital which is currently unavailable to us. In this work we focus on the optimization of overtime assignment policies within the ED as a strategy to quickly and efficiently increase the capacity temporarily.

The motivation behind this work is primarily practical. In all the ED visited, overtime hours are used frequently as a temporary solution to overcrowding. There is often no strategy to distribute overtime hours although the work constraints imposed on them are very strict. In the case of seasonal epidemics, this situation can lead to days where resources are urgently needed and all overtime hours are spent. The difficulties in dealing with overtime hours are many. No the least of which, working overtime hours is related to adverse effects on the health-care personnel. This is particularly problematic in EDs where the attention and well-being of the staff can have fatal consequences on patients [Bae and Fabry, 2014].

6.2 Literature Review

An early paper by M. I. McManus [McManus, 1977] studied the question of overtime assignment in the context of post offices. Post offices are similar to EDs in that the work demand is time-varying. In the paper, the author defines, under some simplifying assumptions, the optimal balance between staffing levels and overtime hours. However in that paper, overtime hours are not considered as surge mechanisms but merely as work hours that are paid at a premium rate. Similarly, in a paper by Stolletz and Brunner [Stolletz and Brunner, 2012], overtime hours are considered in the context of designing shift schedules and not operationally as in this paper.

Studies to find optimal control policies in the health-care domain have been investigated to reduce the spread of the disease [Lee et al., 2012, Jung et al., 2009]. Contrary to our research, these works study the problem from the perspective of an entire population. As such, the systems in which the control policies are evaluated are population models (epidemic models) and not EDs. For example, in a paper by Lee et al. [Lee et al., 2012] the use of antiviral treatment and isolation strategies are tested as possible alternatives to reduce the epidemic transmission in a population. Similarly, Jung et al. [Jung et al., 2009] investigate the propagation of the avian bird flu to humans and the impact of time-dependent control policies to reduce risks.

Concerning emergency departments, a paper by Chockalingam and Lawley [Chockalingam et al., 2010] presents a stochastic control mechanism to regulate ED resources to prevent overcrowding. The authors first model the ED as a Petri Network and define divert states as states where resources are insufficient to cover the demand. Then, given each possible state the ED system can be in, they calculate a "distance to divert" that is a measure of the risk of entering a divert state from a given state. Having the "distances to divert" at different points in time, the authors then calculate control policy to determine when to increase resources. The control resources they consider are surge staff and beds. While it is true that health-case personnel and beds are critical resources in the ED, it is very difficult in practice to have surge physicians come in on demand. Instead, it is common practice to have physicians stay overtime in particular circumstances.

A paper by Xiao et al. [Xiao et al., 2010] proposes a dynamic scheduling mechanism for emergency department resources. In that paper, the authors define an incremental sliding window framework for scheduling resources at fixed points in time. At each decision point, a genetic algorithm is used to allocate resources to activities in an optimized manner. The performance of their method is then evaluated using a simulation model. The authors give no insight however on how the rescheduling process is applied in a real ED setting given the strict work regulations.

Finally, some articles on dynamic ambulance redeployment problem exhibit many similarities with the problem of dynamic overtime assignment [Maxwell et al., 2009, Maxwell et al., 2010, Schmid, 2012]. The problem of ambulance redeployment is the problem of relocating ambulances so as to respond to urgent calls in the most timely manner. Although the problem is primarily a location problem, we can make a parallel between the ambulances, which are the scarce resources, and overtime hours. In both problems, the demand is unpredictable and varies with time. Due to the large state space, dynamic ambulance redeployment has been mostly addressed using approximate dynamic programming techniques to find near optimal policies. In this chapter, we formulate the problem of dynamic overtime assignment as a Markov decision process (MDP) then present an linear programming (LP) approximation and show that we can obtain good quality solutions in a short time.

6.3 Problem Description

Overtime management strategies can improve the service quality in EDs on a day to day basis. But overtime hours are scarce resources. Physicians can only be assigned a certain number of overtime hours per week. Any strategy to assign overtime hours should take into account work regulations and hospital specific policies. In addition, it is wise when taking a decision to consider the possible demand in the following days. Some modeling assumptions are made :

1. We assume a finite planning horizon of one week. This is motivated by the fact that, in the French labor law, a limit on the number of working hours is imposed per week. Consequently, after that period the number of overtime hours can be reset.
2. We assume overtime assignment decisions are taken once for each shift.We set a single decision point at the end of the shift. This way decision makers can be

sure that overtime hours are going to be necessary before making the decision. Physicians' schedules define the work hours physicians are assigned to.

3. We assume that physicians' shifts are non-overlapping. In reality shifts do overlap but the overlapping period is primarily used to transfer patient information and is not used to treat patients.
4. We assume overtime hours are taken for whole hours. In other words, physicians can only be assigned an integer number of overtime hours. This is logical given the way overtime hours are remunerated.

Overtime hours are scarce resources, if needed, all overtime hours are going to be assigned during the week. This is particularly true during periods of chronic overcrowding where overtime hours are frequently used. Hence, in our decision making process, we do not penalize the action of assigning overtime hours but only of not covering the excess demand. This focuses the study towards analyzing the conditions under which it is favorable to allocate overtime hours.

Overtime management strategies require an assessment of the workload in EDs. EDs workload depends to a large extent on the arrival of patients. However, some patients generate more workload than others and hence require additional resources. To estimate the workload we define arrival patterns of patients that we extract from the hospital's historical data. Based on this estimation, we optimize a strategy to distribute overtime hours to cover the workload at the end of each shift. Next, we describe the various assumptions we make to formalize the problem description.

6.3.1 Demand Representation

We also make some assumptions regarding the demand.

5. We assume that the ED process can be represented as a queuing system following the Markovian properties of Poisson arrivals and exponential service times. Physicians are the servers of our system. For each shift, there is a predefined number of physicians working on predetermined working patterns. The total service offered can be calculated given the service time distribution and the fixed physicians' schedules.

This assumption is common in the ED scheduling literature as has been shown in Chapter 3. For the arrival rates, we obtain from the database an average rate per day and per hour and use a Poisson distribution parametrized with these averages. Each patient arrival adds workload to the shift. We assume that the arrivals are independent from one another. Indeed, in most cases, patients arriving to the ED are not directly related.

Based on the arrival patterns and the service time, we calculate the excess workload at the end of each shift. The excess workload is the amount of workload remaining at the end of the shift. An empirical distribution Ω of the excess workload can be calculated by simulating patient arrivals and service time.The excess workload is estimated in hours of work for physicians and is proportionate to the average service times. If a patient arrives on a certain shift, the physician on duty is most likely to be the one attending the patient even if he stays overtime. Finally :

6. We assume then that the excess workload is either covered by overtime hours or penalized but does not transfer to future shifts. The penalization of excess

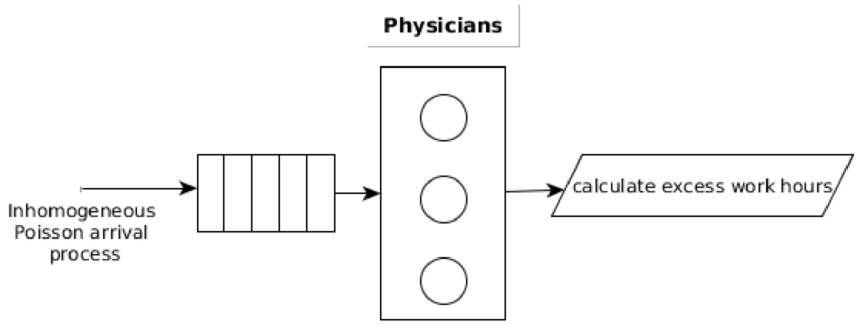


FIGURE 6.1 – M/M/N queue that allows us to calculate the excess workload distribution at the end of each shift

workload counts as additional stress to resources who often adapt to finish treating patients within their working hours. Physicians often stay overtime hours that are not counted as such to continue treating their patients. The penalization can be used to reflect these stressful patterns.

7. We express the excess workload in work hours needed. We obtain this number of work hours by multiplying the number of patients remaining at the end of the shift by the average service time and rounding to the nearest hour.

We illustrate the queue system in Figure 6.1. The number of physicians in the queue depend on the physicians' regular schedule.

6.4 Markov Decision Process

In this section we formally describe the decision process to find an optimal overtime allocation policy π. Let $s_t = \{W_t, h_{i,t}\}$ be the system state. W_t is the excess workload for the current shift. For each physician i, $h_{i,t}$ keeps track of the remaining overtime hours available. The number of overtime hours physicians can take on a given week is limited. Consequently, the variable $h_{i,t}$ is used to make sure that labor laws and hospital specific policies are respected.

As we are working in a finite horizon the period t is important. The period is defined by the day of the week and the shift of the day. The shift corresponds to the hours of the day that the shift covers. For each type of shift, patients arrival patterns are different.

Let M be an upper bound value on the workload.When there are too many patients waiting at the end of the shift, they are redirected to another ED or leave without being seen. The upper bound can be arbitrarily large and is also used to limit the state space of our MDP.

6.4.1 Decision Variables

Assigning overtime hours to physicians is possible if they are working on the current shift, and if their overtime capacity $h_{i,t}$ is not nil. Let $x_i(s_t) \in \{0, 1, .., h_{i,t}\}$ be a decision variable that indicates the number of overtime hours to assign to physician i in state s_t.

Variable	Description	Domain
i	Index of the emergency physician	[1..I]
t	Current Period	[1..T]
W_t	Workload in excess	[1..M]
$h_{i,t}$	Remaining overtime hours per resource	$[1..H]^I$

TABLE 6.1 – Problem variables

In order to determine which physicians are working, the physicians' regular schedule should be available at the time of assigning overtime hours. If physicians are assigned overtime hours, he/she will start his/her overtime hours at the end of the shift. In order to simplify notations in later equations, we also define the vector $\mathbf{x}$ as a combination of decision variables x_i in state s_t. Let $\mathcal{A}(s_t)$ be the set of possible combinations in state s_t.

