Waiting Patiently

An analysis of the performance aspects of outpatient scheduling in health care institutes

BMI - Paper
Anke Hutzschenreuter
Vrije Universiteit Amsterdam

To know how to wait. It is the great secret of success.

Joseph De Maistre **1753-1821, French Diplomat, Philosopher**

Preface

In the second year of the Master's programme in Business Mathematics and Informatics (BMI) at the Vrije Universiteit (Free University) in Amsterdam every student has to write a paper on a self-chosen subject that is related to the curriculum. This means that it should cover at least one of the following areas: business management, mathematics or computer science.

Since I was a child I have been in contact with a medical environment as both my parents are practitioners. My mother as an internal specialist has her own practice. Therefore I was confronted with the difficulties and problems occurring in the daily practice. This is how the question arose for me how to improve the planning in the health care sector.

During my BMI studies I became very interested in the optimization of business processes whereby business problems are tackled with the aid of mathematical solution techniques. Application areas like production planning and call centers were discussed but unfortunately not business processes involving health care. Fortunately, m supervisor in this project, Prof. Dr. Koole, gave me the opportunity to combine these two fields of interest within the scope of the project reported in this BMI paper.

The paper is addressed to anyone interested in the optimization of scheduling systems in health care, especially practitioners and other non-mathematicians. Therefore some of the applied methods are outlined in more detail.

I hope the reader will enjoy this paper.

Amsterdam, January 2004

Anke Hutzschenreuter

Abstract

The goal of this paper was to compare the performance of a selection of appointment scheduling rules under realistic external factors. An appropriate scheduling rule aims at minimizing the physician's idle time and the patients' waiting time. The environmental factors were explored during a literature enquiry and interviews with medical practices. According to the findings patients could be grouped according to service time characteristics. Based on the findings from the literature and the interviews a discrete-event simulation model of a realistic environment was developed.

The scheduling rules were analyzed for different fluctuations of the service times. Due to start-up problems reported in the interviews the performance was measured for both an entire clinic session and the start-up period, which was chosen as the first half an hour.

Although the "best" decision depends on the priorities of a practice, good results could be achieved when scheduling two patients at the start of the clinic session. If patients are grouped by different service durations, it appeared to be best to schedule the patients with short treatment times at the start of the session. Addressing the start-up problems we conclude that the Bailey-Welch rule performs best under all service time fluctuations.

Table of contents

PREFACE **4**

ABSTRACT **5**

TABLE OF CONTENTS **6**

1. INTRODUCTION **8**

2. APPOINTMENT SCHEDULING SYSTEMS **9**

2.1. What is an appointment scheduling system? **9**

2.2. Why is appointment scheduling important? **9**

2.3. Common scheduling methods **10**

2.4. Literature review **11**

2.5. Application of appointment scheduling systems in practice **12**
2.5.1. Interview at an ophthalmic outpatient department 13
2.5.2. Interview at a practice for manual therapy of lymphatic and skin diseases 14
2.5.3. Interview at a joint practice of internal specialists with focus on cardiac diseases 14
2.5.4. Interview at a joint practice of internal specialists with focus on gastroenterological diseases 15
2.5.5. Interview at a general practitioner 16

3. PROBLEM DEFINITION **17**

3.1. Environmental factors **17**
3.1.1. No-shows 17
3.1.2. Punctuality 17
3.1.3. Duration of consult 17

3.2. Performance measures **18**

3.3. Appointment scheduling rules **19**

3.4. Queuing theoretical background **20**

4. METHODS **23**

4.1. Introduction to discrete-event simulation **23**

4.2. Modeling Assumptions **24**

4.3. Simulation model 25

5. RESULTS 27

5.1. Comparing rules over entire clinic session 27
5.1.1. Individual, Bailey-Welch and "Two-at-a-time" ASR 27
5.1.2. Proportional Scheduling 28

5.2. Comparing rules during start-up period of clinic session 30

6. DISCUSSION 32

7. CONCLUSIONS 34

7.1. Suggestions for further research 34

ACKNOWLEDGEMENTS 37

REFERENCES 38

APPENDICES 39

Appendix A: Mean practitioner's utilization over time 39

Appendix B: Tabulated results of simulation experiments 40

1. Introduction

In a modern society like ours it is almost impossible to avoid the daily waiting. We wait for appointments, news, decisions, and for good weather. We wait in the office, in the lecture hall, in front of elevators and on the phone. We are used to waiting at the cinema entrance, the cash desk in the supermarket, traffic lights and at bus stops. And we wait for the doctor.
Everyone is familiar with the situation when entering a waiting room. We are always astonished about the large number of people waiting even though we came on time for our appointment. Waiting times of a forenoon are not unusual in hospitals.
Experienced patients bring their books and businessmen their notebooks with them in order to use the waiting time efficiently. However, the atmosphere in many waiting rooms does not permit to spend the time doing something useful or even pleasant and most people remain waiting patiently until the doctor's assistant comes in and calls their name.
According to a study in 2002 of Statistics Netherlands, an average person consults a general practitioner 3.8 times per year. With a population of 16,258,662 inhabitants this results in 61,782,915.6 patient contacts. If each patient has to wait 10 minutes we talk about 10,297,152.6 man-hours of waiting time.
Of course is it undesirable that patients have to wait longer than necessary. Still many practitioners overestimate the value of their time with respect to the patient's. Nevertheless, the doctor's time has no infinite value and it is desirable to find a reasonable balance between the doctor and the patient.
This essay addresses this problem, analyzing it with the means of applied mathematics, viz. queuing theory.

This paper will proceed as follows. First a description of appointment scheduling systems will be provided in section 2, including a literature review and findings from the interviews with medical doctors. Then the problem is defined and the environmental factors, performance measures and queuing theoretical background will be outlined. This will be followed by a depiction of the research methodology and the results from the simulation study. Finally the results will be discussed in combination with conclusions and suggestions for further research.

2. Appointment scheduling systems

2.1. What is an appointment scheduling system?

An appointment scheduling system is a system used to manage appointment calendars and scheduling appointments for physicians, dentists and other health care providers. It allocates appointments to time slots during the consultation hours. This allocation is done according to so-called appointment scheduling rules (ASRs).
An appropriate scheduling rule should optimize both the facility idle time and the patient waiting time. Minimizing practitioner's idle times will certainly lead to excessive patient waiting times and vice versa. These are contradicting objectives and an individual compromise should be sought.

A formal description according to Ho and Lau (1992) can be given as follows.
Assume that n patients are scheduled for a clinic session. Let A_i denote the arrival time of patient i, where $i = 1,2…n$. This is a deterministic variable with a certain noise resulting from external factors like the tendency to arrive early for an appointment. Let further S_i and b_i be the service time and the time at which the service begins for patient i, respectively. S_i is a random variable following a certain probability distribution. These service times are assumed to be independent and identically distributed. Therefore b_i is also stochastic. Without loss of generality we can assume that $A_1 = b_1 = 0$. For $i > 1$ we get that $b_i = \max (A_i, b_{i-1} + S_{i-1})$. Then the patient waiting time is given by $W_i = \max (0, b_i – A_i)$ for all i. The service facility idle time just before the arrival of patient i is $F_i = \max (0, A_i – b_{i-1} + S_{i-1})$.

2.2. Why is appointment scheduling important?

In the Netherlands, like in other European countries, a debate is taking place on the issue of the public health system. As things are today, the quality of health care is at a very high level. Facing a growing demand for health care due to the increasing aged population, the question arises how to organize and finance the health care system. Financial means and staff are being curtailed, but the budgets cannot be cut endlessly. Thus improvements in efficiency are needed. This starts right from the daily planning of health care institutions, especially hospitals.
But also smaller institutions like practitioners have to meet new requirements. In some areas the doctor patient ratio is very high which means that practitioners have to face a growing competition. Scheduling patients improperly can have a severe impact on the success of a practice. Critical areas like physician's productivity, patient satisfaction and practice profits can be negatively influenced. On many internet pages of hospitals and practices one can find statements like "With a good schedule we provide very short waiting times…" or even promises that no waiting occurs at all.
But also the staff morale can suffer from the feeling of being off-directed and overloaded. Patient scheduling is definitely an area where the following statement holds: If it's broken, fix it now!

2.3. Common scheduling methods

According to A. Soriano (1966) appointment systems can be classified into four general types: (1) Block Appointment Systems, (2) Individual Appointment Systems, (3) Mixed Block-Individual Appointment Systems, and (4) other appointment systems.
A pure Block Appointment Systems assigns a common appointment time at the beginning of a clinic session for all patients to be seen during the session. The patients are seen in order of arrival. Such a system was mainly used in the past by hospitals justifying it with the argument that the length of a consult is highly variable and that patients arrive late on their appointments. This system assures a high productivity of the physician at the costs of extremely long waiting times for the patients. This method can be modified as each session can be divided into smaller units of time, i.e. a "smaller" block of patients is scheduled at the beginning of every hour. Of course this can be further divided which leads to shorter queues and to constantly high productivity of the practitioner.
An Individual Appointment System is a system in which a different appointment time is assigned to each patient. The intervals between two appointments are equally spaced over the day. An interval length of the average consulting time leads to a lower utilization of the doctor's time and to shorter waiting times. Both depend on the variability of the patients' delay and the consulting times. These can be reduced even more by choosing a longer time interval as interarrival time, for instance adding a buffer of five minutes.
A Mixed Block-Individual Appointment System schedules an initial group of patients at the beginning of the clinic session and schedules other patients to arrive at equally spaced intervals. This system combines the advantages of the two appointment systems, which is a relatively high productivity of the physician and shorter queues in the waiting rooms.
Other appointment scheduling systems are modifications of the above-mentioned systems or combinations of them. For example one can think of a system that delays the scheduled arrival of the second, third, etc. patient of the initial group of the Mixed Block-Individual Appointment System. This means that the interarrival time is shorter than the average service time but longer than time 0. Various values for these interarrival times can be chosen and need not be equal.

