In emergency situations where every second counts, the timely presence of an ambulance can be a matter of life or death. This importance justifies research to improve the logistics of Emergency Medical Services (EMS). This thesis introduces several models of EMS processes, and displays a variety of applications of Operations Research techniques to ambulance planning problems. Naturally, our research is focused on reducing response times. We deal with various planning stages in the EMS process, aiming to use a given number of resources (e.g., vehicles or personnel) in either a fair or efficient way. Some classical problems are viewed from a new angle, which leads to new theoretical insights as well as practically applicable solutions. The main contribution of this dissertation lies in the models and methods presented, verified by realistic case studies for a Dutch ambulance provider.

The first part of this thesis deals with EMS dispatching: deciding which ambulance to send to which incident. Many researchers and practitioners use the 'closest idle' policy without questioning it, but this is not necessarily optimal: instead, we could choose an ambulance such that remaining idle vehicles are in a good position with respect to expected incidents in the near future. In Chapter 2 we find such alternative dispatch policies using two methods: a MDP-based solution and a heuristic. The heuristic behaves similarly to the policy obtained from our MDP, but is more scalable. We validate both policies by simulating an urban EMS region and show a significant performance improvement when compared to the closest idle method. This sheds new light on the popular belief that the closest idle policy is near-optimal. Although we do not advise all EMS managers to immediately discard the closest idle dispatch method, we do show that the typical argument – that it would not lead to large improvements in the fraction of late arrivals – should be changed.

While the displayed dispatch policies in Chapter 2 clearly outperform the closest idle policy, the optimal policy remains unknown. Therefore, we continue in Chapter 3 by providing a bound on the performance of an optimal dispatch policy. This is done by introducing a benchmark model (referred to as the offline dispatch model): deciding which ambulance to dispatch when all incidents are known in advance. We show how to calculate the optimal offline dispatch decisions, and the corresponding performance serves as a bound for any - including the optimal - online policy. We perform a worst case analysis which shows that the so-called competitive ratio of the dispatch problem is unbounded; that is, even an optimal online dispatch algorithm can perform arbitrarily bad compared to the offline solution. However, when we consider the average case, the gap between existing solutions and the offline optimum turns out to be much smaller: a case study
for a large ambulance provider in the Netherlands shows that the closest idle policy obtains a fraction of late arrivals that is approximately 2.7 times that of the optimal offline policy. What is perhaps most surprising is that our dispatch heuristic from Chapter 2 manages to reduce this gap to approximately 1.9, i.e., it closes roughly half of the gap between ‘closest idle’ and the offline optimum. This work constitutes the first quantification of the gap between online and offline dispatch policies.

Chapter 4 considers dynamic ambulance repositioning: proactively relocating idle vehicles in order to reduce response times. When an ambulance completes service for a patient, we allow it to be sent to one of the existing base locations. In order to compute where to send these idle vehicles, we propose a heuristic that scales to large EMS regions with many vehicles. Simulations show that this method significantly improves the fraction of late arrivals compared to the scenario in which each vehicle always returns to its home base. Furthermore, not only the performance at the response time threshold is improved, but the whole distribution of response times is shifted to the left. As our method is intuitive and easy to implement, it also serves as a suitable base for extensions. The practical relevance of this heuristic was demonstrated by the implementation in a decision support tool used by the EMS region Flevoland, the Netherlands.

Chapters 5 and 6 introduce several models to improve the fairness in ambulance logistics. Rather than simply maximizing the number of people served, we consider the distribution over the different areas where people live. To that end, we view ambulance optimization models from a social welfare perspective. We analyze existing ambulance planning models and show that they tend to maximize either the number of people served (called utilitarian social welfare) or maximize the service to the person who is worst off (called egalitarian social welfare). We propose a third option: the so-called Bernoulli-Nash social welfare. In Chapter 5, a new facility location model is introduced. This allows us to compute where to open ambulance bases and how to distribute vehicles over those bases, such that the Bernoulli-Nash social welfare is maximized. In several case studies we compare our Bernoulli-Nash optimal solution with the often-used utilitarian optimum. In Chapter 6 we take a different approach: we argue that classical ambulance planning models may have several near-optimal solutions. These have a similar overall performance but differ on a smaller scale, such as individual villages. We propose to avoid the ‘arbitrary’ choice in terms of who gets coverage and who does not, by sharing time between several good ambulance configurations. We formulate an optimization model that computes the time shares such that, again, the Bernoulli-Nash social welfare is maximized. In this chapter we use a combination of simulation and optimization.

Chapter 7 considers a stochastic machine scheduling problem: the scheduling of jobs for which the processing times are not known in advance. Instead, the processing times are governed by independent exponentially distributed random variables. In particular, we analyze the performance of the Weighted Shortest Expected Processing Times first (WSEPT) rule - also known as Smith’s rule -
for minimizing the expected weighted sum of completion times. In this setting, WSEPT has a known upper bound of \((2 - 1/m)\), and in this chapter we prove the first lower bound to be 1.243. This result is particularly surprising when juxtaposed with the deterministic counterpart of this problem: there, Smith’s rule is known to be a \(\frac{1}{2}(1 + \sqrt{2})\) – approximation. Note that 1.243 > \(\frac{1}{2}(1 + \sqrt{2}) \approx 1.207\), hence our result indicates that stochastic scheduling with exponentially distributed processing times has worse worst-case instances than deterministic scheduling.