Omega 116 (2023) 102801

Contents lists available at ScienceDirect

Omega

journal homepage: www.elsevier.com/locate/omega

Dynamic fair balancing of COVID-19 patients over hospitals based on forecasts of bed occupancy[☆]

Sander Dijkstra, Stef Baas, Aleida Braaksma[*], Richard J. Boucherie

Center for Healthcare Operations Improvement and Research (CHOIR), University of Twente, Enschede, the Netherlands

ARTICLE INFO

Article history:
Received 13 September 2021
Accepted 4 November 2022
Available online 16 November 2022

Keywords:
COVID-19
Patient allocation
Queueing theory
Load balancing
Stochastic program
Bed occupancy

ABSTRACT

This paper introduces mathematical models that support dynamic fair balancing of COVID-19 patients over hospitals in a region and across regions. Patient flow is captured in an infinite server queueing network. The dynamic fair balancing model within a region is a load balancing model incorporating a forecast of the bed occupancy, while across regions, it is a stochastic program taking into account scenarios of the future bed surpluses or shortages. Our dynamic fair balancing models yield decision rules for patient allocation to hospitals within the region and reallocation across regions based on safety levels and forecast bed surplus or bed shortage for each hospital or region.

Input for the model is an accurate real-time forecast of the number of COVID-19 patients hospitalised in the ward and the Intensive Care Unit (ICU) of the hospitals based on the predicted inflow of patients, their Length of Stay and patient transfer probabilities among ward and ICU. The required data is obtained from the hospitals' data warehouses and regional infection data as recorded in the Netherlands.

The algorithm is evaluated in Dutch regions for allocation of COVID-19 patients to hospitals within the region and reallocation across regions using data from the second COVID-19 peak.

1. Introduction

Confronted with a pandemic or the outbreak of a severe and highly contagious disease on a national level, governments and organisations must implement appropriate countermeasures [1]. Accurate estimation of disease prevalence is essential for monitoring and decision making [2], as is optimal distribution of vaccines [3,4]. Despite these efforts, hospitals may be overwhelmed by infected patients. These patients arrive in addition to hospitals' regular patients, increasing the strain on hospital staff and resources [5,6]. Alternative resources such as backup and field hospitals or student nurses may offer additional capacity [7]. Despite such measures, hospitals may have no other option than to temporarily decrease the number of regular patients treated [8,9]. A reduction in regular care has serious consequences, in particular for oncology patients and others whose condition may worsen irreversibly if treatment is postponed [10], but also in other medical specialties healthy life years are lost due to fewer treatments [9].

Hospitals may defer infected patients to other hospitals facing a less severe surge in infected patients, either inside or outside their region, to avoid exceeding their maximum capacity [11]. Next to being unavoidable in certain cases, redistributing patients may also balance the staff work pressure across regions and may avoid potentially unethical differences in the accessibility of regular care across regions. Redistributing infected patients according to a fair balancing policy provides the opportunity to share and thus reduce the burden of the pandemic on regular patients as well as hospital staff.

When the COVID-19 pandemic reached the Netherlands, the Dutch government erected a national coordination centre for patient reallocation (in Dutch: 'Landelijk Coördinatie-centrum Patiëntenspreiding', LCPS) with exactly these aims [12]. To fulfil its mission, LCPS cooperates with the twelve 'ROAZ' regions of the country, where each ROAZ (in Dutch: 'Regionaal Overleg Acute Zorgketen') region has its own consultative body for the acute care chain [13]. When a hospital requests to reallocate one of its COVID-19 patients, other hospitals within the same region are considered first, as an intra-regional reallocation is the least burdensome for the patient, his or her relatives, and the ambulance transportation service. If an intra-regional reallocation is not possible, the patient is reallocated to another region in the Netherlands, or to Germany

[☆] Area - Production Management, Scheduling and Logistics This manuscript was processed by Associate Editor Singh.

[*] Corresponding author.

E-mail address: a.braaksma@utwente.nl (A. Braaksma).

https://doi.org/10.1016/j.omega.2022.102801
0305-0483/© 2022 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

as a last resort. As outlined in Bekker et al. [14], reallocation of patients is based on separate forecasts of the COVID-19 bed occupancy at the ward and Intensive Care Unit at the regional level, but neither a detailed model to fairly balance patients across regions taking into account these predictions, nor a model to fairly allocate patients to hospitals within a region taking into account bed occupancy predictions per hospital are included.

To facilitate the reallocation process, hospitals are required to report their bed surplus that is available for COVID-19 ward and ICU patients daily at 10 AM. Physicians report the bed surplus by observing the current situation and likely incorporating some form of safety margin. Bed surplus and bed shortage are influenced by several factors. The surplus of COVID-19 ward beds, for example, increases due to discharges and transfers to the ICU, while it decreases by transfers from the ICU, possible reallocations from other hospitals and new admissions, where the latter depends on the number of COVID-19 infections in the region some days ago. It is infeasible for a physician to gauge and incorporate the combined effect of all these factors when reporting the bed surplus. Consequently, the number of patient reallocations will likely be higher than necessary. For example, a hospital may report ICU bed surplus today and thus receive a COVID-19 ICU patient from another hospital, while that results in ICU bed shortage two days from now, necessitating to reallocate a COVID-19 ward patient that needs ICU care.

This paper presents mathematical models and resulting decision rules that support fair balancing of COVID-19 patients over hospitals in a region and across regions. In these models, the flow of COVID-19 patients is captured in a network of infinite server queues. The first model, at the regional level, is a load balancing model that supports dynamic fair balancing of COVID-19 patients over hospitals in a region. The second model, at the inter-regional level, is a stochastic program that minimises the costs of patient reallocations across regions. Input for the models is the inflow of patients, their Length of Stay (LoS) in the ward and ICU and transfer of patients between these units. To this end, our method is augmented by accurate statistical methods to predict patient arrivals, estimate LoS and transfer probabilities, and forecast the number of COVID-19 patients hospitalised in the ward and ICU of a hospital. Our results are cast in real-time decision rules for patient allocation to hospitals in a region or reallocation across regions based on safety levels that determine the bed surplus or bed shortage for each hospital or region during the next couple of days.

1.1. Literature

Dynamic load balancing Our fair balancing of COVID-19 patients over hospitals within a region falls in the class of load balancing methods that are well-known in communication systems, see, e.g., Ross [15], Zachary and Ziedins [16], van der Boor et al. [17] for an introduction to basic load balancing scenarios, that states: “Load balancing can be broadly categorised as static, dynamic, or some intermediate blend, depending on the amount of state information that is taken into account”. Below, in line with our approach, we focus on dynamic load balancing. Dynamic load balancing algorithms aim at improving the system throughput and reducing the job response time by relocating application tasks among the nodes using information on the instantaneous system load to decide how to relocate the jobs [18]. Sender-initiated strategies (congested nodes push work to lightly loaded nodes) outperform receiver-initiated strategies (lightly loaded nodes pull work from highly loaded nodes) at light to moderate system loads, whereas receiver-initiated strategies are preferable at high system loads [18,19]. Observe that dynamic load balancing may be hindered by incomplete state information [20,21]. Exact performance analysis of dynamic load balancing policies is argued to be difficult due to multidimensional state spaces, see [22,23].

Dynamic load balancing algorithms are developed for systems with stationary arrival and service rates, including web servers [24–26] and large (virtual) call centres [27,28]. Typically, these methods involve large Erlang loss systems and servers with multiple skills as in call centres, involve a queue as in web servers, or introduce replicas of jobs that are sent to join the queue at different servers and upon completion of service of the first replica delete all other replicas. Typical ingredients for load balancing approaches are setting and adjusting routing probabilities [29] or routing policies [22,30,31] and job migration algorithms [32]. The Join-the-Shortest-Queue (JSQ) policy is a centralised dynamic load balancing algorithm, where a dispatcher must immediately forward tasks upon arrival to one of the servers. Implementation of JSQ policies becomes difficult when the number of stations becomes large, in which case asymptotic methods may be used [17,20,21,33].

Our dynamic load balancing method is developed for systems under high load that do not allow for queueing, in which case a receiver-initiated strategy is preferable, which is implemented via a centralised approach to allocate patients to hospitals. We assume complete state information and use simulation to assess performance measures. Our fair balancing method across regions is a stochastic program taking into account forecasts of the future bed surplus or shortage. It is related to mixed integer recourse models [34, Chapter 3]. Typically, solving such models requires scenarios or pre-specified input processes and does not include dynamical statistical methods for real-time forecasts of occupancy and safety levels which is included in our approach. We further show that our stochastic program may be well approximated by a mixed integer program (MIP) that facilitates a direct relation with the way hospitals report bed shortages/surpluses.

Forecast and queueing model Our fair balancing method requires accurate forecasts of the patient arrival rates. Several studies have developed prediction models for the number of hospitalised COVID-19 patients. Focusing solely on predicting ICU occupancy, Farcomeni et al. [35], Goic et al. [36], Manca et al. [37], Massonnaud et al. [38] develop prediction models at the regional level, while [39] provides predictions for individual hospitals. Other studies predict both COVID-19 ward and ICU occupancy [14,40–43]. The prediction in Roimi et al. [42], Zhao et al. [43] is based on regression analysis or epidemic models. Queueing models for predicting ward and ICU occupancy in the Netherlands at the regional and national level are developed in Bekker et al. [14]; their models do not incorporate patient transfers from ward to ICU and vice versa. Transfer probabilities between COVID-19 ward and ICU are derived based on a Markov chain analysis in Foucrier et al. [41]. In a previous paper, Baas et al. [40], we have developed forecasts of COVID-19 ward and ICU occupancy at the individual hospital level, incorporating patient transfers between the ward and ICU. That method uses a Richards’ curve [44] to predict the arrival rates of COVID-19 patients, a Kaplan–Meier estimator [45] to estimate the distribution of the LoS in both the COVID-19 ward and ICU, and we sample patient trajectories in the Poisson Arrival Location Model [46] that determines the queue occupancy in a network of infinite server queues representing the COVID-19 ward and ICU. In this paper, we build on our previous work by significantly improving the forecasts of ward and ICU occupancy and by using these forecasts as a basis for decision rules that facilitate fair balancing of COVID-19 patients over hospitals.

1.2. Contribution

Our contribution in this paper is threefold. First, we develop a load balancing method that incorporates bed occupancy forecasts to fairly balance COVID-19 patients over hospitals in a region. These results extend load balancing results in literature to incorporate forecast occupancy and safety levels. Second, we propose a stochastic program taking into account scenarios of the future bed surpluses or shortages to optimally distribute COVID-19 patients that cannot be accommodated within a region over multiple regions taking into account travelling distances and other differences between regions; the currently available COVID-19 bed-capacity in each region; as well as scenarios of the maximum bed-occupancy over several days. Third, as our models require accurate forecasts of the COVID-19 patient arrival rate, we extend the results of Baas et al. [40] on the prediction of the arrival rate in the network of infinite server queues that we use in the bed occupancy forecasts. Originally, in Baas et al. [40], a Richards' curve was fit to data from the hospitals' data warehouse to predict the COVID-19 patient arrival rate, while in this paper we develop a time-delaying and filtration procedure applied to the exponentially weighted moving average of regional infection data, which results in more accurate bed occupancy forecasts.

Section 2 presents our hierarchical modelling and decision approach. We propose decision rules for individual hospitals in Section 2.1, within a region in Section 2.2, and across regions in Section 2.3. Our load balancing procedure is based on arrival rate scaling, see Appendix A. Appendix D summarises notation used in these models. Section 3 presents our prediction of COVID-19 patient arrival rates from regional infection data and considers the accuracy of our bed occupancy forecasts. Section 4 presents numerical results and illustrates the performance of our models and decision rules: Section 4.1 presents the inter-regional patient allocation results and Section 4.2 considers reallocating patients across regions. Section 5 concludes the paper.

2. Model and decision rules

This section presents our hierarchical modelling and decision approach. Section 2.1 briefly reviews the essential elements of our model for forecasts of the occupancy in the ward and ICU at an individual hospital [40], and introduces a decision rule to determine available capacity for patients from other hospitals. Section 2.2 introduces a load-balancing rule for allocation of patients to hospitals in a region and a decision rule when combining hospitals into one regional hospital exploiting the statistical multiplexing gain [15]. Section 2.3 introduces a recourse model to optimally distribute COVID-19 patients that cannot be accommodated within a region over multiple regions taking into account travelling distances, the currently available COVID-19 bed-capacity in each region, as well as the forecast maximum bed-occupancy over several days.

2.1. Bed surplus or shortage for an individual hospital

Consider a hospital with dedicated COVID-19 ward and ICU, indexed by h, that admits a fraction of COVID-19 patients from its service region as determined by the regional number of infected patients (autonomous arrivals). Following [40], we model the hospital as a network of two infinite server queues that records the number of hospitalised COVID-19 patients. The network includes patient-characteristics $c \in C$ (e.g., age, gender, weight) that may affect the hospitalisation rate and the patient journey in the hospital, a time-dependent Poisson arrival process with rate $\lambda_{hc}(t)$, with fraction $p_{hc}(t)$ admitted to the ward, fraction $1 - p_{hc}(t)$ to the ICU, general and time-dependent LoS $L_{hcW}(t)$ and $L_{hcI}(t)$ at ward and ICU, and discharge probabilities $q_{hcW}(t)$, resp. $q_{hcI}(t)$. The assumption of Poisson arrivals was shown to be justified for arrivals to Emergency Departments [47]. The number of patients $N_{hcW}(t)$ and $N_{hcI}(t)$ with characteristics c at time t have a time-dependent Poisson distribution with means $\rho_{hcW}(t)$, $\rho_{hcI}(t)$ that are determined via the Poisson Arrival Location Model (PALM), see [46, Theorem 2.1]. The total number of patients in the ward, $N_{hW}(t)$, and ICU, $N_{hI}(t)$, have time-dependent Poisson distributions with means $\rho_{hW}(t) = \sum_{c\in C} \rho_{hcW}(t)$, and $\rho_{hI}(t) = \sum_{c\in C} \rho_{hcI}(t)$, see [40].

The Poisson distributions for $N_{hW}(t)$ and $N_{hI}(t)$ allow us to explicitly evaluate relevant performance measures. Let $\mathbf{L}_h(s)$ be tuples of the location and realised LoSs (up to time s) of all patients residing in hospital h. The expected occupancy at time $s+t$ is:

$$\mathbb{E}[N_{hW}(s+t) \mid \mathbf{L}_h(s) = \ell_h], \quad \mathbb{E}[N_{hI}(s+t) \mid \mathbf{L}_h(s) = \ell_h]. \tag{1}$$

The expected maximum occupancy in $[s, s+t]$ is:

$$\mathbb{E}\left[\max_{u\in[s,s+t]} N_{hW}(u) \mid \mathbf{L}_h(s) = \ell_h\right], \quad \mathbb{E}\left[\max_{u\in[s,s+t]} N_{hI}(u) \mid \mathbf{L}_h(s) = \ell_h\right]. \tag{2}$$

The α_{hW}-quantile, $n_{hW,\alpha_{hW}}(s,t)$, and the α_{hI}-quantile, $n_{hI,\alpha_{hI}}(s,t)$, for respectively the maximum occupancy in the ward and ICU at hospital h in $[s, s+t]$,

$$\begin{aligned} n_{hW,\alpha_{hW}}(s,t) &= \min\left\{n :\ \mathbb{P}\left[\max_{u\in[s,s+t]} N_{hW}(u) \le n \,\middle|\, \mathbf{L}_h(s) = \ell_h\right] \ge \alpha_{hW}\right\}, \\ n_{hI,\alpha_{hI}}(s,t) &= \min\left\{n :\ \mathbb{P}\left[\max_{u\in[s,s+t]} N_{hI}(u) \le n \,\middle|\, \mathbf{L}_h(s) = \ell_h\right] \ge \alpha_{hI}\right\}, \end{aligned} \tag{3}$$

determine the required capacity $n_{hW,\alpha_{hW}}(s,t)$, $n_{hI,\alpha_{hI}}(s,t)$ in the ward and ICU to accommodate all autonomous arrivals in $[s, s+t]$ with probability at least α_{hW}, α_{hI}, respectively. We will refer to α_{hW}, α_{hI} as the *safety levels*.

Let $n^*_{hI}(s,t)$ be the number of beds in the ICU of hospital h in the time-interval $[s, s+t]$. If $n_{hI,\alpha_{hI}}(s,t) < n^*_{hI}(s,t)$ we may argue that at safety level α_{hI} a number of beds $\tilde{n}_{hI,\alpha_{hI}}(s,t) = n^*_{hI}(s,t) - n_{hI,\alpha_{hI}}(s,t)$ may be considered unoccupied in $[s, s+t]$. This *bed surplus* may then be allocated to COVID-19 patients from other hospitals. If $n_{hI,\alpha_{hI}}(s,t) \ge n^*_{hI}(s,t)$, hospital h may be confronted with *bed shortage* (at safety level α_{hI}), and will not offer any beds to ICU patients from other hospitals. A similar reasoning applies to the ward. Note that we do not assume a fixed number of beds, but include dependence on s, t in the number $n^*_{hW}(s,t)$, $n^*_{hI}(s,t)$ of beds in $[s, s+t]$. This allows the number of beds to be scaled up or down over time. We arrive at the following decision rule to determine the bed surplus. We introduce the following notation. Let $[x]^+ = \max\{0, x\}$, $[x]^- = \max\{0, -x\}$, for all $x \in \mathbb{R}$.

Decision Rule 1 (Individual hospital)**.** Consider hospital h, with $n^*_{hW}(s,t)$, $n^*_{hI}(s,t)$ beds in the ward resp. ICU in $[s, s+t]$. Consider safety levels α_{hW}, α_{hI}. Let $n_{hW,\alpha_{hW}}(s,t)$, $n_{hI,\alpha_{hI}}(s,t)$ be the α_{hW}-, α_{hI}-quantiles for the maximum occupancy at hospital h, (see (3)). At safety levels α_{hW}, α_{hI}, hospital h has *bed surplus* of

$$\begin{aligned} \tilde{n}_{hW,\alpha_{hW}}(s,t) &= \left[n^*_{hW}(s,t) - n_{hW,\alpha_{hW}}(s,t)\right]^+, \\ \tilde{n}_{hI,\alpha_{hI}}(s,t) &= \left[n^*_{hI}(s,t) - n_{hI,\alpha_{hI}}(s,t)\right]^+, \end{aligned} \tag{4}$$

available beds for COVID-19 patients from other hospitals in $[s, s+t]$.

