Appendix A: Notation

Sets	Description
$\mathcal{N}$	Set of ambulances ($n \in \mathcal{N} = \{1 \dots N\}$)
$\mathcal{C}$	Set of ambulance categories ($c_n \in \mathcal{C}\ \forall n \in \mathcal{N}$)
$\mathcal{I}$	Set of incident types ($i \in \mathcal{I}$)
$\mathcal{K}$	Set of infection types ($k \in \mathcal{K}$)
$\mathcal{I}_c$	Set of incident types served only by ambulance category $c \in \mathcal{C}$
$\mathcal{A}$	Set of all possible combinations ($\mathcal{A} = \{\{A_1, ..., A_{\lvert\mathcal{C}\rvert}\} : \sum_{c \in \mathcal{C}} A_c = N$)
$\mathcal{J}$	Set of nodes ($j \in \mathcal{J}$)
Parameters	**Description**
$\mathcal{A}_c$	Number of ambulances of category c
λ_{ikj}	Arrival rate of incident type i with infection type k at node j ($\lambda_{ij} = \sum_{k \in \mathcal{K}} \lambda_{ikj}$)
d_{cj}	Share of incidents at node j that can be served by ambulance category c
$l(n)$	Location of ambulance n
τ_{ikjn}	Service time of ambulance n for incident of type i and infection type k at node j
$\tau_{l(n)j}$	Driving time from location of ambulance n to node j
τ_D	Dispatching time
t^D	Time threshold for driving
t^R	Time threshold for response
Θ	Cutoff level applied to ambulance reservation strategy
γ_{ic}	Indicator parameter: 1 if ambulance category c can serve incident type i, else 0
1^{st}-stage Variables	
a_{ijr}	r^{th} preferred ambulance of incident type i at node j
r_{ijn}	Rank of ambulance n for incidents of type i at node j
c_n	Category of ambulance n
y_{in}	Indicator variable: 1 if ambulance n serves incident type i, else 0
2^{nd}-stage Variables	
ρ	Average system utilization
ρ_n, ρ'_n	Fraction of time ambulance n is busy
b_{nm}	Number of preference lists in which ambulance m is direct backup for ambulance n
ω_{nm}	Workload shifted from ambulance n to ambulance m
f_{ijn}, f'_{ijn}	Probability: ambulance n is dispatched to incident type i at node j
g_{ijn}	Probability: ambulance n is dispatched to an unqueued incident of type i at node j
ζ^D	Share of arrivals for which driving time exceeds t^D
ζ^R	Share of arrivals for which response time exceeds t^R
P_v	Probability: exactly v ambulances are busy
$P(B_n)$	Event: ambulance n is busy
$P(F_n)$	Event: ambulance n is idle
$P(B_{ikjn})$	Probability: ambulance n is busy with incident and infection type i and k at node j
P^I_n	Infection probability of ambulance n
$\bar{P}^I$, ($\bar{P}^I_c$)	Mean infection probability (per ambulance category c)
$\hat{P}^I$, ($\hat{P}^I_c$)	Maximum infection probability (per ambulance category c)
$Q(N, \rho, j)$	Correction factor to account for independent ambulances
r	Average response time
d	Average driving time
w	Patients' average waiting time in queue before being served
n_q	q^{th} percentile of the queuing time distribution

Table 6 Notation: Sets, Indices and Parameters

Appendix B: Derivation of the dispatching probabilities

We approximate the probability of dispatching ambulance n as the r^{th} favored ambulance to an unqueued emergency call of incident type i at node j, denoted by g_{ijn}, by assuming an $M/M/N/\infty$ queuing system in which we randomly draw ambulances from the set of ambulances $\mathcal{N}$ without replacement. Ambulance n is dispatched in the case it is idle, denoted by F_n, and all better-ranked ambulances are unavailable. $B_{a_{ijr}}$ denotes that the r^{th} ranked ambulance, for incident type i at node j, is busy. r_{ijn} is the rank of ambulance n for incident type i at node j. Thus, we extend Larson's (1975) AHQM by having preference lists not only dependent on node j, but also on incident type i (Jarvis 1985, Larson 1975).

$$g_{ijn} = P\left(F_n \cap \left\{\bigcap_{l=1}^{r_{ijn}-1} B_{a_{ijl}}\right\}\right) \qquad \forall i \in \mathcal{I}, j \in \mathcal{J}, n \in \mathcal{N} \tag{29}$$

where $P(E)$ denotes the probability of event E.