6.4.2 Marginal Cost

Given a state s_t, we call marginal cost the desirability to be in that state. Favorable states are ones where the workload is adequately absorbed. Let $C(.)$ be the marginal cost function defined as follows.

$$C(s_t, \mathbf{x}) = (W_t - \sum_i x_i)^+ \quad \forall\, \mathbf{x} \in \mathcal{A}(s_t) \tag{6.1}$$

Where $(.)^+$ is the positive part of the function. We assume that the demand in workload is expressed in the same unit as overtime hours.

6.4.3 Value Function

Our goal is to find the optimal overtime assignment policy that minimizes the marginal cost plus the expected cost of the decisions in the remaining horizon. Let $\mathbb{P}(s_t, \mathbf{x}, s'_{t+1})$ denote the transition probabilities of going from state s_t to state s'_{t+1} given the decisions $\mathbf{x}$. Once the decision is set, the transition probabilities are obtained from the distribution function of the excess workload in each work shift. The value function to the problem can be expressed as follows.

$$V(s_t) = \min_{\mathbf{x} \in \mathcal{A}(s_t)} C(s_t, \mathbf{x}) + \mathbb{E}[V(f(s_t, \mathbf{x}))] \tag{6.2}$$

where the function $f(s_t, \mathbf{x})$ generates the possible next state s'_{t+1} given the current state s_t and the current decision $\mathbf{x}$. In Algorithm 2 we describe the dynamic programming algorithm used to calculate the optimal policy for our assignment problem. In our model, the support of the excess workload is finite and hence we replace the expected value with a sum over all possible next states s'_{t+1}.

Algorithm 2 Calculate the optimal overtime allocation policy

```
function OPTIMAL_POLICY(schedule)
    for all t : T → 1 do
        for all s_t do
            for all x ∈ A(s_t) do
                G(s_t, x) = C(s_t, x) + Σ_{s'_{t+1} ∈ f(s_t, x)} p_{s'_{t+1}} V(s'_{t+1})
            end for
            x* = min_{x ∈ A(s_t)} G(s_t, x)
            V(s_t) = G(s_t, x*)
        end for
    end for
end function
```

As can be expected, solving the MDP directly is difficult because of the state space explosion. The dimension of the problem increases exponentially with the number of physicians and the number of overtime hours allowed in the planning horizon. In the next section, we describe alternative formulations that allow us to calculate approximate assignment policies that are then compared to the solutions obtained using the MDP.

6.5 Policy Approximations

In this section, we propose three approximation models to the MDP formulation presented above. First in Section 6.5.1, we calculate an upper bound on the value of the function V using a policy π_{upper}. Then, in Section 6.5.2 we calculate a lower bound by making the hypothesis that the future can be know in advance and we determine a corresponding policy π_{lower}. Finally, in Section 6.5.3 we use the upper bound policy in a dynamic setting to have an approximation scheme which gives us results at least as good as the upper bound approximation and that can be used in a real world-setting.

We use the Sample Average Approximation (SAA) method defined in previous chapters (c.f Chapter 4) to solve the lower and upper bound LP described in the following. Since the SAA method is itself an approximation scheme, we are not guaranteed that the approximations are actually lower bounds and upper bounds. However, we note in the experimental design section the confidence interval around these two values.

6.5.1 Upper Bound

Let $X_{t,i}$ be decision variables that assign to physician i at period t, a given number of overtime hours. For the first period, at the moment the decision is taken, the excess workload is well defined. For the remaining periods of the horizon, the value of the excess workload can only be approximated. The objective function of the LP can be expressed as follows.

$$V(s_t) \leq \min_{\mathbf{X} \in \mathcal{A}(s_t)} (W_1 - \sum_i X_{1,i})^+ + \mathbb{E}[\sum_{t>1} (W_t - \sum_i X_{t,i})] \tag{6.3a}$$

Subject to :

$$\sum_t X_{t,i} \leq h_{i,t} \ \forall i \tag{6.3b}$$

$$X_{t,i} = 0 \ \forall t \notin P_i \ \forall i \tag{6.3c}$$

$$X_{t,i} \geq 0 \ \forall t \ \forall i \tag{6.3d}$$

$$X_{t,i} \in \mathbb{N} \ \forall t \ \forall i \tag{6.3e}$$

Constraints (6.3b) impose an upper bound on the number of overtime hours assigned to employee i over the planning horizon. Constraints (6.3c) make sure that only employees working on the current shift can be assigned overtime hours. The variable P_i represent symbolically the schedule of each physician i. Finally, constraints (6.3d) are positivity constraints. Model (6.3) can be solved using a Sample Average Approximation method using sample realizations $\xi \in \Xi$ of excess workload. The model can then be written as :

$$\min_{\mathbf{X} \in \mathcal{A}(s_t)} (W_1 - \sum_i X_{1,i})^+ + \frac{1}{|\Xi|} \sum_{\xi \in \Xi} \sum_{t>1} (W_t(\xi) - \sum_i X_{t,i})^+ \tag{6.4}$$

Equation (6.4) can be linearized as follows :

$$\min_{\mathbf{x} \in \mathcal{A}(s_t)} Y_1 + \frac{1}{|\Xi|} \sum_{\xi} \sum_{t>1} Y_t(\xi) \tag{6.5a}$$

$$Y_1 \geq W_1 - \sum_i X_{1,i} \tag{6.5b}$$

$$Y_t(\xi) \geq W_t(\xi) - \sum_i X_{t,i}, \ \forall \ t \tag{6.5c}$$

$$Y_1, Y_t(\xi) \geq 0, \ \forall \ t \tag{6.5d}$$

$$Y_1, Y_t(\xi) \in \mathbb{N}, \ \forall \ t \tag{6.5e}$$

Subject to constraints (6.3b)-(6.3d).

6.5.2 Lower Bound

Similarly, we can calculate a lower bound on our value function V and an over-optimized policy π_{lower}. The lower bound can be obtained by using a decision variable specific for each realization ξ. Having different decision variables $X_{t,i}(\xi)$ for each realization ξ is tantamount to assuming that decisions are taken with full knowledge of future events. Of course, this assumption is unrealistic given the uncertain nature of the arrival patterns of patients. Nonetheless, using this assumption we can have an indication on the quality of the solution obtained using the MDP. The lower bound formulation can be expressed as follows :

$$V(s_t) \geq \min_{\mathbf{X} \in \mathcal{A}(s_t)} (W_1 - \sum_i X_{1,i})^+ + \frac{1}{|\Xi|} \sum_{\xi \in \Xi} \sum_{t>1} (W_t(\xi) - \sum_i X_{t,i}(\xi)) \tag{6.6}$$

Under constraints (6.3b)-(6.3d) but with variables $X_{t,i}(\xi)$ instead of $X_{t,i}$. Similarly to Section 6.5.1, we linearize equation (6.6) to solve the LP.

6.5.3 Dynamic Approximation Scheme

The policy π_{upper} obtained with the upper bound strategy can be expected to behave poorly in many scenarios. Indeed the variables $X_{t,i}$ are assigned values so that the solution is minimized in an average sense. However, for borderline cases, the solution can be far from desirable. What is more, in an ED setting, there is no need for a priori assignment of overtime hours to be calculated. We can use the upper bound approximation once the excess demand is known at the end of each shift. In Figure 6.2 we describe a dynamic setting where the decision to assign overtime hours is calculated once for every period.

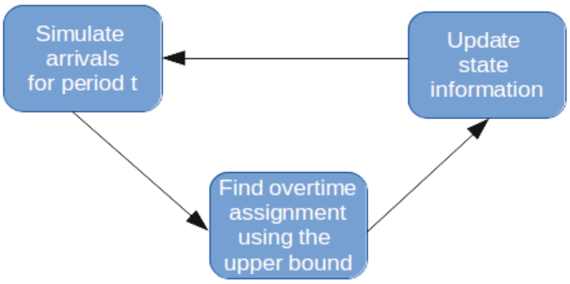


FIGURE 6.2 – A framework for using the lower bound approximation in a dynamic setting

In our framework, we simulate patient arrivals for period t, then based on the excess workload remaining at the end of this period W_t, we calculate the optimal assignment of overtime hours to physicians $\mathbf{X}$. After the decision is made, the state information is updated and the simulation continues to calculate the excess demand for the next period $t+1$. This goes on until we reach the end for the planning horizon. The solutions obtained in this dynamic setting are at least as good than those using the upper bound approximation.

6.6 Experimental Results

To evaluate the performance of the aforementioned policies, we generate test scenarios and compare the values of the objective functions under different work settings. The test scenarios consist of schedules for physicians and workload realizations. The schedules determine the shifts physicians are working on across the week while the workload realizations correspond to possible workload scenarios that physicians can encounter in the ED. Patients' arrival rates from the ED of Lille are used to simulate the excess workload across a week. The average hourly arrival rates λ_h are defined differently for week days and week-ends.

In Section 6.6.2 we compare the solution obtained with the MDP and those obtained with the LP approximations. We also compare the value obtained with the MDP to the value obtained using a simple heuristic policy for the distribution of overtime hours that we tailor to our objective function.

Because of the curse of dimensionality, running the MDP is only possible on small instances. To illustrate the quick increase in problem size, we note in Table 6.2 the number of state variables in the MDP as a function of the number of overtime hours allowed by physician h_i.

h_i	$\|s_t\|$
2	13,440
3	153,090
4	860,160
5	3,281,250
6	9,797,760

TABLE 6.2 – The number of MDP states as a function of the overtime hours allowed by physician with 6 physicians

Solutions for longer planning horizons can give us insights on how to manage resources appropriately during epidemic periods. In Section 6.6.3 we assume that the resources are to be distributed on longer planning horizon and use the approximation scheme of Section 6.5.3 to test our solution policies. All the experiments were run on an Intel(R) Xeon(R) 4 cores CPU E5520 @ 2.27GHz with 8 MB of cache memory and 8 GB of RAM.