Next to these V. Giacolone (2003) mentions the Proportional Appointment System and the Clinical Appointment Systems as popular patient scheduling methods.
In a Proportional Appointment System patients are classified into examination categories and then given appointments with an interarrival time equal to the expected treatment duration. The classification criterion is the expected consulting time. One commonly used classification is that a short consult takes up to 15 minutes, an intermediate consult has duration of 15 to 30 minutes and a long consult exceeds 30 minutes. Factors to be considered are the working speed of the practitioner, the number of examination rooms and the competence of the support staff.
A Clinical Appointment System refers to grouping patients of a specific type of patient (i.e. coronary disease, hematological disease, allergies, etc. for an internal specialist). These groups are seen during one block of time of the session. Like in the proportional appointment system the number of time slots for each group is the key to a successful use. The reason for grouping patients according to their disease is that the doctor can work more efficiently when he sees patients who require similar diagnosis techniques, tests, evaluation and other services. One can think of "setup"-

times of the doctor as opening another software program, changing the examination rooms or even changing his or her "state of mind" when the doctor sees clinically different patients.

2.4. Literature review

Based on their studies of appointment scheduling rules in health care clinics, Bailey (1952, 1954), Bailey and Welch (1952) and Welch (1964) name punctuality and the consulting time as the two main factors affecting the design of an appointment system. They refer to time studies performed at several clinics and state that most of the patients are early and only some are late but that the medical staff in general does not arrive on time. From field studies Bailey determined that most practitioners hold clinics with less than 30 patients, which means that the queuing system will never be in a steady state. In order to balance between the waiting times of the practitioners and the patients Bailey and Welch propose to schedule *n* patients at the start of the clinic and then schedule patients at intervals equal to the average consulting time. They recommend beginning the session with 2 present patients with the purpose of preventing excessive waiting for the patients.

Soriano (1966) compared an individual and a "Two-at-a-time" appointment scheduling system deriving expressions for the steady-state waiting time distributions of the patients under both systems. He analyzed the behavior of the two systems under different load scenarios under the assumption that the patients arrive on time and that the consulting times follow a gamma distribution. One of the results is that the distribution function for the first system always remains between the distribution functions belonging to the first and second treated patients in the second system. He concluded furthermore that for increasing load factor the waiting time of the patient who is treated first in the "two-at-a-time" system decreases compared to the patient scheduled according to the individual appointment system and that the waiting time of the patient treated second in the corresponding arrival pair increases. The differences between the distribution functions are shown to be relatively little and decreasing as the load factor increases. All waiting times increase for increasing load factor. Therefore he proposes to schedule blocks of two patients at a time with an interval of twice the consulting time between two consecutive appointments.

Ho and Lau (1992) studied 50 appointment-scheduling rules and concluded that no single rule outperforms the others under all environment conditions and recommended several rules of thumb under specific environments. Furthermore they proposed the so-called variable-interval appointment scheduling rules as to correct the problem of long waiting times of customers in the late sessions. These rules modify the individual ASR by requiring that before the *K*th patient the intervals between two appointments are shorter than the average service-time interval in order to reduce the facility's idle time. The *(K + 1)*th to the last patient arrive later than the expected service time so the patient's waiting time will be reduced.

Klassen and Rohleder (1996) published the first paper that addresses the performance of scheduling rules under the environmental condition that patients have different characteristics in the service time. They chose the standard deviation of the treatment time as characteristic. Furthermore emergencies are taken into account. Next to the common primary performance measure in the literature, which is the cost of both server and clients' waiting time, they also used other measures like the

relation between the costs of client and server idle time and the end time of a clinic session. They conclude that under these conditions it is best to schedule patients with treatment times with large standard deviation towards the end of the consulting hours. As for the positioning of time slots left open for urgent patients they identified options for some clinic goals.

Ho and Lau (1999) evaluated the impact of environmental conditions on the performance of selected appointment scheduling rules. The factors considered are the probability of no-shows, the coefficient of variation of the consulting times and the size of the daily clinic. As measure of the impact they consider the expected cost of patients' waiting time and the expected cost of service facility's idle time. They selected nine rules out of 50 possible ASRs, which can be roughly classified as the Bailey-Welch rule with modifications, the block appointment rule and the variable-interval rule with modifications. For infrequently occuring no-shows they conclude that one of the variable-interval rules outperforms the other rules for both measuers of costs. As the variability of the consulting times increases the patients' waiting time increases if the Bailey-Welch rule or its modifications are used. The expected costs for the doctor's idle time remain almost constant.
The variable-interval rule and its modification perform well to reduce customers waiting times and intend to increase the facility idle time. For the investigation of the impact of different numbers of patients per clinic they use the average waiting time per patient and the average facility idle time per patient as performance measures. Their simulation results show that the variable-interval and the "block-of-two" rules give almost identical average patient's waiting time for different numbers of patients, the "block-of-two" rule also gives almost identical average facility utilization. The variable-interval rules tend to increase facility idle time. The Bailey-Welch rule and the modifications of it deteriorate the facility idle time as well but show a steady character in terms of the average patient waiting time. They conclude that the choice of an ASR is very situation-specific and depends on the desired combination of doctor's and patients' waiting time.

LaGanga and Lawrence (2003) studied eight appointment scheduling rules based on data recorded at an outpatient mental health clinic. They conclude that the average service time and its distribution, the total number of patients, the length of the clinic session and the no-show probability impact the performance of these rules. Furthermore they considered modifications of the scheduling rules obtained by permutations of the number of patients and the time slots of a session. Also the application of overbooking and metaheuristic techniques is analyzed. Using overbooking the number of additionally scheduled patients depends on the desired number of patients per day and the no-show rate. These patients are incorporated into the arrival pattern of the scheduling rule. Other modifications are formed by crossover and mutations of existing rules.

2.5. Application of appointment scheduling systems in practice

In order to obtain insight in the way hospitals and practices assign their appointments for their daily schedule, interviews were held with a few health care institutions. To achieve a broad spectrum of information on this issue small, medium and large practices were asked to describe their appointment systems and potential problems they encountered. Next to the size also the kind of practice was taken into consideration. As the support staff like assistants or receptionists usually does the

scheduling, the doctors were asked for their permission in advance and sometimes were directly concerned with the interviews.

A predefined list of questions helped to structure the interviews and to achieve homogeneous information. First the interlocutor was asked to describe the principles according to which appointments are assigned to the patients, and then the environmental factors like no-shows and delay of patients were discussed. Finally the persons were asked for their view of the doctor's workflow including the doctor's average idle time and average waiting time of the patients.

This resulted in the findings presented below in the form of case descriptions. First a short description of the practice is provided which is then followed by the content of the answers.

2.5.1. Interview at an ophthalmic outpatient department

This outpatient department consists of a number of oculists, medical assistants, specialists for contact lenses, etc. and provides medical care to ambulatory patients. No surgeries are performed in the department; patients being in need of a surgery or special care are referred to the hospital.
The daily consulting hours start at 8:00 a.m. and end at 3:30 p.m. Appointments are given between 8:00 a.m. and 11:45 a.m. and 1:30 p.m. and 3:30 p.m. Among the scheduled doctors one is assigned to be supervisor who can be asked for advice by his colleagues. His consulting hours are therefore interrupted. Also the consulting hours of other doctors can be interrupted by phone calls, etc.
The waiting period for an appointment is in general three months. However time slots in the afternoon sessions are reserved for urgent patients who get an appointment within one day. Less acute cases are assigned to an appointment within two or three weeks, are scheduled as overbooking. This means that an appointment is given which is also assigned to another patient. Obviously this causes longer waiting queues.
Patients who already had an appointment are assigned to the same doctor as the previous time; the doctor who has the first free capacity in his schedule treats new patients.
Arriving patients first see the receptionist. The appointment is scheduled 15 minutes before the actual consult, which takes between 10 and 15 minutes. Thus, buffers of waiting patients are created which guarantee a more efficient workflow for the doctors. Sometimes preliminary examinations are required and the goal is that the patients should not wait longer than necessary. An individual appointment system is used which assigns an individual time to every patient.
Encountered problems are a no-show rate of about 10% and that the appointments in the early hours are not very popular. The appointments between 10 and 12 a.m. are usually fully booked. Mainly the no-shows and delayed arrivals in these early hours cause start-up problems as there is no buffer built up and the doctors have to wait for their patients.
The outpatient department has a scheduled level of utilization of 98% that causes an immense pressure on the staff.

2.5.2. Interview at a practice for manual therapy of lymphatic and skin diseases

This practice is run by a specialist for lymph and skin therapy. With her work she supports among others the cancer therapy of her patients and cures skin diseases like acne, etc. Female cancer patients whose breast(s) has/have been amputated often suffer from postoperative scars and “thick” arms where the lymph liquid can no longer flow off and is banked up in the vessels.
Her daily consulting hours start at 7:30 a.m. and end at 5:30 p.m. or sometimes at 6:30 p.m. Appointments are given with an interarrival time of 30 or 45 minutes which is equal to the treatment time she calculates for the different forms of therapy. Thus an individual appointment system is applied. As she has no assistant or receptionist she is interrupted by phone calls, etc. during consulting hours.
She says that she “trained” her patients to be on time for their appointments as she sticks to her time scheme herself. If patients arrive too late she tries to influence their behavior. However, if patients arrive later than scheduled the treatment time is reduced by the delay and the patients have to pay for the full treatment.
As her workload is very high, even too high in her opinion, patients rarely cancel their appointments. No-shows occur very seldom because she hands out cards to her patients with the practice conditions. On the card is written that the patients are obliged to call at least 24 hours in advance if they cancel their appointment. If they fail to inform her the consult is still charged. Still, it is not always possible to fill these time slots but she uses these “breaks” for other things like writing letters to doctors or to take care of her organizational duties for the professional union of skin therapists.
Due to the therapeutic nature of her treatments there is almost no variation in the duration of the consulting time. For elderly patients more time, e.g. for dressing, is calculated than for younger patients. If there is the possibility that an appointment takes longer than scheduled she informs the succeeding patients in advance. Consequently her patients have to wait less than 5 minutes.
According to her preferences no clinical scheduling is used. In her opinion it is more interesting to alternate the therapeutical treatments.
In the beginning of her practice she had some problems with her appointment schedule but now she is pleased with her time planning even though she has very few breaks.