Ambulances do not operate independently. On the contrary, the EMS operator considers the status of each ambulance at the time of dispatching. Thus, Larson (1975) introduces a correction factor Q to adjust the individual ambulance workloads. The value of Q is derived by applying laws of probability to the calculation of g_{ijn} in (29). Further, Q depends on the average system utilization ρ, the number of ambulances N, and the ambulance's rank j in the preference list that was drawn from set $\mathcal{N}$. As we consider an $M/M/N/\infty$ queuing system, the system's utilization ρ is $\lambda\tau/N$ (Tijms 2003, pp. 187-188, Jarvis 1985, Larson 1975).

$$Q(N,\rho,j) \equiv \sum_{k=j}^{N-1}\left(\frac{(N-j-1)!\,(N-k)}{(k-j)!}\right)\cdot\frac{N^k}{N!}\rho^{k-j}\left[(1-\rho)\sum_{i=0}^{N-1}\left(\frac{N^i\rho^i}{i!}\right)+\frac{N^N\rho^N}{N!}\right]^{-1} \tag{30}$$

The probability that ambulance n is busy, $P(B_n)$, corresponds to its workload ρ_n, stated in (31). Thus, the probability that ambulance n is idle, $P(F_n)$, is calculated by $(1-\rho_n)$. In the case that the first drawn ambulance n is the most preferred ambulance such that it is ranked at position $j=0$ in the preference list, the correction factor $Q(N,\rho,0)$ is 1, defined in (32). Based on these assumptions, we approximate the probability that ambulance n is dispatched as the r^{th} favored ambulance to an unqueued emergency call of type i at node j in (33) by inserting the definitions of $P(B_n)$ and $P(F_n)$ in (29). Thus, we multiply the availability factor $(1-\rho_n)$ of ambulance n by the workloads of all better ranked ambulances (ρ_n) and amend the result by the correction factor Q (Jarvis 1985, Larson 1975). To account for the different incident types in the pandemic context, we consider incident-dependent preference lists.

$$P(B_n) = \rho_n \qquad \forall n \in \mathcal{N} \tag{31}$$

$$P(F_n) = 1-\rho_n \Leftrightarrow Q(N,\rho,0) = 1 \qquad \forall n \in \mathcal{N} \tag{32}$$

$$g_{ijn} \approx Q\left(N, \rho, r_{ijn} - 1\right)\left(1 - \rho_n\right) \prod_{l=1}^{r_{ijn}-1} \rho_{a_{ijl}} \qquad \forall i \in \mathcal{I}, j \in \mathcal{J}, n \in \mathcal{N} \tag{33}$$

We approximate the probability of an ambulance being dispatched to a queued call by P_N/N, i.e., in the case that all ambulances are busy, each ambulance has an equal probability of becoming idle and dispatched to a queued call. We add this term to the probability that ambulance n is dispatched to an unqueued call, g_{ijn}, to obtain the probability that ambulance n is dispatched to any incident of type i at node j, denoted as f_{ijn}. We refer to f_{ijn} as "dispatching probability".

$$f_{ijn} = Q\left(N, \rho, r_{ijn} - 1\right)\left(1 - \rho_n\right) \prod_{l=1}^{r_{ijn}-1} \rho_{a_{ijl}} + \frac{P_N}{N} \qquad \forall i \in \mathcal{I}, j \in \mathcal{J}, n \in \mathcal{N} \tag{34}$$