6.6.1 Model Parameters

The first workload scenarios are generated from the arrival rates in our partner's hospital averaged across two years. Tables 6.3-6.5 shows these hourly averages for week days and week-ends. After patients enter the ED, they are assessed by a physician. We use exponential service times for the assessment phase. The parameter of the exponential distribution is set to $\mu = \frac{1}{30}$ based on discussions with health care practitioners and our own observations. For the later scenarios, we let α be the percentage increase or decrease of the arrival rates. We vary α by 10% and 20% and simulate the ensuing excess workload.

As we have said previously, we use a limit on the excess workload that is to be performed by physicians. We set this limit M to 10 for the following tests. Taking into account this upper limit, we can calculate the actual increase in ensuing workload. The 10% increase in the arrival rate corresponds to a 48% increase in the excess workload and the 20% increase to a 100% increase in the excess workload on average. Similarly, the 10% decrease in arrival rates corresponds to a 40% decrease in excess workload and the 20% decrease to a 69% decrease in excess workload on average.

Days	1	2	3	4	5	6	7	8
Week	1.62	3.50	3.85	4.09	3.76	3.97	4.18	3.78
Week-end	1.30	2.61	3.90	4.52	4.39	4.16	6	8

TABLE 6.3 – Morning shift (8-16h) arrival averages

We consider an ED with 6 physicians that can be assigned overtime hours. Each physician can work 2 overtime hours per week. Overtime hours are either assigned on the same shift or separately across the week. To ensure that our results are not schedule-dependent we run our experiments on 3 types of physicians' schedules. In this first type, that we refer to as "fixed" schedule, there is the same number of physicians on each of the three

Days	1	2	3	4	5	6	7	8
Week	3.55	4.01	4.90	5.29	4.88	3.68	2.63	1.85
Week-end	4.31	4.45	4.74	4.67	4.30	3.90	3.14	2.03

Table 6.4 – Afternoon shift (16-24h) arrival averages

Days	1	2	3	4	5	6	7	8
Week	1.27	0.94	0.72	0.57	0.52	0.36	0.38	0.60
Week-end	1.69	1.30	0.92	0.52	0.48	0.66	6	8

Table 6.5 – Night shift (24h-8h) arrival averages

shifts of the day across the week. In other words, 2 physicians work morning shifts (M : 8h-16h), 2 others work afternoon shifts (A : 16h-24h) and the last 2 work night shifts (N : 24h-8h). In the second type, "adjusted" schedule, the number of physicians per shift is adjusted according to the expected demand. We then have 2 physicians working morning shifts, 3 working afternoon shifts and 1 working on a night shift. Finally, we a third "manual" schedule is constructed with a more disorderly work pattern. We illustrate this last schedule in Table 6.6. The important thing to note is that in this third schedule, physicians 2 and 3 have individual work patterns that no other physician has.

6.6.2 Policy Approximation

In the first set of experiments, we compare for several test instances the value of the objective function obtained with the parameters detailed in Section 6.6.1. Tables 6.8, 6.7, and 6.9 note the results respectively for the "fixed" schedule, the "adjusted" schedule, and the "manual" schedule. The first column α is the percentage change in the patients' arrival rates compared to those found in the Lille's database. The second column "lower" is the value of the objective function for the lower bound LP approximation. The column "upper" denotes the value of the objective function obtained with the upper bound approximation. The "dynapprox" column is the value of the objective function with the dynamic approximation scheme of Section 6.5.3. Finally, the column "heuristic" is the value of the objective function obtained with a heuristic strategy tailored to our objective function. One intuitive strategy that comes to mind when assigning overtime hours is to assign hours to physicians that have the most leftover hours first and continue in that

Physician	M	A	N	M	A	N	M	A	N	M	A	N	M	A	N	M	A	N	M	A	N
0	1	0	0	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0	0	0	1
1	1	0	0	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0	0	0	1
2	0	1	0	0	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0	0	0
3	0	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1
4	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0	0
5	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0	0

Table 6.6 – Physicians schedule with preferences considerations

order until either no more physicians can be assigned hours or all the excess demand has been covered. We test this "heuristic" strategy on the three types of schedules and compare the results to the value obtained with the optimal policy. After each column, the running time in the format HH :MM :SS is noted. In the tables, the values after the objective function define the 95% confidence interval. We set the number or realizations to $|\Xi| = 500$ and calculate the confidence interval as follows :

$$\bar{V} \pm t_{0.95} * \frac{\sigma}{\sqrt{|\Xi|}} \tag{6.7}$$

Where $\bar{V}$ is the mean and σ is the standard variation of our sample objective function values.

α	lower	time	upper	time	obj	time	dynapprox	time	heuristic	time
20	17.11±0.53	00 :00 :08	19.96±0.52	00 :00 :12	17.68	00 :00 :14	17.47±0.22	00 :12 :05	17.68	00 :00 :08
10	11.70±0.70	00 :00 :05	13.68±0.77	00 :00 :01	11.88	00 :00 :14	12.03±0.16	00 :11 :47	11.88	00 :00 :08
0	6.46±0.36	00 :00 :05	9.72±0.38	00 :00 :04	6.42	00 :00 :14	6.60±0.1	00 :11 :59	6.42	00 :00 :07
-10	2.56±0.24	00 :00 :03	5.28±0.28	00 :00 :03	2.79	00 :00 :14	2.87±0.05	00 :11 :56	2.79	00 :00 :07
-20	0.95±0.13	00 :00 :02	2.86±0.21	00 :00 :01	0.93	00 :00 :14	1.54±0.02	00 :11 :58	0.93	00 :00 :08

TABLE 6.7 – Comparison between the optimal policy and the LP approximations for the “adjusted” schedule

α	lower	time	upper	time	obj	time	dynapprox	time	heuristic	time
20	19.09±0.53	00 :00 :06	21.36±0.53	00 :00 :06	19.66	00 :00 :17	19.45±0.22	00 :12 :02	19.66	00 :00 :06
10	13.8±0.45	00 :00 :04	16.12±0.45	00 :00 :04	13.77	00 :00 :17	13.96±0.16	00 :12 :02	13.77	00 :00 :06
0	8.11±0.38	00 :00 :04	10.58±0.39	00 :00 :03	8.04	00 :00 :17	8.28±0.11	00 :12 :02	8.04	00 :00 :06
-10	3.64±0.27	00 :00 :02	5.90±0.30	00 :00 :02	3.93	00 :00 :17	4.05±0.06	00 :11 :57	3.93	00 :00 :06
-20	1.51±0.17	00 :00 :01	3.23±0.22	00 :00 :01	1.51	00 :00 :17	2.41±0.03	00 :12 :10	1.51	00 :00 :06

TABLE 6.8 – Comparison between the optimal policy and the LP approximations for the “fixed” schedule

α	lower	time	upper	time	obj	time	dynapprox	time	heuristic	time
20	16.34±0.5	00 :00 :04	19.02±0.5	00 :00 :03	17.22	00 :00 :14	17.14 ± 0.2	00 :13 :35	17.43	00 :00 :06
10	11.84±0.42	00 :00 :01	14.10±0.43	00 :00 :01	11.84	00 :00 :14	12.16 ± 0.14	00 :13 :05	12.05	00 :00 :06
0	7.04±0.33	00 :00 :03	9.21±0.37	00 :00 :03	6.99	00 :00 :14	7.33 ± 0.09	00 :13 :44	7.15	00 :00 :06
-10	3.49±0.24	00 :00 :02	5.02±0.28	00 :00 :01	3.64	00 :00 :14	3.92 ± 0.05	00 :13 :20	3.74	00 :00 :06
-20	1.73±0.16	00 :00 :02	2.73±0.20	00 :00 :01	1.71	00 :00 :14	2.8 ± 0.03	00 :13 :54	1.75	00 :00 :06

TABLE 6.9 – Comparison between the optimal policy and the LP approximations for the “manual” schedule

In all the test cases, the dynamic approximation scheme improves significantly on the value of upper bound approximation. Furthermore, on average there is a 3.6 % difference in the objective function of the optimal policy and the dynamic approximation scheme for the arrival rates of our partner ED. The average is taken over the values obtained with the three different schedules. These results confirm the fact that the dynamic approximation scheme can find good overtime assignment strategies in real-time.

With the heuristic strategy and the schedules "adjusted" and "fixed"the optimal solutions are found. It is not hard to see why this is the case. Indeed, our objective function only penalizes the shortage in capacity regardless of the day, the shift,or physician that is assigned overtime hours. In the two schedules "adjusted" and "imposed", physicians are interchangeable so the optimal policy actually follows the strategy described in this Section.

However, when the schedule is not as regular, the optimal policy beats our heuristic strategy. Indeed we can see in Table 6.9 that we have better solution in all the test cases. For our objective function, this simple heuristic approach to assigning overtime hours beats the time consuming dynamic LP approximation we described in Section 6.5.3. However, more complex objective functions and different physician schedules can significantly complicate the task for finding a good heuristic strategy. Next we try to apply the dynamic approximation scheme to more realistic test cases to see if it can be applied in practice.

6.6.3 Long Planning Horizon

As we have said, the optimal policy can not easily be applied to large problem instances because of the state space explosion. In this section we solve larger problem instances using our dynamic approximation scheme described in Section 6.5.3 and compare the results with the lower bound approximation described in Section 6.5.2. Indeed from the results of Section 6.6.2 we notice that the lower bound is very close to the optimal value and as such, serves as a good indicator for the quality of our dynamic approximation scheme. The instances we ran in this Section span a 4 weeks horizon and each physician has 8 overtime hours to be assigned during this planning period.

α	lower	time	dynapprox	time
20	69.05±1.17	00 :00 :39	69.34±0.18	03 :19 :07
10	43.88±1.04	00 :00 :29	44.17 ±0.18	03 :18 :26
0	23.28±0.86	00 :00 :20	23.44±0.17	03 :17 :10
-10	6.47±0.53	00 :00 :16	6.60±0.16	03 :00 :26
-20	0.68±0.16	00 :00 :10	0.77±0.17	03 :13 :21

TABLE 6.10 – Results obtained with the approximation scheme for a 4 weeks planning horizon

As we see from Table 6.10, in all the test cases, the dynamic approximation scheme proposes solutions that are very close to the lower bound. The running time of the approximation scheme depends to a great extent on the number of periods in our planning horizon. Indeed, at each iteration, the approximation scheme solves a linear program. On the other hand, contrary to the MDP, up to a certain extent the number of physicians and the number of overtime hours available per physician do not have a big impact on the running time. Indeed, these are parameters in the LP model and do not affect the total number of iterations.