2.5.3. Interview at a joint practice of internal specialists with focus on cardiac diseases

In this joint practice patients arrive only by reference from general practitioners or other internal specialists. They offer highly specialized treatments for cardiac and oncological diseases. This large practice counts over a thousand patients per quarter of a year.
The consulting hours start at 8:00 a.m. and end according to schedule at 4:00 p.m. but usually take half an hour longer. Appointments are given within this time interval according to the individual rule. Patients are grouped according to their specific treatment or diagnosis, which can be cardiac, internal, hematological, dermatological, etc. For these groups of fixed size time slots are reserved which are marked as colored squares in the schedule. These time slots differ every day, but remain the same every week. Within these groups they have time slots for emergency cases left that are only filled if an urgent case arrives. This appointment system can be identified as clinical scheduling with buffer. They apply this method because they appreciate that they have the same “mind set” during a specific group of patients.

Each doctor has his own patients, which has the advantage that the case history is known and clinical pictures can be analyzed easily. Another advantage is that the mostly elderly patients are used to their doctor and have confidence in his abilities. The doctor who has the first free capacity in his schedule treats new patients.
First a patient gets to see the doctor who orders tests that are mostly done by his assistants. In the meantime the doctor can see one or more other patients. Afterwards the patient is called back to the doctor's office the consult can take place and the test results are discussed and the further proceeding is explained.
The planning assures a steady workflow for the doctor and there are seldom no patients in the practice. However, patient satisfaction is a very important aspect of their planning. They hang up an information poster in the waiting room that the patients can come to reception if they are not seen by the doctor within 30 min. after their arrival. In the past it happened that patients were forgotten and came out after more than three hours to ask when they will consult the practitioner. Despite of this measure one of the doctors has an average waiting time of one hour while the others let their patients wait on average 15 minutes, at most 60 minutes but usually between 10 and 20 minutes. A waiting time of an hour occurs if an emergency patient is processed into the schedule. Then the other patients have to wait. They also make differences between employees and pensioners by assigning them the appointments for which the least waiting time is expected and by giving them priority in the queue.

The patients have to wait a few weeks for an appointment. They encountered that if the time before their appointment exceeds three weeks one or two patients do not show up per day. If the time is less than three weeks no-shows occur very rarely. The open time slots due to the no-shows are not taken into consideration in the appointment scheduling.
Patients being late are not a considerable problem because most of the patients are elderly and arrive half an hour early, only the younger patients occasionally come half an hour late.
Following an estimation of the assistant the doctor sees patients 410 minutes per day, the rest of the time is used for writing letters. This results in a utilization of about 85% when assuming that only the direct contact with patients is the primary process.
They chose a special format of their agenda, as they wanted to have an overview of all the appointments of all the doctors. This system was developed by the whole team and was improved regularly in the course of time.

2.5.4. Interview at a joint practice of internal specialists with focus on gastroenterological diseases

The practice is specialized in gastroenterological diagnosis and treatment. This includes diagnosis techniques as gastroscopy and colonoscopy, which means that an endoscope is inserted in the throat or rectum in order to make diagnosis, e.g. stomach ulcer or polyps in the colon. They are known beyond the region for their professionalism and their patient-friendly treatment.
The consulting hours originally start at 8:00 a.m. and end according to schedule at 5:00 p.m. but as they have many requests they extended the time starting now to 7:30 a.m. Two times a week they also have to run over time over time. As they give different appointments for the examinations and for preliminary talks, the working day is split into two parts. In the morning every 45 minutes a colonoscopy and every 30 minutes a gastroscopy is scheduled. After an examination the endoscope and the other devices need to be cleaned which takes about 30 minutes. Additionally a high fluctuation in the examination times has to be considered as for example a

colonoscopy can take between 25 and 60 minutes. In the afternoon “consulting” hours take place where 30 minutes are reserved for new patients and 15 minutes for known ones.
During this consulting hour the examination is explained to the patient who is then asked to give his informed consent. All appointments are given individually. In the early afternoon there is some space reserved for emergency cases of which at least one per day occur.
In order to maintain a good physician-patient relation each doctor has his own patients. Together they see 60 patients per day. This results in a very high but steady workload, so it is rather appreciated if a patient cancels an appointment on time. However, due to long waiting times before the appointments this seldom occurs. For a non-urgent examination the waiting time is three months, for a preventive medical checkup a patient has to wait almost half a year. The no-show rate is therefore one patient per week, which is a fraction of 1:300.
As mentioned before their aim is to be very patient-friendly, not only in the way the examinations are preformed but also in the waiting times. According to their own estimations patients have to wait 5 minutes in the morning sessions and 10 to 15 minutes during the afternoon. This is surprising if one remembers the high fluctuation in the examination durations.

2.5.5. Interview at a general practitioner

This medium-sized practice counts mainly local residents as patients. In general they see the doctor for minor discomfort and check-ups like colds or blood pressure measurements.
The doctor-patient contacts are therefore rather short. On average the doctor spends five to seven minutes with one patient with a fluctuation over an interval of 2 to 35 minutes. The appointments are scheduled according to a block appointment system that assigns one appointment to a block of two patients and an interarrival time of 15 minutes. This is maintained over all different treatments and only if the doctor says explicitly that more time is needed the receptionist deviates from the scheduling principle. The planning is set up so as to make sure that always one patient sits in the waiting room while the doctor is treating another patient. Naturally this causes longer waiting times than scheduling patients according to the individual rule for instance with an interval of 10 minutes between two arrivals. They experienced that their patients only have to wait 5 to 10 minutes, which suggests the possibility that a “two-at-a-time” appointment system is preferable in such a practice. However, in the worst case the waiting time builds up to 45 minutes which is quite long in relation to the duration of the consult.
Typically for a general practitioner many patients drop in without appointment. In this case one third of the patients just walk in and are worked into the schedule. An interesting point is that these walk-ins differ seasonally. This means that during the winter months, due to colds and waves of influenza, almost two third of the patients she sees per day drop in. They more than make up for the two or three patients per week who do not show up for their appointments. The walk-ins also affect the length of the session (9:00 – 12:00 a.m.), which has to be prolonged every day during the winter season.
Patients usually do not arrive late but will be compensated by a walk-in patient or another patient waiting.

3. Problem definition

3.1. Environmental factors

The aim of this essay is to study the performance of appointment scheduling rules under realistic operating factors. Information from previous studies and from interviews with receptionists at a few practices served as basis for the modeled environmental factors.

3.1.1. No-shows

No-shows occur in practices due to external unexpected events and in general are not related to the schedule. Only one practice reported that no-shows appear to be related to the waiting period for an appointment, see section 2.5.3. An average no-show rate of 10% was reported by the ophthalmic outpatient department, which fluctuated during the day with a higher probability in the early time slots. Ho and Lau (1999) ran simulation experiments using no-show rates of 0%, 10% and 20%. LaGanga and Lawrence (2003) considered a no-show probability as high as 30%.
On the one hand no-shows increase the doctor's idle time as time slots remain open and would only be filled by walk-in patients. On the other hand problems with high patients' waiting times can be alleviated through these open appointments. As Giacolone (2003) states, a high no-show rate can have a severe impact on the success of a practice. Therefore one should strive to minimize the probability by means of reviewing the scheduling system.
Mathematically speaking this can be modeled as a Bernoulli process. Formally one defines $\{X_n, n \geq 1\}$, a sequence of independent Bernoulli random variables with the same success probability p. This sequence can be used then to represent a two-state system in which a state occurs randomly and independently in discrete time. In our situation a success can be seen as a no-show patient. There are more ways how a Bernoulli process can be modeled, but will not be discussed in this paper.

3.1.2. Punctuality

In previous appointment scheduling studies by Bailey (1952 and 1954), Soriano (1966), Ho and Lau (1992 and 1999), Klassen and Rohleder (1995) and LaGanga and Lawrence (2003), patients were assumed to arrive on time. However, from the interviews held at various practices it appeared that patients in general arrive early for their appointments. In the ophthalmic outpatient department a study on patients' waiting times was performed and recorded an average arrival time of 10 minutes earlier than scheduled.
Operationally, early arriving patients help to decrease the doctor's idle time as he can draw forth the next patient. However, patients' waiting times can increase as the patient may have to wait the time he or she arrived early additionally to the pure time.

3.1.3. Duration of consult

In the literature, general gamma distributions have been used for the service time duration. Bailey (1952) and Welch (1964) empirically found gamma distributed treatment times with coefficients of variation (C_v) ranging from 0.51 to 0.62. Also Soriano (1966) assumed gamma distributed consulting times needed for the analytical approach. Ho and Lau (1992) showed that only the coefficient of variation

has influence on the performance of ASRs and not skewness and kurtosis of the probability distribution. In their article from 1999 they considered uniform and exponential service time distributions. LaGanga and Lawrence (2003) based their study on data recorded in an outpatient mental clinic. They used gamma and normal distributions as approximations for treatment times.
In the interviews it appeared that practitioners have very little information about the service times. They could give estimations for the average consulting time and its minimum and maximum. In such cases a triangular distribution of the service times is commonly used.
Of course the duration of the consulting times also determines the number of patients per session which Ho and Lau (1992) identified as an important operating factor. They found that one appointment system might be suitable for only a certain number of patients per session. This is therefore considered as an important factor for the evaluation of an ASR.