Appendix C: Ambulance Workloads and Infection Probabilities

To obtain further insights in the distribution of workloads among ambulances, we present the individual workloads per ambulance in Figures 6 and 7. Each circle represents an ambulance. The circle's location corresponds to the ambulance's depot location. The radius of the circle represents the ambulance's workload. Figure 6 presents the workloads observed for the Covid-19 case study examined in Section 6.2 when applying *no split*. The average workload among all ambulances amounts to 38.50% corresponding to the time the ambulances are busy serving incidents. The workloads range from 22.59% to 57.07% with a standard deviation of 9.76%. Figure 7 shows the results observed for the Covid-19 case study presented in Section 6.2 for the optimal *fixed split*. The *fixed split* results in a slightly higher average workload of 39.78% with a higher standard deviation of 14.14% compared to applying *no split*. We observe the lowest workload of 12.68% for three ambulances co-located at the same depot serving only unsuspicious cases. The highest workload of 72.95% is observed for an ambulance dedicated to suspected and known cases which underlines the superiority of not splitting. Such a high workload can be explained by the high infection risk posed to personnel operating in ambulances assigned to this group of patients. Noticeable are the differences in workloads among ambulances of the same category. In the city centre, ambulance workloads are rather low. The workload of ambulances serving suspected and known cases differ between 32.09% and 72.95%. We observe workloads between 12.68% and 57.90% for ambulances dedicated to unsuspicious cases. These results indicate that the calculation of dispatching probabilities is paramount to obtain realistic approximations for the ambulances' workloads and infection probabilities and, consequently, the system's performance.

Figure 8 presents the mean infection probabilities observed for different *flexible splits*. The results show that the relationship between the number of ambulances per category and the mean infection probability is non-linear and not convex. For a high number of ambulances serving suspected and known cases we observe the smallest difference in the mean infection probabilities comparing the

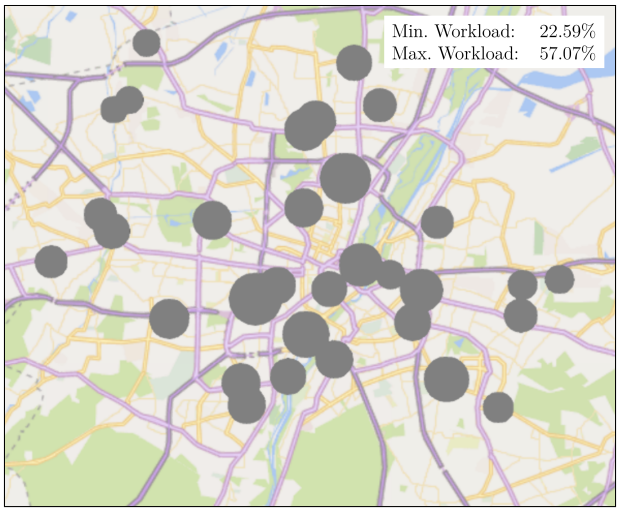

Figure 6 **Covid-19: Comparison of Ambulance Workloads (No/Flexible Split)**

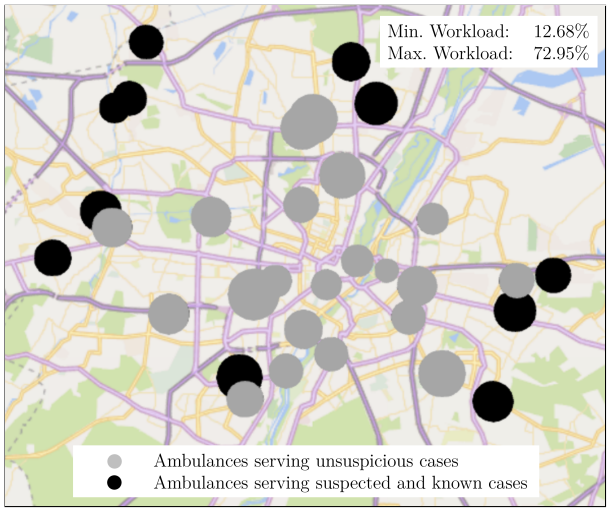

Figure 7 **Covid-19: Comparison of Ambulance Workloads (Fixed Split)**

two ambulance categories. One reason for this observation could be that ambulances dedicated to suspected and known cases often serve as backup ambulance for a high load of unsuspicious cases such that the infection probabilities are more equally distributed among all ambulances. In the case that more than five ambulances are dedicated to unsuspicious cases, the ambulance category with less ambulances faces a higher mean infection probability. The more ambulances are assigned to a category, the lower is its mean infection probability. This can be explained as the risk of infection is distributed among more ambulances. Figures 9 and 10 present the mean infection probabilities for all valid *fixed splits* which do not result in an overloaded system. Comparable to *no split*, there is no linear relation between the number of ambulances per category and its mean infection probability. For both ambulance categories, the mean infection probability decreases when increasing the number of assigned ambulances indicating that the total risk of infection is distributed among more ambulances.