Furthermore, the times noted in Table 6.10 for the dynamic approximation scheme are not representative of the actual time it would take for decision makers to use the program on a daily basis. Indeed, the 3 hours running time includes the simulation of an entire epidemic horizon. While on a daily basis, decision makers would only need to run the program for the current period to obtain quality solutions. In the last Section, we propose a heuristic strategy to assign overtime hours that is specifically tailored for the objective function used. If the conditions are appropriate, this heuristic strategy can be used as a efficient way to distribute overtime hours without requiring computational work.

6.7 Conclusion

In this chapter we have proposed a framework for defining and testing different overtime assignment strategies in the context of EDs. We first modeled the problem as a MDP with the aim of finding an optimal control policy. Then we have developed three approximation schemes that can be used in real time settings with more realistic problem instances.

Our results show that the dynamic approximation scheme finds good quality solutions even for large time horizons. Indeed, in the simulations we ran, there is a 3% average difference between the values obtained with dynamic approximation scheme and those obtained with the lower bound scheme. This indicates that the dynamic approximation scheme can be used in real world-settings efficiently distributing overtime hours regardless of the epidemic scenario and of the physician's schedule used in the ED.

Finally, we have tested a naive heuristic strategy that assigns overtime hours in priority to the physician with the most left over hours available. This heuristic is tailored for our problem as the objective function only penalizes the shortage of capacity. As such, for two types of schedules, this heuristic strategy is optimal. However, if the physicians' work hours are not interchangeable, this strategy is not optimal. Furthermore, the strategy would be sub-optimal if the objective function includes other criteria such as physicians' preferences or time considerations.

During epidemic seasons, when resources are scarce and capacity is critical, we think that using an efficient overtime assignment strategy can help reduce overcrowding in EDs. What's more, from our results we have shown that a simple LP programs that can be used on a daily basis, can significantly improve the assignment of overtime hours. We think that the next step to validate these results is to incorporate the solutions found in this chapter in a simulation framework for the entire ED. This would allow to study the interaction between the physicians using overtime hours and other health care professionals that have different schedules. Furthermore, it would allow us to test the robustness of some of the simplifying hypotheses we made in this chapter such as the independence of the shifts.

Chapitre 7

Synthèse et Conclusions

Dans cette thèse, nous avons étudié le problème des tensions aux urgences pendant les périodes épidémiques en utilisant principalement les méthodes de recherche opérationnelle. Grâce à un travail de terrain et une revue de la littérature, nous avons d'abord identifié la problématique de la gestion des ressources humaines comme étant un goulot d'étranglement majeur. Cette problématique a été traitée sous différents points de vue partant d'une caractérisation stratégique au Chapitre 3 jusqu'à la définition d'une politique opérationnelle au Chapitre 6.

À la fin de chaque chapitre, une conclusion est présentée avec une synthèse des principaux résultats trouvés. Nous faisons ici un retour sur une vision globale des résultats en prenant du recul par rapport à chaque problème traité. Nous évoquons également quelques points que nous n'avons pas eu le temps d'aborder dans cette thèse mais que nous estimons néanmoins être des suites pertinentes au travail présenté.

7.1 Des modèles non-exhaustifs qui doivent capter l'essentiel du système à optimiser

Comme nous l'avons vu dans le Chapitre 2, la problématique des tensions aux urgences ne se limite pas à celle de la gestion des ressources humaines. De plus, dans un système aussi complexe que celui des urgences, il est impossible de prétendre étudier un problème particulier en faisant abstraction du fonctionnement global du processus de soins. De ce fait, tout un travail en amont de chaque chapitre a été nécessaire afin de proposer des hypothèses qui sont en accord avec le fonctionnement global des hôpitaux. Par exemple, nous avons interrogé plusieurs centres hospitaliers sur la faisabilité des postes en astreinte dans les services d'urgences avant d'étudier le problème mathématiquement. En effet, actuellement en France, les postes en astreinte sont uniquement utilisés par des médecins spécialistes aux urgences et à la lecture des textes de lois, il est impossible de déterminer la raison de cette réticence de la part du personnel urgentiste. De même, au Chapitre 4 dans la détermination des postes de médecins, toutes les tâches annexes ont été prises en compte après un entretien avec le personnel urgentiste. Ainsi, des durées aléatoires pour les examens complémentaires ont été considérées même si les médecins n'effectuent pas les examens eux-mêmes. Néanmoins, il est certain que des abstractions doivent être faites pour rendre le problème traitable, nos hypothèses simplificatrices s'appuient donc sur des résultats de la littérature et une validation sur des cas d'études réels et des scénarios réels et réalistes.

7.2 Synthèse de la Gestion des Ressources Humaines

D'un point de vue stratégique, la modélisation proposée au Chapitre 3 permet de répondre au problème de répartition de ressources limitées sur un horizon épidémique. Cette politique pourrait ainsi servir d'indicateur du niveau de tensions à prévoir sur la période épidémique étant donné un niveau de ressources prédéfini. En effet, quand il s'agit de ressources humaines, il est difficile pour les hôpitaux d'augmenter leurs capacités rapidement. Il est donc nécessaire de pouvoir évaluer et de quantifier a priori un manque de ressources potentielles afin de palier à ce manque avant que l'épidémie ne commence.

Sur un plan plus tactique, nous avons optimisé dans le Chapitre 4 la distribution des postes de travail des médecins et des infirmiers aux urgences. Les postes de travail définissent les heures de travail et de repos au cours de la journée. Dans notre cas d'étude, une disparité existe entre le nombre de ressources en poste et le nombre moyen de patients présents aux urgences. En effet, il y a une baisse de l'effectif médical aux heures de pointe. L'optimisation des postes de travail a permis donc de mieux les répartir dans le but de réduire le temps d'attente des patients. Contrairement à la majorité des études sur le sujet, l'optimisation des postes se fait conjointement pour les médecins et les infirmiers. De ce fait, la coordination de ces deux ressources est prise en compte. De plus, l'évolution du nombre de patients au cours de la journée est modélisée de façon rigoureuse et n'utilise pas d'approximations numériques. Finalement, les hypothèses effectuées ont été vérifiées grâce à un modèle de simulation à événements discrets.

Cependant, l'optimisation des postes de travail journaliers est une décision qui engage un changement important dans le fonctionnement des urgences. Il n'est pas pratique de faire des changements réguliers sur les postes de travail. Toutefois, il est nécessaire d'avoir des mécanismes tels que les astreintes ou les heures supplémentaires pour faire face aux imprévus au jour le jour. L'utilisation des astreintes a été évaluée au Chapitre 5. Le contexte épidémique nous a servi de nouveau comme cas d'étude et nous avons quantifié grâce à un plan d'expérience les points forts et les points faibles liés aux astreintes. Les résultats obtenus suggèrent que les astreintes peuvent servir à garantir un meilleur niveau de couverture de la demande tout en gardant le nombre d'heures de travail effectives des ressources constant. De plus, nos résultats montrent que l'utilisation des heures en astreintes est particulièrement bénéfique pendant les périodes où la charge de travail est très variable. Ce résultat confirme l'utilité des astreintes pendant les périodes épidémiques où la charge de travail d'une journée à une autre est fortement variable. Toujours en vue de proposer des solutions opérationnelles au problème des tensions, nous avons étudié dans le Chapitre 6 l'affectation dynamique des heures supplémentaires aux médecins du service. L'affectation des heures supplémentaires se fait au jour le jour mais des contraintes du code du travail imposent une limite sur le nombre d'heures supplémentaires qu'un médecin peut faire par semaine. Les politiques d'affectation sont importantes dans la mesure où elles prennent en compte de potentiels besoins futurs.

7.3 Évaluation par Scénarios

Dans la mesure du possible, nous avons essayé tout au long de la thèse d'évaluer nos résultats en vue de définir au mieux les situations où ils sont applicables ainsi que de quantifier les impacts sur différents indicateurs de performance. Pour ce faire, nous avons utilisé des données réelles d'un service d'urgence à Lille. Par rapport à la littérature existante, notre travail est particulier dans la mesure où nous considérons le cas des épidémies saisonnières. Il est vrai que les résultats obtenus pourraient être utilisés dans d'autres situations plus générales mais les épidémies saisonnières ont quand même quelques caractéristiques propres. Pour commencer, nous savons que le contexte de l'épidémie est limité dans le temps. La planification des ressources se fait donc sur un horizon fini. Les stratégies proposées comme l'utilisation des postes en astreinte et l'utilisation des heures supplémentaires sont pertinentes dans ce contexte particulier puisque ce sont des recours d'urgence qui ne peuvent pas être utilisés de façon régulière au risque de détériorer la qualité et les conditions de travail des ressources humaines. De plus, la demande pendant les épidémies saisonnières a un profil bien défini : elle augmente petit à petit pendant les premières semaines pour arriver à un pic et diminue ensuite progressivement. Cette particularité des saisons épidémiques a fait que des scénarios types de demande ont pu être générés et utilisés dans nos plans d'expérience. Ces scénarios de demande ont été générés grâce à des modèles épidémiques. Cela nous a également donné l'avantage de pouvoir calibrer facilement nos modèles pour simuler des cas d'épidémie plus ou moins violents et relier la demande à des caractéristiques épidémiques tel que le taux d'infection. Finalement, à la différence des pandémies, les épidémies saisonnières n'affectent pas systématiquement la majorité de la population. Cette particularité fait que des plans d'urgence à grande échelle comme le plan blanc ne sont pas adéquats. De plus, la prise en compte de l'impact de l'épidémie sur le personnel médical n'est pas essentielle.