3.2. Performance measures

The purpose of this study is to analyze the performance of scheduling rules measured in the mean patient waiting time and the mean physician's utilization (MPU). Let n be the total number of patients in a clinic session.
The mean patient waiting time (MWT) is given by

$$MWT = \frac{1}{n}\sum_{i=1}^{n} W_i \ ,$$

where W_i is the waiting time of patient i.
The physician's utilization (PU) of a session is the total busy time divided by the actual session length

$$PU = \frac{A_n + W_n + S_n - \sum_{i=1}^{n} F_i}{A_n + W_n + S_n}$$

where F_i is the physician's idle time just before the arrival of patient i, A_n is the arrival time, W_n the waiting time and S_n is the treatment duration of the last patient. The physician's utilization is averaged over all simulated sessions. This is in accordance to previous studies by Bailey (1952), Ho and Lau (1992 and 1999) and Klassen en Rohleder (1995). For an outline of an average utilization over the time the reader is referred to Appendix A.
In this aspect this paper differs from previous studies because in the past the expected total cost of patients' and doctor's idle time is chosen as performance measure. In the author's opinion the doctor's utilization is a better measure in order to receive an impression of the performance of an appointment scheduling rule. The facility's idle time is only a meaningful variable if it is linked to the session length. A total idle time of 30 minutes on a session of 3 hours is certainly less efficient than the same idle time in relation to a session of 6 hours. As reported in the paper by LaGanga and Lawrence (2003) different scheduling methods can result in different session durations.
Apart from this, Ho and Lau (1992) assumed a linear relationship between the patients' waiting costs and the patients' waiting time. However, as Klassen and Rohleder (1996) state one patient with an excessive waiting time may have different

costs than many patients with short waiting times and may still lead to the same total waiting costs. Also the costs can differ from patient to patient. One could think of employees and pensioners.
Nevertheless, this relationship between the waiting costs exceeds the scope of this paper and will therefore not be taken into consideration.

3.3. Appointment scheduling rules

In this research problem the following scheduling rules are considered: the individual, the Bailey-Welch and the two-at-a-time scheduling rule.
From the interviews, as presented in section 2.5, one can conclude that the individual appointment system is a popular scheduling method. In four out of five practices patients are assigned individual appointments. Therefore this scheduling principle deserves further investigation. Also the “Two-at-a-time” method is used in one of the interviewed practices, for a summary see section 2.5.5. It forms a special case of a block appointment system, which is a common scheduling method described in section 2.3. The Bailey-Welch rule is a mixed block-individual appointment rule that is identified as one of the basic scheduling principles in the literature, see a literature review in section 2.4. This rule schedules two patients at the start of a clinic session and the following patients according to the individual appointment system.

The arrival patterns of the individual, the Bailey-Welch and the “Two-at-a-time” appointment system are shown in Figure 1.

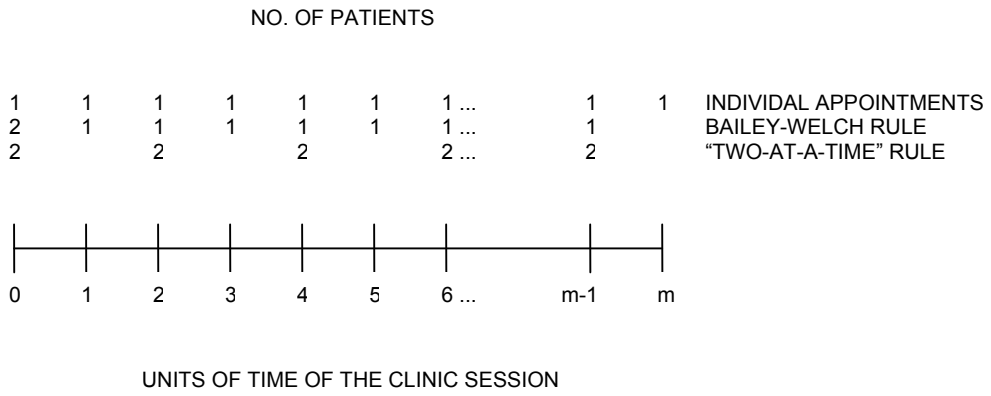

Fig. 1: Arrival patterns of scheduling systems over the clinic session

Furthermore, the effect of proportional scheduling was studied which means that patients are grouped according to equal expected treatment time. In previous studies this scheduling principle has only been addressed by Klassen and Rohleder (1996) but with different standard deviations instead of different treatment times. However, one of the interviewed practices encountered different consulting durations in connection with the medical characteristics. They use the method of proportional scheduling successfully, so it was considered as a scheduling method worth researching. Also possible sequences for these groups are analyzed in comparison to a variant where the patients are scheduled alternately, so one patient of type *i* after a patient with treatment time according to type *j*.
The studied appointment scheduling rules are listed in the following table.

Rule	Method	Arrangement
LBEGIndiv	Individual appointment scheduling system	First patients with long expected service time
SBEGIndiv	Individual appointment scheduling system	First patients with short expected service time
SBEGBW	Mixed Block-Individual appointment scheduling system	First patients with short expected service time and apply Bailey-Welch to first group
2SBEGAPIndiv	Individual appointment scheduling system	Alternate patients with short and long service times, starting with two patients with short treatment durations
AP2SNDIndiv	Individual appointment scheduling system	Alternate patients with short and long service times and schedule two patients at the end with short treatment durations

Table 1: Definitions of scheduling rules

For the rule LBEGIndiv, the schedule begins with patients with a long treatment duration, after which patients with short appointment time are scheduled. The patients within the groups are assigned to individual appointment times.
The rule SBEGIndiv is the opposite of LBEGIndiv, first patients with short treatment times are scheduled and then the time consuming patients are seen by the doctor. The appointments are given individually to the clients.
Contrary to this, SBEGBW schedules the two patients with short treatment time at the beginning of the session on the same time slot and then patients arrive in regular time intervals in the same sequence as for SBEGIndiv. Thus the Bailey-Welch rule is applied to the first group of patients.
For 2SBEGAltIndiv, the schedule begins with two short treatment time patients, after which patients are alternated throughout. According to this rule, all patients are assigned to individual appointments with an interarrival time equal to the mean service time.
According to a schedule following the Alt2SNDIndiv rule, patients are alternated starting with a patient with short expected treatment time. The last two time slots are assigned to patients with short appointment length. Therefore this rule forms the counterpart of the 2SBEGAltIndiv rule.

3.4. Queuing theoretical background

In the following paragraph we assume that the reader is acquainted with the basic concepts of probability theory such as random variables.
The model considered in this essay is the single-server model. This means that there is only one service facility to serve the arriving customers. From the findings from the interviews in section 2.5 one can assume that in many clinics each doctor has his own patients. Despite the fact that clinics often have more than one doctor, they do not share the workload. So this is a realistic modeling assumption.

The single-server model is the most basic model of queuing theory. Queuing theory forms one of the most important branches of probability theory, which has many applications in technology and management. In general a queuing situation can be described as every situation in which a service facility is provided, where waiting and queuing are encountered. In the terminology of queuing theory the service facility is called the *server* who provides service to the *customer*. The customer's behavior in this system has two main features: the *arrival process* and the *waiting process*.
The arrival process can often be described as a stochastic process, for instance as a Poisson process where the interarrival times of the single arriving customers in disjoint time intervals are independent and the number of customers arriving in a unit of time follows a Poisson distribution. On arrival the customer can find the server either busy or idle. If he finds the server idle the server will immediately start to serve. If he finds the server busy, which means that the server is serving another customer, the arriving customer has to wait and the waiting process starts. This process can be as follows: first the customer can wait until he is served; second the customer does not want to wait, goes away and never returns; third he could behave like in the second possibility but return later; fourth the customer joins the queue and waits for a certain amount of time and if he is not served within this time he also disappears.
Usually waiting customers are served in order of arrival, however also variants are possible: service in random order, last-come-first-serve or priority service. The latter is a realistic assumption if emergency patients are considered. Then an urgent patient will be treated before other waiting clients or even the consult of another patient would be interrupted.
One server serves one customer at a time with service duration, called the *service time*, which is usually modeled as a random variable.
The investigation of queuing systems is often used to obtain insight into the effectiveness of the work of a service facility. Most commonly used measuring tools are the waiting time distribution of the customers and the distribution of the busy periods of the server.

The situation in medical practices is such that they have a special arrival pattern. The arrival times are deterministic due to the appointments that are given to the patients. However, the deterministic character of the arrival process shows noise because the patients may be early or late or do not show up at all. According to the interviews held in the first phase of this project, patients wait patiently until they are called to see the doctor. As the treatment of a patient is focused on the whole person the staff can influence the patients not to leave the practice before being treated. In some of the practices priority is given to urgent cases but because many general practitioners do not encounter this situation too often the first-come-first-serve principle dominates.
The characteristics of the treatment process are as follows: the process often consists of many short activities with several activities per process where the kind and duration of the treatment form the important elements of uncertainty.

The state of a queuing model at time t depends on t and on the state of the system at time $t = 0$. Performing an analysis of a queuing system one usually is interested in the system's behavior in the long run, for $t \rightarrow \infty$, the behavior of the *stationary system*. The analysis of a stationary system is much simpler than the investigation of a system with finite values for t, the so-called *transient system*. It is clear that for our specific problem a stationary system is not a realistic model as clinic sessions are of limited duration. Nevertheless, a stationary system can be a good approximation if

one considers a clinic session to be the continuation of the previous session. Then the process can be seen as a process with infinite duration.