Remark 1 (Bed surplus; safety levels)**.** *Observe that the bed surplus $\tilde{n}_{hW,\alpha_{hW}}(s,t)$, $\tilde{n}_{hI,\alpha_{hI}}(s,t)$ takes autonomous patient arrivals, transfers between ward and ICU and patient discharges for hospital h in $[s, s+t]$ into account. Admitting a patient from another hospital to a bed in the ward then assumes that this additional patient will not be transferred to the ICU between time s and $s+t$, which may be reasonable if t is small. Including this possible transfer requires that an additional bed is available in the ICU too, i.e., that also $\tilde{n}_{hI,\alpha_{hI}}(s,t) \ge 1$.*

The bed surplus is determined at safety levels α_{hW}, α_{hI}. These levels may include the hospital's policy for bed allocation, treatment protocols, or case mix, but may also incorporate the size of the hospital. As an extreme case, we may set $\alpha_{hW} = 0$ to indicate that we only focus on overcrowding of the ICU at hospital h.

2.2. Patient reallocation within an individual region

This section extends the decision rule for an individual hospital to a decision rule for an individual region to allocate COVID-19 patients among the hospitals in that region during a time interval $[s, s+t]$. Consider a region, indexed by r, containing H_r hospitals, indexed $h = 1, \dots, H_r$, where each hospital is modelled as described in Section 2.1. We will first determine the autonomous arrival rate for each hospital as fraction of the regional patients in Section 2.2.1. Then, in Section 2.2.2, we determine for each hospital the bed surplus according to Decision Rule 1 and subsequently present a load-balancing Decision Rule 2 based on the bed surplus for the region. In Section 2.2.3, we assume that hospitals disclose all information on the number of beds and hospitalised COVID-19 patients, which allows viewing the COVID-19 wards and ICUs in a region as single merged ward and ICU. This may be viewed to correspond to a regional coordination centre that optimally assigns patients to hospitals, resulting in a lower bound on the number of patients reallocated out of a region.

2.2.1. Autonomous arrival rates

Let $\Lambda_{rc}(u)$ denote the time-dependent rate at which COVID-19 patients with characteristics $c \in C$ request to be hospitalised in region r, and let $\Lambda_r(u) = \sum_{c \in C} \Lambda_{cr}(u)$ be the total arrival rate of COVID-19 patients in region r. Consider hospital h, with $n^*_{hW}(s,t)$, $n^*_{hI}(s,t)$ beds in the ward and ICU in $[s, s+t]$. At safety levels α_{hW}, α_{hI}, hospital h may admit a fraction $\theta_{h,\alpha_{hW},\alpha_{hI}}(s,t)$ of these regional patients in the ward and ICU such that the ward can accommodate all autonomous arrivals with rate $\theta_{h,\alpha_{hW},\alpha_{hI}}(s,t)\Lambda_r(u)$ in $[s, s+t]$ with probability at least α_{hW}, and similarly for the ICU. Let $\mathbb{P}_{\theta,h}$ denote the distribution of the number of patients for hospital h in $[s, s+t]$ given arrival rates $\lambda_{hc}(u) = \theta\Lambda_{rc}(u)$, $u \in [s, s+t]$, $c \in C$. Then $\theta_{h,\alpha_{hW},\alpha_{hI}}(s,t)$ may be determined as

$$\theta_{h,\alpha_{hW},\alpha_{hI}}(s,t) = \max\left\{\theta : \quad \mathbb{P}_{\theta,h}\left[\max_{u \in [s,s+t]} N_{hW}(u) \le n^*_{hW}(s,t) \;\middle|\; \mathbf{L}_h(s) = \ell_h\right] \ge \alpha_{hW}, \right.$$
$$\left. \mathbb{P}_{\theta,h}\left[\max_{u \in [s,s+t]} N_{hI}(u) \le n^*_{hI}(s,t) \;\middle|\; \mathbf{L}_h(s) = \ell_h\right] \ge \alpha_{hI}\right\}. \tag{5}$$

Let $\theta_{r,\alpha_r}(s,t) = \sum_{h=1}^{H_r} \theta_{h,\alpha_{hW},\alpha_{hI}}(s,t)$, with $\alpha_r = \{\alpha_{hW},\ \alpha_{hI} :\ h = 1, \dots, H_r\}$ be the set containing all safety levels of region r.

If $\theta_{r,\alpha_r}(s,t) < 1$, region r has insufficient capacity to accommodate all arrivals during $[s, s+t]$ at safety levels α_r. Hospital h admits the fraction $\theta_{h,\alpha_{hW},\alpha_{hI}}(s,t)$ of autonomous regional patients arriving in $[s, s+t]$ corresponding to its safety levels α_{hW}, α_{hI}. Thus, the autonomous arrival rate of patients with characteristics c for hospital h is

$$\lambda_{hc}(u) = \theta_{h,\alpha_{hW},\alpha_{hI}}(s,t)\Lambda_{rc}(u), \qquad h = 1, \dots, H_r. \tag{6}$$

At safety levels α_r, the remaining fraction of patients arriving at rate $[1 - \theta_{r,\alpha_r}(s,t)]\Lambda_r(u)$ must be accommodated in hospitals outside region r.

If $\theta_{r,\alpha_r}(s,t) \ge 1$, region r has sufficient capacity to accommodate all autonomous arrivals. A fair or load-balancing distribution of COVID-19 patients over the hospitals in the region according to their safety levels α_r is obtained by admitting the fraction $\theta_{h,\alpha_{hW},\alpha_{hI}}(s,t)/\theta_{r,\alpha_r}(s,t)$ of patients in hospital h. Thus, the autonomous arrival rate of patients with characteristics c for hospital h is

$$\lambda_{hc}(u) = \frac{\theta_{h,\alpha_{hW},\alpha_{hI}}(s,t)}{\theta_{r,\alpha_r}(s,t)}\Lambda_{rc}(u), \qquad h = 1, \dots, H_r. \tag{7}$$

In Section 2.2.2, we use (6) and (7) to determine the region's ward and ICU bed shortage or bed surplus.

Remark 2 (Fraction of admitted regional patients)**.** *Observe that $\theta_{h,\alpha_{hW},\alpha_{hI}}(s,t)$ in (5) determines the fraction of patients admitted in hospital h irrespective of admittance to ward or ICU. This is a natural choice, as in most cases hospitals first admit regional patients and then perform triage (i.e., determine whether the patient is admitted to the ward or ICU). Our model includes transfers between ward and ICU that may occur in $[s, s+t]$, so that $\theta_{h,\alpha_{hW},\alpha_{hI}}(s,t)$ yields the fraction of admitted patients at the safety levels of the ward and ICU.*

Observe that the arrival rates $\lambda_{hc}(u)$ in (6) and *(7) imply that the probability that hospital h cannot accommodate its autonomous arrivals in $[s, s+t]$ at its ward, resp. ICU, is at most $1 - \alpha_{hW}$, resp. $1 - \alpha_{hI}$. Clearly, if $\theta_{r,\alpha_r}(s,t) \ge 1$, with arrival rates $\lambda_{hc}(u)$ in (7), these probabilities will most likely be even smaller, as the arrival rates $\lambda_{hc}(u)$ are smaller than the arrival rates $\theta_{h,\alpha_{hW},\alpha_{hI}}(s,t)\Lambda_{rc}(u)$ that hospital h can accommodate at safety levels α_{hW}, α_{hI}.*

If $\theta_{r,\alpha_r}(s,t) < 1$, it may occur that hospital h could admit patients to its ward at higher rates than under $\lambda_{hc}(u)$ (from (6)), *but not to its ICU, or vice versa. This follows from (5), as, at some point, either the ward or the ICU is the bottleneck to further increase $\theta_{h,\alpha_{hW},\alpha_{hI}}(s,t)$ (i.e., one of the two inequalities in (5) is tight). It then follows from Decision Rule 1 that hospital h has bed surplus of $\tilde{n}_{hW,\alpha_{hW}}(s,t) > 0$ ward beds or $\tilde{n}_{hI,\alpha_{hI}}(s,t) > 0$ ICU beds. Note that accepting additional patients in the ward in case $\tilde{n}_{hW,\alpha_{hW}}(s,t) > 0$ might result in overcrowding of the ICU due to these patients transferring from ward to ICU.*

We may extend our model to include different $\theta_{hW,\alpha_{hW}}(s,t)$ and $\theta_{hI,\alpha_{hI}}(s,t)$ for a hospital's ward and ICU, respectively, defined as

$$\theta_{hW,\alpha_{hW}}(s,t) = \max\left\{\theta : \mathbb{P}_{\theta,h}\left[\max_{u \in [s,s+t]} N_{hW}(u) \le n^*_{hW}(s,t) \;\middle|\; \mathbf{L}_h(s) = \ell_h\right] \ge \alpha_{hW}\right\},$$
$$\theta_{hI,\alpha_{hI}}(s,t) = \max\left\{\theta : \mathbb{P}_{\theta,h}\left[\max_{u \in [s,s+t]} N_{hI}(u) \le n^*_{hI}(s,t) \;\middle|\; \mathbf{L}_h(s) = \ell_h\right] \ge \alpha_{hI}\right\}. \tag{8}$$

This may allow more flexibility in accepting COVID-19 patients in the ward when the ICU has reached its capacity, and vice versa, but may also result in overcrowding of either the ward or the ICU.

2.2.2. Load balancing rule to fairly allocate patients to hospitals

Section 2.2.1 has established whether or not the hospitals in a region r may admit all autonomous arrivals at safety levels α_r. If $\theta_{r,\alpha_r}(s,t) < 1$, the autonomous arrival rates (6) determine the region's bed shortage in $[s, s+t]$ at safety levels α_r. If $\theta_{r,\alpha_r}(s,t) \ge 1$, the rates (7) determine the hospitals' bed surplus in $[s, s+t]$ and hence the bed surplus of region r. This section presents a load balancing rule for fair allocation of patients to the hospitals in a region.

First, if $\theta_{r,\alpha_r}(s,t) \ge 1$, we invoke Decision Rule 1 with patient arrival rates (7) to determine the bed surplus for each hospital h in region r and add these numbers to obtain the region's bed surplus in ward, resp. ICU, in $[s, s+t]$ at safety levels α_r as

$$\tilde{n}_{rW,\alpha_r}(s,t) = \sum_{h=1}^{H_r} \tilde{n}_{hW,\alpha_{hW}}(s,t), \qquad \tilde{n}_{rI,\alpha_r}(s,t) = \sum_{h=1}^{H_r} \tilde{n}_{hI,\alpha_{hI}}(s,t).$$

Second, consider the case $\theta_{r,\alpha_r}(s,t) < 1$. At safety levels α_r, the remaining fraction of patients arriving at rate $[1 - \theta_{r,\alpha_r}(s,t)]\Lambda_r(u)$ must be accommodated in hospitals outside region r for each $u \in [s, s+t]$. Our model requires discrimination between the number of patients admitted at the ward and the ICU. To this end, observe that $\theta_{h,\alpha_{hW},\alpha_{hI}}(s,t)/\theta_{r,\alpha_r}(s,t)$ is the fraction of patients that would be admitted to hospital h if all hospitals would have ample capacity. For each $u \in [s, s+t]$ the fraction of patients with characteristics c admitted to all wards of hospitals in region r at safety levels α_r may be obtained as

$$p_{rc}(u) = \sum_{h=1}^{H_r} \frac{\theta_{h,\alpha_{hW},\alpha_{hI}}(s,t)}{\theta_{r,\alpha_r}(s,t)} p_{hc(u)}. \tag{9}$$

If the fractions $p_{hc}(u)$ for all hospitals in the region coincide, say $p_{hc}(u) = p_c(u)$, then (9) reduces to $p_{rc}(u) = p_c(u)$.

Let $M_{rW,\alpha_r}(s,t)$, resp. $M_{rI,\alpha_r}(s,t)$, be the regional bed shortage in ward, resp. ICU, in $[s, s+t]$ at safety levels α_r. Then, $M_{rW,\alpha_r}(s,t)$, resp. $M_{rI,\alpha_r}(s,t)$, are Poisson distributed random variables with means $m_{rW,\alpha_r}(s,t)$, resp. $m_{rI,\alpha_r}(s,t)$:

$$m_{rW,\alpha_r}(s,t) = [1 - \theta_{r,\alpha_r}(s,t)] \int_s^{s+t} \sum_{c\in C} p_{rc}(u)\Lambda_{rc}(u)du, \tag{10}$$

$$m_{rI,\alpha_r}(s,t) = [1 - \theta_{r,\alpha_r}(s,t)] \int_s^{s+t} \sum_{c\in C} (1 - p_{rc}(u))\Lambda_{rc}(u)du, \tag{11}$$

so that the expected regional bed shortage in ward, resp. ICU, in $[s, s+t]$ at safety levels α_r is $m_{rW,\alpha_r}(s,t)$, resp. $m_{rI,\alpha_r}(s,t)$.

Combining the results above with those of Section 2.2.1 we obtain the following load balancing decision rule to allocate patients to hospitals in a region and determining the regional bed shortage or surplus.

Decision Rule 2 (Individual region; load balancing; hospital safety levels)**.** Consider region r with $n^*_{hW}(s,t)$, $n^*_{hI}(s,t)$ beds in the ward and ICU in $[s, s+t]$ at hospitals $h = 1, \ldots, H_r$. Consider safety levels $\alpha_r = \{\alpha_{hW},\ \alpha_{hI},\ h = 1, \ldots, H_r\}$.

If $\theta_{r,\alpha_r}(s,t) \geq 1$, allocate a fraction

$$\frac{\theta_{h,\alpha_{hW},\alpha_{hI}}(s,t)}{\theta_{r,\alpha_r}(s,t)}$$

of the regional patients to hospital h and report bed surplus of

$$\tilde{n}_{rW,\alpha_r}(s,t), \quad \tilde{n}_{rI,\alpha_r}(s,t),$$

beds in ward and ICU in $[s, s+t]$ for reallocation of COVID-19 patients from other regions.

If $\theta_{r,\alpha_r}(s,t) < 1$, allocate a fraction

$$\theta_{h,\alpha_{hW},\alpha_{hI}}(s,t)$$

of the regional patients to hospital h and report bed shortage of

$$m_{rW,\alpha_r}(s,t), \quad m_{rI,\alpha_r}(s,t)$$

beds in $[s, s+t]$ at ward and ICU for reallocation of COVID-19 patients to other regions.

Remark 3 (Dynamic load balancing algorithm)**.** *Decision Rule 2 allocates patients to hospitals using a dynamic rule based on the maximum occupancy in the wards and ICUs in $[s, s+t]$ as determined by $\theta_{h,\alpha_{hW},\alpha_{hI}}(s,t)$, $h = 1, \ldots, H_r$, and may therefore be viewed as a dynamic load balancing algorithm. For a recent overview of load balancing algorithms, see [17], and see [16] for a general reference on load balancing for loss networks.*

2.2.3. Merging all wards and all ICUs in a region

This section introduces regional control of COVID-19 beds. Merging the ICU bed-capacity of individual hospitals into a regional ICU may considerably reduce the number of patients reallocated out of the region, see [48]. We will exploit the so-called statistical multiplexing gain [15], and merge all wards, resp. ICUs, into a single (virtual) regional ward, resp. ICU.

Assume that the hospitals $h = 1, \ldots, H_r$ in region r agree on regional safety levels α_{rW} and α_{rI} for their wards and ICUs and (virtually) merge their COVID-19 wards, resp. ICUs, into a single regional ward and ICU, with capacities

$$n^*_{rW}(s,t) = \sum_{h=1}^{H} n^*_{hW}(s,t), \quad \text{resp.} \quad n^*_{rI}(s,t) = \sum_{h=1}^{H} n^*_{hI}(s,t),$$

in $[s, s+t]$. We may now view the region as a single hospital (as described in Section 2.1) with autonomous arrival rates $\Lambda_{rc}(u)$, $c \in C$, where a fraction $P_{rc}(u)$, resp. $1 - P_{rc}(u)$ is admitted to the regional ward, resp. ICU. Let $N_{rW}(u)$, $N_{rI}(u)$ record the number of patients present in the (virtual) regional COVID-19 ward and ICU at time u, respectively. The α_{rW}-quantile, $n_{rW,\alpha_{rW}}(s,t)$, and the α_{rI}-quantile, $n_{rI,\alpha_{rI}}(s,t)$, for the maximum occupancy in $[s, s+t]$ follow by analogy with the single hospital model of Section 2.1.

For region r, (5) determines the fraction of autonomous regional arrivals that may be accommodated by hospital h, $h = 1, \ldots, H_r$, at its safety levels. Cooperation among the hospitals allows a more refined rule to distribute patients over the hospitals of a region. To this end, first observe that if region r accepts autonomous arrivals to its ward, resp. ICU, at safety levels α_{rW}, resp. α_{rI}, then the hospitals in the region must have sufficient beds to accept these patients in their wards and ICUs. We may, therefore, at safety levels α_{rW}, α_{rI}, distribute patients according to the fractions $\theta_{hW,\alpha_{rW}}(s,t)$, $\theta_{hI,\alpha_{rI}}(s,t)$ defined in (8), while still avoiding overcrowding of the wards and ICUs, recall Remark 2. Region r allocates the fractions

$$\widehat{\theta}_{hW,\alpha_{rW}}(s,t) = \frac{\theta_{hW,\alpha_{rW}}(s,t)}{\sum_{h=1}^{H_r} \theta_{hW,\alpha_{rW}}(s,t)}, \quad \widehat{\theta}_{hI,\alpha_{rI}}(s,t) = \frac{\theta_{hI,\alpha_{rI}}(s,t)}{\sum_{h=1}^{H_r} \theta_{hI,\alpha_{rI}}(s,t)} \tag{12}$$

of the autonomous arrivals that are hospitalised in region r to the ward, resp. ICU, of hospital h, $h = 1, \ldots, H_r$, if the denominators are positive. If $\sum_{h=1}^{H_r} \theta_{hW,\alpha_{rW}} = 0$, we determine safety level $\alpha'_{rW} = \sup\left\{\alpha < \alpha_{rW} : \max_{h\in\{1,\ldots,H_r\}} \theta_{hW,\alpha}(s,t) > \epsilon\right\}$ for small $\epsilon > 0$. Patients are then allocated uniformly at random to hospitals where $\theta_{hW,\alpha'_{rW}} > 0$, and similarly for the ICU.