7.4 Perspectives : Vers des Résultats Pratiques

Malgré toute la complexité pouvant être introduite dans un modèle mathématique et toute la rigueur prise à l'évaluer, il existe toujours des disparités entre la théorie et la pratique. Ces disparités sont inhérentes au processus de modélisation comme nous l'avons expliqué au cours du Chapitre 1. Le but de cette thèse n'est donc pas de présenter des solutions finales à toutes les problématiques de la gestion des ressources humaines. Notre but est aussi de soulever des questions pertinentes et de formaliser les problèmes de façon cohérente et rigoureuse. Nous pensons que la modélisation mathématique se prête bien à cet exercice puisqu'elle permet de poser des hypothèses explicitement et de circonscrire le problème rigoureusement. Par exemple, dans tous nos modèles, les lois d'arrivée des patients suivent une distribution aléatoire qui peut être calibrée selon le contexte étudié. En particulier dans le premier chapitre, où la distribution même des variables aléatoires n'est pas fixée à l'avance. Les résultats obtenus sont ainsi validés quelle que soit cette distribution. De même les modèles épidémiques ont été utilisés dans nos plans d'expériences parce que c'était un outil naturel dans notre contexte d'épidémies saisonnières. Cependant, d'autres méthodes plus ou moins complexes de génération de demande peuvent être utilisées et potentiellement améliorer la qualité de la solution si la prévision est meilleure.

Mis à part au chapitre 4, nous n'avons pas pu vérifier nos résultats avec des modèles de simulation. Cependant, nous pensons qu'avant la mise en place de nouvelles stratégie aux urgences, il est important de passer par une étape de simulation. La simulation permet de vérifier plus en détail la validité des hypothèses et de mettre en lumières les points faibles de la modélisation mathématique. Particulièrement, au chapitre 6 les modèles proposés font abstraction de l'interaction entre les différentes ressources. Une étude de simulation prenant en compte ces interactions permettrait de confirmer nos résultats.

Bibliographie

[Afilalo et al., 2004] Afilalo, J., Marinovich, A., Afilalo, M., Colacone, A., Leger, R., Unger, B., and Giguere, C. (2004). Nonurgent emergency department patient characteristics and barriers to primary care. *Academic Emergency Medicine*, 11(12) :1302–1310.

[Akkerman and Knip, 2004] Akkerman, R. and Knip, M. (2004). Reallocation of beds to reduce waiting time for cardiac surgery. *Health care management science*, 7(2) :119–26.

[Allen and Burgin, 2000] Allen, L. J. and Burgin, A. M. (2000). Comparison of deterministic and stochastic SIS and SIR models in discrete time. *Mathematical biosciences*, 163(1) :1–33.

[Andersen, 1995] Andersen, R. M. (1995). Revisiting the behavioral model and access to medical care : Does it matter ? *Journal of Health and Social Behavior*, 36(1) :1–10.

[Arora et al., 2010] Arora, H., Raghu, T., and Vinze, A. (2010). Resource allocation for demand surge mitigation during disaster response. *Decision Support Systems*, 50(1) :304–315.

[Bae and Fabry, 2014] Bae, S.-H. and Fabry, D. (2014). Assessing the relationships between nurse work hours/overtime and nurse and patient outcomes : Systematic literature review. *Nursing outlook*, 62(2) :138–156.

[Beaulieu et al., 2000] Beaulieu, H., Ferland, J. a., Gendron, B., and Michelon, P. (2000). A mathematical programming approach for scheduling physicians in the emergency room. *Health care management science*, 3(3) :193–200.

[Bekker and De Bruin, 2010] Bekker, R. and De Bruin, A. M. (2010). Time-dependent analysis for refused admissions in clinical wards. *Annals of Operations Research*, 178(1) :45–65.

[Bester et al., 2007] Bester, M., Nieuwoudt, I., and Van Vuuren, J. H. (2007). Finding good nurse duty schedules : a case study. *Journal of Scheduling*, 10(6) :387–405.

[Bienstock and Zenteno, 2012] Bienstock, D. and Zenteno, A. C. (2012). Models for managing the impact of an epidemic. Technical report, Working paper, Department of Industrial Engineering and Operations Research, Columbia University. Version Sat.

[Brandeau et al., 2004] Brandeau, M. L., Sainfort, F., and Pierskalla, W. P. (2004). *Operations research and health care : a handbook of methods and applications*, volume 70. Springer.

[Britton, 2010] Britton, T. (2010). Stochastic epidemic models : a survey. *Mathematical biosciences*, 225(1) :24–35.

[Bruin et al., 2009] Bruin, a. M., Bekker, R., Zanten, L., and Koole, G. M. (2009). Dimensioning hospital wards using the Erlang loss model. *Annals of Operations Research*, 178(1) :23–43.

[Brunner et al., 2009] Brunner, J. O., Bard, J. F., and Kolisch, R. (2009). Flexible shift scheduling of physicians. *Health Care Management Science*, 12(3) :285–305.

[Brunner and Edenharter, 2011] Brunner, J. O. and Edenharter, G. M. (2011). Long term staff scheduling of physicians with different experience levels in hospitals using column generation. *Health care management science*, 14(2) :189–202.

[Burke et al., 2004] Burke, E. K., De Causmaecker, P., Berghe, G. V., and Van Landeghem, H. (2004). The state of the art of nurse rostering. *Journal of scheduling*, 7(6) :441–499.

[Carter and Lapierre, 2001] Carter, M. W. and Lapierre, S. D. (2001). Scheduling emergency room physicians. *Health care management science*, 4(4) :347–60.

[Chockalingam et al., 2010] Chockalingam, A., Jayakumar, K., and Lawley, M. A. (2010). A stochastic control approach to avoiding emergency department overcrowding. In *Proceedings of the 2010 Winter Simulation Conference*, pages 2399–2411.

[Chowell et al., 2008] Chowell, G. M. A. M., Miller, M. A., and Viboud, C. (2008). Seasonal influenza in the united states, france, and australia : transmission and prospects for control. *Epidemiology and infection*, 136(06) :852–864.

[Cochran and Roche, 2009] Cochran, J. K. and Roche, K. T. (2009). A multi-class queuing network analysis methodology for improving hospital emergency department performance. *Computers & Operations Research*, 36(5) :1497–1512.

[Daskin and Dean, 2004] Daskin, M. S. and Dean, L. K. (2004). Location of health care facilities. In Brandeau, M., Sainfort, F., and Pierskalla, W. P., editors, *Operations Research and Health Care*, volume 70 of *International Series in Operations Research & Management Science*, pages 43–76. Springer US.

[des Comptes, 2014] des Comptes, C. (2014). Les urgences hospitalières : une fréquentation croissante, une articulation avec la médecine de ville à repenser. Online ; https ://www.ccomptes.fr.

[Edie, 1954] Edie, L. C. (1954). Traffic delays at toll booths. *Operations Research*, 2(2) :107–138.

[Ernst et al., 2004] Ernst, A., Jiang, H., Krishnamoorthy, M., and Sier, D. (2004). Staff scheduling and rostering : A review of applications, methods and models. *European Journal of Operational Research*, 153(1) :3–27.

[Flahault et al., 2006] Flahault, A., Vergu, E., Coudeville, L., and Grais, R. F. (2006). Strategies for containing a global influenza pandemic. *Vaccine*, 24(44) :6751–6755.

[Garg et al., 2012] Garg, L., McClean, S. I., Barton, M., Meenan, B. J., and Fullerton, K. (2012). Intelligent Patient Management and Resource Planning for Complex, Heterogeneous, and Stochastic Healthcare Systems. *IEEE Transactions on Systems, Man, and Cybernetics - Part A : Systems and Humans*, 42(6) :1332–1345.

[Gonnet and Lucas, 2003] Gonnet, F. and Lucas, S. (2003). *L'hôpital en question(s) : un diagnostic pour améliorer les relations de travail.* Wolters Kluwer France.

[Green and Soares, 2006] Green, L. and Soares, J. a. (2006). Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine*, 13(1) :61–68.

[Green et al., 2001] Green, L. V., Kolesar, P. J., and Soares, J. a. (2001). Improving the Sipp Approach for Staffing Service Systems That Have Cyclic Demands. *Operations Research*, 49(4) :549–564.

[Guo and Harstall, 2006] Guo, B. and Harstall, C. (2006). Strategies to reduce emergency department overcrowding. Technical report, Alberta Heritage Foundation for Medical Research. Health Technology Assessment (HTA) Database.

[Gutjahr and Rauner, 2007] Gutjahr, W. J. and Rauner, M. S. (2007). An ACO algorithm for a dynamic regional nurse-scheduling problem in Austria. *Computers & Operations Research*, 34(3) :642–666.

[Hajipour et al., 2014] Hajipour, V., Rahmati, S. H. A., Pasandideh, S. H. R., and Niaki, S. T. A. (2014). A multi-objective harmony search algorithm to optimize multi-server location-allocation problem in congested systems. *Computers & Industrial Engineering*, 72 :187–197.

[Harper, 2002] Harper, P. R. (2002). A framework for operational modelling of hospital resources. *Health care management science*, 5(3) :165–173.

[Hoot and Aronsky, 2008] Hoot, N. R. and Aronsky, D. (2008). Systematic review of emergency department crowding : causes, effects, and solutions. *Annals of emergency medicine*, 52(2) :126–136.

[Howard et al., 2005] Howard, M. S., Davis, B. A., Anderson, C., Cherry, D., Koller, P., and Shelton, D. (2005). Patients' perspective on choosing the emergency department for nonurgent medical care : a qualitative study exploring one reason for overcrowding. *Journal of Emergency Nursing*, 31(5) :429–435.

[Howell et al., 2008] Howell, E., Bessman, E., Kravet, S., Kolodner, K., Marshall, R., and Wright, S. (2008). Active bed management by hospitalists and emergency department throughput. *Annals of internal medicine*, 149(11) :804–810.

