Real-time scheduling in outpatient clinics

Summary

Waiting time is an important indicator of the quality of service offered by outpatient clinics (OPCs). In spite of its growing importance, OPCs are still ill-equipped to reduce the waiting time effectively. The current applications of operations management often match patients' demand with supply. The OPC system is a push system that plans and schedules patients and resources ahead of time and is often not based on the actual demand. Additionally, OPCs are functionally organised to manage and optimise resources, patients and activities of the departments locally. In order to achieve efficiencies at the departmental level, the departments take crucial decisions with limited knowledge about other departments. Further, OPCs utilise feedback only for long term planning and scheduling.

However, in reality, OPC systems are open loop systems that have dynamic interactions with their environments. The unpredictable patient demand (due to a large number of walk-in patients, especially in Indian OPCs), unstable patient conditions and complex patient pathways result in high variability and uncertainty in OPCs. The following are the major reasons that make the OPCs' operating systems inefficient and result in prolonged waiting and cycle times, in spite of applying the existing knowledge of operating systems:

1. The planning and scheduling are performed in advance.
2. The planning and scheduling are not based on the actual patient demand.
3. The feedback is used only for future planning.
4. The entire OPC is not viewed as a single interdependent system.

Ideally, OPCs should provide services to patients when they ask for it (pull system) and to achieve this, OPCs should plan, schedule and control their resources, patients and activities in real-time with a global perspective. Therefore, rather than simply responding in an ad-hoc and individualistic manner, there is a need to systematically manage the OPC system with respect to the prevailing situation. The present thesis aims to design an operating system to minimise the waiting and the cycle times. The current thesis tests the hypothesis that "A hospital system with disparate subsystems cannot minimise cycle/waiting time by separately optimising subsystems, or by scheduling that does not adjust in real-time". In the present research, real-time scheduling and global perspectives are applied to each subcomponent of the operating system, namely, resource planning, patient scheduling and resource coordination. The real-time workflow optimisation uses the feedback in the same time horizon, thus reducing the latency. This transforms the OPC system from an open loop system to a closed loop system. The optimisation models and the results of all studies described in this thesis are summarised.

Chapter 1 gives an introduction to OPC systems (functional and operational structure) and explains the causes of operational problems, such as waiting and cycle times. It explains the push and the pull operating systems that are used in manufacturing industries, like Toyota Production System (TPS), albeit from the hospital context. The chapter ends with the thesis

outline and a list of operational definitions of the terms used in the thesis. The overall research question of the present thesis "Does the overall (global) optimisation along with real-time scheduling improve the wait time/cycle time and hospital performance?" is introduced in this chapter. The research question leads to four sub-questions on resource planning, patient scheduling, resource coordination and an optimal mix of these sub-components.

Chapter 2 describes the literature survey done in the present study. This chapter examines different studies in this field and identifies the important variables required to design an operating system that improves the performance of OPCs. The chapter reviews various methods and techniques used in resource planning, scheduling of patients and resources and resource coordination. The literature review forms the basis of the present study, building on what is already known and identifying the research gaps that can be explored in this study.

Chapter 3 presents the research methodology, the design variables which were identified through the literature survey and the research strategies and design, which describes the research location (Aravind Eye Hospital (AEH), Madurai, India, one of the largest eye care providers in the world), the participants and the data collection method. The chapter ends with a presentation of the research flow.

Chapter 4 describes how a robust predictive resource planning reduces the waiting and the cycle times in an OPC. As a preliminary step, the patterns of patients' arrival and resource scheduling in AEH were studied in detail. The OPC in AEH scheduled the resources once in a month, ahead of time and on the basis of the average demand, with local control in the departments. A simulation model was built to understand the baseline functioning of the variables. In the proposed optimisation model, the predictive resource planning considered a global perspective (entire OPC) and short-term demand variability. The model employed the patient demand, the number of resources available and precedence constraints as inputs and the daily resource plan of the OPC system as output. The global perspective of resource planning was implemented through Takt time management, which is widely used in TPS.

Resource plans were obtained from the proposed optimisation model for different patient demands. The resources were accordingly scheduled both in the simulation model and in the actual model of AEH. The results confirmed that the average waiting time in the OPC was reduced by 43.4% during the simulation study and by 41.1% during its actual implementation. The reaction time is the time taken to respond to the change that is triggered by external or internal factors. Along with the demand variability and planning methods, the reaction time was also found to influence the waiting and cycle times. Therefore, the resource plans of the optimisation model were analysed with different reaction times. The present study demonstrates the matching of supplies according to short-term demand with a global perspective that enables a better planning, thereby reducing the waiting and cycle times.