Appendix D: Additional Performance Measures and Experimental Results

To evaluate the performance of an EMS in detail, we analyze additional performance measures.

Computation of the Percentiles of the Waiting Time Distribution

We investigate the time patients have to wait in the queue before an ambulance becomes idle and is dispatched. Therefore, we introduce additional performance measures to obtain more information about the queuing time distribution. Thus, we calculate the qth percentiles of the queuing time distribution, denoted by η_q. q denotes the steady-state probability that a delayed patient has to wait less than η_q time units (Tijms 2003, p. 199).

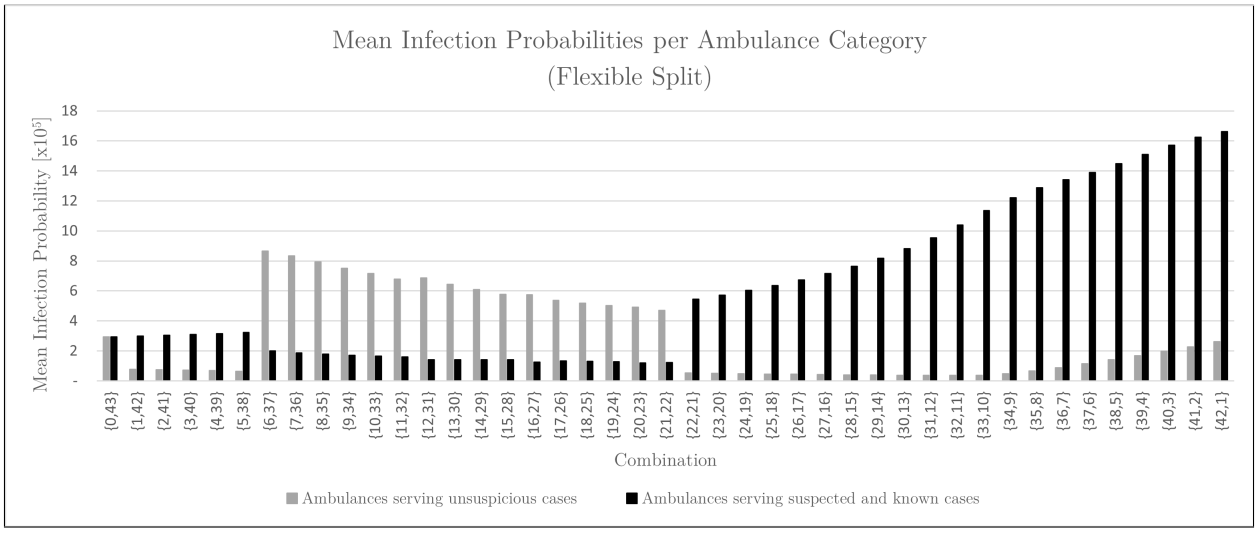

Figure 8 **Covid-19: Mean infection probabilities per combination (Flexible Split)**