[Huang, 1995] Huang, X.-M. (1995). A planning model for requirement of emergency beds. *Mathematical Medicine and Biology*, 12(3-4) :345–353.

[Ingolfsson et al., 2002] Ingolfsson, A., Amanul Haque, M. D., and Umnikov, A. (2002). Accounting for time-varying queueing effects in workforce scheduling. *European Journal of Operational Research*, 139(3) :585–597.

[Insee, 2014] Insee (2014). Évolution des naissances, de la natalité et de la part des naissances hors mariage. Online ; http ://www.insee.fr.

[Izady and Worthington, 2012] Izady, N. and Worthington, D. (2012). Setting staffing requirements for time dependent queueing networks : The case of accident and emergency departments. *European Journal of Operational Research*, 219(3) :531–540.

[Jaumard et al., 1998] Jaumard, B., Semet, F., and Vovor, T. (1998). A generalized linear programming model for nurse scheduling. *European journal of operational research*, 107(1) :1–18.

[Jennings, 1996] Jennings, O. B. (1996). Server staffing to meet time-varying demand. *Management Science*, 42(10) :1383–1394.

[Jia et al., 2007] Jia, H., Ordóñez, F., and Dessouky, M. (2007). A modeling framework for facility location of medical services for large-scale emergencies. *IIE transactions*, 39(1) :41–55.

[Jones et al., 2009] Jones, S. S., Evans, R. S., Allen, T. L., Thomas, A., Haug, P. J., Welch, S. J., and Snow, G. L. (2009). A multivariate time series approach to modeling and forecasting demand in the emergency department. *Journal of biomedical informatics*, 42(1) :123–39.

[Jones and Thomas, 2008] Jones, S. S. and Thomas, A. (2008). Forecasting daily patient volumes in the emergency department. *Academic Emergency Medicine*, 15(2) :159–170.

[Jung et al., 2009] Jung, E., Iwami, S., Takeuchi, Y., and Jo, T.-C. (2009). Optimal control strategy for prevention of avian influenza pandemic. *Journal of theoretical biology*, 260(2) :220–229.

[Kleywegt et al., 2002] Kleywegt, A. J., Shapiro, A., and Homem-de Mello, T. (2002). The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2) :479–502.

[Lee et al., 2012] Lee, S., Golinski, M., and Chowell, G. (2012). Modeling optimal age-specific vaccination strategies against pandemic influenza. *Bulletin of mathematical biology*, 74(4) :958–980.

[L'Express, 2014] L'Express (2014). Comment l'économie française s'est transformée en trente ans ? Online ; http ://lexpansion.lexpress.fr.

[Marmor et al., 2012] Marmor, Y. N., Golany, B., Israelit, S., and Mandelbaum, A. (2012). Designing patient flow in emergency departments. *IIE Transactions on Healthcare Systems Engineering*, 2(4) :233–247.

[Maxwell et al., 2009] Maxwell, M. S., Henderson, S. G., and Topaloglu, H. (2009). Ambulance redeployment : An approximate dynamic programming approach. In *Winter Simulation Conference*, pages 1850–1860.

[Maxwell et al., 2010] Maxwell, M. S., Restrepo, M., Henderson, S. G., and Topaloglu, H. (2010). Approximate dynamic programming for ambulance redeployment. *INFORMS Journal on Computing*, 22(2) :266–281.

[Mazier, 2010] Mazier, A. (2010). *Optimisation Stochastique pour la gestion des lits d'hospitalisation sous incertitudes*. PhD thesis, Ecole Nationale Supérieure des Mines de Saint-Etienne.

[McCarthy et al., 2008] McCarthy, M. L., Zeger, S. L., Ding, R., Aronsky, D., Hoot, N. R., and Kelen, G. D. (2008). The challenge of predicting demand for emergency department services. *Academic Emergency Medicine*, 15(4) :337–46.

[McManus, 1977] McManus, M. I. (1977). Optimum use of overtime in post offices. *Computers & Operations Research*, 4(4) :271–278.

[Menchine and Baraff, 2008] Menchine, M. D. and Baraff, L. J. (2008). On-call specialists and higher level of care transfers in california emergency departments. *Academic Emergency Medicine*, 15(4) :329–336.

[Mestre et al., 2015] Mestre, A. M., Oliveira, M. D., and Barbosa-Póvoa, A. P. (2015). Location–allocation approaches for hospital network planning under uncertainty. *European Journal of Operational Research*, 240(3) :791–806.

[Miles, 2008] Miles, I. (2008). Patterns of innovation in service industries. *IBM Systems journal*, 47(1) :115–128.

[Miller et al., 1976] Miller, H. E., Pierskalla, W. P., and Rath, G. J. (1976). Nurse scheduling using mathematical programming. *Operations Research*, 24(5) :857–870.

[Mobasher et al., 2011] Mobasher, A., Lim, G., Bard, J. F., and Jordan, V. (2011). Daily scheduling of nurses in operating suites. *IIE Transactions on Healthcare Systems Engineering*, 1(4) :232–246.

[Omar et al., 2014] Omar, E.-R., Garaix, T., Augusto, V., and Xie, X. (2014). A stochastic optimization model for shift scheduling in emergency departments. *Health care management science*, pages 1–14.

[Parvin et al., 2012] Parvin, H., Goel, P., and Gautam, N. (2012). An analytic framework to develop policies for testing, prevention, and treatment of two-stage contagious diseases. *Annals of Operations Research*, pages 707–735.

[Patel and Derlet, 2006] Patel, P. B. and Derlet, R. W. (2006). Ambulance diversion reduction : the Sacramento solution. *The American journal of emergency medicine*, 24(2) :206–213.

[Patrick and Puterman, 2007] Patrick, J. and Puterman, M. L. (2007). Improving resource utilization for diagnostic services through flexible inpatient scheduling : A method for improving resource utilization. *Journal of the Operational Research Society*, 58(2) :235–245.

[Patrick et al., 2008] Patrick, J., Puterman, M. L., and Queyranne, M. (2008). Dynamic multipriority patient scheduling for a diagnostic resource. *Operations research*, 56(6) :1507–1525.

[Rais and Viana, 2011] Rais, A. and Viana, A. (2011). Operations research in healthcare : a survey. *International Transactions in Operational Research*, 18(1) :1–31.

[Ricard Hibon and Petit, 2012] Ricard Hibon, A. and Petit, P. (2012). Les 1ères Assises de l'Urgence Samu-Urgences. de France Comment garantir l'accès à des soins médicaux de qualité en urgence ? Technical report, SAMU-Urgences, http ://www.sante.gouv.fr.

[Rico et al., 2007] Rico, F., Salari, E., and Centeno, G. (2007). Emergency departments nurse allocation to face a pandemic influenza outbreak. In *Simulation Conference, 2007 Winter*, pages 1292–1298.

[Rico, 2009] Rico, F. A. (2009). *Emergency department capacity planning for a pandemic scenario : Nurse allocation*. PhD thesis, University of South Florida.

[Rogers et al., 2004] Rogers, A. E., Hwang, W. T., Scott, L. D., Aiken, L. H., and Dinges, D. F. (2004). The Working Hours Of Hospital Staff Nurses And Patient Safety. *Health Affairs*, 23(4) :202–212.

[Rudkin et al., 2004] Rudkin, S. E., Oman, J., Langdorf, M. I., Hill, M., Bauché, J., Kivela, P., and Johnson, L. (2004). The state of ed on-call coverage in california. *The American Journal of Emergency Medicine*, 22(7) :575–581.

[Saghafian et al., 2012] Saghafian, S., Hopp, W. J., Van Oyen, M. P., Desmond, J. S., and Kronick, S. L. (2012). Patient streaming as a mechanism for improving responsiveness in Emergency Departments. *Operations Research*, 60(5) :1080–1097.

[Sanchez et al., 2006] Sanchez, M., Smally, A. J., Grant, R. J., and Jacobs, L. M. (2006). Effects of a fast-track area on emergency department performance. *The Journal of emergency medicine*, 31(1) :117–120.

[Sasser, 1976] Sasser, W. E. (1976). Match supply and demand in service industries. *Harvard Business Review*, 54(6) :133–140.

[Schmid, 2012] Schmid, V. (2012). Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. *European Journal of Operational Research*, 219(3) :611–621.

[Schuur and Venkatesh, 2012] Schuur, J. D. and Venkatesh, A. K. (2012). The growing role of emergency departments in hospital admissions. *New England Journal of Medicine*, 367(5) :391–393.

[See et al., 2009] See, B., Liu, S.-P., Lu, Y.-W., and Pang, Q. (2009). Staffing a pandemic urgent care facility during an outbreak of pandemic influenza. In *Proceedings of the 2009 Winter Simulation Conference (WSC)*, pages 1996–2007.

[Sinclair, 2007] Sinclair, D. (2007). Emergency department overcrowding - implications for paediatric emergency medicine. *Paediatrics & child health*, 12(6) :491–494.

[Sinreich and Jabali, 2007] Sinreich, D. and Jabali, O. (2007). Staggered work shifts : a way to downsize and restructure an emergency department workforce yet maintain current operational performance. *Health Care Management Science*, 10(3) :293–308.

[Sinreich et al., 2012] Sinreich, D., Jabali, O., and Dellaert, N. P. (2012). Reducing emergency department waiting times by adjusting work shifts considering patient visits to multiple care providers. *IIE Transactions*, 44(3) :163–180.

[Stang et al., 2010] Stang, A. S., McGillivray, D., Bhatt, M., Colacone, A., Soucy, N., Léger, R., and Afilalo, M. (2010). Markers of overcrowding in a pediatric emergency department. *Academic emergency medicine : official journal of the Society for Academic Emergency Medicine*, 17(2) :151–156.

[Stolletz and Brunner, 2012] Stolletz, R. and Brunner, J. O. (2012). Fair optimization of fortnightly physician schedules with flexible shifts. *European Journal of Operational Research*, 219(3) :622–629.

[Sven et al., 2011] Sven, O., Håkan, J., Jon, R., and Katarina, G. (2011). A systematic review of triage-related interventions to improve patient flow in emergency departments. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, 19(43) :1–9.