Chapter 5 examines how integral patient scheduling reduces the waiting and cycle times. Integral patient scheduling combines the concepts of a global perspective and real-time patient scheduling. The global perspective is implemented through path optimisation and the real-time patient scheduling is implemented through the actual status system of the OPC. A hybrid ant agent algorithm was developed, which considered the actual status of all departments in the OPC. The algorithm identified an optimal pathway for patients that minimised the waiting and cycle times. The proposed model was integrated into the simulation model of AEH and was implemented in their Hospital Management System. Integrated patient scheduling reduced the average waiting time in the OPC by 33% during the simulation study and by 26.5% during its actual implementation. On an average, all patients spent the same amount of time in waiting and the variability with the waiting time was observed to be reduced, as indicated by the standard deviation. Patient scheduling in real-time also depends on the IT infrastructure, which stores and retrieves knowledge of the OPC at regular and expected intervals. This study confirms that integrated patient scheduling reduced the waiting and cycle times in OPCs.

Chapter 6 describes how a real-time coordination mechanism for rescheduling resources reduces the waiting and cycle times. The individual goal of the patients is to reduce their waiting time, whereas the goal of the resources is to improve their utilisation. The goal of the OPC system is to reduce the waiting and cycle times. Therefore, the OPC system is modelled as a multi-agent system to address both individual and OPC goals. In the present study, patients are passive agents, whereas resources are active agents. The real-time scheduler schedules the patients according to the optimal pathway, depending on the actual status of all departments in the OPC. The scheduler maintains a pool to identify the available resources in the OPC and enable the rescheduling of resources. The coordination mechanism uses the Bayesian game and auction bidding methods for real-time rescheduling of resources. The resources are transferred from either the resource pool or other departments, as and when required. The coordination mechanism was analysed with different reaction times as they affect the waiting time as well. The findings of the study confirm that the real-time coordination mechanism for real-time scheduling reduced the waiting time by 55.8% during simulation and by 51.6% during actual implementation. The resource utilisation increased by 8.3% during actual implementation.

Chapter 7 discusses the main findings of each of the research questions of the present thesis. All optimisation models developed in this thesis were employed in different combinations, intended for designing the operating system. The results show that the optimal mix was the combination of predictive resource planning (push), integral patient scheduling (pull) and real-time coordination mechanism (pull). The research confirms that an OPC system with disparate subsystems cannot minimise the waiting time by separately optimising the subsystems or by scheduling that does not adjust in real-time. This chapter reflects the contribution and limitations of the present study and enlists the potential avenues for future research.

Samenvatting

Wachttijd is een belangrijke indicator voor de kwaliteit van de diensten aangeboden door poliklinieken. Ondanks het toenemende belang ervan zijn poliklinieken nog steeds niet goed uitgerust om wachttijden substantieel te reduceren. De huidige toepassingen van operations management in de gezondheidszorg zijn meestal gericht op het matchen van de patiëntenvraag met het aanbod. Het polikliniek-systeem is een push-systeem dat plannen en afsprakenschema's van patiënten ruim voorafgaand aan de uitvoering maakt. Deze zijn vervolgens vaak niet gebaseerd op de werkelijke vraag. Poliklinieken zijn bovendien functioneel georganiseerd en gericht op het lokaal managen en optimaliseren van middelen, patiëntenstromen en diensten van afdelingen. Om efficiënter te werken nemen afdelingen cruciale beslissingen terwijl zij maar beperkte kennis over elkaar hebben. Vaak gebruiken poliklinieken alleen feedback informatie voor lange termijn planning en niet voor de (hele) korte termijn.

In werkelijkheid zijn poliklinieken open-loop systemen die dynamisch interacteren met hun omgeving. De onvoorspelbaarheid van de patiëntenvraag (als gevolg van een groot aantal walk-in patiënten, vooral in India, patiënten met instabiele condities en complexe patiënttrajecten) leidt tot hoge variabiliteit en onzekerheid in poliklinieken. Dit leidt tot wachttijden en inefficiënt gebruik van capaciteit. Hieronder worden de belangrijkste oorzaken genoemd van het inefficiënt functioneren van poliklinieken met als gevolg lange wacht- en doorlooptijden:

1. De planning en het maken van afspraakschema's vinden ruim voor de uitvoeringsperiode plaats.
2. De planning en afsprakenschema's zijn niet gebaseerd op de werkelijke patiëntenvraag.
3. De feedback wordt alleen gebruikt voor toekomstige planning (en niet op de huidige in uitvoering zijnde planning).
4. De poliklinieken van een ziekenhuis worden niet gezien als één enkel onderling afhankelijk systeem.