4. Methods

4.1. Introduction to discrete-event simulation

In many application areas such as production planning, mathematical models need to be built as they offer the possibility to study an encountered phenomenon and to analyze a real-world situation, to forecast and to optimize under certain criteria. This can be achieved by using a number of tools and techniques one of which is simulation. The Oxford English Dictionary describes simulation as:

> *"The technique of imitating the behavior of some situation or system (economic, mechanical, etc.) by means of an analogous model, situation, or apparatus, either to gain information more conveniently or to train personnel."*

Simulation represents a good alternative to direct experimentation when this is not feasible or expensive. By running “what if” experiments cause and effect relationships of the studied system can be approximated and used for further analysis.
The advantages of performing a simulation study are that simulation allows for testing of every aspect of a proposed change before changing the real system and committing any resources. Once a system is installed it can be very expensive to change or correct it. An example of this is a production line of a big factory that would need to be stopped in order to change it. The company would surely experience a great financial loss if the production was shut down.
New policies or operating procedures cannot be explored without expenses and would interrupt the system. Using simulation the analyst does not have to face this problem. Another advantage is that it allows compressing or expanding time. This means that a whole clinical month can be examined within a few minutes. Furthermore, many simulation packages offer an animation feature, which facilitates the communication between problem owner and problem solver and allows the participants to see what the planned system will look like.
Simulation models also allow the complete control of single parameters of a system. Often a large number of parameters play a role in the system and their influence can not be explored separately of each other. Using a sensitivity analysis one can get insight into this influence.
Disadvantages are clearly that the usage of simulation requires some degree of training to build a model. Also the construction may be time-consuming as the usage of simulation tools is not always intuitive or the software offers too many features. In addition, the simulation model needs to be validated to ensure that the results are reliable.

When using a simulation study the first step is to collect data and perform a data analysis on it. Then the model can be built and simulation experiments can be run. The last step is the analysis of the results.

Discrete-event simulation is one special simulation technique that enables one to observe the time based behavior of a system. There are formal methods for building such a model. Concepts such as verification and validation ensure that the simulated system is reliable.

4.2. Modeling Assumptions

A model is a description of a (part of a) real-world system such that it allows an analysis of aspects the model builder is interested in. In order to keep the model simple, assumptions on the system need to be made.
In this simulation study the arrival process is modeled as a stochastic process. Patients may arrive early, late or not at all. Taking the pattern from the study performed at the ophthalmic outpatient department, the earliness (or delay) is modeled as a normal distributed random variable with a mean equal to 10 minutes. As a simplification the standard deviation is set to 10 minutes. This means that about one sixth of the patients arrive more than 10 minutes late for their appointment or more than half an hour early. The no-show probability was fixed at 10%, which is in accordance with the study that took place at the ophthalmic outpatient department. In the paper by Ho and Lau (1992), they confirm the author's opinion that varying this probability does not lead to new findings. The no-show patients are immediately disposed of the system.
Furthermore, we assumed that patients are seen in order of arrival. This means that it is possible that patients arriving very early are seen to before the actually scheduled patients. This is unfair in the sense that patients who arrive on time for their appointment may have to wait because another patient arrived early and therefore was served before his actual appointment time. This is an assumption, which does not completely correspond to real situations. But in order to keep our model simple we allow this "unfairness".
It is common in the literature to simulate half-day sessions. A scheduling period of three hours was chosen which corresponds to the simulation period in the paper of LaGanga and Lawrence (2003). All simulations were made for 300 sessions which equals the average number of half-day sessions a practice is open per year.
For service time distributions triangular and gamma distributions were chosen in accordance with the interviews and previous studies referred to in the literature overview. The coefficients of variation of the consulting times were chosen as 0.2, 0.7 and 1.2. With these values a broad spectrum of possible distribution shapes was covered and their influence could be studied. Two different scenarios were simulated with an average consulting time of 10 and 30 minutes. This results in 18 and 6 time slots per session respectively. In previous work only a certain number of time slots was modeled and the corresponding length of the simulation period was not taken into account. In none of the practices we encountered that patients were sent away unseen. This results in a consulting hour that takes longer than planned. Therefore, a terminating condition was introduced to ensure that the doctor sees all patients.
It is assumed that the patients do not have to wait when entering the practice but go directly to the waiting room if the doctor is busy. After the consultation, the patient leaves the practice immediately. This is in accordance to the situation we encountered when the interviews with the receptionists were held. Usually patients do not need to see the receptionist before the consult so no delay results from it. However, after the consult the patients often see the receptionist, for instance to make a new appointment or to receive a prescription. As for the consultation process, it ends when the patients leaves the doctor's consultation room. Therefore one can consider this as the point of time when the patient leaves the system. If any waiting occurs at the reception it is not related to the primary process of the patient-doctor contact.

4.3. Simulation model

We performed an analysis comparing different scheduling policies using a simulation model developed on the Arena software platform (Rockwell Software Inc.).
The following figure gives an example of a scheduling rule modeled in Arena 5.0.

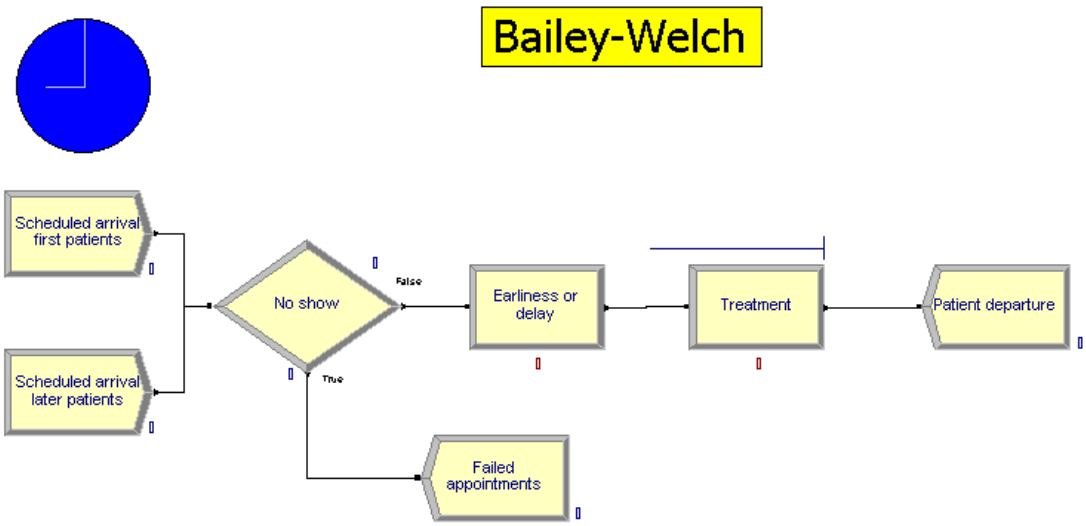

Fig. 2: Model of the Bailey-Welch scheduling rule

The basic process template provides all modules used in this model. The arrival process was split into two parts for the patients scheduled at the start of the clinic and for the later patients. The next table shows how the arrival pattern of the scheduling rule can be modeled for the case of an expected treatment duration of ten minutes.

	Name	Units	Entities per Arrival	Max Arrivals	First Creation
1	Scheduled arrival first patients	Minutes	2	1	0
2	Scheduled arrival later patients	Minutes	1	16	10

Table 2: Arrival pattern of Bailey-Welch rule

The randomly occurring no-shows of the patients can be represented as a two-way decision with probability of 10% for “false”, which means in this case that a patient shows up for his appointment.
If the patients show up they may arrive early or late. As outlined in section 4.2 we assume that patients arrive on average ten minutes early. In order to implement this assumption the doctor was given a daily schedule according to which his capacity was planned. The whole practice was then shifted 30 minutes earlier and the

earliness was modeled as a process with a normally distributed delay with a mean of 20 minutes. According to the schedule, the doctor arrived half an hour later than the “start” of the clinic session, which leads to the desired “earliness” effect.

5. Results

5.1. Comparing rules over entire clinic session

5.1.1. Individual, Bailey-Welch and "Two-at-a-time" ASR

The experimental results of the interaction effects between ASR and environmental factors for the mean utilization and the mean waiting time of the patients are tabulated in Appendix B, Table 3. It is apparent that fluctuations in the service times and other external factors lead to a poor performance of the ASR.
In the following figures, the mean doctor's utilization (Fig. 3) and the mean patients' waiting time (Fig. 4) are shown for six combinations of mean service times and their variability.
They reveal that if the variability of the service time increases, the waiting time for the patients increases and the practitioner's utilization decreases.

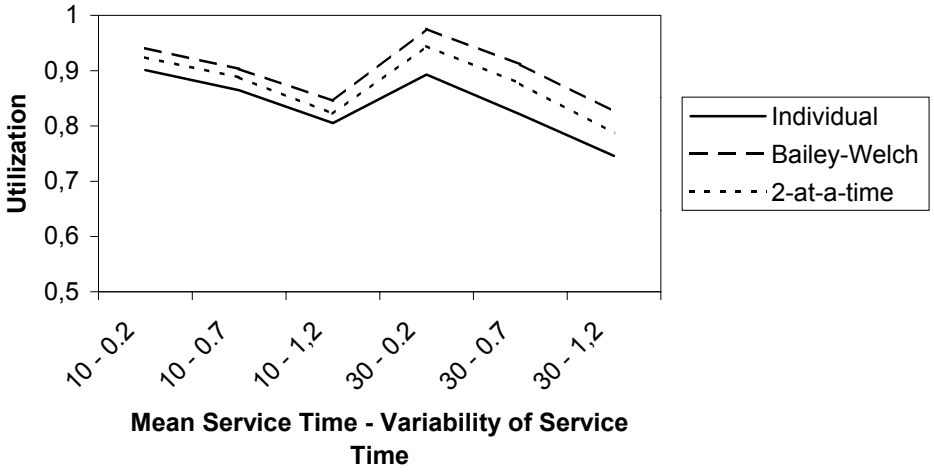

Fig. 3: Mean doctor's utilization for combinations of mean service time and its variability.