[Tan et al., 2013] Tan, K. W., Tan, W. H., and Lau, H. C. (2013). Improving patient length-of-stay in emergency department through dynamic resource allocation policies. In *IEEE International Conference on Automation Science and Engineering (CASE)*, pages 984–989.

[Tien and Kamiyama, 1982] Tien, J. M. and Kamiyama, A. (1982). On manpower scheduling algorithms. *Siam Review*, 24(3) :275–287.

[Valdmanis et al., 2010] Valdmanis, V., Bernet, P., and Moises, J. (2010). Hospital capacity, capability, and emergency preparedness. *European Journal of Operational Research*, 207(3) :1628–1634.

[van de Vrugt and Boucherie, 2015] van de Vrugt, M. and Boucherie, R. J. (2015). Assigning treatment rooms at the emergency department. *Operations Research for Health Care*.

[Van den Bergh et al., 2013] Van den Bergh, J., Beliën, J., De Bruecker, P., Demeulemeester, E., and De Boeck, L. (2013). Personnel scheduling : A literature review. *European Journal of Operational Research*, 226(3) :367–385.

[Vile, 2013] Vile, J. (2013). *Time-dependent Stochastic Modelling for Predicting Demand and Scheduling of Emergency Medical Services*. PhD thesis, Cardiff University.

[Vilke et al., 2004] Vilke, G. M., Brown, L., and Skogland, P. (2004). Approach to decreasing emergency department ambulance diversion hours. *The Journal of emergency medicine*, 26(2) :189–192.

[Wargon et al., 2009] Wargon, M., Guidet, B., Hoang, T. D., and Hejblum, G. (2009). A systematic review of models for forecasting the number of emergency department visits. *Emergency medicine journal : EMJ*, 26(6) :395–9.

[Warner, 1976] Warner, D. M. (1976). Scheduling nursing personnel according to nursing preference : A mathematical programming approach. *Operations Research*, 24(5) :842–856.

[Xiao et al., 2010] Xiao, J., Osterweil, L. J., and Wang, Q. (2010). Dynamic scheduling of emergency department resources. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 590–599. ACM.

[Yankovic and Green, 2008] Yankovic, N. and Green, L. V. (2008). A queueing model for nurse staffing. *Under review at Operations Research*, 7.

[Yankovic and Green, 2011] Yankovic, N. and Green, L. V. (2011). Identifying Good Nursing Levels : A Queuing Approach. *Operations Research*, 59(4) :942–955.

[Yarmand et al., 2014] Yarmand, H., Ivy, J. S., Denton, B., and Lloyd, A. L. (2014). Optimal two-phase vaccine allocation to geographically different regions under uncertainty. *European Journal of Operational Research*, 233(1) :208–219.

[Zeltyn et al., 2011] Zeltyn, S., Carmeli, B., Greenshpan, O., Mesika, Y., Wasserkrug, S., Vortman, P., Marmor, Y. N., Mandelbaum, A., Shtub, A., Lauterman, T., Schwartz, D., Moskovitch, K., Tzafrir, S., Basis, F., and Health, R. (2011). Simulation-Based Models of Emergency Departments : Operational, Tactical and Strategic Staffing. *ACM Transactions on Modeling and Computer Simulation*, 21(4) :24 :1–24 :25.

Annexe A

Compte Rendu de la visite du CHRU de Lille

(Dr) Jean-Marie Renard, (Professeur) Alain Martinot, (Dr) Laurent Happiette, (Dr) Dorkenoo Aimée, Omar El-Rifai (21/11/2012)

A.0.1 Principales Maladies

A part les pics de tension créés par les grippes et bronchiolites, certaines maladies hivernales récurrentes participent à la création de tensions qui apparaissent avec une évolution moins aiguë. Entres autres on dénombre les gastro-entérites et les infections respiratoires ; intestinales et fébriles. Nous pouvons voir à la Figure A.1 le taux d'incidence des syndromes grippaux en France pour l'année (2011).

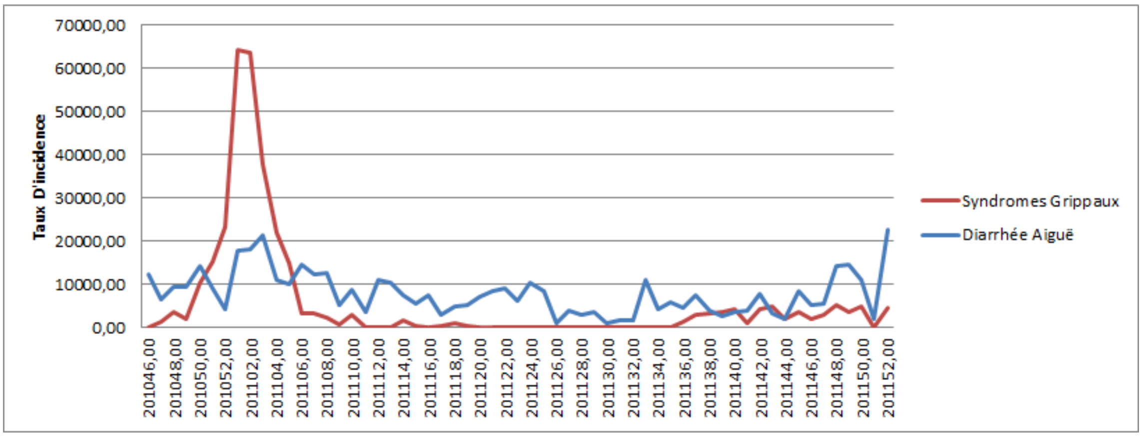


FIGURE A.1 – Évolution des syndromes grippaux et des diarrhées aigues pour l'année 2010(fin)/2011.Source : Réseau Sentinelles

A.0.2 Le services des urgences

Le service des urgences pédiatriques du CHRU de Lille a une configuration typique d'un service d'urgence mis à part quelques subtilités. Les chambres d'hospitalisation de courte durée et les box de consultation sont à proximité. De plus, les patients passent par un parcours typique d'urgence sauf pour les patients atteint de poly-traumatimes qui vont dans une salle de déchocage adulte gérée par des anesthésistes adultes et non par des pédiatres (taux très faible).

Les principaux composants du service des urgences :

— Accueil Administratif (partagé avec les urgences adultes) Pour le recueil des informations personnelles et des documents de prise en charge des soins (carte vitale, attestation de sécurité sociale et attestations de mutuelles)

— Accueil à l'entrée des urgences pédiatriques assuré en journée par une hôtesse et de nuit et les WE par l'ensemble du personnel médical et para-médical avec création d'un dossier papier sur lequel sont notés le motif de consultation, le nom et l'adresse du médecin/ du pédiatre traitant, les accompagnant et leurs coordonnées téléphoniques, les antécédents, les allergies, les prises de médicament

— Accueil par les infirmières d'accueil et d'orientation qui prennent les constantes (pouls, TA, fréquence cardiaque et respiratoire, poids) et évaluent en fonction de grille pré-établies le niveau de gravité du patient sous forme de code couleur : rouge, bleu ou vert. Les patients les plus graves devant être vus en priorité, les autres étant vus en fonction de leur niveau de priorité et de leur heure d arrivée aux urgences.

— Salle de déchocage : pour les patients très grave nécessitant une réanimation immédiate.

— Salle d'attente (commune aux patients en attente d'être vus et ceux déjà vus en attente de réévaluation, de résultats sanguins ou d'imagerie).

— 4 box de consultation :servant à la consultation et aux soins (prise de sang, lavements, pose de médicaments intra-veineux,...). Un des box étant réservé pendant les périodes épidémiques de gastro-entérites à l examen des patients atteints de gastro. Ils sont en général saturés de 18h à 22h.

— 1 salle de suture et une salle de plâtre (également utilisés pour les consultations et les soins)

— 10 chambres HCD : servant à la fois pour les hospitalisations de courte durée, de chambre d'attente pour les patients en attente d'une hospitalisation dans un service, de box d'isolement pour les patients contagieux ou fragiles (patients immunodéprimés, enfants < 3 mois, ...) et en période épidémique 1 à 2 box sont transformés en salle d'attente pour les patients atteints de bronchiolites.

A.0.3 Les ressources humaines

Au sein des urgences le personnel se compose

Pendant la journée : 3 séniors (2 pédiatres et 1 urgentiste chargé de la traumatologie), 2 internes de pédiatrie, 2 internes de médecine générale, 1 interne de chirurgie et 3 externes

Pendant le soir :

1 sénior (pédiatre) qui fait la nuit complète, 1 interne de pédiatrie qui fait la nuit complète, 1 interne de médecine générale jusqu'à minuit, 1 interne de chirurgie et 2 externes.

Pendant la période hivernale (épidémique des bronchiolites) :

Un renfort est assuré avec : 1 sénior de pédiatrie qui reste jusqu à 20h et 1 interne de pédiatrie de 18h jusqu'à minuit.

L'emploi du temps est créé pour un mois mais peut-être modifié par un accord commun entre les médecins. Des pédiatres spécialisés qui ne travaillent pas normalement aux urgences peuvent également intervenir.

A.0.4 Les Outils informatiques

Le personnel a à sa disposition un outil informatique (en fonction des heures d'entrée des patients) resUrgence qui donne en temps réel pour chaque patient enregistré auprès de l'infirmière d'accueil et d'orientation :

— L'emplacement actuel du patient
— Le temps écoulé depuis l'entrée aux urgences
— Les médecins en charge du patient (sénior, interne, externe) représentés par leurs initiales
— Le motif de consultation
— Les examens complémentaires (prise de sang, radio, etc...) et soins demandés fait et à faire
— D'autres informations spécifiques à l'identité du patient (date de naissance, constantes, adresse personnelle, coordonnées téléphoniques, constantes, allergies, antécédents,etc...)

Ces informations sont sauvegardées dans une base de données et pourraient être utilisées ultérieurement ; après un processus d'anonymisation pour nos travaux.