Combining the results above we obtain the following decision rule for allocation of patients to hospitals in a region and determining the regional bed shortage or bed surplus.

Decision Rule 3 (Individual region; regional safety levels)**.** Consider a region r with $n^*_{rW}(s,t)$, $n^*_{rI}(s,t)$ beds in the (virtual) ward and ICU in $[s, s+t]$, and safety levels α_{rW}, α_{rI}. Let $n_{rW,\alpha_{rW}}(s,t)$, $n_{rI,\alpha_{rI}}(s,t)$ be determined based on (3), and $\widehat{\theta}_{hW,\alpha_{rW}}(s,t)$, $\widehat{\theta}_{hI,\alpha_{rI}}(s,t)$ according to (12).

If $n_{rW,\alpha_{rW}}(s,t) > n^*_{rW}(s,t)$, report a bed shortage

$$n_{rW,\alpha_{rW}}(s,t) - n^*_{rW}(s,t)$$

in the wards of region r in $[s, s+t]$. Otherwise, report bed surplus

$$n^*_{rW}(s,t) - n_{rW,\alpha_{rW}}(s,t)$$

in the wards of region r in $[s, s+t]$. In both cases, allocate a fraction

$$\widehat{\theta}_{hW,\alpha_{rW}}(s,t)$$

of the regional autonomous arrivals that are hospitalised in the wards in region r to hospital h.

For the ICU, the rules above apply with W replaced by I.

Remark 4 (Statistical multiplexing gain; comparison of Decision Rules 2, 3)**.** *Merging the wards, resp. ICUs, of the hospitals into a regional ward, resp. ICU, exploits the so-called statistical multiplexing gain, see [15,48,49]. In particular, it avoids that one hospital has bed shortage, while another hospital has a bed surplus.*

Decision Rule 2 uses the safety levels of the individual hospitals to determine the bed surplus for each hospital, whereas Decision Rule 3 uses the safety levels of the merged hospitals to determine the regional bed surplus. A merged hospital requires fewer beds to accommodate patients at the same safety level, see [15,49] and therefore the bed surplus under Decision Rule 3 exceeds the bed surplus under Decision Rule 2.

2.3. Patient reallocation across multiple regions

This section builds on the bed surplus and bed shortage for individual regions as determined under Decision Rule 2 or 3 to obtain a decision rule for (part of) a country consisting of R regions with $n^*_{rW}(s,t)$, resp. $n^*_{rI}(s,t)$, beds in the ward, resp. ICU, in $[s,s+t]$, $r=1,\dots,R$. This decision rule determines the number of patients to be reallocated across the regions at each decision epoch s, taking into account the current number of hospitalised patients as well as the maximum number of patients hospitalised in the regions in $[s,s+t]$. Patients that cannot be accommodated in the regions may be reallocated to an external region.

Let $\tilde{n}_{rW}(s)$, resp. $\tilde{n}_{rI}(s)$, be the current (at epoch s) bed shortage ($\tilde{n}<0$) or bed surplus ($\tilde{n}\ge 0$) in the ward, resp. ICU, of region r. Let the random variables $\tilde{N}_{rW}(s,t)$, resp. $\tilde{N}_{rI}(s,t)$, be the bed shortage ($\tilde{N}<0$) or bed surplus ($\tilde{N}\ge 0$) in the ward, resp. ICU, of region r from decision epoch s up to time $s+t$ taking into account the patients reallocated out of region r at decision epoch s determined by

$$\tilde{N}_{rW}(s,t) = n^*_{rW}(s,t) - N_{rW}(s,t) - [\tilde{n}_{rW}(s)]^-,$$
$$\tilde{N}_{rI}(s,t) = n^*_{rI}(s,t) - N_{rI}(s,t) - [\tilde{n}_{rI}(s)]^-.$$

In the above, $N_{rW}(s,t)$, $N_{rI}(s,t)$ represent the maximum occupancy at the ward and ICU for region r respectively, during the period $[t,s+t]$ under the regional hospital model introduced in Section 2.2.3. For computational tractability reasons, we propose to employ the regional model of Section 2.2.3 with corresponding Decision Rule 3 within the model for patient reallocation across multiple regions. The more detailed and computationally more intensive load balancing model of Section 2.2.2 with corresponding Decision Rule 2 can subsequently be used within each region for allocating the patients assigned to that region to hospitals. The external region has ample capacity, i.e., $\tilde{n}_{R+1,W}(s)=\tilde{n}_{R+1,I}(s)=\infty$ as well as $\tilde{n}_{R+1,W}(s,t)=\tilde{n}_{R+1,I}(s,t)=\infty$.

Reallocation of a patient from region r to region r' incurs cost $\gamma_{r,r'}$ incorporating, for example, travel distance for reallocation of the patient or travel distance for his or her relatives, and differences between regions r and r' with respect to the number or size of hospitals. We may impose the (strict) triangle inequality on the costs $\gamma_{r,r'}$:

$$\gamma_{r_1,r_3} < \gamma_{r_1,r_2} + \gamma_{r_2,r_3} \quad \forall r_1,r_2,r_3 \in \{1,2,\dots,R+1\}, \quad r_1 \ne r_2 \ne r_3, \tag{13}$$

to avoid that region r_2 functions as an intermediate stop for reallocation of patients from region r_1 to region r_3, which may happen, for example, if $\gamma_{r_1,r_2}=\gamma_{r_2,r_3}=2$ and $\gamma_{r_1,r_3}=5$, which is excluded by (13).

We will now develop a recourse model with objective to

> minimise the costs of patient reallocations across regions at the current decision epoch s as well as during $[s,s+t]$

such that

(i) patients are distributed over regions such that the current bed shortages are resolved, taking into account the bed shortages in $[s,s+t]$, and
(ii) the relative remaining bed surplus (detailed below) is balanced over the regions.

The *here-and-now* decision variables $f_{W,r,r'}(s)$, resp. $f_{I,r,r'}(s)$, are the number of ward, resp. ICU, patients to reallocate from region r to region r' at decision epoch s. The *wait-and-see* decision variables $F_{W,r,r'}(s,t)$, resp. $F_{I,r,r'}(s,t)$, are the number of potentially required additional reallocations among the wards, resp. ICUs, of regions r and r' in $[s,s+t]$ based on the bed surplus or shortage $\tilde{N}_{rW,\alpha_r}(s,t)$, $\tilde{N}_{rW,\alpha_r}(s,t)$ in $[s,s+t]$ for all regions r. To penalise the imbalance in bed surplus in the wards and ICUs across the regions, we introduce a penalty function $g(\cdot)$ (e.g. $g(x)=x^2$ or $g(x)=x$). The optimal number of patients $f_{W,r,r'}(s)$, resp. $f_{I,r,r'}(s)$, reallocated from the ward, resp. ICU, of region r to the ward, resp. ICU, of region r', $r,r'=1,\dots,R+1$, is determined as the *argmin* of the following recourse model:

$$\min \sum_{r,r'} \gamma_{r,r'}\left(f_{W,r,r'}(s)+f_{I,r,r'}(s)+\mathbb{E}[F_{W,r,r'}(s,t)+F_{I,r,r'}(s,t)]\right) \tag{P}$$
$$+\sum_{r=1}^{R} g(\delta_{W,r}(s))+g(\delta_{I,r}(s))+\mathbb{E}[g(\Delta_{W,r}(s,t))+g(\Delta_{I,r}(s,t))]$$

s.t.

(Here-and-now constraints, resolve current shortages in original ward and ICU)

$$\sum_{r'\ne r} f_{W,r,r'}(s) - [\tilde{n}_{rW}(s)]^- = 0, \qquad \forall r$$
$$\sum_{r'\ne r} f_{I,r,r'}(s) - [\tilde{n}_{rI}(s)]^- = 0, \qquad \forall r$$

(Here-and-now constraints, dont́ cause shortages in destination ward and ICU)

$$[\tilde{n}_{rW}(s)]^+ - \sum_{r'\ne r} f_{W,r',r}(s) \ge 0, \qquad \forall r$$
$$[\tilde{n}_{rI}(s)]^+ - \sum_{r'\ne r} f_{I,r',r}(s) \ge 0, \qquad \forall r$$

(Here-and-now constraints, transport to external region only if strictly necessary)

$$\sum_{r\neq R+1} f_{W,r,R+1}(s) - \left[\sum_{r\neq R+1} \tilde{n}_{rW}(s)\right]^{-} = 0, \qquad \forall r$$

$$\sum_{r\neq R+1} f_{I,r,R+1}(s) - \left[\sum_{r\neq R+1} \tilde{n}_{rI}(s)\right]^{-} = 0, \qquad \forall r$$

(Wait-and-see constraints, anticipate on future shortages in ward and ICU)

$$\tilde{N}_{rW}(s,t) + \sum_{r'\neq r}\{[f_{W,r,r'}(s) - f_{W,r',r}(s)] + [F_{W,r,r'}(s,t) - F_{W,r',r}(s,t)]\} \geq 0, \qquad \forall r$$

$$\tilde{N}_{rI}(s,t) + \sum_{r'\neq r}\{[f_{I,r,r'}(s) - f_{I,r',r}(s)] + [F_{I,r,r'}(s,t) - F_{I,r',r}(s,t)]\} \geq 0, \qquad \forall r$$

(Domain constraints)$f_{W,r,r'}(s),\ f_{I,r,r'}(s),\ F_{W,r,r'}(s,t),\ F_{I,r,r'}(s,t) \in \mathbb{N}_0$, $\forall r, r'$,

where we have used the additional notation for the current and future relative remaining bed surplus $\delta_{W,r}(s)$, $\delta_{I,r}(s)$, $\Delta_{W,r}(s,t)$, $\Delta_{I,r}(s,t)$ in region r,

$$\delta_{W,r}(s) = \frac{\tilde{n}_{rW}(s) + \sum_{r'\neq r}[f_{W,r,r'}(s) - f_{W,r',r}(s)]}{n^*_{rW}(s,t)},$$

$$\delta_{I,r}(s) = \frac{\tilde{n}_{rI}(s) + \sum_{r'\neq r}[f_{I,r,r'}(s) - f_{I,r',r}(s)]}{n^*_{rI}(s,t)},$$

$$\Delta_{W,r}(s,t) = \frac{\tilde{N}_{rW}(s,t) + \sum_{r'\neq r}\{[f_{W,r,r'}(s) - f_{W,r',r}(s)] + [F_{W,r,r'}(s,t) - F_{W,r',r}(s,t)]\}}{n^*_{rW}(s,t)},$$

$$\Delta_{I,r}(s,t) = \frac{\tilde{N}_{rI}(s,t) + \sum_{r'\neq r}\{[f_{I,r,r'}(s) - f_{I,r',r}(s)] + [F_{I,r,r'}(s,t) - F_{I,r',r}(s,t)]\}}{n^*_{rI}(s,t)}.$$

Note that the wait-and-see and domain constraints hold point-wise on the sample space, i.e., for every possible outcome of $\tilde{N}_{rW}(s,t), \tilde{N}_{rI}(s,t)$. The above program is a mixed integer recourse model (see e.g. [34, Chapter 3]), where all possible outcomes for the maximum occupancy are considered at the second decision stage at time $s+t$.

The stochastic program does not directly incorporate Decision Rule 2 or 3 that are amenable for practical implementation. To incorporate these rules in our optimisation approach, we approximate (P) by the integer program described below that is based on the safety levels α_{rW}, α_{rI}, and the expected shortages or surpluses reported by the regions.

From Decision Rule 2 or 3, let $\tilde{n}_{rW}(s,t)$, resp. $\tilde{n}_{rI}(s,t)$, be the forecast bed shortage ($\tilde{n} < 0$) or bed surplus ($\tilde{n} \geq 0$) in the ward, resp. ICU, of region r from decision epoch s up to time $s+t$ taking into account the patients reallocated out of region r at decision epoch s. Thus, under Decision Rule 2:

$$\tilde{n}_{rW}(s,t) = \begin{cases} \tilde{n}_{rW,\alpha_r}(s,t) - [\tilde{n}_{rW}(s)]^-, & \text{if } \tilde{n}_{rW,\alpha_r}(s,t) > 0,\\ -m_{rW,\alpha_r}(s,t) - [\tilde{n}_{rW}(s)]^-, & \text{otherwise,}\end{cases}$$

$$\tilde{n}_{rI}(s,t) = \begin{cases} \tilde{n}_{rI,\alpha_r}(s,t) - [\tilde{n}_{rI}(s)]^-, & \text{if } \tilde{n}_{rI,\alpha_r}(s,t) > 0,\\ -m_{rI,\alpha_r}(s,t) - [\tilde{n}_{rI}(s)]^-, & \text{otherwise,}\end{cases}$$

and under Decision Rule 3:

$$\tilde{n}_{rW}(s,t) = n^*_{rW}(s,t) - n_{rW,\alpha_{rW}}(s,t) - [\tilde{n}_{rW}(s)]^-,$$
$$\tilde{n}_{rI}(s,t) = n^*_{rI}(s,t) - n_{rI,\alpha_{rI}}(s,t) - [\tilde{n}_{rI}(s)]^-.$$

The forecast shortages are resolved by the wait-and-see variables $f_{W,r,r'}(s,t)$, $f_{I,r,r'}(s,t)$ according to wait-and-see constraints:

(Wait-and-see constraints, anticipate on future shortages in ward and ICU)

$$\tilde{n}_{rW}(s,t) + \sum_{r'\neq r}\{[f_{W,r,r'}(s) - f_{W,r',r}(s)] + [f_{W,r,r'}(s,t) - f_{W,r',r}(s,t)]\} \geq 0, \qquad \forall r,$$

$$\tilde{n}_{rI}(s,t) + \sum_{r'\neq r}\{[f_{I,r,r'}(s) - f_{I,r',r}(s)] + [f_{I,r,r'}(s,t) - f_{I,r',r}(s,t)]\} \geq 0, \qquad \forall r.$$

The wait-and-see outcomes then induce future relative remaining bed surpluses

$$\delta_{W,r}(s,t) = \frac{\tilde{n}_{rW}(s,t) + \sum_{r'\neq r}\{[f_{W,r,r'}(s) - f_{W,r',r}(s)] + [f_{W,r,r'}(s,t) - f_{W,r',r}(s,t)]\}}{n^*_{rW}(s,t)},$$

$$\delta_{I,r}(s,t) = \frac{\tilde{n}_{rI}(s,t) + \sum_{r'\neq r}\{[f_{I,r,r'}(s) - f_{I,r',r}(s)] + [f_{I,r,r'}(s,t) - f_{I,r',r}(s,t)]\}}{n^*_{rI}(s,t)}.$$

The objective is now reformulated as

$$\min \sum_{r,r'} \gamma_{r,r'}(f_{W,r,r'}(s) + f_{I,r,r'}(s) + f_{W,r,r'}(s,t) + f_{I,r,r'}(s,t)) + \sum_{r=1}^{R} g(\delta_{W,r}(s)) + g(\delta_{I,r}(s)) + g(\delta_{W,r}(s,t)) + g(\delta_{I,r}(s,t)) \qquad (P')$$

All constraints in the optimisation model are automatically satisfied for external region $R+1$ because $n^*_{R+1,W}(s) = n^*_{R+1,I}(s) = \infty$ and $n^*_{R+1,W}(s,t) = n^*_{R+1,I}(s,t) = \infty$. Due to the first set of here-and-now constraints, in the optimisation model it is optimal to set $f_{W,R+1,r'}(s) = f_{I,R+1,r'}(s) = 0$ for all r', i.e., patients are not reallocated from the external region $R+1$ to a region r'. Moreover, the optimum cannot include both $f_{W,r,r'}(s) > 0$ and $f_{W,r',r}(s) > 0$ or both $f_{I,r,r'}(s) > 0$ and $f_{I,r',r}(s) > 0$, i.e., patients are not exchanged between regions. A feasible solution to the above program is guaranteed since the number of patients reallocated to the external region is not restricted.

Decision Rule 4 (Multiple regions; reallocation based on. *(P')*) Consider (part [4] of) a country consisting of R regions with $n^*_{rW}(s,t)$, resp. $n^*_{rI}(s,t)$, beds in the ward, resp. ICU, in $[s, s+t]$, $r = 1, \ldots, R$, and safety levels α_{rW}, α_{rI}, augmented with an external region $R+1$, that has ample capacity. Let $\tilde{n}_{rW}(s)$, $\tilde{n}_{rI}(s)$ be the observed bed shortage / bed surplus at decision epoch s in region r, $r = 1, \ldots, R$. Then, as obtained from (P'), at decision epoch s reallocate

$$f_{W,r,r'}(s), \text{ resp. } f_{I,r,r'}(s),$$

patients from the ward, resp. ICU, of region r to the ward, resp. ICU, of region r', $r, r' = 1, \ldots, R+1,\ r \neq r'$.

Remark 5 (Patient reallocation at decision epoch. s) *Under Decision Rule 4 patients are reallocated across regions at decision epoch s taking into account bed shortage and bed surplus in all regions in $[s, s+t]$. Decisions on patient reallocation at a later epoch, e.g., $s+1$, in the interval $[s, s+t]$ will be taken at that epoch taking into account the state at $s+1$ and the interval $[s+1, s+t+1]$.*

Remark 6 (Generalisations)**.** *We have presented the optimisation model of Decision Rule 4 in a relatively simple form. The model may readily be generalised to include, e.g., different costs or penalty functions for current and future reallocations, different costs for ward and ICU reallocations, or different loss functions for ward and ICU.*

3. Predicting arrival rates and forecasting bed occupancy

The effectiveness of the decision rules developed in Section 2 relies on an accurate real-time forecast of the COVID-19 bed occupancy, and therefore on an accurate prediction of the arrival rate and estimation of the LoS. This section considers prediction of arrival rates for a region and a single hospital, as well as generation of bed occupancy forecasts.