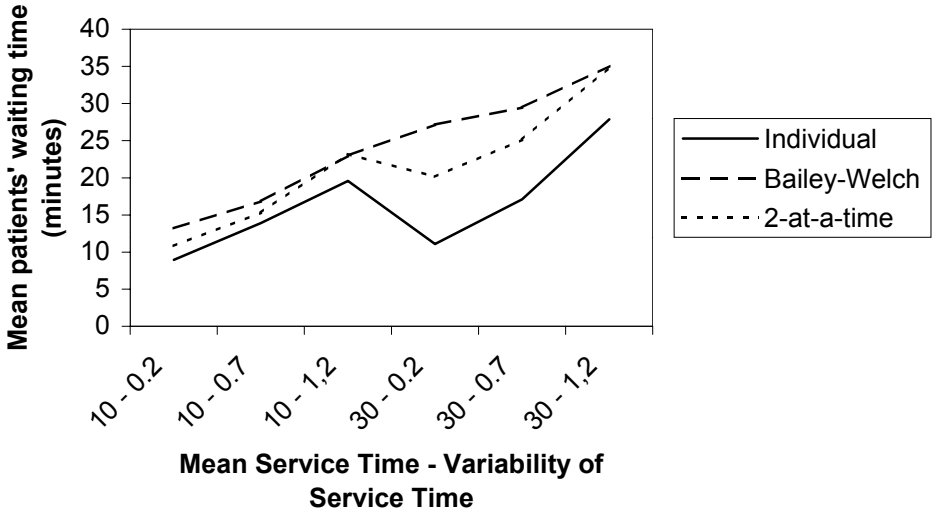

Fig. 4: Mean patients' waiting time for combinations of mean service time and its variability

Generally speaking, when the service times show higher fluctuation the patients tend to wait longer in the queue. There are, however, differences between the ASRs.
From Fig. 4, it is obvious that the performance deteriorates considerably for high fluctuations in the service time. When the variability increases the waiting times almost double for the individual and the "Two-at-a-time" rule. The Bailey-Welch rule shows the same behavior for short treatment times. The results differ very little. Patients wait about 4 minutes longer when scheduled according to Bailey-Welch than for the individual scheduling method. However, for long treatment times, the Bailey-Welch rule underperforms the other rules and leads to relatively long waiting times for low variability. The difference of about 17 minutes to the best rule, the individual rule, is severe. This is not surprising, because one client always has to wait. For high variance the "Two-at-a-time" rule results in almost the same long waiting times. It seems that the "inventory" effect gradually diminishes when C_v increases.
For high variance (C_v = 1.2) the "Two-at-a-time" rule results in as long waiting times as the Bailey-Welch rule. For small variance (C_v = 0.2) it ranges between the individual and the BW rule. For increasing variance the waiting times grow longer and tend towards the results of BW. The gradient is higher than for the BW rule but are almost as steep as for the individual scheduling method.
Based on the mean practitioner's utilization, the individual ASR is clearly inferior to the Bailey-Welch and the "Two-at-a-time" scheduling rules. This may be explained by this rule being designed to reduce patients' waiting time and therefore increases the idle time of the doctor. No "inventory" of waiting patients is built up, which would guarantee a steadier workflow under the external conditions of no-shows, punctuality and variability in service times. This difference is even more obvious for longer treatment times where the difference is almost 10% to the top scoring rule, the Bailey-Welch rule. It holds for all rules that the doctor's utilization decreases when the variability of the service time becomes volatile.
The Bailey-Welch rule outperforms the other rules for all combinations of mean service times and their variability. Not surprisingly the utilization under this scheduling rule is higher due to the fact that "inventory" is built at the start of the session. For long treatment times and a low coefficient of variance (C_v = 0.2) the utilization is even higher than for short treatment times with the same variability. However for high variability the utilization deteriorates considerably but remains above 80%.
The "Two-at-a-time" rule shows an almost parallel trend to the one from the Bailey-Welch rule, but utilization remains below the one from BW. The difference is about 2% and can be explained by the fact that only "inventory" is build at the appointment time. After the first patient is served and the doctor sees the second client there is no client waiting to be served. With higher variability of the service time this leads to higher facility idle time.
It appears that the practitioner's utilization can be increased by scheduling several patients (in these ASR 2 patients) at the start of a session. Therefore, if the facility utilization is the main criterion to evaluate ASRs, the individual rule should be avoided.

5.1.2. Proportional Scheduling

A summary of the simulation results is shown graphically in Fig. 5 and 6 and in tabular form in Appendix B, Table 4.

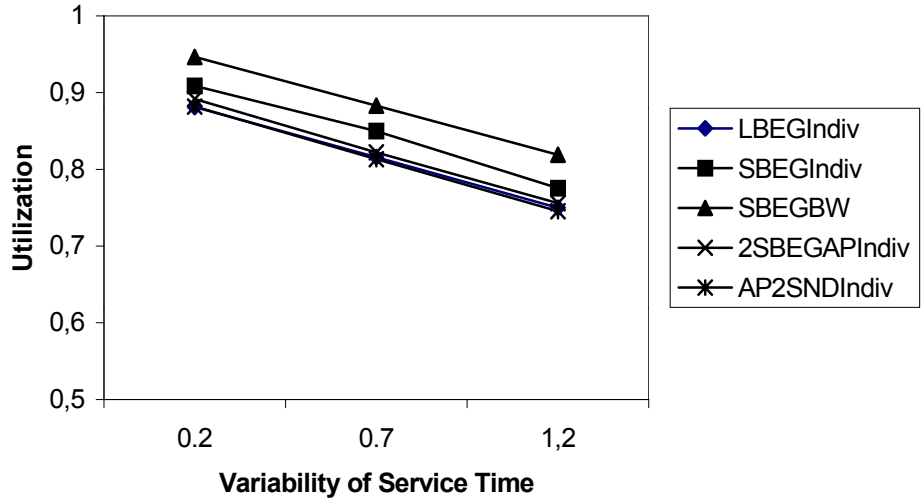

Fig. 5: Mean practitioner's utilization using proportional scheduling

In general, if the variability of the service time increases, the practitioner's utilization decreases. The impact is almost equal for all scheduling rules.
Some points are obvious from Fig. 5. The highest doctor's utilization is achieved by the SBEGBW rule. In this case, the doctor is on average busy during almost 95% of a session if there is low variability in the service time. For high variance the performance deteriorates, but still the idle time of the practitioner is less than 20% of the total time of consulting hours.
For the practitioner, there is little difference between scheduling first the patients with long service time and alternating "long" and "short" patients. Both possibilities show poor performance in the physician's utilization in comparison to the other rules.
Scheduling two "short" patients at the beginning and then alternating the patients tends to result in slightly lower facility idle time than the other variants (the doctor's utilization is increased by 1%).
The LBEGIndiv and the AP2SNDINdiv rule show very little difference in their performance (about 0,05%).
Thus if it is known that a difference between patients exists, scheduling the patients with short treatment time first is the best method. Additionally applying the Bailey-Welch rule to the first group leads to even better results.

Fig. 6 shows that patients with short and long treatment times have different waiting times. If the patients with long examination time are scheduled first, patients with shorter treatment time tend to wait longer than the other patients. This effect is just the other way round if "short" patients are scheduled first. However, these differences are more obvious for the LBEGIndiv scheduling method than for SBEGIndiv. For high variance the "short" patients have to wait more than 17 minutes longer than the "long" patients. Especially when considered with respect to their expected treatment duration (10 min.) this is "unfair". For low variance both types of patients have to wait about 11 minutes. Waiting times for the SBEGIndiv rule differ more for high variance. Then "short" patients have to wait about 10 minutes less than "long" patients.

For 2SBEGAPIndiv, waiting times increase with increasing variability of treatment times while for each coefficient of variance, “short” patients wait approximately 6 minutes longer than “long” patients.

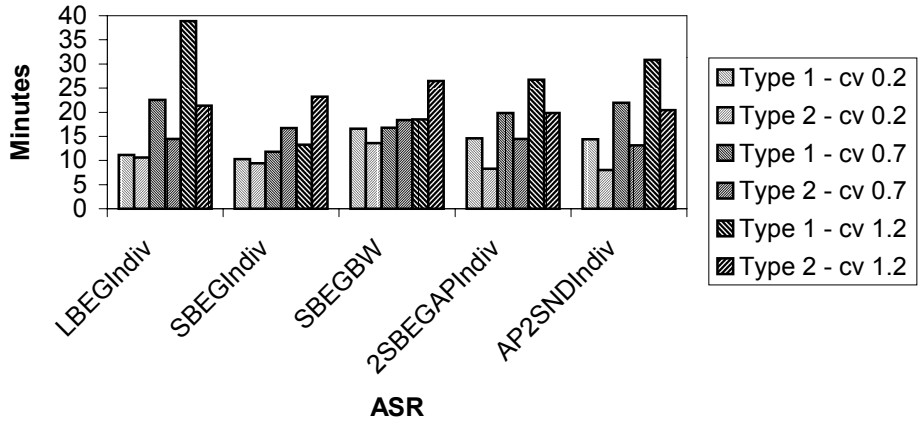

Fig. 6: Mean waiting time of patients with short (type 1) and long (type 2) treatment times

With respect to average waiting times, SBEGIndiv is the best rule. The weighted average waiting time is approximately 10 (C_v = 0.2), 15 (C_v = 0.7) and 19 minutes for high service time variability (C_v = 1.2). Additionally applying the BW rule on the first group leads to waiting times that are on average about 3 minutes longer but increases utilization of the doctor.
The two alternating rules show rather similar performance. For middle and high variabilities of service time, AP2SNDIndiv shows slightly longer waiting times for both types of patients. The differences in the waiting times between the two types of patients remain almost the same if two “short” patients are scheduled at the beginning or at the end of the session.
Therefore, scheduling short patients first is better than scheduling long patients first or alternating patient types. For further performance improvement the groups can be considered separately and rules like Bailey-Welch or “Two-at-a-time” can be applied and their performance studied.