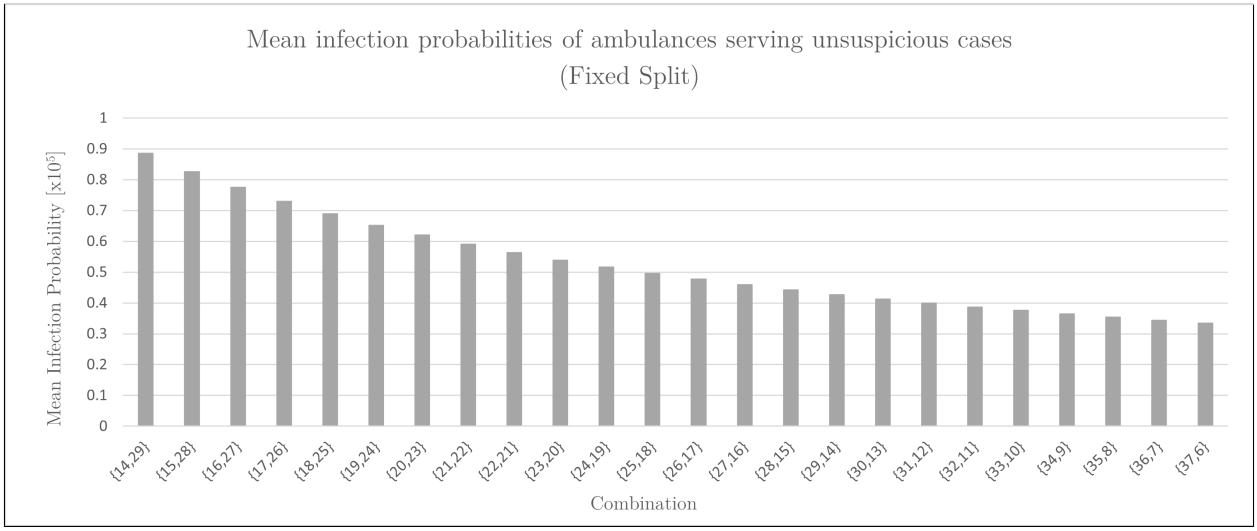

Figure 9 **Covid-19: Mean infection probabilities per combination for ambulances serving only unsuspicious cases (Fixed Split)**

$$\eta_q = \frac{-ln(1-q)\tau}{N(1-\rho)} \tag{35}$$

We calculate η_q of the waiting time distribution $\forall q \in \{0.25, 0.50, 0.75\}$.

Computation of the Maximum Infection Probability

In the case study, when applying a *fixed split*, we observe significant differences in the mean infection probabilities for ambulances and their personnel dependent on the patient categories the ambulance is allocated to. Thus, we are interested in the highest infection risk faced by an ambulance, denoted

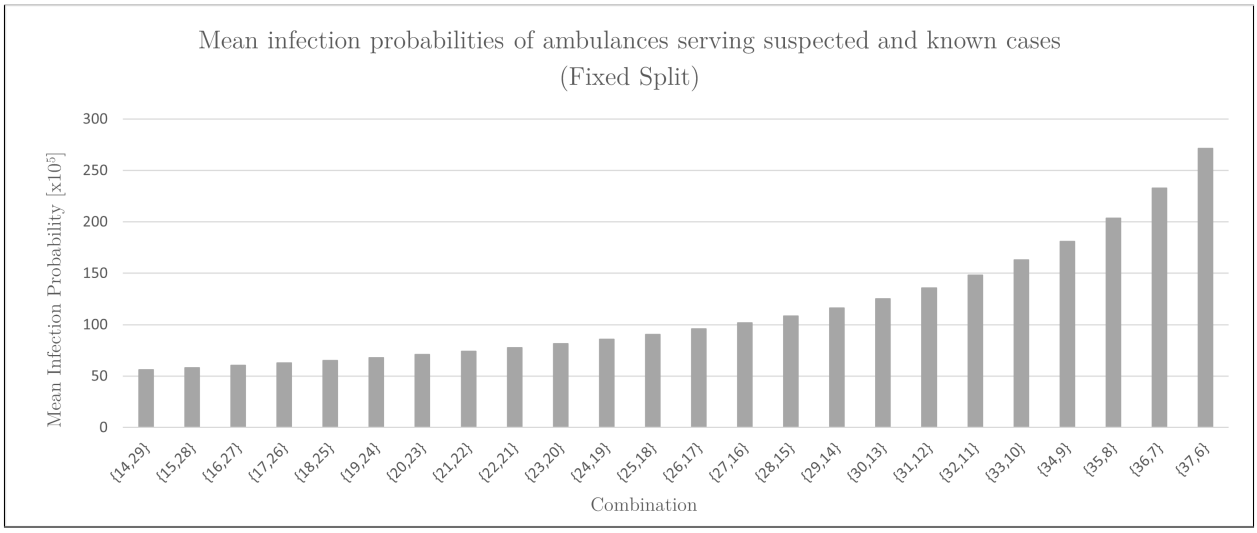

Figure 10 **Covid-19: Mean infection probabilities per combination for ambulances serving only suspected and known cases (Fixed Split)**

as $\hat{P}^I$. Here, we make use of the ambulances' individual infection probabilities, P_n^I, and return the highest observed value.