For each ROAZ region, the number of positive COVID-19 tests is available on a daily basis on the website of the Dutch National Institute for Public Health and the Environment (RIVM) [50]. In this data set, the number of infections on a given day represents the number of people that (retrospectively) tested positive for COVID-19 on that day. In addition to infection data, national hospital admission data per region, collected by the Dutch foundation for National Intensive Care Evaluation (NICE) is available on the website of the RIVM [51]. The number of admissions per day represents the number of patients who have tested positive for COVID-19 and are admitted to a hospital in the respective region. For our arrival rate prediction, we focus on data of the ROAZ region *Netwerk Acute Zorg (NAZ) West*, containing the hospitals Groene Hart Ziekenhuis (GHZ), HagaZiekenhuis (Haga) and Leiden University Medical Center (LUMC). To evaluate the accuracy of our prediction, in this section we focus on Haga, a 600-bed hospital in The Hague that admits approximately 29,000 inpatients per year. For this hospital, relevant data is available to us from September 4, 2020 until January 31, 2021.

3.1. Prediction of the arrival rates

In this section, we develop a new arrival rate predictor based on regional infection data, and compare its performance to the predictor in Baas et al. [40] that was shown to result in accurate bed occupancy forecasts. The new predictor makes explicit use of the delay between COVID-19 infection and hospitalisation, and enables us to predict the arrival rate at the (virtual) merged regional ward and ICU.

An increase in the number of COVID-19 infections results some days later in an increasing number of hospital admissions. As a consequence, we may expect a filtration and time-delay between the number of infections and the number of hospitalised patients. We assume the number of infections to be a Poisson process with time-dependent rate, so that the resulting autonomous regional arrival rate of COVID-19 patients is again a Poisson process, see, e.g., Chiu et al. [52]. We estimate the time-delay and filtration between COVID-19 infections and hospitalisations directly on the data for NAZ West in the RIVM data set [50]. The time-delay that gives the best fit is defined as the value between two and fourteen days resulting in the minimal mean squared error (MSE) when performing ordinary least squares (OLS) between a weekly moving average of the regional infections and a delayed exponentially weighted moving average (coefficient 0.1) of the regional hospitalisations [51]. The filtration that gives the best fit is the coefficient resulting from the OLS procedure corresponding to the optimal time-delay.

The result of this time-delaying and filtration procedure applied to NAZ West data is shown in the top graphs in Fig. 1. The black dots in the top left, resp. top right, graph of represent the realised number of infections, resp. hospitalisations, per day in NAZ West. The red line in the top left graph shows the 7-day moving average (MA) of the number of infections and the red line in the top right graph shows an exponentially weighted moving average (EWMA; coefficient 0.1) of the number of hospitalisations, both to indicate the trend. These trend lines already reveal the time-delay and filtration between the number of infections and hospitalisations. The purple line in the top right graph is the result of the time-delaying and filtration procedure applied to the (averaged) daily number of infections displayed as the red line in the top left graph. The best-fit time-delay equals 7 days, which is in accordance with a recent study performed in Belgium [53], where an average time-delay of 5.74 days is estimated, with medians ranging from 3 to 10.4 days, depending on patient characteristics. The best-fit filtration factor is found to be 3.1%. The extremes of the purple line in the top right graph of Fig. 1 coincide with the extremes of the number of hospitalisations (red line), but additional fine-tuning is required since the extremes over- and undershoot those of the number of hospitalisations.

To this end, we develop an t-days ahead prediction of the number of hospitalisations displayed (for $t = 3$) as the purple curve in the bottom left graph in Fig. 1. The static predictor (purple line in the top right graph) is corrected by estimating the scaling factor of the infections using weighted least squares between delayed infection and arrival data up to time s. The weights used for this least squares procedure are normalised exponential weights with base 1.2 so that errors in the fit for recent hospitalisations are penalised more than those for earlier hospitalisations. The effect of the weighted least squares procedure is that for each time s the t-days ahead prediction starts around the trend in the number of hospitalisations at time s. Hence, in our t-days ahead prediction the daily number of infections up to time s have a larger influence on the slope of the fine-tuned purple curve in the bottom left graph than on the starting point for the prediction determined by the number of hospitalisations (orange). This is motivated by the observation that the regional fraction of hospitalised patients

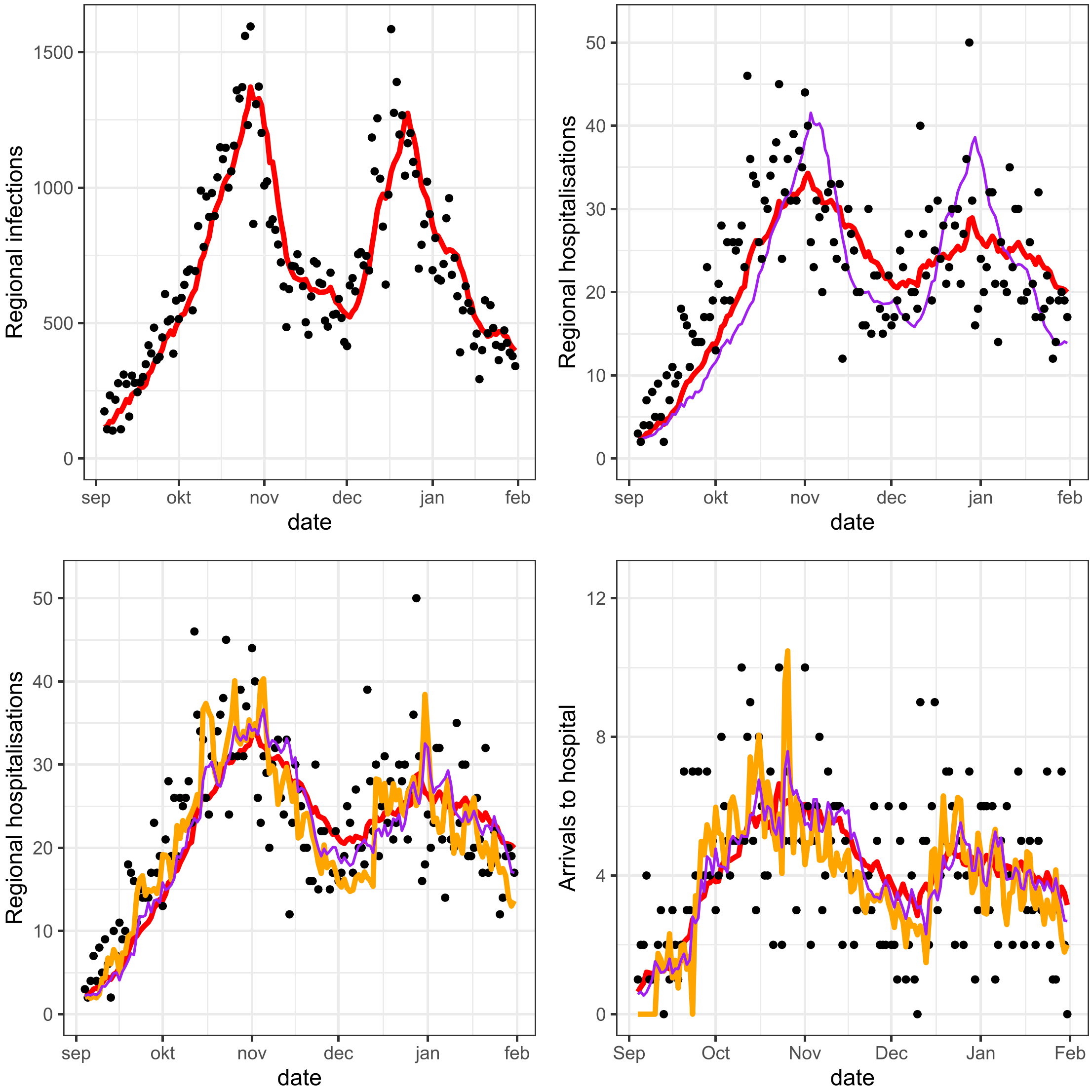

Fig. 1. ROAZ region NAZ West, hospital Haga, September 4, 2020 until January 31, 2021. Top left: Regional infections (black) and 7-day moving average (MA) of the realisations (red). Top right: Regional hospitalisations (black), exponentially weighted moving average (EWMA; coefficient 0.1) of realised regional hospitalisations (red) and filtered/scaled 7-day moving average of infection data (purple). Bottom left: Regional hospitalisations (black), EWMA (coefficient 0.1) of realised regional hospitalisations (red), 3-days ahead expanding window predictions of the arrival rate by the Richards' curve model (orange) and 3-days ahead expanding window predictions of the arrival rate from regional infections (purple). The purple line is made thinner to distinguish it from the other two (this is also the case in Fig. 3). Bottom right: Autonomous arrivals to hospital Haga (black), EWMA (coefficient 0.1) of realised autonomous arrivals (red), 3-days ahead expanding window predictions of the arrival rate by the Richards' curve model (orange) and 3-days ahead expanding window predictions of the arrival rate from regional infections (purple). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

is determined to a large extent by unknown factors (e.g. current regional shortage/surplus), while daily infections remain a good indicator for whether hospitalisations will go up or down. In addition, we have implemented an improved version of our t-days ahead prediction using a 5-parameter Richards' curve [40]. The improvements include the possibility of estimating a mixture of multiple Richards' curves, and to return a logistic, exponential or linear fit through (early-stage) data if this results in a smaller mean squared error (similar to the procedure described in Lee et al. [54]). The Richards' curve predictions were quite sensitive to outliers occurring around the time of prediction. To account for this, the Richards' curve was estimated on the arrival data up to point s, augmented with 7 days of arrival data set equal to the weighted average of the number of arrivals at time s.

The bottom row of Fig. 1 displays 3-days ahead expanding window predictions [40] of the arrival rate for NAZ West (left) and Haga hospital (right) in the period September 4, 2020 until January 31, 2021. The bottom left graph also includes the hospitalisations (dots) and the trend (red) from the top right graph. The orange line corresponds to the expanding window prediction of the arrival rates by the Richards' curve estimator. The purple line corresponds to the predictions generated from the daily number of infections. Observe that the predictions resulting from the infection data are less extreme when compared to those made by the Richards' curve predictor. This sensitivity of Richards curve forecasts to the prediction date (especially before the inflection point) was also seen in Wu et al. [55]. Furthermore, observe that the daily infections predictor shows an earlier increase around mid December, when the number of hospitalisations starts to increase again. We conclude that the prediction based on daily infections outperforms the prediction based on the Richards' curve using patient arrivals in the Hospital Information System.

Table 1
Accuracy measures for our forecasting method based on arrivals predicted by a Richards' curve, regional infections and the oracle for Haga hospital. CR: coverage rate of occupancy – 95% prediction interval; bias: estimated by averaging errors; MAE: mean absolute error.

		Ward			ICU		
Method	Type	CR	Bias	MAE	CR	Bias	MAE
Richards' Curve	1 day ah.	0.95	−0.09	2.03	0.97	0.15	0.62
	2 days ah.	0.91	−0.50	3.02	0.98	0.28	0.84
	3 days ah.	0.89	−0.88	3.80	0.96	0.40	0.98
	5 days ah.	0.83	−1.64	4.51	0.96	0.61	1.30
	7 days ah.	0.75	−3.85	5.30	0.96	0.70	1.53
	Max. 3d. ah.	0.69	1.63	3.39	0.87	0.44	0.87
	Max. 5d. ah.	0.74	1.31	3.77	0.91	0.58	1.07
	Max. 7d. ah.	0.75	0.92	3.90	0.93	0.67	1.23
Reg. Infections	1 day ah.	0.97	−0.17	2.12	0.97	0.14	0.62
	2 days ah.	0.93	−0.57	2.92	0.98	0.27	0.85
	3 days ah.	0.92	−0.79	3.32	0.95	0.40	0.99
	5 days ah.	0.92	−0.98	3.49	0.95	0.64	1.28
	7 days ah.	0.88	−2.64	3.91	0.97	0.78	1.51
	Max. 3d. ah.	0.72	1.42	3.21	0.87	0.42	0.88
	Max. 5d. ah.	0.81	1.16	3.38	0.91	0.58	1.06
	Max. 7d. ah.	0.81	1.00	3.42	0.93	0.70	1.21
True Arrivals	1 day ah.	0.93	−0.03	1.80	0.98	0.14	0.53
	2 days ah.	0.87	−0.31	2.37	0.96	0.25	0.73
	3 days ah.	0.90	−0.47	2.74	0.96	0.35	0.87
	5 days ah.	0.89	−0.65	3.02	0.97	0.48	1.08
	7 days ah.	0.87	−0.76	3.32	0.96	0.55	1.29
	Max. 3d. ah.	0.53	2.26	2.92	0.87	0.39	0.74
	Max. 5d. ah.	0.61	2.07	2.95	0.91	0.48	0.81
	Max. 7d. ah.	0.62	2.02	3.05	0.93	0.53	0.90

3.2. Forecast of bed occupancy

This section investigates the forecasting power of bed occupancy of our method and compares it to the improved version of the Richards' curve predictor of [40]. To evaluate the maximum possible gain in forecasting power, we compare our results with an oracle predictor of the arrivals, that uses the actual realised patient arrivals to forecast bed occupancy. The forecasts of bed occupancy used in this section are obtained by the sampling method presented in Baas et al. [40] for patient trajectories in the Poisson Arrival Location Model (PALM). We consider 3-days ahead expanding window forecasts for the daily and maximum occupancy for the Haga hospital in the period September 4, 2020 until January 31, 2021 as displayed in Fig. 2, and include Table 1 containing coverage rates, bias and MAE for the forecast occupancy.

The top row of Fig. 2 contains daily occupancy forecasts made by the oracle forecaster (cyan) and the daily infections forecaster (purple) for the ward (left) and ICU (right). The forecasts for the ICU lie closer to each other (Pearson correlation coefficient 0.96) than those for the ward (Pearson correlation coefficient 0.91). This might be due to the smaller number of direct autonomous arrivals to the ICU (28%) than to the ward (90%). Hence, the difference in predicted arrival rates is expected to have the largest influence on occupancy forecasts for the ward. Note that the pattern seen in the oracle forecast

for the ward is also seen in the daily infections predictions with a delay of three days, which makes sense as we are considering a 3-days ahead expanding windows forecast.

The middle row of Fig. 2 shows forecasts for the Richards' curve (orange) and daily infections (purple) along with the realised occupancy (red) in the ward (left) and ICU (right). The reported occupancy excludes patients reallocated from/to other hospitals. The two forecasts lie very close to each other (Pearson correlation coefficient 0.97 for the ward and 1.00 for the ICU), and close to the oracle. In accordance with the predictions for the arrival rates, the Richards' curve forecasts have a more fluctuating behaviour in comparison to the daily infections forecasts. The forecasts are close to the realised daily occupancy with a delay of 3 days and have a higher forecasting power when the occupancy decreases. A detailed comparison is included in Table 1. The difference between occupancy forecasts is again largest for the ward, which can be explained by the larger fraction of direct arrivals at the ward. Observe that the fluctuations in the arrival rate predictions in Fig. 1 are dampened in the forecast in Fig. 2, which may be partially explained as the load is obtained as integral over the arrival rates [46, Theorem 1.2].

The bottom row of Fig. 2 shows forecasts of the maximum occupancy over 3 days and their realisations in the ward (left) and ICU (right). The forecasts for maximum occupancy lie very close to each other for both ward (Pearson correlation coefficient 0.99) and ICU (Pearson correlation coefficient 1.00). Clearly, fluctuations in the maximum occupancy over 3 days are lower than those for daily occupancy predictions.

Table 1 presents a detailed overview and comparison of the quality of the forecasting results for our Richards' curve, daily infections and oracle forecasting methods for COVID-19 bed occupancy. We show results for coverage rate (CR), bias and mean absolute error (MAE), of which MAE is the most important measure as it captures the average absolute difference (distance) between forecast and realisation. For the ICU, we observe that the Richards' curve and regional infections forecast show similar results for all measures, and are outperformed by the oracle forecaster as is to be expected since the oracle forecaster uses the exact values for the number of patients in the hospital. For the ward, with respect to MAE an ordering can be found in the quality of the forecast: the oracle forecast outperforms the daily infections forecast, which in turn outperforms the Richards' curve forecast. This is not seen when looking at the 1-day ahead MAE for the Richards' curve and daily infections forecasts, as the Richards' curve is explicitly designed to extrapolate the current trend. The regional infections forecast also outperforms

the Richards' curve forecast for CR and bias when forecasting the maximum 3−, 5− and 7-days ahead. Note that the CR increases and bias decreases in the horizon for the ward for all forecasts.