5.2. Comparing rules during start-up period of clinic session

One of the interview partners reported high idle times of the physicians during the start-up period. This chapter will investigate the physician’s utilization during this period when different scheduling rules are applied. Proportional alternating scheduling showed a bad total performance. Proportional block wise scheduling will be covered by scenarios where only patients with the same expected treatment time are scheduled. Therefore, performance of proportional scheduling will not be considered in this section.
In Figure 7 the practitioner’s utilization is shown in the first half hour. For tabulated results see Appendix B, Table 5.

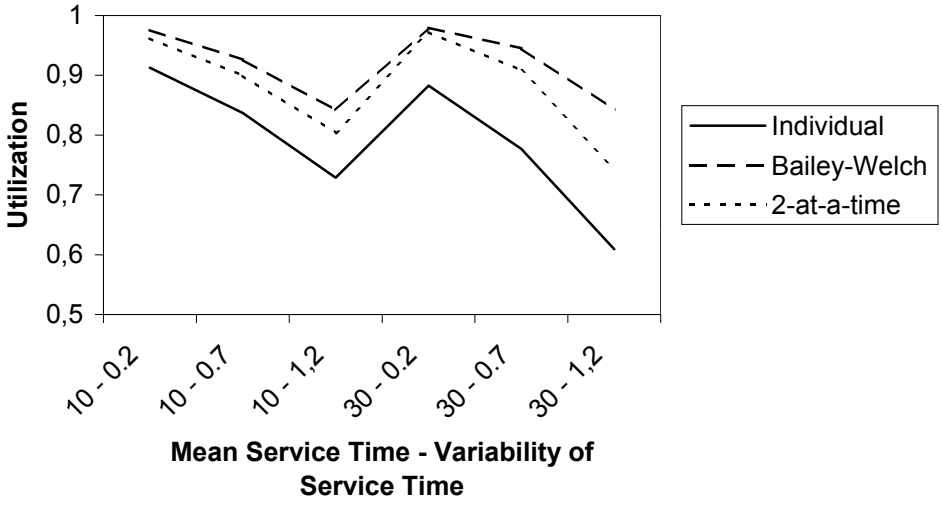

Fig. 7: Mean practitioner's utilization in the first half hour of clinic session

From Fig. 7, it is obvious that for all scheduling rules the utilization decreases for higher variability of the service time.
The Bailey-Welch rule is superior to the other rules, leading to the highest utilization for the doctor. However, it deteriorates for short service times with middle and high variance and for long treatment times with high variance (10 – 0.7, 10 – 1.2 and 30 – 1.2). For service times with high variability, the utilization reaches only about 85%. The difference between the Bailey-Welch and the "Two-at-a-time" rule becomes clear as the performance of the latter is several percent worse than the Bailey-Welch rule. Long treatment times with low variance pose an exception where the two rules show almost the same result. However, for the first appointment both rules are equal. For higher variance the Bailey-Welch rule is better because a "buffer" is built up due to the arrival scheme (Fig. 1).
The individual scheduling method improves its performance for low variability of service time. For long treatment times, this results in a difference of about 10%. On the other hand, it deteriorates for high variance. For the 30 – 1.2 combination, it leads to a utilization of only 61%. This means that in 61% of the sessions the doctor was busy in the first half hour. Compared to the results of the Bailey-Welch rule, the difference is about 23%. It is obvious that this rule should be avoided if the utilization of the practitioner is the main focus in the start-up period.
Further we can conclude that the results for the whole clinic session are in accordance to the results during the start-up period. This means that, for the physician's utilization, the "best" ASR during the start-up phase is also the "best" rule over the whole session, i.e. the Bailey-Welch rule. This also holds for the other ASRs studied.

6. Discussion

Like in previous approaches to this subject the impact of external factors like no-shows and variability of service time are studied in this paper. Additionally to these factors the punctuality of patients is taken into account. In all papers studied, it was assumed that patients arrive on time for their appointment. Therefore the results are not directly comparable to former results. However, certain results also hold under the conditions considered in this study. Like Ho and Lau (1992), one can conclude that the Bailey-Welch rule performs best if the practitioner's utilization is the main criterion for evaluation. This rule turns out to be robust under the different scenarios studied in this paper. The individual scheduling principle, which was studied here in its basic form, appeared to be the "best" method if a practice makes the patients' waiting time a top priority. The simulation results indicate that the "Two-at-a-time" rule as advocated by Soriano (1966) shows a "poor" performance for both the doctor's utilization and the patients' waiting time. The patient to whom the second appointment is given always has to wait as long as the treatment of the first patient which is clearly not a good strategy in order to minimize the patients' waiting time. But also with respect to the physician's productivity this rule is unsatisfactory. The "inventory" that is needed for a steady workflow for the doctor is not build up in a constant manner but is instead depleted after each "block" of two patients. This means that only at the scheduled arrival a "buffer" is produced but due to the fluctuations of the treatment times this is not sufficient for a high and constant doctor's productivity. Besides, the waiting time is not equally distributed among the patients. The second patient always has to wait longer than the first one. If both patients arrive on time this means that the second one has to wait the whole duration of the treatment of the first patient.
The simulation results are in accordance with the results of Klassen and Rohleder (1996). In the present study it appeared that it was better to schedule patients with a long treatment time at the end of the clinic session. In their paper, they conclude that patients with high standard deviation should be assigned to later appointments. Although we controlled the coefficient of variance, this implicitly affects the standard deviation and results can be compared.
Moreover, we showed that the performance further improves by combining different scheduling principles. This has not been considered in literature before.

The goal of this study is to obtain insight in the performance of various scheduling methods under realistic conditions. Therefore, also the punctuality of the patients was taken into account for the simulation model. According to waiting time studies held at an outpatient department for eye care, patients arrive on average about 10 minutes early (Wing's Consultancy, 1999). The time the patients arrive early is also assumed to be waiting time. On the one hand, one can argue that this time should not be considered as waiting time, which is the time between appointed time and actual start of the treatment. For the doctor, only scheduling rules are controllable, whereas patient behavior is out of his control. A certain waiting time can therefore never be avoided. On the other hand, patients satisfaction should also be taken into account as it is also important for the practitioner. The patient feels the time he arrived early as waiting time just as the time he has to wait after the appointed time. He or she will certainly be displeased if the waiting time exceeds 30 minutes even if he or she arrived 10 minutes early.

The busy time of the doctor was assumed to be concentrated on the patients, other activities as paperwork or telephone calls were not considered. But these tasks are, of course, closely related to the patients and necessary for operating a successful practice. In one of the practices the doctor is busy for one and a half hours per session (session length: 8 hours) writing letters to other practitioners, see section 2.5.3. for the interview. These activities can fill up gaps caused by no-shows or late arrivals, increasing the doctor's utilization. However, the main interest of this study was to optimize the scheduling to increase the efficiency of the direct contact with the patients. Another aspect could be at what time to do paperwork. This would also enable more reliable planning of these administrative tasks, e.g. at the end of the whole session or as an extension of treatment directly after seeing the patient.

7. Conclusions

This study has presented a new perspective on the patient scheduling problem in outpatient departments. Its aim was to explore the performance of different scheduling methods under realistic environmental conditions.
We have shown that scheduling patients with short expected treatment duration is the best scheduling principle if patients can be characterized according to their service times. Furthermore, the effect of combining scheduling rules was studied, i.e. proportional scheduling and Bailey-Welch ASR. Over all coefficients of variance of treatment time, this method led to the highest average physician's utilization. The difference in the waiting time between patients with long and short expected treatment duration can also be seen. However, the combination of proportional scheduling with Bailey-Welch leads to a fair distribution of waiting times between the two types of patients, while average waiting time of all patients increased only slightly. The shortest waiting times were achieved when patients were assigned to individual appointments.
In contrast to other studies, the utilization of the doctor during the start-up period of the clinic session was analyzed. We know that scheduling two patients at the start of the clinic insures that the doctor is busy during the the first half an hour in most of the sessions. Especially the Bailey-Welch rule guarantees a high utilization of the doctor under different coefficients of variance. Although the "Two-at-a-time" rule has the same arrival pattern in the considered period, it appears that this method yields worse performance for high variability of service times.
Next to the performance over the start-up period, also the whole clinic session was simulated. These experiments reveal that, if the variability of treatment durations increases, both patients and doctor tend to have longer waiting or idle times, respectively. For high variance (C_v = 1.2), both the "Two-at-a-time" and the Bailey-Welch rule lead to high waiting times for the clients. The "inventory" effect of the latter seems to diminish when the variance increases. The individual appointment rule is clearly superior in this aspect and is the rule at hand if the patients' waiting time is the main evaluation criterion. On the other hand, if the performance is measured from the practitioner's perspective the Bailey-Welch rule is the best method. This rule outperforms the other rules for all combinations of mean service duration and their variability. As the patients have to wait only few minutes longer for short treatments, this could be a very good solution for standard application. However, for long treatment durations the waiting times differ considerably which should be taken in order to improve the appointment scheduling system.