A.0.5 Les goulots d'étranglement principaux et les pics d'affluence

Au niveau de la journée, Les heures de pointes sont généralement vers midi et le soir après 18h en semaine Les tensions se créent les jours de congés et les week-ends. les tensions se créent également logiquement à cause de la diminution du personnel effectif le soir.

Les médecins pensent à l'unanimité que les principales causes de tensions sont dues aux :

— Consultations requérant de la médecine de ville et ne nécessitant pas de consultation aux urgences
— Engorgements au niveau du flux de sortie (notamment pour trouver des lits d'aval).

A.0.6 Dispositifs mis en place pour désengorger les urgences

Les stratégies mises en place pour faire face au situation de tensions pendant la période épidémique -12/15 Novembre à 12/15 Mars- ne suivent pas de protocole mais sont plus ou moins les suivantes :

— Redistribuer des chambres vides le matin du même jour ou des salles de services auxiliaires pour une longue durée afin de créés des boxs de consultations d'épidémies hivernales
— Programmer les rendez-vous de bilan annuel vers des saisons moins tendues.

- Faire appel à des infirmières du pool de remplacement (avec engagement d'au moins deux jours)
- Quelques seniors restent en continuité de soin (bénévolement)
- Renfort d'une demie garde d'interne de pédiatrie aux urgences pour le créneaux (18 :00 à Minuit)

D'autres mécanismes sont considérés également tout au long de l'année comme la réorientation des patients nécessitant une hospitalisation vers d'autres centres hospitaliers par manque de lits et le déroutement des ambulances vers d'autres hôpitaux moins engorgés.

A.0.7 Autres stratégies envisagées

Quelques stratégies éventuelles qui ont été discutées avec les points positifs et les points négatifs :

Stratégies	**Points Positifs**	**Points Négatifs**
Avoir un médecin à l'accueil	— Triage plus précis — Éviter les consultations ne requérant pas de l'urgence	— Consommation d'un médecin à l'accueil qui n est pas au sein des urgences à assurer des consultations pour les patients en ayant réellement besoin — L'évaluation à l'accueil par un médecin est plus complète et donc plus chronophage — Risque paradoxal de voir une augmentation des flux car les patients voyant le médecin d'accueil sortent avec un diagnostic plus rapide.

Ouvrir un circuit court	— Accélération des flux des patients ne requérant pas des urgences et de la bobologie. — Diminution des tensions car la séparation des patients graves des autres permet d'éviter des incompréhensions liées au fait de voir un autre patient passer avant eux (car mauvaise compréhension des patients et de leur entourage de la priorité faite de la gravité sur l'heure d'arrivée)	— Effet paradoxal possible car la création d'une offre de soin rapide expose au risque d'augmentation des flux des patients ne requérant pas des urgences et de la bobologie.
Mise en place d'une caution qui sera non restituée si la consultation du patient est considérée comme ne requérant pas des urgences	— Diminution des flux des patients ne requérant pas des urgences et de la bobologie.	— Risque d'inadéquation et donc de conflits entre les patients et les médecins car les patients ne se sentent souvent pas aptes à évaluer ce qui requiert réellement de l'urgence. — Risque de retard voire de non consultation de patients requérant pourtant de l'urgence par peur de payer cette caution (notamment parmi les familles défavorisées qui n ont pas toujours les moyens d'avancer la caution
Report des hospitalisations des patients non urgents en dehors des périodes épidémiques	— Augmentation des lits d'aval pour les urgences	— Augmentation des temps d'attente avant hospitalisation des patients non urgents en période épidémique avec risque de mécontentement de ces derniers.

Recours à du personnel de l'extérieur (médecins généralistes et pédiatres) en soirée et WE	— Amélioration des flux aux urgences. — Diminution de la charge de travail pour le personnel des urgences — Désengorgement des urgences des consultations ne requérant pas de l'urgence — Intérêt financier pour les médecins extérieurs qui ne voient que des enfants déjà dont les constantes ont été déjà prises et ne requérant pas des urgences (donc pas grave avec des consultations plus courtes) avec possibilité de diriger l'enfant vers les urgences en cas de mauvais triage initial	— Difficulté à trouver du personnel extérieur — Risque paradoxal d'augmentation de flux de patients non grave venant aux urgences consulter le médecin extérieur pour des raisons pratiques, leur médecin ne pouvant pas les prendre le jour même ou étant fermé le soir et les WE — Perte financière pour les urgences car les consultations ne requérant pas de l'urgence sont plus rapides et donc souvent plus rentables.
Adaptation de l'emploi du temps du personnel des urgences de manière optimale à l'affluence	— Diminution des temps d'attente aux urgences aux heures d'affluence.	— Les médecins ne travailleraient quasiment plus en journée mais principalement les soirs et les WE. — Risque paradoxal que l'augmentation de personnel en soirée et WE conduise à une augmentation du nombre de consultation aux urgences.

A.0.8 Services Auxiliaires

Principaux services auxiliaires :

- Les services d'hospitalisation et de consultation disponibles sur l'hôpital
- Le Service de Biologie (analyses de biologie médicale)
- Le pôle de l'urgence qui comprend en sus des urgences pédiatriques
  - Un service d'urgence médical adulte
  - Un service d'urgence chirurgical adulte
  - Un service d'hospitalisation de courte durée adulte (hébergement)
  - Un Service de radiologie des urgences comprenant :

1. Un IRM
2. Un Scanner
3. 2 salles de radiologie conventionnelle
4. Une salle d'échographie

— Un déchocage adulte couplé à un service de soins intensifs
— Des blocs opératoires et une salle de réveil dédiés à l urgence (adulte et pédiatrique pour les enfants > 1 an et > 10 kg)
— Un Centre des brûlés pour les adultes et les enfants de plus de 3 ans
— Un centre d'appel SAMU (Service d'Aide Médicale Urgente) : le 15, associé à un SMUR (Service mobile d'Urgence et de Réanimation). Le SAMU a pour fonctions :

1. De répondre aux appels urgents (arrêts cardiaque, accidents cérébraux, infarctus, accidents de la voie publique, incendies, ...) avec déclenchement d'un SMUR
2. Organisation de transferts de patients instables ne pouvant être transportés par une ambulance conventionnelle
3. Une activité conseil assurée par des médecins généralistes (conseils relatifs à des médicaments ou des prescriptions, indication ou non à consulter aux urgences,...) avec possibilité en cas de surcharge d'un service d'urgence de dévier les flux vers d'autres hôpitaux.

— Le service de médecine nucléaire
— La faculté dentaire (accessible pour les soins dentaires aux heures ouvrées)
— Le centre anti-poison

École Nationale Supérieure des Mines
de Saint-Étienne

NNT : 2015 EMSE 0802

Omar EL-RIFAI

HUMAN RESOURCE CAPACITY PLANNING OF AN EMERGENCE DEPARTMENT DURING EPIDEMIC SEASON

Speciality: Industrial Engineering

Keywords : Emergency Departments; Epidemics; Human Resources; Stochastic Optimization; Scheduling; Operations Research; Combinatorial Optimization

Abstract:

In France, the problem of overcrowding in Emergency Departments (ED) is particularly relevant today because of increasing admissions and budget restrictions in health facilities. Formally, overcrowding can be defined as a situation where the demand surpasses the service capacity. Studies that have dealt with ED overcrowding have mostly dealt with issues of patient flow management and resource management.

Our work focuses on resource management and more particularly on the impact that human resources have on ED overcrowding. In the first part of the study, we formulate the capacity allocation problem in a generic form. As such, we examine the capacity management problem and derive interesting properties for a general demand distribution and a normal demand distribution. Then, we examine the cyclic shift scheduling problem as it exists in ED. This research allows us to examine different cyclic scheduling strategies and answer the question of whether lack of flexibility in the schedules has an impact on the waiting time of patients. We propose an original stochastic linear formulation for the problem that accounts for the non-stationary work demand. After that we evaluate two scheduling mechanisms to reduce overcrowding in EDs: on-call duties and overtime hours. We study the conditions under which these mechanisms can be beneficial. The on-call duties problem is modeled as a two-stage stochastic optimization problem and the overtime management problem as a Markov decision problem.

École Nationale Supérieure des Mines
de Saint-Étienne

NNT : *2015 EMSE 0802*

Omar EL-RIFAI

GESTION DES RESSOURCES HUMAINES D’UN SERVICE D’URGENCE EN PERIODE EPIDEMIQUE

Spécialité: Génie Industriel

Mots clefs : Urgences, Epidémie, Planification, Ressources humaines, Optimisation stochastique, Recherche Opérationnelle, Optimisation combinatoire

Résumé :

Cette thèse s'inscrit dans le cadre du projet ANR HOST (Hôpital : Optimisation, Simulation et évitement des Tensions) qui vise à étudier et résoudre le problème de tension aux urgences hospitalières. Le projet cherche premièrement à définir rigoureusement la notion de tension et puis d'utiliser des outils mathématiques pour proposer des solutions qui prennent en compte la complexité du système de santé. Malgré la récurrence des phénomènes de tension, ils sont limités dans le temps et sollicitent par conséquent des solutions à court terme. La difficulté de trouver des solutions efficaces est principalement due à l'incertitude et au dynamisme caractéristique du service des urgences. D'une année à une autre, l'intensité de l'épidémie, et donc le nombre de patients se présentant aux urgences peut varier drastiquement. De plus, pour un même nombre de patients, il est difficile d'estimer correctement la charge de travail qu'il représente. L'objectif de cette thèse est donc de proposer des solutions stratégiques, tactiques et opérationnelles de gestion des ressources en prenant en compte les aléas caractéristiques des urgences. D'un point de vue stratégique nous étudions la distribution optimale de la capacité de travail sur une période épidémique. Ensuite, nous traitons le problème de confection des postes de travail journalier en fonction de la dynamique de la demande. Nous étudions également la possibilité d'affecter des postes d'astreinte aux médecins pendant les périodes épidémiques. Finalement, nous traitons le problème d'affectation des heures supplémentaires aux ressources dans un contexte plus opérationnel.