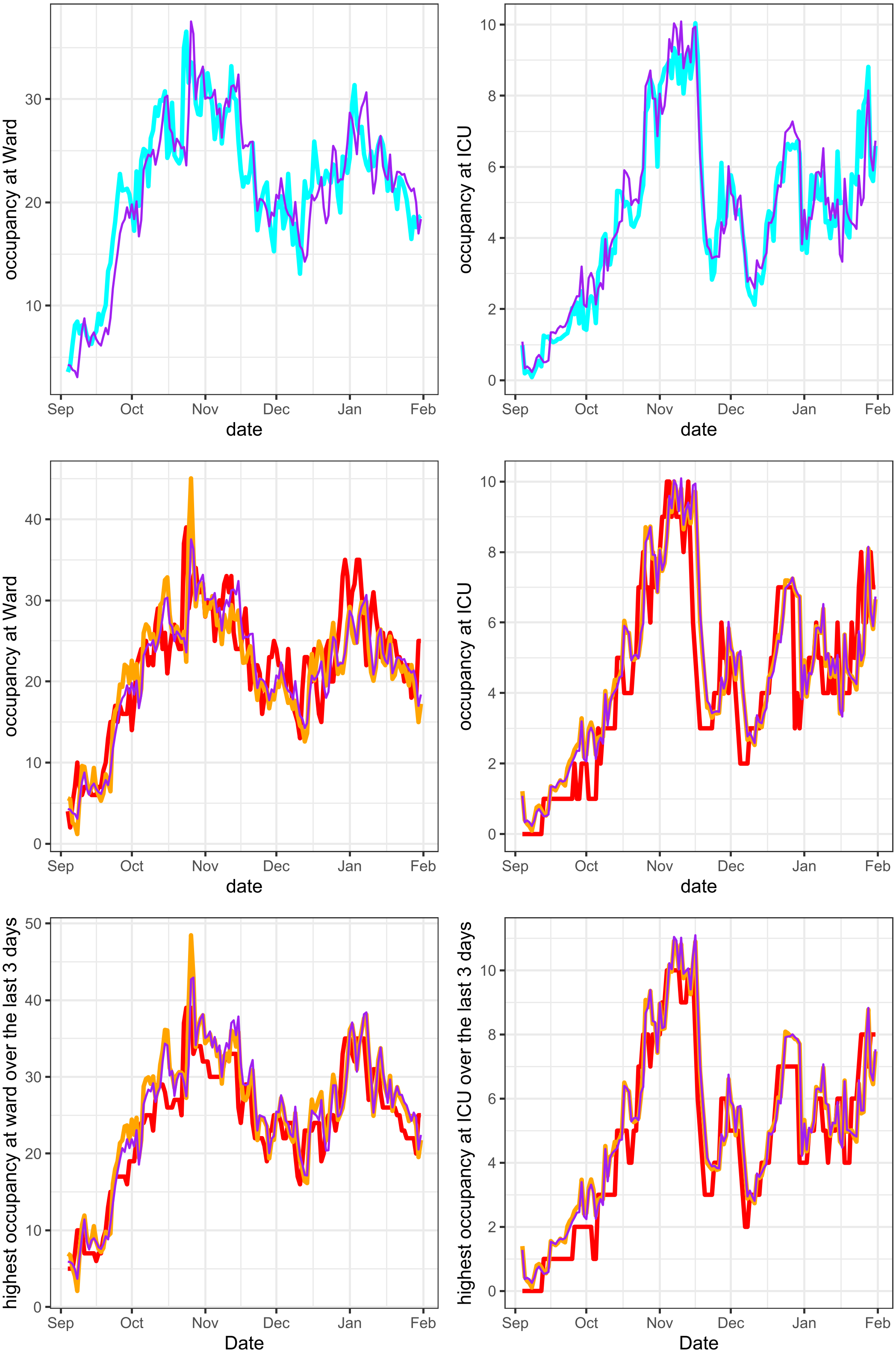

Fig. 2. Haga hospital, September 4, 2020 until January 31, 2021. Top row: 3 day ahead forecasts of the COVID-19 occupancy for the oracle (cyan) and daily infections (purple) forecasters. Middle row: 3 day ahead forecasts vs realised occupancy (red) for the Richards' curve (orange) and daily infections forecasters (purple). Bottom row: forecasts and realisations (red) of the maximum occupancy over the last 3 days for the Richards' curve (orange) and daily infections forecasters (purple). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

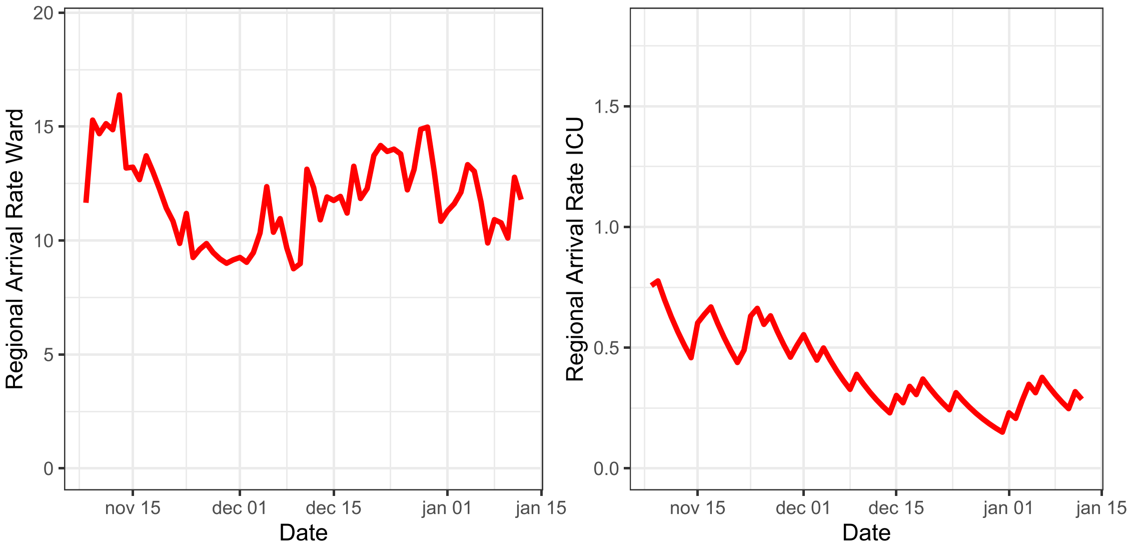

Fig. 3. Daily regional arrival rates of ward (left) and ICU (right) patients in the region for the simulation study of Section 4.1.

Apparently, the maximum occupancy forecast tends to overestimate the realisation for shorter horizons, while this bias lessens for longer horizons. Observe that the oracle forecast shows a larger bias in the expected maximum occupancy forecasts. This is because maximum occupancy forecasts are always higher than or equal to the current occupancy. Hence forecasts for the maximum often point in the right direction (less negative bias) when occupancy increases and in the wrong direction (more positive bias) when occupancy decreases. This effect occurs to a larger extent for the oracle, as it can better forecast increases in occupancy. As we are mainly interested in forecasts of the maximum occupancy, we conclude that the daily infections forecast is to be preferred over the Richards' curve forecast and will be used in subsequent sections.

Remark 7 (Possibility of Overfitting)**.** *A question that might arise is whether there might be an overfitting issue under the proposed method. While this is indeed a possibility when fitting the arrival predictor, this cannot be the case for the occupancy forecasts as we do not directly use occupancy data in the forecasts. Furthermore, all our evaluations are expanding window out-of-sample forecast evaluations, meaning that we never train and evaluate the model on the same data. Looking at the outcomes for the Haga data, the Richards' curve and regional infections forecasts for the realised arrivals are likely not an overfit on that data set as the forecasts follow the trend and clearly have less variance than the actual realisations, which would be the case with an extreme overfit.*

4. Numerical results

This section presents numerical results illustrating the performance of our hierarchical model that fairly balances COVID-19 patients over hospitals in a region and across regions. Section 4.1 considers allocation of patients to hospitals within a region, and Section 4.2 considers optimal reallocation of patients across regions.

4.1. Allocating regional COVID-19 patients to hospitals in a region

This section compares the impact of the three levels of regional coordination described in Section 2.2: individual hospitals (Section 2.2.1), load balancing (Section 2.2.2), and merging all hospitals into a regional hospital (Section 2.2.3) on the allocation of patients to hospitals in a region via a simulation study.

Table 2
Capacity and initial occupancy for each hospital used for the simulation study of Section 4.1. The capacity chosen here is not directly based on the actual capacity at the hospitals and merely chosen as such for the simulation study.

	Capacity		Init. occupancy	
Hospital	Ward	ICU	Ward	ICU
LUMC	29	11	22	7
Haga	41	17	30	13
GHZ	31	13	25	8

We consider occupancy by COVID-19 patients in the period 8 November 2020 until 7 January 2021 (60 days) at three hospitals in ROAZ region NAZ West: Groene Hart Ziekenhuis (GHZ), HagaZiekenhuis (Haga) and Leiden University Medical Center (LUMC). Each hospital is modelled as described in Section 2.1. The probability of class assignment and class-dependent LoS distributions are estimated using the estimation procedure from Baas et al. [40] from pooled data collected from the hospital data warehouses of the three hospitals. For each patient class, the same fixed allocation probability and LoS distribution is assumed for each hospital as well as for the merged regional hospital. Table 2 presents the number of beds and initial occupancy for each hospital that are used in our simulation study. The number of beds chosen here is of the order of magnitude of the actual capacity at the hospitals, but is chosen fixed to reveal the differences in the allocation methods. In practice, the number of beds may have fluctuated over the days.

The arrival process of patients to the ward and ICU is found to be a non-homogeneous Poisson process with arrival rates displayed in Fig. 3. These arrival rates are obtained using an exponentially weighted moving average (coefficient 0.3 for the ward and 0.1 for the ICU), and scaling (0.95) of the daily hospitalisations, obtained as described in Section 3.1. In the considered period, on average around 12 ward patients arrive per day with a fluctuating rate, and one ICU patient arrives every three days, with fewer patients arriving towards the end of the considered period.

Using the arrival processes, LoS distributions and transition probabilities, the PALM of the system of infinite server queues corresponding to each hospital can now be simulated on a day-to-day basis according to three levels of regional coordination, using the

Table 3
Average number of over-bed days and over-beds (with 95% confidence interval) for safety level 0.9 for each hospital.

Hospital	KPI	Indiv. hospitals	Load balancing	Reg. hospital
LUMC	Over-bed Days Ward	11.80 ± 2.35	5.70 ± 0.76	–
	Over-beds Ward	1.41 ± 0.41	0.40 ± 0.10	–
	Over-bed Days ICU	15.51 ± 2.53	4.43 ± 0.73	–
	Over-beds ICU	1.22 ± 0.29	0.13 ± 0.04	–
Haga	Over-bed Days Ward	9.37 ± 1.92	3.80 ± 0.55	–
	Over-beds Ward	1.17 ± 0.35	0.24 ± 0.04	–
	Over-bed Days ICU	14.70 ± 2.37	4.43 ± 0.84	–
	Over-beds ICU	1.15 ± 0.26	0.16 ± 0.05	–
GHZ	Over-bed Days Ward	12.51 ± 2.24	6.58 ± 0.85	–
	Over-beds Ward	1.22 ± 0.32	0.47 ± 0.09	–
	Over-bed Days ICU	9.32 ± 1.98	3.10 ± 0.57	–
	Over-beds ICU	0.57 ± 0.16	0.09 ± 0.02	–
Region	Over-bed Days Ward	30.58 ± 2.34	13.14 ± 1.23	2.07 ± 0.55
	Over-beds Ward	3.80 ± 0.52	1.12 ± 0.17	0.22 ± 0.10
	Over-bed Days ICU	36.62 ± 2.02	9.91 ± 1.17	2.56 ± 0.60
	Over-beds ICU	2.95 ± 0.33	0.38 ± 0.07	0.12 ± 0.03

method proposed in Baas et al. [40]. The time of evaluation for each day is set to 10 AM, in accordance with the time that Dutch hospitals report their occupancy. The three levels of regional coordination are as follows:

- **Individual Hospitals:** Patients are randomly allocated to hospitals according to fixed probabilities equal to fraction of COVID-19 patients allocated to that hospital over the evaluation period: 0.30 for LUMC, 0.43 for Haga and 0.27 for GHZ.
- **Load Balancing:** Patients are allocated to hospitals according to the load balancing Decision Rule 2. The estimation procedure for $\theta_{h,\alpha_{hW},\alpha_{hI}}(s,t)$ is given in Appendix A and is based on the daily infections predictor, see Section 3. The safety levels α_{hW}, α_{hI} were set to 0.9 for each hospital. When $\theta_{h,\alpha_{hW},\alpha_{hI}}(s,t) = 0$ for each hospital the allocation probabilities are set equal to those under the rule “Individual Hospitals”.
- **Regional Hospital:** All wards and ICUs are merged into a regional ward and ICU as described in Section 2.2.3. Patients are distributed over the hospitals according to Decision Rule 3.

In the simulation study, all patients are distributed over beds at hospitals in the region, and hence cannot be reallocated outside the region. If the capacity of a hospital is exceeded, patients stay at a so-called over-bed until the hospital’s bed shortage is resolved. An over-bed is an originally unequipped, non-staffed bed which is forcefully brought into operation.

For the three coordination levels, Table 3 presents average Key Performance Indicators (KPIs) with 95% confidence intervals based on Student’s t-distribution. The averages and confidence intervals are calculated based on 250 independent simulation replications under each policy, generated as described above. For each hospital, the number of over-bed days (days with a bed shortage) was determined for the ward and ICU. This KPI does not reveal the number of over-beds. To this end, the average daily number of over-beds (averaged number of occupied over-beds per day averaged over the evaluation period) was also determined for both departments at each hospital. Under the rules “Individual Hospitals” and “Load Balancing” the total number of over-bed days and over-beds for the region was determined by summing the KPIs for the individual hospitals. The last column of Table 3 includes only the number of over-bed days and the number of over-beds for the region as under Decision Rule 3 patients are allocated to an over-bed only if none of the hospitals in the region has a bed surplus.

We observe a clear ordering in the performance of the allocation rules. At all hospitals, Load Balancing yields a significant reduction of the number of over-bed days of around 50% for the ward and around 60–75% for the ICU when compared to Individual Hospitals. This also holds for the average number of over-beds with an approximate 60–70% reduction for the ward and 80–90% reduction for the ICU. Regional Hospital further significantly improves performance, reducing the number of over-bed days to around 2 for the regional ward and ICU, with on average only 0.22 (ward) and 0.12 (ICU) over-beds needed per day. Appendix B contains a plot displaying the evolution of the occupancy under the regimes Load Balancing and Individual Hospitals, as well as a plot of the evolution of the allocation probabilities under Decision Rule 2. These results show the clear advantage of regional collaboration.

4.2. Reallocating COVID-19 patients across regions

This section considers four policies for reallocating patients across regions to alleviate bed shortages. We consider occupancy in the period 8 November 2020 until 7 January 2021 (60 days) for four ROAZ regions: Acute Zorgregio Oost (region 1), Netwerk Acute Zorg Brabant (region 2), Netwerk Acute Zorg West (region 3) and Traumazorgnetwerk Midden-Nederland (region 4), augmented with an external region. The LoS and transition probabilities are estimated from the data warehouses of a representative large hospital in each region. To focus on the reallocation across regions, each region is modelled as a (virtual) regional ward and ICU, as described in Section 2.2.3.

Patients arrive autonomously to either the ward or ICU in each region according to the arrival rates displayed in Fig. 4. These arrival rates are predicted using the procedure outlined in Section 3.1 using an exponentially weighted average (coefficient 0.3) of the hospital admissions reported in the data set with daily infections for each ROAZ region [50]. This regional arrival rate is then multiplied with the historical fraction (from the hospital data warehouses) of patients allocated to the ward and ICU to obtain the autonomous arrival rates to the wards and ICUs. To take into account that COVID-19 beds are occupied by both COVID-19 confirmed and non-confirmed COVID-19 patients, the hospitals’ arrival rates are scaled to obtain autonomous arrival rates close to those reported in (for instance) the dashboard [56] for the evaluation period. The capacity and initial occupancy, as well as the scaling factor used for the arrival rate for both departments at each of the regions is given in Table 4.

If the occupancy in a region exceeds capacity (measured at 10 AM, the evaluation time), a patient has to be reallocated to another region. We consider the following policies:

- **Any Surplus:** A region has sufficient surplus at a department (ward or ICU) when the surplus of beds exceeds a safety threshold k_{as}, which in our experiments is set to $k_{as} \in \{0, 1, 2\}$.

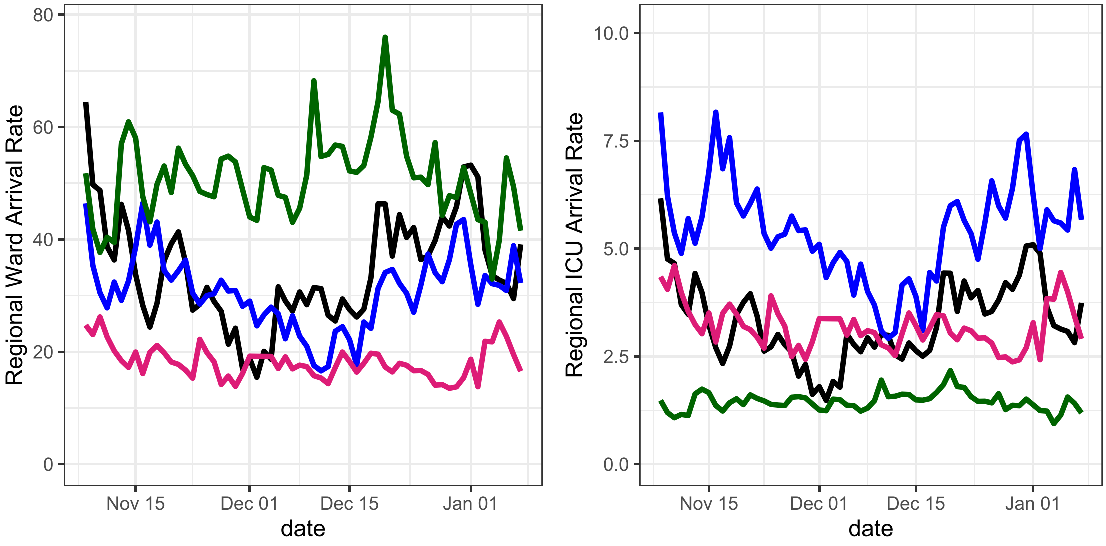

Fig. 4. Daily regional arrival rates to the ward (left) and ICU (right) to region 1 (black), 2 (green), 3 (blue) and 4 (dark red) in the period 8 November 2020 until 7 January 2021. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Capacity, initial occupancy and scaling factor for each region.