7.1. Suggestions for further research

Given the results of the work presented here, there are many interesting areas available for further research.
As outlined in Appendix A, also the practitioner's utilization taken on average over time is a possible performance measure that has not been taken into account in this work. One can expect that the performance differs when this measurement tool is used. Therefore it would be interesting to run the simulation experiments presented above again for this mean utilization.
It is felt that the waiting time increases for the later appointments when the schedule is done using the individual ASR. One could counter this if the interarrival times increase in the course of the consulting hours. Thus it could be worth while to study

this effect using simulation. Of course this represents a special case of a variable-interval scheduling system.

This study focused mainly on three appointment scheduling rules: the individual, the Bailey-Welch and the "Two-at-a-time" rule. When applying the proportional scheduling method, it was shown that it is better if first the patients with short expected treatment times are scheduled and then the patients with long expected service times. Next, we saw that combining the Bailey-Welch rule with proportional scheduling can further improve the results for the physician's utilization. Also combinations with other ASRs would be interesting but were not investigated. In Ho and Lau (1992), see section 2.4, a number of scheduling rules is described which could be studied in combination with the proportional scheduling method. Like for instance the variable-interval rule which is designed to reduce the patients' waiting time.

The findings from the interviews showed that walk-ins could be another environmental factor for a medical practice. Therefore, a simulation study including this factor could lead to different results than those presented here. But next to the fact that there are patients that arrive without appointment also the seasonal fluctuations of these arrivals could be taken into account. As described in section 2.5.5, the rate of walk-ins can double during the winter months. For a specialist in allergic diseases, these "seasons" can be different as some allergies break out according to the bloom of trees and bushes and others during midsummer when the insect bites can cause allergic shocks.

The assumption was made in section 2.1 that the service times are independent and identically distributed. However, it is felt that this assumption may not hold for all situations. One can think of a practice where the number of patients waiting influences the treatment times. The doctor could reduce the scheduled duration of the consults if the queue exceeds a certain threshold. Statistical data analysis would be necessary in order to examine the correctness of the assumption.

I may not have gone where I intended to go, but I think I have ended up where I intended to be.

Douglas Adams

Acknowledgements

Many people contributed to the success of this paper and I would like to mention some of them at this time.

In this sense I would like to thank Prof. Dr. G. M. Koole, the supervisor of this study, for his support and encouragement. I am very pleased to have worked with him and I learned a lot from his helpful remarks.

I would also like to thank Dr. van Rijn, Mrs. Feenstra - Verbruggen and the practice of Dr. Grewe, Dr. Hansi, Dr. Nissen, the practice Dr. Hammann, Dr. Klatt and the practice Dr. Kreuzer for their help. They either spent time themselves on an interview or made their receptionists available for it. Their experience and information made an important contribution to this study.

My special thanks go out to my dear parents for their inspiration and support. They shared their knowledge and experience with me and helped me to get in contact with several practices named above.

References

N. T. J. Bailey, 1952, “A study of queues and appointment systems in hospital outpatient department, with special reference to waiting times”, Journal of the Royal Statistical Society B, 14, pp. 185 – 199

N. T. J. Bailey, J. D. Welch, 1952, “Appointment systems in hospital outpatient departments”, Lancet, Vol. 259, Issue 6718, pp. 1105 - 1108

N. T. J. Bailey, 1954, “Queuing for medical care”, Journal of the Royal Statistical Society C, 3, 137 – 145

J. D. Welch, 1964, “Appointment systems in hospital outpatient departments”, Operations Research, Vol. 15, No. 3, pp. 224 – 232

A. Soriano, 1966, “Comparison of two scheduling systems”, Operations Research, Vol. 14, No. 3, pp. 388 - 397

C. -J. Ho, H. -S. Lau, 1992, “Minimizing total cost in scheduling outpatient appointments”, Management Science, Vol. 38, No.12, pp. 1750 - 1764

K. J. Klassen, T. R. Rohleder, 1996, “Scheduling outpatient appointments in a dynamic environment”, Journal of Operations Management 14, pp. 83 - 101

C. -J. Ho, H. -S. Lau, 1999, “Evaluating the impact of operating conditions on the performance of appointment scheduling rules in service systems”, European Journal of Operations Research 112, pp. 542 – 553

Wings Consultancy/Bedrijfscase BWI, project report, 1999, “Met het oog op wachttijden. Wachttijdensimulatie bij de polikliniek oogheelkunde van het Academisch Ziekenhuis Vrije Universiteit”

L. LaGanga, S. Lawrence, 2003, “Appointment scheduling and patient flow in a clinical practice”, http://ucsu.colorado.edu/~laganga/

V. Giacolone, 2003, “Common scheduling methods”, http://www.bsmconsulting.com/archives/CommonSchedulingMethodsPart2.pdf

J. W. Cohen, “The single server queue”, North Holland Series in Applied Mathematics and Mechanics, 8, 1982

W. D. Kelton, R. P. Sadowski, D. A. Sadowski, “Simulation with Arena”, 3rd ed., McGraw – Hill, 2000

Introduction to Discrete-event simulation http://www.dmem.strath.ac.uk/~pball/simulation/simulate.html

D. Adams, “A hitchhikers guide to the galaxy”, Ballantine Books, Reissue edition (September 27, 1995)

Appendices

Appendix A: Mean practitioner's utilization over time

Next to the utilization taken on average over all sessions one can also study the performance of a scheduling rule according to the utilization taken on average over the time. Suppose for example that two sessions are simulated with a length of three and two units of time, respectively. In the first and second session the doctor is busy during two unites of time. This results in an utilization of $^{2}/_{3}$ and 1, respectively. Taking the average utilization over the two sessions would give a MPU of $^{5}/_{6}$. On the other hand one could consider the utilization over the whole time simulated which is equal to five units of time. Then the doctor has been busy during four units, which results in a mean utilization of $^{4}/_{5}$.
As discussed in section 3.2, the average per session is the standard measure in the literature. However, renewal theory, a part of the theory of stochastic processes, focuses on the second view as performance measure. From the perspective of the practitioner this would have the advantage that not the busy time of a session is considered but the whole busy time over a certain period of time.
Statistically speaking, the second measure is the weighted average of the utilizations per session. It gives a better image of the performance as sessions with longer duration have a bigger influence on the average utilization than in the other case.
A session can be modeled to start right after the previous session ends. Then, the whole time the doctor is busy is the sum over all treatment times over the whole period. Formally this is given by

$$\sum_{k=1}^{M}\sum_{i=1}^{n_k} S_{i_k}$$

where M is the number of sessions and n_k is the number of patients of session k, $k=1,2.M$. According to the definition given in section 2.1, S_{ik} is the service time of patient i in session k. This is divided then by the actual length of the period. As we already saw in section 3.2 the length of a session is determined by the arrival time, the waiting time and the treatment duration of the last patient. In a formal way this can be written as

$$\sum_{k=1}^{M} \max\{A_{n_k}, A_{n_k-1} + S_{n_k-1}\} + W_{n_k} + S_{n_k}$$

where A_{nk}, W_{nk} and S_{nk} denote the arrival time, the waiting time and the service time of the last patient in session k, $k=1,2…M$, respectively. The utilization is the result of the division of the total busy time by the duration of the whole period.

Appendix B: Tabulated results of simulation experiments

Table 3

Appointment scheduling rule	Mean service time - variability of service time											
	10 - 0.2		10 - 0.7		10 - 1,2		30 - 0.2		30 - 0,7		30 - 1,2	
	MPU	MWT	MPU	MWT	MPU	MWT	MPU	MWT	MPU	MWT	MPU	MWT
Individual	0,9013	8,96	0,8645	13,905	0,8053	19,583	0,893	11,094	0,821	17,109	0,7457	27,865
BW	0,9409	13,213	0,9031	16,842	0,8454	22,974	0,9757	27,171	0,9114	29,472	0,8261	35,036
2-at-a-time	0,9241	10,845	0,8879	15,324	0,8219	23,197	0,9445	20,137	0,8765	25,143	0,787	34,871

Table 3: Simulation results for individual, Bailey-Welch and "Two-at-a-time" scheduling rules over the entire clinic session.
MPU: mean physician's utilization, MWT: mean patient's waiting time.

Table 4

Appointment Scheduling Rule	Variability of service time							
	0.2			0.7			1,2	
	MPU	MWT 1	MWT 2	MPU	MWT 1	MWT 2	MPU	MWT 1	MWT 2
LBEGIndiv	0,8813	11,128	10,627	0,8159	22,574	14,437	0,75	38,899	21,345
SBEGIndiv	0,9089	10,266	94,214	0,8498	11,833	16,729	0,7755	13,248	23,192
SBEGBW	0,9466	16,61	13,574	0,8829	16,764	18,389	0,8189	18,533	26,482

Table 4: Mean practitioner's utilization and mean waiting time for patients with short (10 min) and long (30 min) expected treatment times, scheduled according to the proportional scheduling method.
MPU: mean physician's utilization, MWT: mean patient's waiting time (1: short, 2: long treatment time)

Table 5

Appointment scheduling rule	Mean Service Time - Variability of Service Time											
	0:00 - 0:30 hours						0:00 - 0:30 hours					
	10 - 0.2		10 - 0.7		10 - 1,2		30 - 0.2		30 - 0.7		30 - 1,2	
	MPU	MWT	MPU	MWT	MPU	MWT	MPU	MWT	MPU	MWT	MPU	MWT
Individual	0,9127	92.099	0,837	84.932	0,7287	78.188	0,8828	93.021	0,7768	84.632	0,6089	76.421
BW	0,9757	14.916	0,9262	13.333	0,8416	12.296	0,9797	16.268	0,9452	14.496	0,8424	12.371
2-at-a-time	0,9618	12.791	0,8994	11.475	0,8036	10.727	0,9722	16.274	0,9081	14.863	0,7439	13,5

Table 5: Mean physician's utilization and mean patients' waiting time for the start-up period of the clinic session. MPU: mean physician's utilization, MWT: mean patient's waiting time.