In het ideale geval leveren poliklinieken hun diensten aan patiënten op het moment dat hen daar naar wordt gevraagd (pull), en om dit te kunnen doen moeten poliklinieken plannen, roosteren en monitoren in real-time met een globaal (organisatiebreed) perspectief. Daarom, in plaats van te reageren op ad hoc en individualistische wijze, zouden poliklinieken als systeem moeten reageren.

Dit promotie-onderzoek is gericht op het ontwerpen van een besturingssysteem dat wacht- en doorlooptijden minimaliseert. De huidige thesis toetst de volgende hypothese: "Een ziekenhuis systeem met ongelijksoortige subsystemen kan wacht-/doorlooptijden niet minimaliseren door subsystemen afzonderlijk te optimaliseren, of door niet real-time te plannen, afspraken te maken en te monitoren." In de huidige studie, worden real-time scheduling en globale optimalisatie toegepast op elk subonderdeel van het besturingssysteem. Deze subonderdelen zijn: resourceplanning, patiëntenplanning en resourcecoördinatie. Real-time workflow optimalisatie maakt gebruik van feedback informatie uit dezelfde

uitvoeringsperiode. Dit transformeert de polikliniek van een open loopsysteem naar een gesloten kringloop.

Hoofdstuk 1 geeft een inleiding tot polikliniek systemen m.b.t. tot hun functionele en operationele structuur en de oorzaken van operationele problemen, zoals wacht- en doorlooptijden. Het bespreekt de push- en pull besturingssystemen die worden gebruikt in de industrie, zoals het TPS, in de context van het ziekenhuis. De algemene onderzoeksvraag van het huidige proefschrift 'Doet totale (globale) optimalisatie samen met real-time scheduling wacht-/doorloop tijden in poliklinieken reduceren en prestatie maximaliseren?' wordt in dit hoofdstuk geïntroduceerd. De onderzoeksvraag leidt tot vier sub-vragen. Deze hebben betrekking op resourceplanning, patiëntenplanning, resource coördinatie en het optimaliseren van de mix van deze subonderdelen. Het hoofdstuk eindigt met een overzicht van het proefschriftoverzicht en een aantal definities van de termen die worden gebruikt.

Hoofdstuk 2 beschrijft het literatuuronderzoek uitgevoerd en de centrale variabelen die nodig zijn voor het ontwerpen van een besturingssysteem dat de prestaties van poliklinieken verbetert. Het hoofdstuk beoordeelt diverse methoden en technieken die worden gebruikt in de planning van patiënten en resources, en de coördinatie hiervan.

Hoofdstuk 3 presenteert de onderzoeksmethodologie, de variabelen en de strategieën van onderzoek en ontwerp, de methode van dataverzameling en de locatie van het onderzoek. Het ontworpen systeem is ontwikkeld voor en getest in het AEH te Madurai, India. AEH is één van de grootste aanbieders van oogzorg ter wereld.

Hoofdstuk 4 beschrijft de methodiek van resourceplanning die wacht- en doorlooptijden in een polikliniek reduceert. Als eerste stap werden de aankomstenpatronen van patiënten en de planning van resources in AEH in detail geanalyseerd. De poliklinieken in AEH plannen de middelen een maand vooruit en op basis van de gemiddelde vraag, met daarnaast lokale besturing door de afdelingen. Een simulatiemodel werd gebouwd om de huidige situatie te analyseren. In het voorgestelde model voor resource planning wordt geoptimaliseerd vanuit een globaal perspectief (de gehele polikliniek) en rekening houdend met korte termijn variabiliteit. Het optimalisatiemodel gebruikt gegevens omtrent de verwachte patiëntenvraag, het aantal beschikbare resources en de volgorde regels als input. De output van het optimalisatiemodel is de resourceplanning op dagbasis van het gehele polikliniek-systeem. Het globale perspectief van de resourceplanning werd geïmplementeerd via Takttimemanagement.