	Capacity		Init. occupancy		Scaling factor	
Region	Ward	ICU	Ward	ICU	Ward	ICU
1	183	61	155	55	8.0	20
2	371	165	305	92	5.0	2
3	212	114	162	63	2.5	10
4	170	80	151	58	2.5	10

This minimal surplus is included to guarantee bed-capacity for autonomous arrivals in the region during the day. If a region r has bed shortage in ward and/or ICU at the evaluation time and other regions have sufficient bed surplus at the respective departments, a patient from region r is reallocated with equal probability to any of the regions having sufficient surplus at the respective departments. Decisions for reallocation of patients are taken one-by-one, such that for the next patient reallocated the occupancy of the regions takes into account the previous reallocation decisions. If after patient reallocation to a region the surplus of the department in that region is no longer sufficient (i.e., less than k_{as}), the department at that region is no longer considered for reallocation. If there is no region with a sufficient surplus for the respective department, the patient is reallocated to the external region.

- **Number of Beds:** This policy differs from the Any Surplus policy in that patients are reallocated with equal probability to any of the available (surplus) beds in the regions with a sufficient surplus. The effect of this is that regions with a larger bed surplus at the respective department have a larger probability to receive patients.
- **Stochastic Program:** This policy reallocates patients to regions according to the here-and-now decision coming from Program (P), which is a stochastic program. To approximate the optimal solution to the stochastic program, sample average approximation is used for the second stage. The scenarios are determined from Decision Rule 3 using 1000 samples of the maximum occupancy obtained from the PALM of the system of infinite server queues for the daily infections forecast and a horizon of 3 days. The costs $\gamma_{r,r'}$ of patient reallocation from region r to r' are given in Table 5. This cost matrix indicates that regions 1, 4 and regions 2, 3 lie close to each other, i.e., the regions are clustered in clusters of two. Next, we take $g = x \mapsto x^2$.
- **Integer Program:** This policy reallocates patients to regions according to the here-and-now decision coming from Program (P′), which is an integer program. The forecasts are determined from Decision Rule 3 using 1000 samples of the maximum occupancy obtained from the PALM of the system of infinite server queues for the daily infections forecast and a horizon of 3 days. The costs of patient reallocation and g are the same as in the stochastic program.

Given the initial occupancy, arrival rates, LoS distribution and probabilities of transfers to other departments, the PALM of the system of infinite server queues corresponding to each region modelled as a single hospital is simulated according to the method described in Baas et al. [40] on a day-to-day basis for each day in the evaluation period. The results of the simulation study are reported in Tables 6 and 7. The policies are evaluated on the average (over 250 replications) total amount of reallocated patients (Total), amount of patients reallocated to the external region, amount of patients reallocated across regions within the clusters (across regions 1 and 4 and across regions 2 and 3, In clusters), between clusters (Btw. clusters) and the average cost per reallocation, defined as $\sum_{s,r,r'} \gamma_{r,r'} f_{W,r,r'}(s) / \sum_{s,r,r'} f_{W,r,r'}(s)$ for the ward and by analogy for the ICU. To evaluate the significance of differences in the averages, the 95% confidence interval based on Student's t-distribution is also shown in the tables.

Table 6 includes results for $k_{as} = 2$, i.e., for regions that reserve 2 ward and ICU beds for autonomous same-day admissions. The four policies do not show a significant difference in the average total number of reallocated patients in the wards and ICUs (confidence intervals are overlapping), but do show a clear difference in number of patients reallocated to the external region, and most importantly, both optimisation models considerably reduce the number of patients reallocated out of a cluster (by roughly 50%) compared to the other methods and thereby also the average reallocation cost. Most of the KPIs do not differ significantly between the stochastic program (Stoch. Prog.) and the integer program (Int. Prog.). It is seen that under the stochastic program, more patients are reallocated across regions in the clusters, and on average, more patients are reallocated in total. Further, there is a signif-

Table 5
Cost of patient reallocation across regions.

From/To	Region 1	Region 2	Region 3	Region 4	External region
Region 1	10	50	50	10	100
Region 2	50	10	10	50	100
Region 3	50	10	10	50	100
Region 4	10	50	50	10	100
External region	100	100	100	100	100

Table 6
Average number of patient reallocations across regions (with 95% confidence interval) for $k_{as} = 2$.

KPI	Dept.	Any Surplus	Num. Beds	Stoch. Prog.	Int. Prog.
Total	Ward	186.14 ± 7.60	178.43 ± 7.43	179.92 ± 7.48	173.80 ± 6.37
	ICU	96.91 ± 5.25	92.66 ± 5.13	101.76 ± 5.65	100.40 ± 5.92
External	Ward	8.74 ± 1.55	6.44 ± 1.45	4.67 ± 1.19	3.50 ± 1.05
	ICU	21.10 ± 2.83	17.92 ± 2.56	12.70 ± 2.07	14.14 ± 2.31
In clusters	Ward	63.37 ± 2.81	61.20 ± 2.66	134.53 ± 5.10	122.11 ± 4.26
	ICU	27.09 ± 1.54	27.20 ± 1.59	55.96 ± 2.68	50.90 ± 2.83
Btw. clusters	Ward	114.04 ± 4.91	110.80 ± 4.91	40.72 ± 3.72	48.20 ± 3.75
	ICU	48.72 ± 2.24	47.54 ± 2.38	33.10 ± 2.46	35.37 ± 2.32
Average cost	Ward	38.51 ± 0.59	37.69 ± 0.58	20.51 ± 0.77	22.25 ± 0.77
	ICU	46.92 ± 1.28	45.56 ± 1.25	31.22 ± 1.36	33.98 ± 1.25

Table 7
Average number of patient reallocations across regions (with 95% confidence interval) for $k_{as} = 0, 1, 2$.

		$k_{as} = 0$	$k_{as} = 1$	$k_{as} = 2$	
KPI	Dept.	Num. Beds	Num. Beds	Num. Beds	Int. Prog.
External	Ward	4.34 ± 1.11	6.04 ± 1.40	6.44 ± 1.45	3.50 ± 1.05
	ICU	12.65 ± 2.25	16.95 ± 2.39	17.92 ± 2.56	14.14 ± 2.31
In clusters	Ward	62.33 ± 2.94	63.18 ± 3.01	61.20 ± 2.66	122.11 ± 4.26
	ICU	31.32 ± 1.87	29.64 ± 1.60	27.20 ± 1.59	50.90 ± 2.83
Btw. clusters	Ward	115.14 ± 5.35	116.02 ± 4.98	110.80 ± 4.91	48.20 ± 3.75
	ICU	55.84 ± 2.92	51.12 ± 2.55	47.54 ± 2.38	35.37 ± 2.32
Average cost	Ward	37.12 ± 0.49	37.76 ± 0.57	37.69 ± 0.58	22.25 ± 0.77
	ICU	41.56 ± 0.97	44.22 ± 1.12	45.56 ± 1.25	33.98 ± 1.25

icantly lower average reallocation cost. This is due to the stochastic program being less conservative than the integer program.

Table 7 presents the results of bed reservation via the thresholds k_{as} for the Number of Beds policy (that outperforms the Any Surplus policy) and the integer program. With increasing k_{as} the number of patients reallocated to the external region increases, as is to be expected. Our optimisation model slightly outperforms the Number of Beds policy. The Number of Beds policy may be a good heuristic if hospitals and regions may be convinced to not use safety beds, that are not required in our collaborative optimisation policy.

In conclusion, the integer program is a good approximation of the stochastic program, that avoids reallocations to the external region and is able to reallocate patients to closer regions, while keeping the total number of reallocations at roughly the same level.

Sensitivity analyses, evaluating the results presented in this and the previous subsections under different horizons and safety levels are presented in Appendix C.

5. Discussion and conclusion

This paper has introduced mathematical models and decision rules for dynamic fair balancing of COVID-19 patients over hospitals in a region and across regions. Patient flow is captured in the Poisson Arrival Location Model (PALM) and the corresponding network of infinite server queues for the ward and Intensive Care Unit (ICU) of a single hospital. The model includes transfers between ward and ICU and allows determining safety levels for ward and ICU bed occupancy and corresponding forecasts of bed surplus or bed shortage in the ward and ICU of each hospital or region. The dynamic fair balancing approach within a region is based on a dynamic predictive load balancing model incorporating a forecast of the occupancy based on publicly available regional infection data and Length of Stay (LoS) and transfer probabilities obtained from the Hospital Information System (HIS). This model extends load balancing models in literature to include real-time estimations of the arrival process, service and routing processes and their impact on forecast occupancy. The dynamic fair balancing model across regions is a stochastic program that may be accurately approximated by a mixed integer program taking into account forecasts of the future bed surpluses or shortages. It hence takes into account both the current occupancy and the forecast maximum occupancy over the next couple of days.

Our mathematical model is augmented by accurate statistical methods to predict patient arrivals, estimate LoS and transfer probabilities. For LoS and transfer probabilities, we have used the Kaplan-Meier estimators for censored data as developed in Baas et al. [40]. For patient arrivals, we have both improved prediction of patient arrivals based on the HIS and Richards' curves that was developed and shown to be very accurate in Baas et al. [40] and developed prediction of patient arrivals based on regional infection data. We have found that the latter provides better results as it captures changing trends in hospitals' arrival rates a few days earlier than the HIS data. In addition, for our dynamic load balancing model, we have developed an estimator of the load balancing dynamic allocation fractions of patients to hospitals in a region. Our forecasting method for bed occupancy is based on simulation of the PALM as developed in Baas et al. [40] using the estimated

LoS and transfer probabilities and predicted arrivals based on regional infection data.

Our dynamic fair balancing models and statistical methods yield implementable decision rules for patient allocation to hospitals in a region or reallocation across regions based on safety levels and forecast bed surplus or bed shortage for each hospital or region. We have tested accuracy of our forecast using HIS data from September 4, 2020 until January 31, 2021 of hospitals in the ROAZ region *Netwerk Acute Zorg West*, containing the hospitals Groene Hart Ziekenhuis (GHZ), HagaZiekenhuis (Haga) and Leiden University Medical Center (LUMC). Our forecast of bed occupancy and of maximum bed occupancy over the next couple of days are shown to be very accurate. Using these forecasts, we have investigated the benefits of three levels of regional collaboration: individual hospitals (or no collaboration among hospitals), dynamic load balancing and merging all hospitals into a (virtual) regional hospital. The regional hospital exploits the statistical multiplexing gain and clearly makes optimal use of available beds, but may include patient transfers from the ward of one hospital to the ICU of another and requires hospitals to give complete control over patient admission to a regional dispatcher. Load balancing allows hospitals to govern their own policy and has clear and substantial benefits with respect to levelling the load over hospitals in the region.

The intra-regional load balancing decision rule may be developed into a decision support tool and incorporated in the ROAZ dashboard for allocating patients to hospitals. First steps in this direction have been set in collaboration with ROAZ region *Netwerk Acute Zorg West*. We have explored optimal reallocation of patients across regions based on current and forecast load in the regions and found that our decision rule that takes into account reallocation costs across regions and the current and forecast load in the regions results in fewer reallocations to regions far away. This inter-regional reallocation rule requires the same information as shared with the Landelijk Coördinatie-centrum Patiëntenspreiding (LCPS) and may be developed into a decision support tool for patient reallocation.

In addition to developing our results into decision support tools, several points for further research or improvement may be addressed. In our simulation study we considered a fixed decision epoch at 10 AM each day. As a consequence, a patient arriving in-between two decision epochs is admitted to an over-bed until the next decision epoch. Immediate reallocation of this patient may be included in our simulation approach. However, this requires a real-time update of new admissions, discharges and reallocations among hospitals for all hospitals or all regions. As this results in increased dependence among decision epochs, a Markov decision process approach might also be investigated. In our hierarchical model, we split the decisions for inter-regional reallocations and load balancing within the region. As long as we consider the region as a single hospital, this does not influence the number of inter-regional reallocations since the bed surplus/shortage of a region is independent of (and determined prior to) the load-balancing allocation of patients to hospitals. Integrating the intra-regional and inter-regional decision levels is an open question for further research.

Given the quality of our forecasts, the significant reduction in reallocations to distant regions and the significant improvement of balanced load among hospitals within a region, we are confident that our decision rules provide an important step towards practical implementation of a decision support tool for real-time reallocation of COVID-19 patients. Moreover, our methodology may also be beneficial for patient reallocation during future pandemics or national outbreaks, with fine-tuning of the statistical methods. Our mathematical models are generic and not specific to COVID-19; all they require is data from which patient arrival rates can be predicted as well as in-hospital data on patient transfers and discharges. Lastly, we envision our dynamic load balancing procedure to be applicable well beyond the scope of patient reallocation. We aim to unlock this potential in future research, by incorporating our dynamic load balancing procedure in a generic queueing framework.

Data availability

Data will be made available on request.

CRediT authorship contribution statement

Sander Dijkstra: Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. **Stef Baas:** Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – review & editing. **Aleida Braaksma:** Formal analysis, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing. **Richard J. Boucherie:** Conceptualization, Funding acquisition, Formal analysis, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

Acknowledgements

The authors would like to thank Fieke J. Snijders, Lars Tiemessen and Plom van Rooij for providing patient record data of Leiden University Medical Center, HagaZiekenhuis, Groene Hart Ziekenhuis, Elisabeth-TweeSteden Ziekenhuis and Rijnstate and helpful discussions to clarify patient reallocation protocols in the Netherlands.

Appendix A. Dynamic scaling of arrival rates under load balancing

Allocation of patients to hospitals in a region under Decision Rules 2 and 3 requires estimation of the allocation fractions $\theta_{h,\alpha_{hW},\alpha_{hI}}$. This appendix presents an estimation procedure to obtain $\theta_{h,\alpha_{hW},\alpha_{hI}}$ from the sample paths of the PALM.

Consider day s and forecast horizon t. Recall the notation introduced in Section 2. Fix $h, \ell_h, \alpha_{hW}, \alpha_{hI}$, $n^*_{hW}(s,t), n^*_{hI}(s,t)$ and let

$$\pi_{hW}(\theta) = \mathbb{P}_{\theta,h}\left[\max_{u\in[s,s+t]} N_{hW}(u) \le n^*_{hW}(s,t) \mid \mathbf{L}_h(s) = \ell_h\right] - \alpha_{hW}$$

and define $\pi_{hI}(\theta)$ similarly.

If $N_{hW}(s) > n^*_{hW}(s)$ or $N_{hI}(s) > n^*_{hI}(s)$, we set $\theta_{h,\alpha_{hW},\alpha_{hI}} = 0$. Otherwise, we estimate $\theta_{h,\alpha_{hW},\alpha_{hI}}$ as described below. Given an initial value $\chi_{h,0}$, consider the sequence $(\chi_{h,n})_n$:

$$\chi_{h,n+1} = \chi_{h,n} + a(n)\min\left(\tilde{\pi}_{hW,n}(\exp(\chi_{h,n})), \tilde{\pi}_{hI,n}(\exp(\chi_{h,n}))\right), \tag{14}$$

with step-size $a(n)$ satisfying the Robbins-Monro conditions [57] and

$$\tilde{\pi}_{hW,n}(\exp(\chi_{h,n})) = \left[\frac{1}{M_{\text{in}}}\sum_{i=1}^{M_{\text{in}}} \mathbb{1}_{\left[\max_{u\in[s,s+t]} N^{(n,i)}_{hW}(u) \le n^*_{hW}(s,t)\right]}\right] - \alpha_{hW}$$

$$\tilde{\pi}_{hI,n}(\exp(\chi_{h,n})) = \left[\frac{1}{M_{\text{in}}}\sum_{i=1}^{M_{\text{in}}} \mathbb{1}_{\left[\max_{u\in[s,s+t]} N^{(n,i)}_{hI}(u) \le n^*_{hI}(s,t)\right]}\right] - \alpha_{hI}.$$

In the above, $\left(N^{(n,i)}_{hW}(u), N^{(n,i)}_{hI}(u)\right)_{u\in[s,s+t]}$ is a sampled trajectory of the occupancy in the period $[s, s+t]$ given the current state ℓ_h

and scaling factor $\theta = \exp(\chi_{h,n})$ for the arrival rate. The trajectories are independent over i, n and can be sampled according to the PALM simulation method given in Baas et al. [40]. The parameter M_{in} denotes a number of inner simulations, which are performed for each iterate in (14). From the above, it can be seen that $\tilde{\pi}_{hW,n}(\exp(\chi_{h,n}))$, $\tilde{\pi}_{hI,n}(\exp(\chi_{h,n}))$ are bounded random variables with expectation equal to $\pi_{hW}(\exp(\chi_{h,n}))$, $\pi_{hI}(\exp(\chi_{h,n}))$. The sequence defined by (14) is a Robbins-Monro sequence [57], for which we verify (almost sure) convergence below.

Let

$$\eta(x) = \mathbb{E}[\min\left(\tilde{\pi}_{hW,1}(\exp(x)), \tilde{\pi}_{hI,1}(\exp(x))\right)].$$

If $\eta(x) = 0$ for some x and $\pi_{hW}(0), \pi_{hI}(0) \geq 0$, all assumptions in Blum [58] are satisfied, so that $(\chi_{h,n})_n$ converges (almost surely) to a constant limit χ_h such that $\eta(\chi_h) = 0$.

Next, consider the case $\eta(x) \neq 0$ for all $x \geq 0$. It can be shown that η is a decreasing, differentiable function with $\lim_{x\to\infty} \eta(x) = \min(-\alpha_{hW}, -\alpha_{hI})$. Hence if $\eta(x) \neq 0$ for all $x \geq 0$, we have that η is negative everywhere. The minimum in (14) can be decomposed in $\eta(\chi_{h,n})$ and a martingale difference $\epsilon_{h,n}(\chi_{h,n})$, from which it follows that $\chi_{h,n}$ is the sum of an almost surely converging martingale (bounded in L^2 as $\sum_n \alpha(n)^2 < \infty$) and a deterministic series with negative increments. From this, it follows that $\chi_{h,n} \to -\infty$ almost surely.