$$\hat{P}^I = \max_{n \in \mathcal{N}} \{P_n^I\} \tag{36}$$

To compare the maximum risk for an ambulance to be taken out of service for the different ambulance categories $c \in \mathcal{C}$, we determine the maximum infection probability $\hat{P}^I$ for each ambulance category, separately:

$$\hat{P}_c^I = \max_{n \in \mathcal{N}: c_n = c} \{P_n^I\} \qquad \forall c \in \mathcal{C} \tag{37}$$

Supplementary results

We summarize the results derived from the numerical case study for an Ebola, Covid-19 and Influenza A outbreak based on data of Munich in Table 7. The average response time, the average driving time and the average patients' queuing time are denoted by r, d and w, correspondingly. The share of late arrivals, ζ^D and ζ^R, present the percentage of incidents for which the driving time threshold t^D or response time threshold t^D is exceeded. $\bar{P}^I$ is the mean infection probability over all ambulances $n \in \mathcal{N}$. $\bar{P}^I_{\{U\}}$ denotes the mean infection probability for ambulances allocated to unsuspicous cases, $\bar{P}^I_{\{S,K\}}$ the mean infection probability for ambulances designated to serve suspected and known cases.

The results have partially been described and analyzed in Section 6.2 for Covid-19 and Section 6.4.2 for Ebola and Influenza A. Thus, we concentrate on the performance measures n_q, $\hat{P}^I$ and $\hat{P}_c^I$ introduced in this section. For the three data instances, applying a *fixed split* approximately

	Covid-19		Ebola		Influenza A	
	Flexible & No Split	Fixed S.	Flexible & No Split	Fixed S.	Flexible & No Split	Fixed S.
Combination		{32,11}		{32,11}		{28,15}
r [min]	7.27	8.72	7.38	9.06	7.80	9.87
d [min]	3.50	4.85	3.62	4.89	4.03	5.57
w [min]	0.00	0.10	0.00	0.41	0.02	0.21
ζ^R [%]	0.39	7.03	0.53	9.29	1.32	13.25
ζ^D [%]	0.39	6.34	0.53	6.56	1.32	9.53
$\eta_{0.25}$ [min]	0.30	0.61	0.31	0.59	0.32	0.57
$\eta_{0.50}$ [min]	0.72	1.46	0.74	1.41	0.76	1.37
$\eta_{0.75}$ [min]	1.44	2.92	1.49	2.82	1.52	2.75
$\bar{P}^I_{\{U\}}$ [‰]		0.00		0.00		0.01
$\bar{P}^I_{\{S,K\}}$ [‰]		1.48		1.27		3.70
$\bar{P}^I$ [‰]	0.03	0.38	0.03	0.33	0.10	1.30
$\hat{P}^I_{\{U\}}$ [‰]		0.01		0.01		0.02
$\hat{P}^I_{\{S,K\}}$ [‰]		2.18		1.78		4.98
$\hat{P}^I$ [‰]	0.04	2.18	0.04	1.78	0.13	4.98

Table 7 Case Study results applying a flexible, fixed and no ambulance split

doubles the 25%-, 50%- and 75%-percentiles of the waiting time distribution. This can be explained by the higher average waiting times observed for the *fixed split*. To give an example for the Covid-19 instance, with a probability of 75%, patients wait less than 2.92min when applying the *fixed split*. In the case of *no split*, patients need to wait less than 1.44min until an ambulance is dispatched. These results underline the drawbacks from applying a *fixed split* for patients. However, for all data sets, the *fixed split* is beneficial for the share of personnel designated to unsuspicious cases as the mean infection probabilities are lower than when applying *no split*. Nevertheless, personnel serving suspected and known cases face higher mean infection probabilities for the *fixed split*. We observe comparable results for the maximum infection probabilities. While personnel serving unsuspicious cases benefit from a *fixed split*, personnel serving suspected and known cases face the highest infection risk.