De resourceplannen werden geproduceerd op basis van verschillende vraagpatronen. Deze plannen werden vervolgens getest in een simulatiemodel en op bepaalde patiëntenstromen in het AEH. De resultaten toonden aan dat de gemiddelde wachttijd in de polikliniek was verlaagd met 43.4% volgens de simulaties en met 41.1% in de implementatiestudie in het

AEH. Samen met de variabiliteit van de vraag en de planningsmethode bleken ook de reactietijden de wacht- en doorlooptijden te beïnvloeden. De reactietijd is de tijd die nodig is om te reageren op veranderingen (bijvoorbeeld de aankomst van nieuwe patiënten of het wegvallen van resources). Daarom werden de resourceplannen die door het optimalisatie-model werden geproduceerd, geanalyseerd met verschillende reactietijden. Aangetoond wordt dat het aanpassen van resources op de korte termijn, dus dicht tegen het werkelijk optreden van de patiëntenvraag, in combinatie met een globaal perspectief mogelijk is en betere plannen oplevert.

Hoofdstuk 5 onderzoekt hoe integrale patiëntplanning de wacht- en de doorlooptijden vermindert. Integrale patiëntenplanning combineert het perspectief van het geheel van polikliniekken met real-time patiëntenplanning. Dit globale perspectief wordt geïmplementeerd via padoptimalisatie en real-time scheduling door het voortdurend monitoren en reageren op de status van het polikliniekssysteem. Een hybrid ant agent algoritme werd ontwikkeld die de werkelijke status van alle afdelingen in de kliniek beschouwt. Het algoritme identificeert de optimale route voor patiënten die de wacht- en doorlooptijden minimaliseert. Het voorgestelde model werd geïntegreerd in het simulatiemodel van AEH en werd geïmplementeerd in het Ziekenhuis Managementinformatie en planningssysteem van AEH. De geïntegreerde patiënt planning vermindert de gemiddelde wachttijd in de polikliniek met 33% tijdens de studie van de simulatie en met 26,5% tijdens de eigenlijke uitvoering. De variatie in wachttijd daalde en de doorlooptijd van patiënten vertoonde daardoor ook minder variatie. Deze studie bevestigt dat de geïntegreerde patiëntenplanning de wacht- en de doorlooptijden in poliklinieken aanzienlijk reduceert. De mate waarin dit model succesvol kan worden toegepast is sterk afhankelijk van de IT-infrastructuur. Deze moet immers steeds de werkelijke situatie in de poliklinieken kunnen waarnemen en communiceren.

Hoofdstuk 6 beschrijft hoe een real-time coördinatiemechanisme voor het herplannen en realloceren van resources de wacht- en doorlooptijden vermindert. De patiënten wensen hun wachttijden te minimaliseren en de resources willen hun bezettingsgraad maximeren. Het doel van het polikliniek-systeem is het verminderen van de wacht- en doorlooptijden. Daarom is de polikliniek gemodelleerd als een multi-agent systeem om zowel individuele als polikliniek doelen te realiseren. In de huidige studie zijn patiënten passieve agenten, terwijl resources actieve agenten zijn. De real-time scheduler plant de patiënten volgens het optimale pad rekening houdend met de actuele status van alle afdelingen in de polikliniek. De scheduler beheert een overzicht van resources in de poliklinieken die beschikbaar zijn om opnieuw ingepland te worden. Het coördinatiemechanisme gebruikt Bayesiaanse spelen de veilingmethoden voor real-time resource planning en reallocatie. Het coördinatiemechanisme werd onderzocht met verschillende reactietijden omdat deze van invloed zijn op wachttijden. De bevindingen van de studie bevestigen dat de real-time coördinatiemechanismen voor real-time planning en reallocatie de wachttijd met 55,8% reduceren tijdens

simulatie experimenten en met 51,6% tijdens de feitelijke implementatie. De bezettingsgraad is 8,3% gestegen tijdens de feitelijke uitvoering.

Hoofdstuk 7 bespreekt de belangrijkste bevindingen van dit onderzoek. Alle optimalisatiemodellen ontwikkeld in dit proefschrift kunnen in verschillende combinaties worden getest en geïmplementeerd. De vraag is dan welke combinatie het beste ontwerp is voor een besturingssysteem van een polikliniek. Het blijkt dat de optimale mix een combinatie betreft van voorspellende resourceplanning (push), integrale patiënt (pull-) planning en real-time coördinatie (pull). Het onderzoek bevestigt dat een polikliniek-systeem met ongelijksoortige subsystemen de wachttijden niet kan minimaliseren door subsystemen afzonderlijk te optimaliseren of door niet real-time te coördineren.