Now, set $\theta_{h,n} = \exp(\chi_{h,n})$. By the above discussion it follows that either $\theta_{h,n} \to 0$ in the case of bed shortage or

$$\begin{aligned}\min&\left(\mathbb{E}\left[\tilde{\pi}_W(\theta_{h,n})\right], \mathbb{E}\left[\tilde{\pi}_I(\theta_{h,n})\right]\right)\\ &\geq \mathbb{E}[\min\left(\tilde{\pi}_W(\theta_{h,n}), \tilde{\pi}_I(\theta_{h,n})\right)] = \eta(\chi_{h,n}) \to 0 \qquad (15)\end{aligned}$$

in the case of bed surplus at hospital h. Hence, in the latter case, the limit of the sequence $\theta_{h,n}$ satisfies the condition given in (5). We estimate $\theta_{h,\alpha_{hW},\alpha_{hI}}$ as the almost sure limit of $\theta_{h,n}$ in the manner described below.

For each day s, we sample the sequence defined in (14) with $\chi_{h,0}$ the logarithm of $\theta_{h,\alpha_{hW},\alpha_{hI}}$ obtained for day $s-1$. For day 0, we set $\chi_{h,0}$ equal to the logarithm of the historical fraction of COVID-19 patients allocated to the hospital. We chose step-size $a(n) = n^{-0.51}$ (satisfying the Robbins–Monro conditions [57]) and $M_{\text{in}} = 5$, which was seen to result in fast convergence of the iterates to a stationary point. Convergence of the sampler is assessed by checking for every batch of 300 iterations whether the batch-mean of $\exp(\chi_{h,n})$ is smaller than 10^{-5} or whether the batch-mean of the residuals $\min\left(\tilde{\pi}_{hW,n}(\exp(\chi_{h,n})), \tilde{\pi}_{hI,n}(\exp(\chi_{h,n}))\right)$ is smaller than 0.01. After diagnosing convergence on a batch of iterations, $\theta_{h,\alpha_{hW},\alpha_{hI}}$ is estimated as the mean of samples $\theta_{h,n}$ for that batch.

Appendix B. Sample path of hospitals' occupancy in single region

Fig. 5 presents a sample path of the ward and ICU occupancy for all hospitals in a region for the coordination levels Load Balancing (left) and Individual Hospitals (right) under the simulation study described in Section 4.1. Jumps occur on a daily basis, starting at November 8, 2020, 10:00. The evolution of the allocation probabilities is included in the bottom row of the figure.

Under Load Balancing, an inverse proportional relation may be observed between occupancy at a certain hospital and its allocation probability. Note that as $\theta_{h,\alpha_{hW},\alpha_{hI}}(s,t)$ aims to control both the occupancy in the ward and ICU, $\theta_{h,\alpha_{hW},\alpha_{hI}}(s,t)$ is low (often zero) when one of these departments reaches a bed shortage. It is seen in this sample path that the ICU at LUMC is often full, hence the corresponding allocation probability is also often set to zero. Note that it seems harder to control the occupancy at the ICU using the allocation probability, as most of the patients at the ICU originate from the ward. During periods when every hospital has a crowded department, the allocation probabilities are seen to have a fluctuating behaviour, often sending all patients to one hospital one day, and the next day to another. This can sometimes lead to a large increase in over-beds, for instance in the ward around January 5, for Haga. Around December 21, 2020, there were bed shortages at the ICU of hospitals GHZ and LUMC and in the ward at Haga, as a result the historical allocation probabilities (0.30,0.43,0.27 for hospitals GHZ, Haga and LUMC resp.) were used around this period. Note that as departments at the hospitals are already over-occupied during this period, setting these probabilities larger than zero will lead to an even larger bed shortage. This is a consequence of the setup in Section 4.1, where patients have to be admitted to a hospital in the region. In reality, as is also considered in Section 4.2, patients will be allocated out of the region. From the sample path it can be seen that bed shortages are often

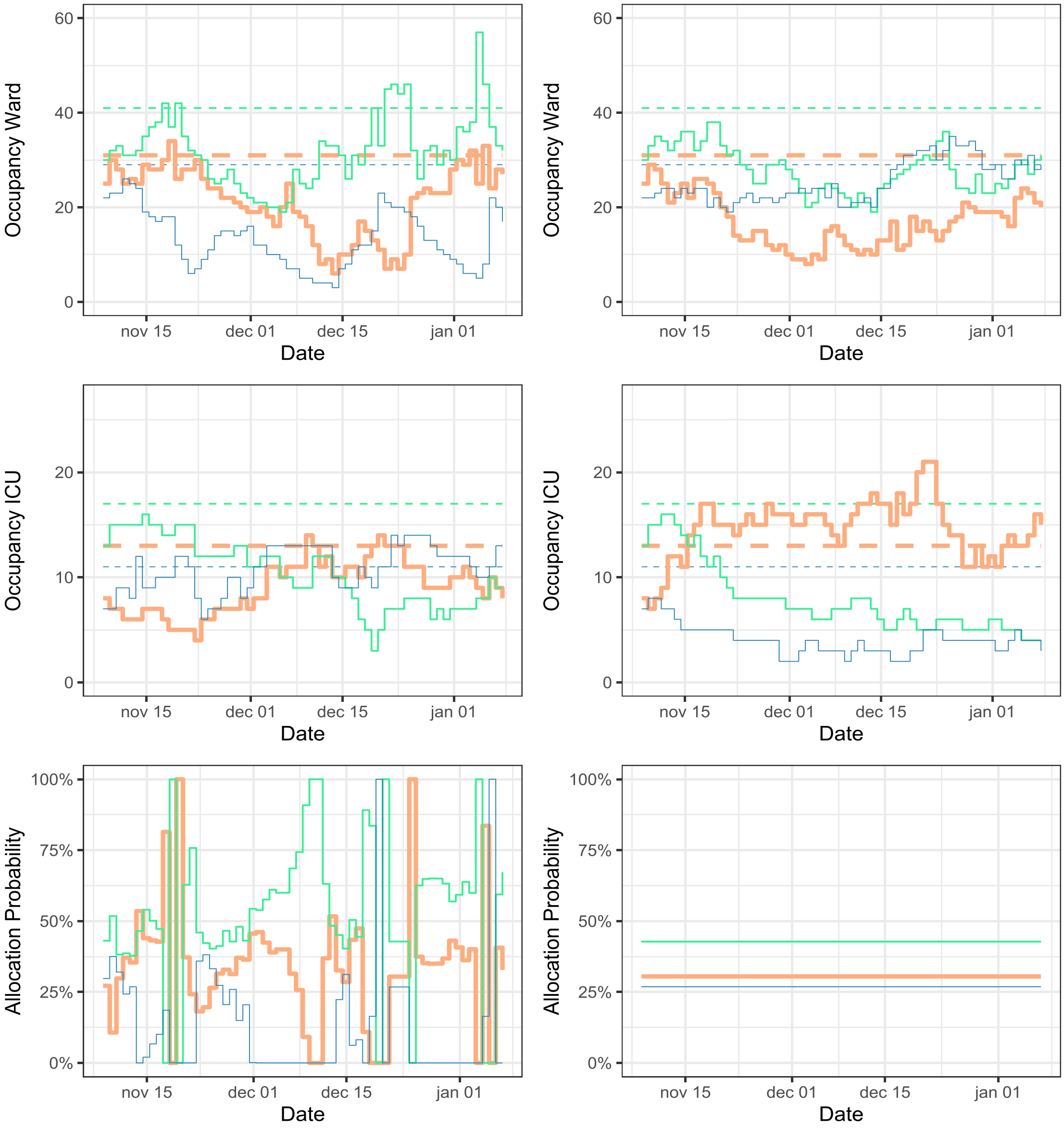

Fig. 5. One sample path of occupancy (solid) in the ward (top) and ICU (middle) over the simulation period of the simulation study from Section 4.1 for hospitals GHZ (orange), Haga (green) and LUMC (blue) under the Load Balancing dynamic allocation rule (left) and the Individual Hospital rule (right). The dashed lines represent the capacity at the respective departments and hospitals and the (dynamic) allocation probabilities are shown in the bottom plot with colours matching those for the hospitals. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

resolved quickly, often in a matter of a few days. Under the Individual Hospitals scenario, the allocation probabilities stay constant over time. The result is that there are long periods of bed shortages, as can be seen from Fig. 5. The ICU is often overcrowded at GHZ, while the ICUs at the other hospitals become almost empty. On a sample path level, the Load Balancing rule indeed seems to show a more balanced behaviour of the occupancy over time.

Appendix C. Sensitivity analyses

In this Appendix sensitivity analyses are performed, showing what the results in Sections 4.1 and 4.2 look like under different horizons and safety levels.

We first evaluate the results for our load balancing method under different scenarios for the safety levels while keeping the rest of the setup in Section 4.1 the same. In scenario 1 we set all safety levels to 0.9, in scenario 2 we set $\alpha_{\text{Haga}} = 0.7$ and all other safety levels to 0.9 and in scenario 3 we set $\alpha_{\text{LUMC}} = 0.99, \alpha_{\text{Haga}} = 0.7, \alpha_{\text{GHZ}} = 0.9$, these safety levels hold both for the ward and the ICU. These scenarios were chosen as Haga (LUMC) is the largest (smallest) hospital out of the three, hence it might be tempted to set a lower (higher) safety level than the other hospitals in practice. The results are given in Table 8, where we have also shown the average occupancy at both the ward and ICU and the average load fraction ($\theta_{h,\alpha_{hW},\alpha_{hI}}(s,t)$).

When comparing scenario 1 with scenario 2, no significant differences are seen for the over-beds and the occupancy. On average, the over-beds, occupancy and load coefficient for Haga are higher, while these measures stay roughly the same for the other hospitals.

When comparing scenario 1 with scenario 3, significant differences are seen, the occupancy at both departments at LUMC is significantly lower, while the KPIs for over-beds stay roughly the same. For Haga, the over-bed (days) and occupancy significantly increase for the ward, while they stay about the same for the ICU as occupancy at the ICU is less sensitive to the direct arrivals. Finally, the number of over-bed days at the ward of GHZ also significantly increases. This could be explained by the fact that the load balancing rule sends patients to Haga most of the time in scenario 3, and if bed shortages occur there almost all patients will go to GHZ as the safety level of LUMC is a lot higher. A significantly lower (higher) load coefficient is seen at LUMC (Haga) for scenario 3 than for scenario 1, while the load coefficient for GHZ decreases slightly. When comparing the KPIs for the region as a whole for the three scenarios, scenario 1 is the preferred choice as it has the lowest amount of over-bed (days) on average.

Table 8
KPIs for the load balancing allocation rule for patients to hospitals. In scenario 1 we set all safety levels to 0.9, in scenario 2 we set $\alpha_{\text{Haga}} = 0.7$ and all other safety levels to 0.9 and in scenario 3 we set $\alpha_{\text{LUMC}} = 0.99, \alpha_{\text{Haga}} = 0.7, \alpha_{\text{GHZ}} = 0.9$.

Hospital	KPI	Scenario 1	Scenario 2	Scenario 3
LUMC	Over-bed Days Ward	5.70 ± 0.76	5.28 ± 0.67	5.26 ± 0.71
	Over-bed Days ICU	4.43 ± 0.73	4.82 ± 0.93	3.82 ± 0.81
	Over-beds Ward	0.40 ± 0.10	0.38 ± 0.08	0.41 ± 0.07
	Over-beds ICU	0.13 ± 0.04	0.16 ± 0.05	0.12 ± 0.03
	Occupancy Ward	20.42 ± 0.50	20.12 ± 0.51	18.31 ± 0.50
	Occupancy ICU	7.95 ± 0.21	7.90 ± 0.23	7.33 ± 0.20
	Load Fraction	0.26 ± 0.01	0.28 ± 0.02	0.17 ± 0.01
Haga	Over-bed Days Ward	3.80 ± 0.55	4.39 ± 0.64	5.81 ± 0.73
	Over-bed Days ICU	4.43 ± 0.84	5.06 ± 0.91	5.00 ± 0.72
	Over-beds Ward	0.24 ± 0.04	0.26 ± 0.05	0.37 ± 0.06
	Over-beds ICU	0.16 ± 0.05	0.17 ± 0.04	0.14 ± 0.03
	Occupancy Ward	30.04 ± 0.50	30.58 ± 0.53	31.59 ± 0.49
	Occupancy ICU	13.02 ± 0.26	13.10 ± 0.29	13.44 ± 0.25
	Load Fraction	0.41 ± 0.02	0.49 ± 0.02	0.44 ± 0.02
GHZ	Over-bed Days Ward	6.58 ± 0.85	6.57 ± 0.95	8.14 ± 0.83
	Over-bed Days ICU	3.10 ± 0.57	3.13 ± 0.58	3.54 ± 0.65
	Over-beds Ward	0.47 ± 0.09	0.48 ± 0.10	0.57 ± 0.08
	Over-beds ICU	0.09 ± 0.02	0.08 ± 0.02	0.10 ± 0.03
	Occupancy Ward	23.51 ± 0.48	23.04 ± 0.52	24.01 ± 0.47
	Occupancy ICU	9.11 ± 0.22	8.87 ± 0.25	9.23 ± 0.23
	Load Fraction	0.29 ± 0.01	0.30 ± 0.02	0.27 ± 0.01
Region	Over-bed Days Ward	13.14 ± 1.23	13.49 ± 1.33	16.11 ± 1.29
	Over-bed Days ICU	9.91 ± 1.17	10.66 ± 1.31	10.13 ± 1.17
	Over-beds Ward	1.12 ± 0.17	1.12 ± 0.17	1.35 ± 0.16
	Over-beds ICU	0.38 ± 0.07	0.41 ± 0.08	0.35 ± 0.06
	Occupancy Ward	73.97 ± 0.97	73.74 ± 1.02	73.92 ± 0.97
	Occupancy ICU	30.07 ± 0.45	29.87 ± 0.49	30.00 ± 0.45

Table 9
KPIs for the load balancing allocation rule for patients to hospitals for forecasting horizons 3 and 5 under the simulation study setup of Section 4.1.

Hospital	KPI	$s = 3$	$s = 5$
LUMC	Over-bed Days Ward	5.70 ± 0.76	5.98 ± 0.82
	Over-bed Days ICU	4.43 ± 0.73	4.54 ± 0.84
	Over-beds Ward	0.40 ± 0.10	0.44 ± 0.14
	Over-beds ICU	0.13 ± 0.04	0.14 ± 0.04
	Occupancy Ward	20.42 ± 0.50	20.60 ± 0.54
	Occupancy ICU	7.95 ± 0.21	7.89 ± 0.21
	Load Fraction	0.26 ± 0.01	0.21 ± 0.01
Haga	Over-bed Days Ward	3.80 ± 0.55	3.89 ± 0.55
	Over-bed Days ICU	4.43 ± 0.84	4.68 ± 0.93
	Over-beds Ward	0.24 ± 0.04	0.25 ± 0.04
	Over-beds ICU	0.16 ± 0.05	0.17 ± 0.07
	Occupancy Ward	30.04 ± 0.50	30.05 ± 0.51
	Occupancy ICU	13.02 ± 0.26	12.97 ± 0.29
	Load Fraction	0.41 ± 0.02	0.33 ± 0.02
GHZ	Over-bed Days Ward	6.58 ± 0.85	6.94 ± 0.97
	Over-bed Days ICU	3.10 ± 0.57	3.47 ± 0.58
	Over-beds Ward	0.47 ± 0.09	0.52 ± 0.13
	Over-beds ICU	0.09 ± 0.02	0.10 ± 0.02
	Occupancy Ward	23.51 ± 0.48	23.41 ± 0.52
	Occupancy ICU	9.11 ± 0.22	9.17 ± 0.22
	Load Fraction	0.29 ± 0.01	0.23 ± 0.01
Region	Over-bed Days Ward	13.14 ± 1.23	14.16 ± 1.35
	Over-bed Days ICU	9.91 ± 1.17	10.20 ± 1.26
	Over-beds Ward	1.12 ± 0.17	1.21 ± 0.22
	Over-beds ICU	0.38 ± 0.07	0.41 ± 0.09
	Occupancy Ward	73.97 ± 0.97	74.07 ± 0.98
	Occupancy ICU	30.07 ± 0.45	30.03 ± 0.47

Table 10
Average number of patient reallocations across regions (with 95% confidence interval) according to the simulation study of Section 4.2 for horizons s set to 3 and 5.

KPI	Dept.	Horizon 3	Horizon 5
Total	Ward	173.80 ± 6.37	177.70 ± 6.90
	ICU	100.40 ± 5.92	99.31 ± 5.36
External	Ward	3.50 ± 1.05	4.93 ± 1.20
	ICU	14.14 ± 2.31	12.27 ± 2.03
In Clusters	Ward	122.11 ± 4.26	124.51 ± 4.63
	ICU	50.90 ± 2.83	52.56 ± 2.60
Btw. Clusters	Ward	48.20 ± 3.75	48.26 ± 3.69
	ICU	35.37 ± 2.32	34.48 ± 2.29
Average Cost	Ward	22.25 ± 0.77	22.66 ± 0.77
	ICU	33.98 ± 1.25	32.38 ± 1.29

Next, in Table 9, results are shown for the simulation study of Section 4.1 for horizons 3 and 5 days. No significant differences (95%) are seen in the KPIs, the load balancing policy is seen to perform slightly worse on average for a horizon of 5 days when looking at the regional numbers. The samples of maximum occupancy over 5 days are larger than over 3 days, hence the load coefficients are significantly smaller for a horizon of 5 days.

Finally, in Table 10, the results for the simulation study of Section 4.2 are shown for horizons 3 and 5. The policies show a similar performance, no significant differences were found. Slight increases are seen on average when looking at the number of patients reallocated in the clusters and the total number of reallocated patients.

Appendix D. List of symbols

General indices	
r	regions, $r = 1, \ldots, R$
h	hospitals (within a region r), $h = 1, \ldots, H_r$
$c \in C$	patient class determined by characteristics c
s, t, u	time; usually, s denotes the current time and u is in the time interval $[s, s+t]$
Individual hospital model	
$\lambda_{hc}(t)$	arrival rate of c patients at hospital h
$p_{hc}(t)$	fraction of c patients admitted to hospital h's ward
$q_{hcW}(t),\ q_{hcI}(t)$	probability that a c patient is discharged
$L_{hcW}(t),\ L_{hcI}(t)$	length of stay (LoS) of c patients admitted to the ward, ICU of hospital h
$\mathbf{L}_h(s),\ \ell_h$	tuples of patients' location and realised LoSs (up to time s) in hospital h.
$\alpha_{hW},\ \alpha_{hI}$	safety levels of the ward, ICU of hospital h
$N_{hcW}(t),\ N_{hcI}(t)$	number of c patients in hospital h
$n^*_{hW}(s,t),\ n^*_{hI}(s,t)$	number of beds in hospital h in $[s, s+t]$
$n_{hW,\alpha_{hW}}(s,t),\ n_{hI,\alpha_{hI}}(s,t)$	α_{hW}-quantile, α_{hI}-quantile for maximum occupancy in ward, resp. ICU of hospital h in $[s, s+t]$
$\tilde{n}_{hW,\alpha_{hW}}(s,t),\ \tilde{n}_{hI,\alpha_{hI}}(s,t)$	bed surplus in the ward, ICU of hospital h in $[s, s+t]$ at safety levels α_{hW}, resp. α_{hI}
Individual region model	
$\Lambda_{rc}(t)$	arrival rate of c patients in region r
$\Lambda_r(t)$	$\sum_{c \in C} \Lambda_{rc}(t)$
$p_{rc}(t)$	fraction of c patients admitted to the ward of hospitals in region r
$P_{rc}(t)$	fraction of (regional) c patients admitted to the virtually merged regional ward of region r
α_r	set containing all safety levels of individual hospitals in region r: $\alpha_r = \{\alpha_{hW},\ \alpha_{hI} : h = 1, \ldots, H_r\}$
$\alpha_{rW},\ \alpha_{rI}$	safety levels of the (virtually merged) ward, ICU of region r
$\theta_{h,\alpha_{hW},\alpha_{hI}}(s,t)$	fraction of regional arrivals that hospital h can accommodate in $[s, s+t]$ at safety levels $\alpha_{hW},\ \alpha_{hI}$
$\theta_{r,\alpha_r}(s,t)$	$\theta_{r,\alpha_r}(s,t) = \sum_{h=1}^{H_r} \theta_{h,\alpha_{hW},\alpha_{hI}}(s,t)$
$\widehat{\theta}_{hW,\alpha_{rW}}(s,t),\ \widehat{\theta}_{hI,\alpha_{rI}}(s,t)$	fraction of regional patients hospitalised in the ward, ICU of hospital h after admittance to the virtually merged regional ward, ICU of region r
$\tilde{n}_{rW,\alpha_r}(s,t),\ \tilde{n}_{rI,\alpha_r}(s,t)$	bed surplus in the ward, ICU of region r in $[s, s+t]$ at individual hospital safety levels α_r.
$n^*_{rW}(s,t),\ n^*_{rI}(s,t)$	number of beds in the virtually merged regional ward, ICU of region r in $[s, s+t]$
$n_{rW,\alpha_{rW}}(s,t),\ n_{rI,\alpha_{rI}}(s,t)$	α_{rW}-quantile, α_{rI}-quantile for maximum occupancy in the virtually merged ward, resp. ICU of region r in $[s, s+t]$
$M_{rW,\alpha_r}(s,t),\ M_{rI,\alpha_r}(s,t)$	regional bed shortage in the ward, ICU of region r in the time-interval $[s, s+t]$ at all individual hospital safety levels α_r
$m_{rW,\alpha_r}(s,t),\ m_{rW,\alpha_r}(s,t)$	mean regional bed shortage in the ward, ICU of region r in $[s, s+t]$ (belonging to $M_{rW,\alpha_r}(s,t)$, resp. $M_{rI,\alpha_r}(s,t)$)
Multiple regions model	
$\tilde{n}_{rW}(s),\ \tilde{n}_{rI}(s)$	bed surplus (≥ 0) or shortage (< 0) in the ward(s), ICU(s) of region r at time s
$\tilde{n}_{rW}(s,t),\ \tilde{n}_{rI}(s,t)$	forecast of the bed surplus (≥ 0) or shortage (< 0) in the ward(s), ICU(s) of region r in $[s, s+t]$.
$\gamma_{r,r'}$	costs for reallocating a patient from region r to region r'
$f_{W,r,r'}(s),\ f_{I,r,r'}(s)$	the number of ward, ICU patients to reallocate from region r to region r' at time s
$f_{W,r,r'}(s,t),\ f_{I,r,r'}(s,t)$	the number of potentially additionally required reallocations of ward, ICU patients from region r to region r' in $[s, s+t]$ based on the bed forecasts
$g(\cdot)$	penalty function to balance/level bed surpluses across regions
$\delta_{W,r}(s),\ \delta_{I,r}(s)$	relative remaining bed surplus in the ward, ICU of region r at (current) time s
$\delta_{W,r}(s,t),\ \delta_{I,r}(s,t)$	forecast relative remaining bed surplus in the ward, ICU of region r in $[s, s+t]$

(*continued on next page*)

References

[1] Li X. A two-level policy for controlling an epidemic and its dynamics. Omega 2023;115:102753. doi:10.1016/j.omega.2022.102753.

[2] Nguyen NT, Bish EK, Bish DR. Optimal pooled testing design for prevalence estimation under resource constraints. Omega 2021;105:102504. doi:10.1016/j.omega.2021.102504.

[3] Mohammadi M, Dehghan M, Pirayesh A, Dolgui A. Bi-objective optimization of a stochastic resilient vaccine distribution network in the context of the COVID-19 pandemic. Omega 2022;113:102725. doi:10.1016/j.omega.2022.102725.

[4] Tang L, Li Y, Bai D, Liu T, Coelho LC. Bi-objective optimization for a multi-period COVID-19 vaccination planning problem. Omega 2022;110:102617. doi:10.1016/j.omega.2022.102617.

[5] da Silva F, Barbosa C. The impact of the COVID-19 pandemic in an intensive care unit (ICU): psychiatric symptoms in healthcare professionals. Prog Neuro-Psychopharmacol Biol Psychiatry 2021;110:110299. doi:10.1016/j.pnpbp.2021.110299. https://www.sciencedirect.com/science/article/pii/S0278584621000580

[6] González-Gil M, González-Blázquez C, Parro-Moreno A, Pedraz-Marcos A, Palmar-Santos A, Otero-García L, et al. Nurses' perceptions and demands regarding COVID-19 care delivery in critical care units and hospital emergency services. Intensive Crit Care Nurs 2021;62:102966. doi:10.1016/j.iccn.2020.102966. https://www.sciencedirect.com/science/article/pii/S0964339720301695

[7] Shaker Ardakani E, Gilani Larimi N, Oveysi Nejad M, Madani Hosseini M, Zargoush M. A resilient, robust transformation of healthcare systems to cope with COVID-19 through alternative resources. Omega 2023;114:102750. doi:10.1016/j.omega.2022.102750.

[8] COVIDSurg Collaborative. Elective surgery cancellations due to the COVID-19 pandemic: global predictive modelling to inform surgical recovery plans. Br J. Surg. 2020;107(11):1440–9. doi:10.1002/bjs.11746.

[9] van Giessen A., de Wit A., van den Brink C., Degeling K., Deuning C., Eeuwijk J., van den Ende C., van Gestel I., Gijsen R., van Gils P., IJzerman M., de Kok I., Kommer G., Kregting L., Over E., Rotteveel A., Schreuder K., Stadhouders N., Suijkerbuijk A.. Impact van de eerste COVID-19 golf op de reguliere zorg en gezondheid: inventarisatie van de omvang van het probleem en eerste schatting van gezondheidseffecten. Rijksinstituut voor Volksgezondheid en Milieu (RIVM); 2020. Report in Dutch; English abstract provided. 10.21945/RIVM-2020-0183

[10] Hanna T, King W, Thibodeau S, Jalink M, Paulin G, Harvey-Jones E, O'Sullivan D, Booth C, Sullivan R, Aggarwal A. Mortality due to cancer treatment delay: systematic review and meta-analysis. BMJ 2020;371. doi:10.1136/bmj.m4087. https://www.bmj.com/content/371/bmj.m4087

[11] Sarkar S, Pramanik A, Maiti J, Reniers G. COVID-19 outbreak: a data-driven optimization model for allocation of patients. Comput Ind Eng 2021;161:107675. doi:10.1016/j.cie.2021.107675.

[12] LCPS. Over het LCPS. 2021. http://lcps.nu/over-ons/ (In Dutch); Last accessed: March 30, 2021.

[13] Landelijk Netwerk Acute Zorg (LNAZ). http://www.lnaz.nl/acute-zorg (In Dutch); Last accessed: May 11, 2021.

[14] Bekker R, uit het Broek M, Koole G. Modeling COVID-19 hospital admissions and occupancy in the Netherlands. Eur J Oper Res 2023;304:207–18.

[15] Ross K. Multiservice loss models for broadband telecommunication networks. Telecommunication networks and computer systems. 1st ed. London: Springer-Verlag; 1995. doi:10.1007/978-1-4471-2126-8.

[16] Zachary S, Ziedins I. Loss networks. In: Boucherie R, van Dijk N, editors. Queueing networks: a fundamental approach. International series in operations re-

search & management science (ISOR), vol. 154. Boston, MA: Springer; 2011. p. 701–28. doi:10.1007/978-1-4419-6472-4_16.

[17] van der Boor M, Borst S, van Leeuwaarden J, Mukherjee D. Scalable load balancing in networked systems: universality properties and stochastic coupling methods. In: Proceedings international congress of mathematicians (ICM 2018), Rio de Janeiro; 2018. p. 3911–42.

[18] Eager D, Lazowska E, Zahorjan JA. A comparison of receiver-initiated and sender-initiated adaptive load sharing. Perform Eval 1986;6(1):53–68.

[19] Minnebo W, Van Houdt B. A fair comparison of pull and push strategies in large distributed networks. IEEE/ACM Trans Netw 2014;22(3):996–1006. doi:10.1109/TNET.2013.2270445.

[20] Cao P, Zhong Z, Huang J. Dynamic routing in a distributed parallel many-server service system: the effect of ξ-choice. Eur J Oper Res 2021;294(1):219–35. doi:10.1016/j.ejor.2021.01.026.

[21] Zhong Z, Cao P. Balanced routing with partial information in a distributed parallel many-server queueing system. Eur J Oper Res 2023;304(2):618–33. doi:10.1016/j.ejor.2022.02.042.

[22] Bonald T, Jonckheere M, Proutiére A. Insensitive load balancing. ACM SIGMETRICS Perform Eval Rev 2004;32(1):367–77. doi:10.1145/1012888.1005729.

[23] Lin H-C, Raghavendra C. Modelling and analyses of dynamic load-balancing policies by state aggregation. Int J Model Simul 1997;17(1):20–8. doi:10.1080/02286203.1997.11760307.

[24] Cardellini V, Colajanni M, Yu P. Dynamic load balancing on web-server systems. IEEE Internet Comput 1999;3(3):28–39. doi:10.1109/4236.769420.

[25] Hellemans T, Bodas T, Van Houdt B. Performance analysis of workload dependent load balancing policies. Proc AMC Meas Anal Comput Syst 2019;3(2). doi:10.1145/3341617.3326150.

[26] Milani AS, Navimipour NJ. Load balancing mechanisms and techniques in the cloud environments: systematic literature review and future trends. J Netw Comput Appl 2016;71:86–98. doi:10.1016/j.jnca.2016.06.003.

[27] van Dijk N, van der Sluis E. To pool or not to pool in call centers. Prod Oper Manag 2008;17(3):296–305. doi:10.3401/poms.1080.0029.

[28] Wallace R, Whitt W. A staffing algorithm for call centers with skill-based routing. Manuf Serv Oper Manag 2005;7(4):276–94. doi:10.1287/msom.1050.0086.

[29] Towsley D. Queuing network models with state-dependent routing. J ACM 1980;27(2):323–37. doi:10.1145/322186.322196.

[30] Jonckheere M. Insensitive versus efficient dynamic load balancing in networks without blocking. Queueing Syst 2006;54:193–202. doi:10.1007/s11134-006-0066-3.

[31] Leino J, Virtamo J. Insensitive load balancing in data networks. Comput Netw 2006;50(8):1059–68. doi:10.1016/j.comnet.2005.09.009. Selected Papers from the 3rd International Workshop on QoS in Multiservice IP Networks (QoS-IP 2005)

[32] Marin A, Balsamo S, Fourneau J-M. LB-networks: a model for dynamic load balancing in queueing networks. Perform Eval 2017;115:38–53. doi:10.1016/j.peva.2017.06.004.

[33] Mukherjee D., Borst S.C., van Leeuwaarden J.S., Whiting P.A.. Asymptotic optimality of power-of-*d* load balancing in large-scale systems. arXiv preprint arXiv:1612007222016.

[34] Klein Haneveld W, Van der Vlerk M, Romeijnders W. Stochastic programming: modeling decision problems under uncertainty. Springer Nature; 2019.

[35] Farcomeni A, Maruotti A, Divino F, Jona-Lasinio G, Lovison G. An ensemble approach to short-term forecast of COVID-19 intensive care occupancy in Italian regions. Biom J 2021;63:503–13.

[36] Goic M, Bozanic-Leal M, Badal M, Basso L. COVID-19: short-term forecast of ICU beds in times of crisis. PLoS One 2021;16(1):e0245272.

[37] Manca D, Caldiroli D, Storti E. A simplified math approach to predict ICU beds and mortality rate for hospital emergency planning under COVID-19 pandemic. Comput Chem Eng 2020;140:106945.

[38] Massonnaud C., Roux J., Crépey P.. COVID-19: forecasting short term hospital needs in France. medRxiv preprint2020;Available at https://doi.org/10.1101/2020.03.16.20036939.

[39] Cheng F-Y, Joshi H, Tandon P, Freeman R, Reich D, Mazumdar M, et al. Using machine learning to predict ICU transfer in hospitalized COVID-19 patients. J Clin Med 2020;9(6). doi:10.3390/jcm9061668. https://www.mdpi.com/2077-0383/9/6/1668

[40] Baas S, Dijkstra S, Braaksma A, van Rooij P, Snijders F, Tiemessen L, et al. Real–time forecasting of COVID-19 bed occupancy in wards and intensive care units. Health Care Manag Sci 2021;24:402–19.

[41] Foucrier A, Perrio J, Grisel J, Crépey P, Gayat E, Vieillard-Baron A, et al. Transition matrices model as a way to better understand and predict intra-hospital pathways of COVID-19 patients. Sci Rep 2022;12:17508. doi:10.1038/s41598-022-22227-8.

[42] Roimi M, Gutman R, Somer J, Ben Arie A, Calman I, Bar-Lavie Y, Gelbshtein U, Liverant-Taub S, Ziv A, Eytan D, Gorfine M, Shalit U. Development and validation of a machine learning model predicting illness trajectory and hospital utilization of COVID-19 patient: a nationwide study. J Am Med Inform Assoc 2021.

[43] Zhao C, Tepekule B, Criscuolo N, Wendel-Garcia P, Hilty M, Fumeaux T, Boeckel TV. *Icumonitoring.ch*: a platform for short-term forecasting of intensive care unit occupancy during the COVID-19 epidemic in Switzerland. Swiss Med Weekly 2020;150:w20277.

[44] Richards F. A flexible growth function for empirical use. J Exp Bot 1959;10(2):290–301.

[45] Kaplan E, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958;53(282):457–81.

[46] Massey W, Whitt W. Networks of infinite-server queues with nonstationary Poisson input. Queueing Syst 1993;13(1–3):183–250.

[47] Whitt W, Zhang X. A data-driven model of an emergency department. Oper Res Health Care 2017;12:1–15.

[48] Litvak N, van Rijsbergen M, Boucherie R, van Houdenhoven M. Managing the overflow of intensive care patients. Eur J Oper Res 2008;185(3):998–1010.

[49] Smith D, Whitt W. Resource sharing for efficiency in traffic systems. Bell Syst Tech J 1981;60(1):39–55. doi:10.1002/j.1538-7305.1981.tb00221.x. http://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1981.tb00221.x

[50] Rijksinstituut voor Volksgezondheid en Milieu (RIVM). COVID-19 aantallen per gemeente per publicatiedatum. https://data.rivm.nl/geonetwork/srv/dut/catalog.search#/metadata/5f6bc429-1596-490e-8618-1ed8fd768427. Last accessed: March 26, 2021.

[51] Rijksinstituut voor Volksgezondheid en Milieu (RIVM). COVID-19 ziekenhuisopnames (volgens NICE registratie) per gemeente per ziekenhuisopnamedatum en meldingsdatum. https://data.rivm.nl/geonetwork/srv/dut/catalog.search#/metadata/4f4ad069-8f24-4fe8-b2a7-533ef27a899f. Last accessed: March 26, 2021.

[52] Chiu S, Stoyan D, Kendall W, Mecke J. Stochastic geometry and its applications. Wiley series in probability and statistics. 3rd ed. John Wiley & Sons, Ltd; 2013. doi:10.1002/9781118658222.

[53] Faes C, Abrams S, Van Beckhoven D, Meyfroidt G, Vlieghe E, Hens NBelgian Collaborative Group on COVID-19 Hospital Surveillance. Time between symptom onset, hospitalisation and recovery or death: statistical analysis of Belgian COVID-19 patients. Int J Environ Res Public Health 2020;17(20):7560.

[54] Lee S, Lei B, Mallick B. Estimation of COVID-19 spread curves integrating global data and borrowing information. PLoS One 2020;15(7):e0236860.

[55] Wu K, Darcet D, Wang Q, Sornette D. Generalized logistic growth modeling of the COVID-19 outbreak: comparing the dynamics in the 29 provinces in China and in the rest of the world. Nonlinear Dyn 2020;101:1561–81. doi:10.1007/s11071-020-05862-6.

[56] NAZB. Dashboard COVID-19 regio Brabant. https://www.nazb.nl/covid-19/dashboard-covid-19-regio-brabant; Last accessed March 29, 2021.

[57] Robbins H, Monro S. A stochastic approximation method. Ann Math Stat 1951;22(3):400–7.

[58] Blum J. Approximation methods which converge with probability one. Ann Math Stat 1954;25(2):382–6.